

*ADVANCING OPERATIONAL INTELLIGENCE: PREDICTIVE PROCESS  
MONITORING FOR THE IMPERFECT ORDER PREDICTION IN THE ORDER LIFE  
CYCLE MANAGEMENT THROUGH MACHINE LEARNING AND DATA SCIENCE*

by

Shreya Tiwari, MTech

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfilment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

<MONTH OF GRADUATION, YEAR>

*ADVANCING OPERATIONAL INTELLIGENCE: PREDICTIVE PROCESS  
MONITORING FOR THE IMPERFECT ORDER PREDICTION IN THE ORDER LIFE  
CYCLE MANAGEMENT THROUGH MACHINE LEARNING AND DATA SCIENCE*

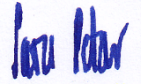
by

Shreya Tiwari

Supervised by

Dr. Anna Provodnikova

APPROVED BY



---

Prof.dr.sc. Saša Petar, Ph.D., Dissertation chair

RECEIVED/APPROVED BY:

---

Admissions Director

## **Acknowledgements**

Throughout the doctoral journey, I have been blessed with the unwavering support of many people who have played important roles in my success.

Firstly, I offer my deepest gratitude to the divine force that has guided me throughout this journey to complete the doctoral degree.

Secondly, I would like to express my deepest appreciation for Professor Anna, who acted as my supervisor and chairperson of the dissertation, and whose constant guidance, precious advice, constructive criticism, and never-ending encouragement forms an integral part of what led me here today. Without your direction and insights into how difficult this research journey might be professor!

I am also indebted to the management and staff at SSBM for giving me chance to study at such a prestigious business school.

I dedicate this doctoral degree to my beloved parents and sister Poorva. Your unswerving dedication, unconditional love and unwavering moral support have been essential pillars upon which I build my academic career on. Overcoming obstacles through determination instilled by you while finishing this doctorate research.

Finally, yet not least importantly is Sri Harsha, my dear husband. Your love, encouragement, and unwavering patience have been my constant companions through every challenge and triumph. Thank you for being my rock, my confidant, and my greatest supporter. This achievement is as much yours as it is mine. Your unwavering belief in me and your endless support have enabled me to pursue and achieve my dreams. I am forever grateful for your presence in my life.

ABSTRACT  
*ADVANCING OPERATIONAL INTELLIGENCE: PREDICTIVE PROCESS  
MONITORING FOR THE IMPERFECT ORDER PREDICTION IN THE ORDER LIFE  
CYCLE MANAGEMENT THROUGH MACHINE LEARNING AND DATA SCIENCE*

Shreya Tiwari  
2024

Dissertation Chair: <Chair's Name>  
Co-Chair: <If applicable. Co-Chair's Name>

**Abstract:** Predictive analytics are vital to modern business processes. Businesses rely on them to analyze, understand, forecast, and make strategic decisions based on future data points. This dissertation focuses on predicting the completion status of ongoing processes within the order life cycle. The order life cycle is a broad series of events that begins with placing an order, all the way through to return or delivery. It's difficult identifying problems within this lifecycle manually, emphasizing the need for predictive techniques.

By using process mining, we may anticipate the outcome of running instances by adding temporal information to previously recognized process models. We construct configurable process models using time-based sampling from older instances. Our main goal is predictive monitoring tasks - more specifically, being able to predict the end output of an ongoing order in an order life cycle. We use various classification techniques to achieve the goal.

Using actual event logs and incorporating time-based sampling methods gives us a good benchmark for predictions performance against reality. According to our research, the best outcomes are obtained when a bag of activities is included as a feature. We demonstrate how elements in event logs affect classifier selection using qualitative analysis.

The assessment of our prediction using F1 score, and support metrics yields valuable information about the factors that lead to order defects. This information may be used to improve operational intelligence and enable pre-emptive actions throughout the order life cycle.

## **Keywords**

Machine Operational Intelligence; Predictive Analytics; Business Process Management; Process Mining; Predictive Monitoring; Order Life Cycle; Machine Learning; Data Science; Imperfect Order Prediction; Classification Techniques; Bag of Activity

## **List of Abbreviations**

OLC: Order Life Cycle

OLCM: Order Life Cycle Management

PM: Process Mining

PPM: Predictive Process Mining

BOA: Bag of Activity

ML: Machine Learning

LM: Last Mile

SAC: Service Authorization Code

LCD: Life Cycle Development

CC: Customer Care

EDA: Exploratory Data Analysis

## TABLE OF CONTENTS

List of Figures .....	xii
<b>Chapter I.....</b>	<b>1</b>
<b>INTRODUCTION.....</b>	<b>1</b>
1.1 Introduction.....	1
1.2 Research Problem .....	5
1.3 Purpose of Research.....	7
1.3.1 Review on Predictive Process Mining .....	9
1.3.2 Identification of Improvement Opportunities .....	10
1.3.3 Analysis of Impact on Operational Excellence .....	11
1.3.4 Guidance Implementation .....	13
1.4 Significance of the Study .....	13
1.4.1 Academic Contribution .....	14
1.4.2 Real World Impact .....	14
1.4.3 Business Strategy .....	14
1.4.4 Operational Excellence .....	15
1.4.5 Customer Happiness Index .....	15
1.5 Research Purpose and Questions .....	18
<b>Chapter II REVIEW OF LITERATURE.....</b>	<b>21</b>
2.1 Theoretical Framework .....	22
2.1.1 Order Life Cycle Management .....	22
2.1.2 Process Mining and Event Logs.....	22
2.1.3 Predictive Process Monitoring.....	22
2.1.4 Combining and Integrating Predictive Process Monitoring and Process Mining.....	23
2.1.5 The Role of Machine Learning (ML) in Predictive Process Monitoring .....	23
2.2 Empirical Studies .....	24
2.3 Conceptual Studies.....	25
2.4 Current Position and Goals .....	26
2.5 Predictive Process Monitoring: Theory and Practice.....	27
2.6 A Predictive Approach for Order Life Cycle Management Performance Improvement .....	28
2.7 Mathematical formalism of the existing literature .....	29
2.7.1. Order life cycle management and predictive process monitoring are not integrated. ....	29
2.7.2. Limited Research on Order Fulfilment Optimisation Using Predictive Process Mining .....	31
2.7.3. Little focus on imperfect order prediction in real time .....	32
2.7.4. Insufficient Research on Machine Learning Methods for Order Processing .....	34



2.7.5. Narrow comprehension of prognostic process tracking regarding the customer .....	35
<b>Chapter III: METHODOLOGY .....</b>	<b>38</b>
3.1 Overview of the Research Problem .....	38
3.2 Operationalization of Theoretical Constructs .....	41
3.2.1 Operationalization of Predictive Process Monitoring .....	41
3.2.2 Operationalization of Order Life Cycle Management .....	42
3.2.3 Integration of Predictive Process Monitoring and Order Life Cycle Management.....	42
3.3 Machine Learning Models used in Predictive Process Mining.....	43
3.3.1. Logistic Regression.....	45
3.3.2. Decision tree .....	53
3.3.3. Random Forest .....	62
3.3.4. XG Boost .....	71
3.4 Research Design.....	82
3.4.1 Pre-Processing.....	82
3.4.2 Model Training .....	89
3.4.3 Model Testing .....	91
3.4.4 Post Processing .....	92
3.5 Population and Sample .....	93
3.5.1 Overview .....	94
3.5.2 Sample Creation.....	95
3.5.3 Inclusion Criteria .....	96
3.5.4 Exclusion Criteria .....	97
3.5.5 Event Log Dataset of Order Life Cycle .....	98
3.5.6 Bag of Activity (BoA) Feature.....	100
3.6 Data Collection Procedures.....	103
3.7 Data Analysis .....	104
3.8 Methodological Insights: Unveiling How Selected Approaches Address Research Questions .....	105
3.8.1 Research Question 1: How efficient are methods for pattern recognition and outcome prediction in order lifecycle management using predictive process mining?.....	105
3.8.2 Research Question 2: What are the main elements affecting the predictability and accuracy of predictive models in terms of identifying imperfection or inefficiencies in order processing? .....	112
3.8.3. Research Question 3: How do various machine learning algorithms stack up in terms of how well they predict order lifecycle management outcomes? .....	122
3.8.4. Research Question 4: How will putting predictive process mining technologies into practice affect customer satisfaction and organizational performance?.....	131
3.8.5 Research Question 5: To increase productivity and lower mistake	

rates, how can companies successfully incorporate predictive process mining into their current order management systems? .....	137
3.9 Research Design Limitations .....	144
3.10 Conclusion .....	145
<b>Chapter IV RESULTS .....</b>	<b>147</b>
4.1 Research Question 1: How do various machine learning algorithms stack up in terms of how well they predict order lifecycle management outcomes? .....	147
4.1.1 Logistic Regression.....	148
4.1.2 Decision Tree .....	150
4.1.3 Random Forest .....	153
4.1.4 XG Boost .....	155
4.2 Research Question 2: How efficient are methods for pattern recognition and outcome prediction in order lifecycle management using predictive process mining?	
4.2.1 Case Study One.....	161
4.2.2 Case Study Two .....	163
4.2.3 Case Study Three .....	165
4.3 Research Question 3: What are the main elements affecting the predictability and accuracy of predictive models in terms of identifying imperfection or inefficiencies in order processing?.....	169
4.4. How will putting predictive process mining technologies into practice affect customer satisfaction and organisational performance? .....	176
4.5 In order to increase productivity and lower mistake rates, how can companies successfully incorporate predictive process mining into their current order management systems?.....	184
4.6. Summary of Findings.....	190
4.7 Conclusion .....	191
<b>Chapter V DISCUSSION.....</b>	<b>193</b>
5.1 Discussion of Results.....	193
5.2 Discussion of Research Question.....	193
5.3 Effectiveness of Predictive Process Mining.....	195
5.4 Integration of Activity-Based Modeling.....	196
<b>Chapter VI SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS.....</b>	<b>198</b>
6.1 Summary .....	198
6.2 Implications.....	202
6.2.1 Practical Implications.....	202
6.2.2 Theoretical Implications .....	204
6.2.3 Methodological Implications .....	205
6.2.4 Managerial Implications .....	207
6.3 Recommendations for Future Research .....	208
6.3.1 Customized Training and Prediction.....	209
6.3.2 Feature Enrichment for Improved Accuracy.....	210
6.3.3 Cross-Country Model Adoption .....	211
6.3.4 Integration with Order Systems .....	213

6.3.5 Scalability and Generalizability .....	214
6.3.6 Evaluation of Real Time Predictive Capabilities .....	215
6.4 Conclusion .....	217
<b>References .....</b>	<b>219</b>

## LIST OF FIGURES

Figure 1: An excerpt of a “spaghetti”-like process model .....	6
Figure 2: A structured, or “lasagna”-like, process model .....	7
Figure 3: General overview of research design.....	82
Figure 4: EDA: Percentage distribution of perfect and imperfect .....	83
Figure 5: Class Distribution before Under-sampling .....	84
Figure 6: Class Distribution after Under-sampling.....	85
Figure 7: Percentage count of unique imperfect orders with a position of occurrence of the first imperfect order activity.....	86
Figure 8: Most frequent imperfect activity that occur in OLC .....	87
Figure 9: Data Pre-processing flow .....	89
Figure 10: Model Training flow.....	90
Figure 11: Model Testing Flow .....	92
Figure 12: Post Processing flow.....	93
Figure 13: Sample event log dataset of an Order Life Cycle.....	101
Figure 14: Sample bag of activity dataset with all perfect activities in sequence.....	102
Figure 15: Sample bag of activity dataset with all imperfect activities in sequence .....	102
Figure 16: Logistic Regression: Training Classification Report.....	150
Figure 17: Logistic Regression: Test Classification Report .....	150
Figure 18: Decision Tree: Training Classification Report.....	153
Figure 19: Decision Tree: Test Classification Report.....	153
Figure 20: Random Forest: Training Classification Report.....	155
Figure 21: Random Forest: Test Classification Report.....	155
Figure 22: XG Boost: Training Classification Report .....	158
Figure 23: XG Boost: Test Classification Report .....	158
Figure 24: Figure 24 SHAP Explainer for case key 1358405196A01422.....	162
Figure 25: Feature Importance Heat map for case key 1358405196A01422 .....	162
Figure 26: SHAP Explainer for case key 1345897825A01288 .....	163
Figure 27: Feature Importance Heat map for case key 1345897825A01288 .....	164

Figure 28: Order Life Cycle of Case Key:1360331646A01288 .....	166
Figure 29: SHAP Explainer : 1360331646A01288 .....	167
Figure 30: Feature Importance Heatmap for case key 1360331646A01288 .....	168

## CHAPTER I

### INTRODUCTION

#### **1.1 Introduction**

In the information age, many entities in different industries are realizing that data can be a source of insights to make decisions and improve operational efficiency. Towards this end, machine learning (ML) is at the forefront and is able to extract actionable knowledge from large volumes of data. In business process management (BPM), which has always been biased on traditional manual data collection procedures, this change heralds a significant paradigm shift towards data-centric methods.

For example, it is anticipated that integrating ML technologies into BPM practices will bring about a significant realignment during runtime phases such as monitoring where predictive insights can be made. This move represents a shift from batch-oriented historical analysis to real-time event-driven decision-making. Thus, organizations are provided with real-time predictive analytics capabilities that helps them in anticipating process behavior, performance trends and outcome probabilities. In this space within BPM domain stand process mining that offers valuable insights derived from its historical event logs and other related processes information.

Predictive monitoring has emerged as an important sub-branch of business process mining because organizations are increasingly relying on predictive insights for proactive decision making. Predictive monitoring enables decision-makers to predict future states of business processes so that they can address possible risks before they materialize or leverage emerging opportunities. However, among many methods developed for predictive monitoring, identifying the most suitable approach for different

application scenarios remains a challenge because there does not exist one best way to deal with all situations with respect to their scale or specific requirements.

Over the years, BPM has changed significantly with more focus on data-driven approaches particularly in runtime phases like monitoring to draw out predictive insights. This shift shows that the usual ways of managing business processes are not adequate in today's dynamic and competitive world. By using data analytics and machine learning technologies, firms can gain a better understanding of their processes by identifying patterns, trends and make real time decisions based on what they have found. The progression towards data driven BPM is a paradigm shift for organizations as they move from reactive to proactive strategies.

Similarly, this changeover to data-driven BPM has been influenced by the growing availability of data and technological advancements. The widespread use of enterprise systems plus the proliferation of digital technologies have given organizations access to enormous amounts of data generated from sources such as sensors, transaction logs or social media platforms. For organizations willing to tap into its potential value chain, it represents both a challenge and an opportunity due to abundant information available. Process mining within BPM field specifically emerged as an important mechanism of dissecting and interpreting this kind of information in order for entities to expose hidden insights that would optimize business performance through process improvements.

The transformational trends in BPM towards becoming more inclined on using data is fundamentally interesting; this is how organizations manage their process now-a-days. With knowledge that's derived from various sources including those mentioned above when put under analysis tools like machine learning algorithms or statistical software programs for decision making purposes one can be able to predict outcomes that

may arise in future thereby being able to take measures before hand any business success could be achieved through procedural knowledge gained after having insight into deeper aspects about them. This development highlights the relevance of embracing these strategies in operating environment that is full competition besides giving hope for continuous innovation so as improve BPM practices further if possible.

Predictive monitoring represents one aspect within process mining aimed at preempting future states in business processes thus enabling decision makers respond beforehand by utilizing such information pro actively . The objective of predictive monitoring is to predict events or results in a business process' life cycle with the help of historical statistics and advanced analysis. Decision-maker's insight into process trajectory, potential disruptions that may occur, as well as preemptive measures taking to minimize danger is therefore enhanced. However, despite the promise of predictive monitoring, selecting suitable techniques for specific application scenarios remains a significant challenge. This difficulty stems from the fact that there are many methods in this area predicting future demands due to processes that vary across industries which are why most relevant context for predictive monitoring cannot be determined . As such there is need for comprehensive evaluations of predictive monitoring techniques to ascertain their suitability and effectiveness in different domains and operational contexts.

This thesis will address this issue by quantitatively assessing predictive-monitoring techniques for business processes, mainly concentrating on forecasting the end results before hand in order management life cycle. In many traditional literature surveys, existing techniques are generally described in broad terms without any detailed quantitative analysis. However, we intend to measure these approaches across multiple scenarios and settings to establish comprehensive benchmarks in our research. Our main objective is developing insights that may be used as a guide to action with regard to this



important aspect of operations management since our research is specifically targeted at the order management lifecycle. And this kind of approach allows us to examine how effective predictive monitoring can be in real life situations where accurate prediction of outcomes has a lot of implications on business success. In view of this, we plan through our quantitative assessment to explore the strengths and weaknesses of different predictive monitoring approaches used by organizations so that they know which one is most suitable for their specific requirements.

The value of this benchmark lies in its help in identifying the most effective methods given a context chosen by decision makers. Predictive monitoring technique options are abundant in a modern fast-paced business environment, all claiming superiority over others from their fellow competitors and marketers. However, given no empirical evidence supporting such claims it becomes problematic selecting the most appropriate method among them. For instance, this benchmark has been done on purpose for decision-makers whereby it seeks at giving them some measurable quantities that could compare and contrast with regards to various predictive monitoring techniques employed. When compared with other means and datasets applied; test these approaches extensively yields greater insights into more robust methods.

Also, experiment-based benchmarks should be used because performance tends to differ significantly across different datasets due to machine learning techniques being idiosyncratic. The result ensures fair comparison since there are experiments conducted under equal conditions that can comprehensively evaluate each technique performance systematically. An inclusion of variety of classifiers and data sets within benchmarking process enhances its usefulness and reliability. For instance, by using these findings in making the most appropriate choice of predictive monitoring technique for their

organizations, decision-makers can optimize their operational processes and enhance business performance.

This thesis is a major contribution to the area of BPM in terms of predictive process monitoring. Based on an empirical analysis of machine learning methods, whose main focus is on whether they are able to predict end results beforehand, this research aims at filling this crucial gap. Thus, such a study seeks to establish a benchmark that could be used for choosing the best machine learning method in relation to order management throughout its lifecycle. Consequently, this benchmark has an important role to play in optimizing operational processes by enabling managers to understand how different techniques function in terms of predictive capabilities. As such our aim is to enable companies choose the right predictive monitoring strategy so as they would be able to optimize their order management systems leading into efficiency and effectiveness within all company activities.

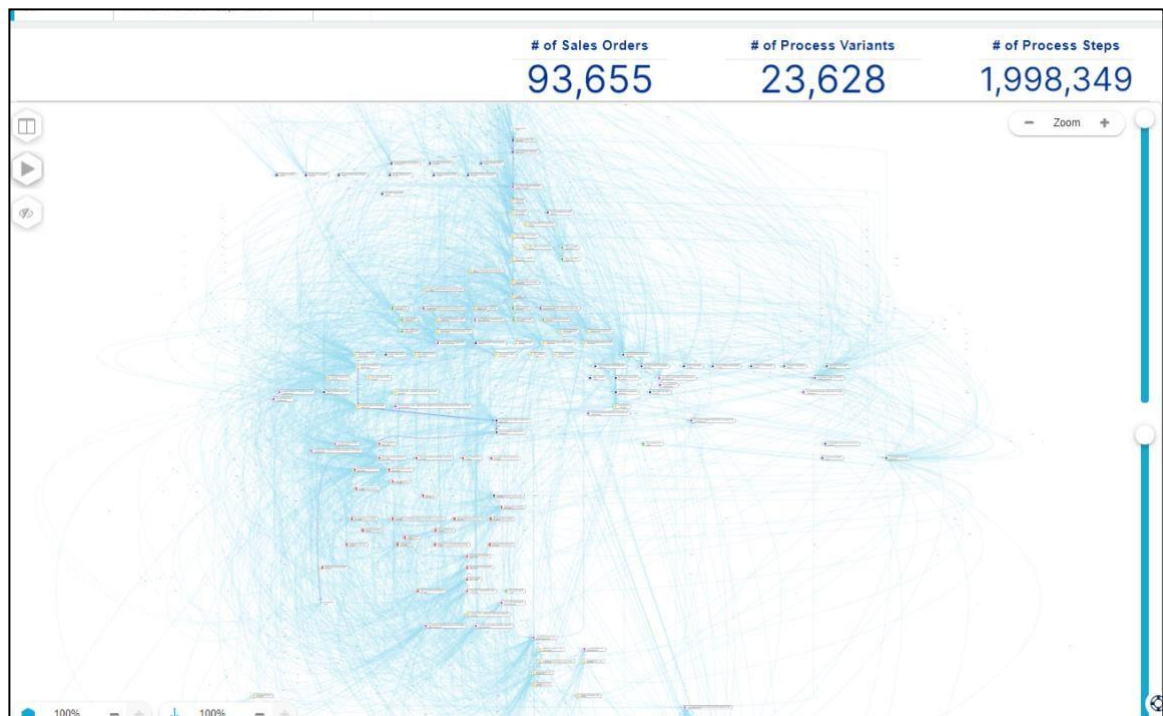
## **1.2 Research Problem**

Event logs are generated every second, and those can provide valuable information about a company's operations and the potential solutions. Nevertheless, it is impossible to manually navigate through such a large pool of data. In this case, predictive process mining – a method that employs machine learning algorithms appears to be the most promising solution that will allow disclosing latent knowledge hidden in these event logs. If businesses identify patterns and places for improvement, they can increase operational efficiency and solve the root causes of their problems.

Although predictive process mining is promising in many areas, there is a particularly interesting possibility of extending it to order management lifecycles due to its complexity. Order management lifecycles encompassing sales, payments, fulfillment,

finance and digital processes are critical in ensuring customer satisfaction. However, the complexity of this lifecycle poses challenges in optimizing operations effectively.

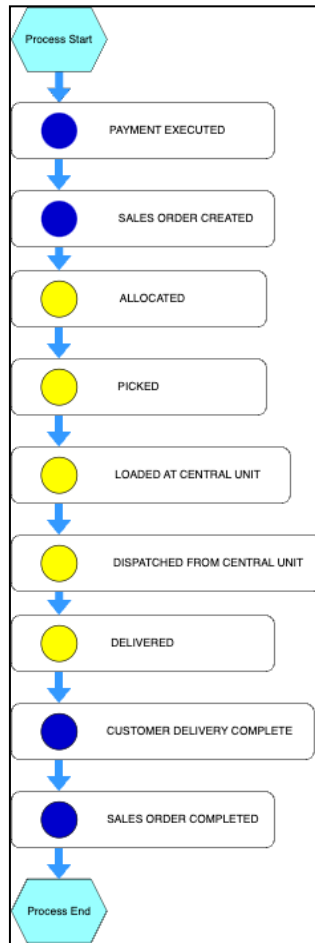
The research problem revolves around eliminating the need of manual efforts towards navigation in ordering Management life cycle's "spaghetti-like" real world process (Figure 1). Decision makers no longer must deal with the complication of this maze-like procedure itself. By using predictive process mining concept businesses can automate analysis of event logs making them actionable without involving any manual activities.



*Figure 1: An excerpt of a "spaghetti"-like process model.*

To illustrate, Figure 1 represents the existing "spaghetti-like" process model for order management lifecycle. The intricate diagram illustrates the interdependencies and complexity present in the real-world order life cycle processes. On the other hand, Figure 2 presents an ideal "lasagna-like" process model depicting a straightforward approach

that business would like to adopt into their systems. Through this research, we aim to relieve businesses from the burden of navigating the complex structure and intricacies of the order management lifecycle.



*Figure 2: A structured, or “lasagna”-like, process model*

### **1.3 Purpose of Research**

The goal of this thesis is to investigate the potential of predictive process mining in transforming organizational processes. In essence, it represents a sophisticated way through which organizations can gain deeper insights into their operational activities, predict potential issues and make informed choices to improve results. This research

therefore focuses on how predictive process mining techniques could be applied in order to gain valuable insights on key organizational processes that take place in areas such as sales, operations, customer support or employee workflows. The purpose here is thus to dissect these operations through the eyes of predictive process mining so as to unravel them and reveal some underlying patterns and trends.

The study tries to go beyond superficial observation or traditional process analysis methods by using predictive process mining to unearth concealed insights and predictions abilities. By examining historical event logs as well as process data, predictive process mining techniques have the potential to discover patterns, correlations and anomalies that may not be easily found out through manual analysis. Its main objective is therefore understanding organization-based processes better; thereby exposing its strengths and weaknesses.

Furthermore, this paper utilizes the benefits of prediction capabilities presented by process mining in anticipation of possible errors and inefficiencies within the organizational set up. Organizations can thus proactively trace opportunities for improvement and optimization before anything goes wrong hence improving operational excellence levels. Another advantage associated with predictive process mining is its ability to improve decision making by availing actionable insights based on predictions models and analyses.

The study would enable organizations to embrace predictable procedures' mine with the aim of making it a strategic tool used for fostering performance enhancement while achieving operational excellence (Onwubiko & Dupont, 2018). The purpose here will therefore be empirical research including data analysis and qualitative analysis aimed at providing knowledge and tools required by organizations to exploit fully value from predictive process mining in their operations so far. The ultimate goal was therefore

contributing towards development regarding what could happen next from an analytical point view if we use analytics thereby enhancing efficiency and effectiveness of an organization. In particular, the study is aimed at achieving the following goals:

### **1.3.1 Review on Predictive Process Mining**

The aim of the research in this regard is to investigate predictive process mining as a method within organizations. The study, therefore, seeks to establish a deeper understanding of crucial organizational processes that relate to diverse areas like sales, operations, customer support or employee workflows through such exploration. By leveraging on predictive process mining techniques, this research aims at discovering valuable insights into the complex dynamics of these processes by unveiling their underlying patterns, trends, and anomalies. This includes looking for possibly meaningful patterns and correlations from event logs or transaction data as well as other process-related information that could indicate future outcomes or behaviors within the organizational context.

Moreover, the study of predictive process mining entails using sophisticated data analytics methods and machine learning algorithms to obtain actionable insights from a large amount of process information. By using these techniques' ability to predict, this research seeks not only to describe but also to predict the future states of organizational processes with some level of accuracy. This investigation therefore aims at giving organizations tools and methodologies for pre-sensing future events, identifying potential risks or opportunities and making informed decisions that will enable them to achieve set objectives.

Similarly, predictive process mining is an analysis approach that identifies factors affecting process outcomes and performances. This involves looking at such things as

contextual variables, environmental considerations and organizational dynamics that influence how a process behaves or performs. The aim of this study is to unravel how these factors interact with each other in order to assist organizations by providing useful information on what needs to be done so as to optimize their processes, avoid exposure to risk among others. Therefore, through exploring predictive process mining in general terms it becomes apparent that this stage is essential for realizing the value locked in organizational data thus maximizing its use towards strategic decisions making and operational excellence.

### **1.3.2 Identification of Improvement Opportunities**

The study is geared towards the identification of potential areas for improvement and optimization in organizational processes. The research takes advantage of predictive process mining techniques to identify those sections amenable to streamlining leading to efficiency, effectiveness and overall performance improvements. Consequently, the investigation will analyze such data as event logs as well as performance indicators so as to find out bottlenecks, inefficiencies, and instances of poor performance.

Moreover, The study would consider standardization within diverse workflows. To achieve this aim, it would determine recurring trends or deviations within processes hence make suggestions on standardizing procedures and streamlining operations. This may comprise elimination of redundant steps, automation of manual tasks or redesigning workflows so as to increase efficiency while reducing cycle times.

In addition, the research will focus on scalability with a view to determining ways through which these processes can be expanded in order that they may be able to accommodate growth and business changes respectively. Among other things, this will involve assessing the ability of the current system to scale up thus suggesting any areas

that need improvement due to increased volume or complexity or geographic extension. By anticipating future needs and problems; the study aims at designing strategies for scaling businesses effectively.

Furthermore, cost-efficiency potentials will be investigated during the course of this research. Among other things like resource utilization inefficiencies in overhead expenses and process costs; such measures lie in where cost savings could be made without affecting quality or productivity adversely. In this regard resource optimization should play its role by restructuring vendor contracts renegotiation; implementing lean practices aiming at waste minimization adding value instead.

The main purpose of this research is therefore based on giving recommendations that can improve organization processes from several aspects such as being efficient, standardized, scalable and cost effective. The study is aimed at providing recommendations about opportunities for optimization and enhancement that could help firms attain their strategic objectives and consequently support sustainable growth options for businesses around them.

### **1.3.3 Analysis of Impact on Operational Excellence**

It is important to analyze how well our business operations perform in terms of operational excellence in relation to predictive process mining. We want, for example; see if there are any positive effects relating the use of predictive process mining techniques on customers' satisfaction as well as overall employee engagement.

Organizations aiming at sustainable growth and attaining competitive advantage must pursue operational excellence as their major objective. Process optimization leading to better operational workflows may be achieved via employing Predictive Process Mining which will help do away with redundant steps throughout the concerned



processes. Consequently, the extent by which an organization can put into practice “predictive” predictions depends upon the degree of knowledge about both types of organization’s structures as well as attributes thereof through empirical examination.

Moreover ,the study examines whether there exists a relationship between predictive process mining outcome and the key performance indicators (KPIs) which represent operational excellence. Examples of such KPIs include; process efficiency, resource allocation trends, cycle time, error frequencies and overall output. The research attempts to draw empirical evidence from these metrics in order to determine the extent to which predictive process mining can drive operational improvement.

On the other hand, this analysis evaluates how knowledge gained from predictive process mining concerning customer satisfaction. In view of revealing patterns, trends as well as potential problems within operational processes companies will be better prepared to meet their customer’s needs early enough while they also minimize risks posed by them thus ensuring improved general consumer experience. Qualitative and quantitative assessments will be used in order to establish a link between the outcomes of predictive process mining and measures for customer satisfaction.

Furthermore, the study investigates how predictive process mining affects employee engagement and satisfaction. Organizations that utilize insight derived from process-mining analyses enable their staff members to make informed decisions on what they should do hence they should use those opportunities for innovation and participate in initiatives leading towards process improvement. Predictive Process Mining is also essential as it helps organizations achieve better job-related results among their employees, enhancing their morale thereby resulting in a higher level of satisfaction.

This part of our research analysis aims at providing insights that are valuable for enterprises about the real advantages associated with adoption of these technologies.

Therefore, through quantifying its effects on operational performance as well as employee engagement or customers' satisfaction this study intends to show how predictable process mining has the potential of making companies successful and competitive in the business sector.

#### **1.3.4 Guidance Implementation**

The research aims to provide a clear and practical guide for organizations interested in adopting predictive process mining. This will be made possible through empirical analysis, data mining, and qualitative studies that will offer concrete strategies capable of facilitating the successful implementation of prediction process mining techniques.

These guidelines are meant to cater to the specific issues and problems encountered by companies in various sectors. Moreover, the results from this study will provide businesses with insights on how predictive process mining can be deployed effectively to facilitate optimization, anticipation of faulty orders and help drive CI activities.

Additionally, it is important that these instructions also address broader organizational objectives as well as aligning them with predictive process mining efforts. All this will require resource allocation, consultation of stakeholders as well as change management tools for a smooth implementation. Finally, the research will highlight best practices and real life cases.

#### **1.4 Significance of the Study**

This theory carries important ramifications for the research community and business:

### **1.4.1 Academic Contribution**

The study contributes to the expanding corpus of information on predictive process mining and its uses in organisational contexts. The study increases knowledge of how predictive analytics can lead to operational excellence and broadens theoretical frameworks by investigating the integration of predictive process mining techniques with conventional process mining methodology.

### **1.4.2 Real World Impact**

The practical ramifications of the study's findings extend to organisations operating in diverse sectors. Organisations can proactively detect and address difficulties within their order management processes, improving consumer satisfaction, operational efficiency, and cost-effectiveness. This can be achieved by showcasing the predictive process mining technology's ability to predict defective orders in advance. The practical ramifications of the study's findings extend to organisations operating in diverse sectors. Organisations can proactively detect and address difficulties within their order management processes, improving consumer retention, operational efficiency, and cost-effectiveness. This can be achieved by showcasing the predictive process mining technology's ability to predict imperfect orders in advance.

### **1.4.3 Business Strategy**

Organisations can use the knowledge gained from this research to guide their strategic decision-making. Businesses can gain a competitive edge by optimising processes, improving overall performance, and anticipating possible issues by utilising predictive process mining skills.

#### **1.4.4 Operational Excellence**

The research emphasizes the importance of using data-driven approaches to drive operational excellence. Another area where predictive process mining can come into play is by streamlining operations, standardizing processes, and scaling businesses effectively, ultimately leading to improved efficiency and productivity.

#### **1.4.5 Customer Happiness Index**

This research's importance also applies to customer experience by enhancing business Customer Happiness Index. Process mining is a predictive tool that could help organizations identify defective orders ahead of time so that they can enhance customer satisfaction and loyalty. One of the main drivers of customer happiness is the reduction in order errors and delays. The ability to detect risks inherent in the order lifecycle leads businesses to act proactively thus reducing fulfillment errors and delays.

Additionally, organizations may better their overall service quality through preemptively examining areas for improvement within the order management process. This includes streamlining, optimizing workflows, and ensuring that all orders are delivered consistently and on time. Consequently, customers experience hassle-free transactions, faster deliveries, as well as greater dependability making them more satisfied than before.

Subsequently, this will not only increase short-term consumer satisfaction but will also promote their loyalty and retention through being able to predict imperfect orders and take proactive measures towards rectifying them. This is why companies continually delivering high-quality products and services without faults or delays have higher chances of keeping their customers loyal. In other words, an organization can effectively raise its Customer Happiness Index by focusing on anticipating flawed orders; hence leading directly to sustainable business growth and success.

In summary, this thesis has delved into how predictive process mining can improve overall management decisions made by organizations concerning processes used in their daily activities such as routine activities based on process historical data analysis patterns components predictions trends from log files methodology paradigm model transformation language method repository logs users records instance class metafor. By applying advanced analytical techniques it was possible to gain meaningful insights from complex event data helping clients with experience enhancement operating excellence anticipation of incomplete orders.

Bridging the gap between theoretical assumptions about predictive process mining applications with practical realities constitute one major contribution of this study. In relation to existing literature on this topic, this paper provides real life examples. Additionally, it was also able to offer a comprehensive understanding of how such techniques could be applied in the real life situations. Another significance of this study is that it gives insights on how predictive process mining can be used to bridge the gap between theory and practice in this field.

The thesis has critically examined perspectives and methodologies towards applying predictive process mining for achieving specific business objectives. For example, results from this research highlight the importance of finding opportunities for improvement within organizational processes with an emphasis on simplification, standardization, scalability, and cost effectiveness. By providing areas where optimizations and improvements may take place, this research assists organizations to become more efficient in their operations.

Also, this study provides valuable information for businesses who would like to leverage predictive analytics as a way of enhancing customer experience compared to their competitors in today's competitive business environment. To achieve operational

excellence and customer satisfaction companies are advised to apply predictions given by these models.

This thesis is a contribution to the existing body of research on predictive process mining by providing strategies, methods, and techniques for embedding predictive analytics into organizational practice. It combines theoretical robustness with practical importance and hence offers an overall framework that shows how organizations can leverage data-based insights towards their strategic objectives and competitive advantage. These findings are useful in driving innovation, efficiency, and success in the contemporary business world where various companies are opting for digital transformation as well as data-driven decision-making.

For these reasons, this study is not only of academic interest but also has practical implications that help to address the complex nature of modern business operations. Companies can outdo their competitors in today's volatile market through anticipation and reduction of operational risks, process optimization, and satisfaction of customers using predictive process mining. This shows how transformative using predictive analytics can be towards organizational success as well as promoting a culture for continuous improvement. The recommendations made are therefore useful in aiding decision makers when integrating predictive process mining into their organizations so as to navigate the complexities surrounding decision making and process optimization. Therefore, what counts most in this research is its ability to enable organizations to explore every bit of available data for purposes such as driving innovation, creating efficiencies and ensuring that there is sustainable development even in an increasingly data-focused world.

## **1.5 Research Purpose and Questions**

In today's business landscape where data is king, the use of predictive process mining techniques can be an effective way to unravel organizational processes complexities. By applying predictive analytics, organizations can have a deeper understanding of their operations by identifying potential challenges as well as making informed decisions to optimise performance and mitigate risks. This study thus seeks to explore the multifaceted applications of predictive process mining within the order management domain with a critical eye on its likely revolutionizing effects on traditional approaches towards order fulfillment. The research aims at exploring the effectiveness of predictive process mining in enhancing operational efficiency, improving customer satisfaction and fostering a culture of continuous improvement as guided by rigorous empirical investigation and analysis. Through answering key research questions and objectives, this study seeks to offer insights that are valuable and practical for businesses seeking to employ predictive process mining in navigating through complexities associated with modern day order management.

The research aims to examine how predictive process mining techniques can be used in order lifecycle management with the aim of detecting defects, streamlining processes and improving organizational performance. The investigation is guided by a number of specific research questions which are geared towards understanding various aspects of predictive process mining and its implications on order management.

One main objective of this research is to evaluate the efficiency of predictive process mining methods in identifying patterns and predicting outcomes throughout the order lifecycle management. The research therefore uses data from event logs and activities done at different stages in an attempt to determine whether or not predictive models can accurately anticipate imperfections, inefficiencies as well as other challenges

that may occur during order processing and fulfillment. This will involve testing different machine learning algorithms against each other based on how well they capture and analyze orders related information for actionable insights.

Another important goal is to investigate what factors affect the accuracy or reliability of predictive models when forecasting flaws or bottlenecks within order processes. In particular it seeks to find out how features such as attributes of an order, sequence followed by activities involved in its execution and timing influence performance in terms of predicting these errors. To achieve this, we examine relationships between these factors vis-à-vis occurrence of process deviations from standard ways so that we can identify key drivers and predictants for them which would then lead into development more robust predictors.

Additionally, we aim to compare various machine learning algorithms with regards their predictability for order life cycle management. By doing strengths weaknesses analysis logistic regression decision trees random forests XGBoost among others; it enables him know which technique(s) should be adopted when trying detect defects during this stage. Such a comparative assessment will be instrumental in providing insights about applicability different types modelling methods vis-à-vis real world scenarios where orders are managed.

Moreover, another objective involves examining what happens organizationally when predictive process mining solutions are implemented? Therefore; according this investigation customer satisfaction levels associated with order processing times completion rates should also measured alongside other key performance indicators like accuracy levels etc., so that we can establish if there is need for any improvements from current situation or not.



Finally, besides offering general suggestions on how predictive process mining can be integrated into existing systems of managing orders, the researcher aims at giving practical advice to business enterprises on this matter. Through synthesis findings out an analysis carried and drawing experiences gained during predictive analytics as well process mining; recommendations will then be made based on what has worked in different places hence organizations which want optimize their order lifetime management processes through use of these two techniques should consider them. Hence, such proposals are intended help firms become more efficient by reducing mistakes through strategic utilization of predictive analytics methods. The research questions that are addressed in a thesis on predictive process mining and order lifecycle management includes:

1. How efficient are methods for pattern recognition and outcome prediction in order lifecycle management using predictive process mining?
2. What are the main elements affecting the predictability and accuracy of predictive models in terms of identifying imperfection or inefficiencies in order processing?
3. How do various machine learning algorithms stack up in terms of how well they predict order lifecycle management outcomes?
4. How will putting predictive process mining technologies into practice affect customer satisfaction and organisational performance?
5. In order to increase productivity and lower mistake rates, how can companies successfully incorporate predictive process mining into their current order management systems?

## CHAPTER II

### REVIEW OF LITERATURE

The fusion of process mining and predictive process monitoring has emerged as a critical area of research in business process management (BPM) in recent years. Process mining, as presented by van der Aalst (2011), is an approach that utilizes event logs to automatically find, evaluate and enhance business processes. Information system generated event logs are very detailed and offer valuable insights into process execution.

Predictive process monitoring takes this a step further with its use of machine learning techniques to predict behavior, performance and outcomes of business processes. Organizations can take a more proactive approach by identifying potential issues before they arise, optimizing their efficiency and decision-making.

By integrating predictive process monitoring with process mining, organizations can not only streamline their operations but also improve resource allocation and customer satisfaction. By analyzing historical event data with advanced analytics, businesses can get an in-depth understanding of how their processes perform. This will enable them to identify bottlenecks in real-time as well as predict future behavior.

This research area has gained significant attention from academia and industry because of its ability to drive operational excellence through data-driven decision-making. However, despite the growing number of papers on this subject matter, there still remains some challenges that need to be addressed.

In this review, our goal is to present a thorough summary of the state of research in process mining and predictive process monitoring, emphasising important contributions, approaches, and directions for further investigation. By combining knowledge from the body of current literature, we seek to contribute to a deeper understanding of this dynamic and evolving field.

## **2.1 Theoretical Framework**

In the review of literature, the theoretical framework encompasses several important ideas related to order life cycle management, predictive process monitoring, process mining, and the incorporation of machine learning (ML) approaches:

### **2.1.1 Order Life Cycle Management**

The order life cycle consists of multiple stages such as order placement, payment processing fulfillment etc (Kang et al., 2015). This framework highlights the importance of managing each stage efficiently in order to meet customer satisfaction deadlines. Comprehending the intricacies of the order life cycle is important in order to detect any obstructions and enhance operational procedures.

### **2.1.2 Process Mining and Event Logs**

Process mining is an analytical technique that extracts insightful information about workflow and processes (van der Aalst, 2011). Event logs record every little activity within each stage of the order life cycle; allowing users to spot patterns, deviations, efficiencies and much more. With process mining techniques businesses are able to identify deeper insights into their order management process which may lead to new opportunities for growth.

### **2.1.3 Predictive Process Monitoring**

Predictive process monitoring extends process mining by using predictive analytics to forecast future process behavior and outcomes (Maggi et al., 2014). This framework enables organizations to anticipate potential issues within the order life cycle, such as delays in payment processing or fulfillment. By applying machine learning

algorithms to historical event data, predictive process monitoring empowers businesses to take proactive measures to mitigate risks and optimize order management processes.

#### **2.1.4 Combining and Integrating Predictive Process Monitoring and Process Mining**

Predictive process monitoring and process mining can be integrated to improve the efficacy and accuracy of prediction models (van der Aalst et al., 2012). Organisations may build more reliable predictive models that can recognise trends and forecast future events in the order management process by fusing insights from process mining with predictive analytics.

#### **2.1.5 The Role of Machine Learning (ML) in Predictive Process Monitoring**

By enabling the creation of prediction models based on past event data, machine learning techniques are essential to predictive process monitoring (Tax et al., 2017). Event logs may be analysed by ML algorithms to spot trends and forecast how a process will behave in the future. Proactive decision-making in order life cycle management is made easier when machine learning techniques are used with predictive process monitoring. This improves the accuracy and dependability of predictive models.

When combined with predictive process monitoring, integrating ML helps increase accuracy and reliability in our models so we can take action before problems even arise during the order life cycle management.

By using these theoretical frameworks together, researchers and workers alike will be able to save their sanity with a smoother approach to order life cycle management. We're talking about leveraging these two methods with machine learning techniques in an effort to optimize operational processes for maximum customer satisfaction.

## 2.2 Empirical Studies

Several empirical studies have contributed in the advancement of predictive process monitoring and process mining. As well as shedding light on their effectiveness, Studies have also shown us their capabilities, along with other potential impacts they could bring about when applied towards organizational performance.

Notable studies by Maggi et al (2014), which focuses on predicting remaining cycle time in business processes using predictive process monitoring techniques. The study evaluated different machine learning algorithms performance when it came to predicting how much longer a given process instance would take to complete. The researchers proved that it is possible to predict remaining cycle time with accuracy by using previous event logs and time-series data. This allows organisations to manage resource allocation and process execution proactively.

Another notable study by Tax et al (2017) had researchers exploring predictive process monitoring with structured and unstructured data. Their focus was on how well integrating these two types of data sources could improve predictive capabilities. For example, combining structured event data with information lifted from the process documentation and logs which led to improved accuracy and robustness in their predictions. This work paved the path for more thorough and perceptive assessments by highlighting the need of using both structured and unstructured data in predictive process monitoring systems.

Additionally, van der Werf et al (2016)'s study on process mining for inter-organizational workflows, which employed a bag-of-tasks approach during analysis. The goal was to try and understand complex interconnected process structures that spanned multiple organizational units at once. By breaking things down into simple individual tasks, they were able to identify some common patterns across different workflows. This

allowed them to get a better look at the challenges and opportunities companies have when trying out inter-organizational process mining techniques.

Several empirical investigations, like this one, have shed light on the ramifications and real-world uses of process mining and predictive process monitoring methods. Researchers have increased our knowledge of predictive models, algorithms, and techniques' capabilities and limits by experimentally testing them. This has paved the way for better informed decision-making and actionable insights in real-world business scenarios.

### **2.3 Conceptual Studies**

In addition to empirical investigations, several conceptual studies have contributed to shaping the theoretical foundations and conceptual frameworks underlying predictive process monitoring and process mining. These studies have explored key concepts, theoretical constructs, and methodological approaches, laying the groundwork for further research and practical applications in the field.

One seminal study by van der Aalst et al (2011) provided a comprehensive overview of process mining. It focused on the discovery, conformance, and enhancement of business processes. The authors described the core principles of process mining, with an emphasis on extracting valuable insights from event logs and improving organizational processes. This conceptual framework then became a basic resource for other researchers interested in this topic.

Another noteworthy study by Di Francescomarino et al. (2019) delved into predictive monitoring of business processes using various techniques for different prediction tasks. The study evaluated metrics, considered algorithmic considerations

while also discussing many other aspects that need consideration when trying to implement predictive monitoring in business process management.

Leontjeva et al (2018) conducted a study on efficient sequence representations which are often used to model sequences in computer science. The study proposed novel methods for representing data which can be leveraged later during modeling or even during analysis.

Among other important contributions to the advancement of our theoretical knowledge of process mining and predictive process monitoring have been these conceptual investigations. These studies have furnished a strong basis for empirical investigation, pragmatic implementations, and more theoretical advancements in the domain by clarifying fundamental ideas, frameworks, and techniques.

## **2.4 Current Position and Goals**

In order life management there is great potential for enhancing efficiency through the adoption of predictive process mining. A good understanding of the existing situation and well-defined objectives are essential to driving predictive process mining adoption in order life management.

In many different sectors, order life cycle analysis and optimisation can be achieved with the help of predictive process mining. Predictive analytics methods are being used by businesses more and more to estimate order fulfilment timeframes, spot any bottlenecks, and expedite operational procedures. Research like Schuh et al (2020b)'s demonstrate how predictive process mining can be used to enhance production systems and show how important it is for order fulfilment process optimisation. Furthermore, research by Kang et al (2015) shows how predictive process mining may be applied in

healthcare settings to optimise order processing and delivery, highlighting its adaptability to a variety of contexts.

Predictive process mining adoption in order life management encompass several key objectives. First and foremost, companies seek to increase order fulfilment efficiency by precisely forecasting order completion dates and spotting possible process bottlenecks. Secondly, the emphasis is on improving customer satisfaction through the prompt delivery of orders and the reduction of mistakes throughout the order fulfilment process. Thirdly, by optimising order management procedures based on predictive insights, organisations aim to maximise resource utilisation and save operating costs. The main objective is to use predictive process mining to make order life management a more customer-focused, nimble, and efficient procedure.

## **2.5 Predictive Process Monitoring: Theory and Practice**

The adoption of predictive process monitoring involves a synthesis of theoretical frameworks and practical implementations to effectively leverage predictive analytics in process management.

Numerous studies have delved into the theoretical foundations and practical applications of predictive process monitoring. Tax et al (2017) explore the theory behind predictive process monitoring, particularly its integration with structured and unstructured data. The study also discusses the impact of different modeling techniques in real-world situations.

Additionally, Burattin et al (2019) provide an extensive analysis of predictive process monitoring methods — evaluating their suitability in a range of contexts — as well as the theoretical underpinnings for different algorithms that predict behavior and outcomes.



In practice, organizations are increasingly embracing predictive process monitoring to enhance operational efficiency and decision-making. Maggi et al (2014)'s empirical study demonstrates the effectiveness of predictive models when forecasting remaining cycle time and optimizing performance in business process management.

It is imperative that businesses understand both theory and practice to fully realize predictive analytics' potential for continuous improvement and innovation in managing processes.

## **2.6 A Predictive Approach for Order Life Cycle Management Performance Improvement**

In recent years, organizations have been turning to predictive process mining to improve order life cycle management's performance. The approach uses advanced analytical techniques to anticipate bottlenecks, inefficiencies, and errors in the order life cycle so that organizations can address them before they become catastrophic problems.

Several studies have investigated this application area. Smirnov et al (2019) research investigated how these techniques can be applied specifically to order-to-cash processes — successfully using models to forecast completion times, fulfillment accuracy, among other crucial metrics like optimization of order management processes.

Similarly, Leontjeva et al. (2018), proposed a new machine learning algorithm that predicts future order statuses based on historical data; it seeks out deviations from expected flows.

Finally, Marrella et al. (2019) Customer-centered research emphasized the importance of understanding customer journey and preferences to optimize management processes and enhance satisfaction.

Overall, the predictive process mining methodology has a significant potential for enhancing order life cycle management performance by empowering businesses to foresee and mitigate operational challenges in real-time.

## **2.7 Mathematical formalism of the existing literature**

The lack of research on prediction process mining meant for optimization of order completion is a vital point that needs more study to realize its full potential. Although predictive process mining has been applied in many sectors to improve operational efficiency and decision-making, it has not been used in the management of orders as required. As one of the critical areas within businesses across different industries, customer satisfaction can only be achieved if orders are fulfilled optimally. This involves order processing, inventory management, logistics among others which foster operational success. Nonetheless, there are complexities associated with managing orders coupled with ever changing consumer demands and market situations thus calling for special research considerations. Therefore, this gap indicates a necessity for custom predictive process mining models and techniques that cater for intricacies involved during order completion so as to enhance effectiveness, cost efficiency and customer service delivery in general. The gap in the existing research is explained in more details below:

### **2.7.1. Order life cycle management and predictive process monitoring are not integrated.**

In the research world, a lot has been left untouched in so far as predictive process monitoring (PPM) is concerned. Order life cycle management (OLCM) should be integrated into predictive process monitoring (PPM). This is a significant gap by itself because all these areas are crucial parts of business process management which is not

reflected in the current literature where there is no study that looks at how they intersect each other and what can be done to make sure that orders are managed efficiently through the application of PPM techniques.

Process mining and order management have always had separate histories thus leading to this gap. In traditional process mining event logs are analysed while looking for process models which will help organizations become efficient operationally besides improving decision making capabilities during different stages of various organizational processes but on the flip side, order life cycle management deals with fulfillment or realization stages involved in meeting customer needs from placing an order until it delivered.

Predictive Process Monitoring (PPM) and Order Life Cycle Management (OLCM) have never been considered as one thing hence limiting their joint exploration among researchers. This has resulted into lack of knowledge on how best predictive process monitoring may be used together with order life cycle management so as to improve efficiency levels within organizations more especially those which deal directly with customers.

Most existing literature tends to take a generic perspective when discussing about applications related to business processes without putting into consideration specific requirements needed for proper functioning of systems handling orders. Thereby failing to recognize peculiarities attached with these types of activities such as lead time variations, product availability differentiation based on location among other things like customer preferences which could impact significantly on service delivery timelines if not handled well.

### **2.7.2. Limited Research on Order Fulfilment Optimisation Using Predictive Process Mining**

In spite of the growing popularity in predictive process mining, little has been done to establish whether it can optimize order fulfillment processes. Various industries such as manufacturing, finance and healthcare have widely researched into this area except for order management which still remains less studied.

Order realization optimization includes different stages like stock control, order processing, logistics among others. All these steps offer unique opportunities for improvement hence making them good candidates for applying predictive process mining techniques.

However, most existing works touch on predictive process mining generally thus focusing more on aspects like performance analysis, conformance checking and process discovery instead of delving into complexities associated with order fulfillment optimization. Consequently there is no enough study that takes care of specific requirements of predictive analytics in relation to order management processes.

One possible explanation for such a gap could be attributed to intricate nature of order completion systems coupled with wide range factors affecting their efficiency or effectiveness. If we have to know so much about what will happen next then there is a need to have all data events required by Predictive Process Mining (PPM). These are not easily available within reach during order management due to data silos, inconsistent formats among other privacy issues.

Additionally real-time decision making during optimization may call for dynamic adjustments which further complicates matters where models are built using PPMs since they do not adequately address temporal aspects associated with it in this context. In

other words the current studies fail consider these two features when dealing with orders hence leading us nowhere nearer towards PPMs used here.

Addressing this requires interdisciplinary research involving business process management (BPM), operations management(Opsmgt) and Predictive Analytics(PA). Future efforts should focus more on developing customized predictive process mining models suitable for addressing challenges brought about by optimizing order completions.

Another thing researchers can do is looking at innovative ways through which advanced analytics coupled with machine learning algorithms can be deployed in trying to predict as well optimize different types of real-time order fulfillment processes. There is still much more that needs to be done if we want things work perfectly well for us in terms of reducing costs and customer satisfaction within various sectors relying heavily on efficient systems for managing orders.

### **2.7.3. Little focus on imperfect order prediction in real time**

One of the biggest problems with studies in order life cycle management is that they do not pay enough attention to predicting imperfect orders in real time. Predictive analytics has been used largely throughout several sectors to make projections and decisions, but it has mainly ignored its use for recognizing faults during order processing and fulfillment.

Delays, errors or deviations from anticipated results can occur at any point within a traditional order management system. Such flaws may result in customer dissatisfaction as well as higher operational costs and reputational risks for enterprises. However, current researches are mostly retrospective analyses or batch processes applied on past

data to create predictions. Because of this fact, there is minimal exploration on models which predict continuously while identifying them at once.

If businesses are to anticipate imperfect orders immediately, then advanced predictive process monitoring methods capable of analyzing streaming information from order management systems in a continuous manner must be developed. This calls for an integration between algorithms used for predictive analysis and frameworks employed in processing real-time information besides event-triggered designs. Such systems can predict potential imperfections based on patterns as well as abnormalities identified within streams of current orders by making use of machine learning models trained using historical records about orders.

Further still, proactive decision-making mechanisms should be put into place so that they automatically activate corrective actions or send alerts when there are variations from the expected outcomes of orders in real time. Dynamic rerouting allocations extra resources while issues are being escalated among others which may need immediate intervention by relevant staffs for rectification purposes forms part of such like structures. Therefore ability to intervene immediately could greatly reduce inefficiency caused by defects during order completion process since customers will not have their needs met promptly hence affecting satisfaction levels

What needs to happen next involves interdisciplinary efforts aimed at tackling this area's neglect on predictive modeling approaches that operate continuously detecting them live should be pursued thus combining expertise from different fields such as; predictive analytics ,real-time data processing among others coupled with knowledge related to order management system. By coming up with strong forecasts and implementing systems which keep track of events as they happen in real time businesses can become more flexible when dealing with imperfections associated with orders

thereby improving overall performance within operation while at same time meeting customer expectations.

#### **2.7.4. Insufficient Research on Machine Learning Methods for Order Processing**

There are still many ways in which the current research could be extended because of the limited exploration into machine learning for order management. Machine learning has proven to be very useful in many areas but there have not been enough studies done on how it can be specifically applied to order processing systems. Most predictive process monitoring and order life cycle management papers only touch on basic statistical methods or simple machine learning techniques like decision trees and logistic regression without going into more complex ones.

Order Management Systems are complex and involve a lot of different types of data sources, which is another reason why people have not been able to apply advanced machine learning techniques here. Data is generated from things like customer service, inventory management, shipping etc., all these processes being interconnected with each other during an order fulfillment cycle. With this in mind one might use various ML algorithms capable of handling such diverse information streams by analyzing them together thus allowing organizations to make data driven decisions while optimizing their Order Management Processes.

Traditional statistics may not do a good job at dealing with non-linear relationships among variables or patterns hidden deep within large datasets as well as high-dimensional data itself – which is where advanced machine-learning algorithms come into play. For instance; deep learning has been successful in many domains due its ability learn multiple levels representations from input while ensemble methods often

improve generalization performance by combining several models predictions together. Reinforcement learning, on the other hand, has proven to be very effective in dealing with situations where an agent needs to learn from interacting with its environment over time and choose actions that maximize some cumulative reward.

However applying these advanced machine learning techniques for order management requires consideration around a number of issues such as; data preprocessing feature engineering model selection performance evaluation among others. Again integrating them into existing systems will raise challenges related to scalability interpretability real time processing.

In conclusion, it would be necessary for researchers from different fields including computer science, operations management and supply chain management to work together so as to address this gap in knowledge about machine learning for order management. To further clarify studies ought to develop custom built ML models which can accurately analyze Order Fulfillment process data while predicting outcomes like time taken or customer satisfaction levels then providing actionable insights that can be used by decision makers. Additionally efforts should be made to overcome technical difficulties associated with implementing such solutions within real world environments thus enhancing efficiency through customer centricity during fulfillment of orders.

#### **2.7.5. Narrow comprehension of prognostic process tracking regarding the customer**

When it comes to order management, one major loophole in this field of study is its failure to take on a customer-centric approach toward predictive process monitoring. In order fulfillment, efficiency and operational optimization are important but so too is



knowing what exactly the customers expect from the organization; meeting those expectations will ensure their satisfaction thereby making them loyal customers for life.

Customer centric predictive process monitoring entails integrating the knowledge obtained about clients into predictive analytics systems which in turn informs how best orders should be managed based on individual customer needs or preferences. This method goes beyond traditional methods where processes are optimized without considering such things as client behavior history records among others that may have an impact on service delivery standards.

However, many studies tend to ignore this particular aspect of predictability within the domain of managing orders. They mostly focus on efficiencies within internal processes and operational metrics without examining their effect on satisfaction levels experienced by consumers themselves as a result of dealing with businesses during their transactional interactions. Thus there exists no understanding about what can be done through prediction monitoring systems towards making order processing more centered around the needs or wants shown by different buyers.

To fill this gap requires shifting attention towards embedding client-centered views into forecast modeling frameworks utilized for monitoring processes over time. What researchers should therefore look into is how data concerning clients could find its way easily into models used for doing predictions at different stages throughout an entire ordering cycle; such information might include things like analyzing feedback given by a purchaser after using certain products among other touch points they had while making various purchases.

Moreover empirical researches need to be conducted so as ascertain whether real world scenarios would support effectiveness associated with these strategies when adopted during management of orders placed within businesses where everything

revolves around consumers' needs being met satisfactorily every time they transact with them . There is hence need for carrying out case analysis which will help show why customercentricity should always be considered during development of frameworks designed to enhance different types of forecasting.

## CHAPTER III: METHODOLOGY

### **3.1 Overview of the Research Problem**

Order life cycle constitutes different stages like order placement, processing, fulfillment, and delivery which offers unique opportunities for optimization as well as posing distinct challenges. In traditional approaches to order management or UAM, there is a tendency to reactively dealing with issues that comes up. Hence, by using predictive analytics organizations can have proactive anticipation and tackling such problems like inefficiencies, bottle necking and errors in the system.

Predictive process mining offers a data-driven approach to understanding and optimizing complex business processes, such as order management. Predictive models can detect patterns that can be useful in future predictions through analysis of historical event logs and process data. This will enable organizations to forecast outcomes in the future taking note of potential challenges thus putting preventive measures which will bring about smoother operation on orders.

Research that starts from a passive position on an issue is what this study seeks to change. Thus, shifting from reactive to proactive paradigm in managing order lifecycle based on predictive insights driven decision-making and process optimization has been pursued by this research. It focuses more on turning order management into an analytical model where energy consumption can be analyzed for efficiency purposes using predictive process mining techniques thus reducing operational costs and increasing customer satisfaction leading better business performance.

This study's goals are:

1. To find out whether predictive models could be used effectively for anticipating challenges within the order life cycle; hence determine if

predictive process mining could identify possible bottlenecks or inefficiency or errors during various phases of the procedure.

2. The intention is therefore to develop models that predict future specific results related to purchasing transactions such as time taken for execution, fulfillment accuracy rate, customer satisfaction score etc., using data analytics. These models are meant to inform user decisions regarding forthcoming events in respect of ordering activities.
3. Examine how key performance indicators around order lifecycle management are influenced by predictive process mining; this may be for example, how well the predictive models can increase the efficiency of the operations or reduce costs incurred in the process. Additionally, it will critically look at all aspects of managing orders that can be improved.
4. Consequently, recommendations and best practices concerning putting into use predictive process mining application in practical situations should be provided by this research. These include data collection and preprocessing; model selection and training; deployment strategies; and monitoring and refining processes with regards to different factors such as data quality issues. By offering guidelines, this work is meant to ease adoption of predictive process mining into organizations.

In line with this, our main aim through this research is making substantial contributions within the field of order life cycle management through leveraging on Predictive Process Mining techniques. Our objective is thus to improve forecasts accuracy while still enhancing efficiency of order management processes hence benefiting both consumers and businesses. This study presents innovative solutions

necessary for predicting some potential problems involved in filling an order like forecasting possible challenges faced by logistics companies when trying to keep up with sales growth rates while optimizing their operational response abilities.

The goal of conducting this research is therefore to make significant contributions to order lifecycle management using predictive process mining techniques. We are aiming at increasing accuracy in forecasting as well as improving overall efficiency in processing orders to benefit both customers and firms concerned. On that note, we are trying out new ideas which will steer organizations toward being proactive rather than reactive in terms of implementing a more integrated fulfillment function capable of anticipating potential problems thereby enabling them manage operations proactively.

This research seeks to develop predictive models that could accurately forecast various elements related to the order life cycle including but not limited to time taken for execution, fulfillment accuracy rate, customer satisfaction score etc., using data analytics approach, hence leading better business performance from a customer perspective. Through employing Predictive Analytics' power, establishments can have insights on what will happen next thus choosing right actions and pre-emptively ensuring that they are optimizing their own operations.

In addition, we are striving towards showcasing the values of these techniques through assessing impacts on key performance indicators such as order life cycle management. From reducing costs and improving operational efficiency to overall order management process optimization, it is possible for predictive process mining to cause significant improvement in organizational performance.

Eventually, this research seeks to provide pragmatic insights and recommendations that can enable businesses to implement predictive process mining systems in real-world settings successfully. Our objective is to guide enterprises on how

they can effectively obtain information, build models, deploy solutions, or monitor their progress with time that will make predictive analytics a way of simplifying their supply chain activities and ensuring better customer experience. We expect these initiatives to make firms more efficient and adaptable while still focusing on the needs of consumers in the current business world where competition for market dominance is intense.

### **3.2 Operationalization of Theoretical Constructs**

To provide a clear understanding of order life cycle management and predictive process monitoring application in the context of this study, we operationalize the theoretical constructs of order life cycle management (OLCM) and predictive process monitoring (PPM) in this section.

#### **3.2.1 Operationalization of Predictive Process Monitoring**

Predictive Process Monitoring (PPM), as the name implies, uses machine learning algorithms to predict the behaviour, performance and outcomes in a business process based on historical event logs. To operationalize PPM, we define key concepts followed by outlining the process implementation.

With PPM, the goal is to know for case the order life cycle during the active business process instances by using models derived from historical event logs. The concept includes being able to predict the outcome of the business and predicting the future path of the process already in the running state.

The operationalization of PPM involves several steps, Data preprocessing, feature selection, model training, evaluation etc. The dataset is an event logs consist of event and its timestamp as two important features. we preprocess the data in a way that we can handle missing values, duplicates, encode categorical variables and scale numerical

features. After that we split the event data into training and testing sets, so we don't get biased results. Multiple machine learning models are then trained on the preprocessed data and at last multiple model performance is compared to select the best fitting model to predict the future events on a real unseen dataset.

### **3.2.2 Operationalization of Order Life Cycle Management**

Order life cycle management (OLCM) is the process of managing the several phases of order processing, starting from order initiation, and ending with order fulfilment. OLCM must be operationalized by identifying its elements and detailing the procedures for putting it into practice.

OLCM has activities such as order creation, processing, payment, fulfillment, and delivery. It aims to ensure timely processing of orders while also meeting customer expectations and organizational objectives.

The process begins with the synthesis of event log data that mimics order life cycle processes. Key stages in the order life cycle are recognized and used to design synthetic logs that capture activities in a sequence complete with timestamps. Predictive process monitoring models then anticipate order outcomes using these logs, as well as identify areas to improve order management.

### **3.2.3 Integration of Predictive Process Monitoring and Order Life Cycle Management**

Applying predictive analytics techniques on OLCM, it becomes possible for organizations to predict potential problems. This leads to an optimization of resource allocation, increasing efficiency.

The integration is based on the theoretical framework of process improvement and optimization. Merging predictive analytics with order management practices brings about better results which boost customer satisfaction.

Applying machine learning algorithms to data from an order's life cycle helps us predict its outcome and discover areas in which we can do better. By addressing issues even before they occur, one can streamline workflows involved during processing.

In conclusion, to optimize business processes and improve organisational performance, the operationalization of order life cycle management and predictive process monitoring include identifying important ideas, laying out implementation procedures, and integrating theoretical frameworks.

### **3.3 Machine Learning Models used in Predictive Process Mining**

Machine learning (ML) models act as necessary tools in this study to understand the intricacies of order life cycle management using predictive process mining techniques. The research methodology employs ML models' ability to analyze event logs, predict faults within the order life cycle and enable proactive decision-making. In this approach, it is important to select the right ML models because they determine how accurate and effective predictive process mining can be.

ML models are useful for discovering hidden patterns and insights within vast amounts of event data generated throughout an order's life cycle. These models use complex algorithms and methods to sift through complicated datasets, detect relevant attributes and forecast future results. This helps organizations gain more knowledge about their processes in managing orders, foresee potential problems and take preventive measures that would reduce risks while enhancing performance.



Moreover, ML models are adaptable and scalable which makes them suitable for dealing with various challenges associated with order life cycle management. For logistic regression used in binary classification tasks or decision trees intended to produce interpretable findings; random forests designed for ensemble learning or XGBoost created specifically for gradient boosting – each machine learning model has its own strengths. By employing different techniques like those mentioned above, companies can customize their approach towards predictive process mining depending on what works best for them considering their goals or objectives during that time.

In general terms machine-learning methods represent a cornerstone for innovation within supply chain management systems where they allow automation at every single stage from creating forecasts based on historical data all through monitoring current performance indicators against pre-set targets until generating real-time alerts which facilitate timely corrective actions whenever deviations occur thus ensuring continuous service excellence levels across entire logistics networks involved in product flow coordination activities such as warehousing transportation distribution among others.

Many Machine Learning (ML) models are carefully chosen and then used to study event data and estimate flaws in the order life cycle. Each of these models is unique, with different strengths and abilities, which helps us to explore predictive process mining techniques more fully. We utilize various algorithms ranging from logistic regression through decision trees/random forests up till XGBoost; each algorithm providing different benefits that make up this whole Predictive Process Mining step towards success story completion, each machine learning algorithm is explained in detail below:

### 3.3.1. Logistic Regression

Logistic regression is a core statistical method that is used in binary classification tasks. It is a way to predict the likelihood that an observation will fall into one of two classes. In contrast to linear regression, which predicts continuous outcomes, logistic regression deals with discrete outcomes, making it particularly useful for situations where the response variable is binary.

Healthcare, finance, marketing, and social sciences are just some of many fields where people choose logistic regression due to its simplicity and interpretability. Being easy to implement while being able to reveal connections between input variables and chances of certain events occurring has led it to be widely adopted.

Logistic regression aims at estimating the probability that the dependent variable (also referred to as target or response variable) takes on a specific category given one or more independent variables (also known as features or predictors). This is done by means of a mathematical function called the logistic function which transforms any real-valued input into a value between 0 and 1 representing probabilities.

The sigmoid function or logistic function can be represented as  $\sigma(z) = \frac{1}{1 + e^{-z}}$ , where  $(z)$  stands for linear combination involving independent variables along with their respective coefficients. Such function yields S-shaped curve in which higher values closer towards 1 indicate greater likelihoods for positive class while lower ones closer towards 0 signify higher probabilities belonging to negative class.

One notable strength of logistic regression lies within its interpretability. Within this model type, each coefficient denotes how much log-odds ratio would change given one unit shift in corresponding predictor variable assuming all other predictors remain constant; hence analysts can know both directionalities as well magnitude size effect exerted by each predictor on odds chances outcome occurring.

Moreover, logistic regression is not easily affected by noise and can work with either categorical or continuous inputs besides providing tests of fit such as likelihood ratio test together with Hosmer-Lemeshow test that give an indication about overall model fit to data.

However, logistic regression has some limitations despite its simplicity. It assumes linearity between independent variables and log-odds of the outcome which might not always hold in practice. In addition, when there are many predictors relative to observations then logistic regression tends towards overfitting. Regularization methods such as L1 or L2 regularization can be used when this happens so as to prevent overfitting and promote generalization ability for a given model.

Logistic regression is a flexible and interpretable machine learning technique that works best for binary classification problems. Its ease of use, interpretability as well ability to handle categorical & continuous variables make it one among many tools in the arsenal of any data scientist worth their salt. Nevertheless, care should be taken by analysts about assumptions made by logistic regressions along with their practical limitations when applying them into real-world scenarios.

### ***Mathematic Formulation***

Logistic regression is a statistical technique used in binary classification problems where one seeks to predict the probability that a given observation belongs to either of two classes. Instead of continuous outcomes as linear regression does, it handles discrete outcomes and therefore can be applied when dealing with binary response variables.

The logistic regression model is built around the mathematical function known as the sigmoid or logistic function. It takes any real input, returning an output value between 0 and 1 which represents a probability. That function is symbolized by  $(\sigma(z) =$

$\frac{1}{1 + e^{-z}}$ ), where ( z ) refers to a linear combination involving independent variables and their respective coefficients. Mathematically, this ( z ) equals dot product between feature vector ( X ) and parameter vector (  $\theta$  ), plus intercept term ( b ):

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n = \theta^T X + b$$

Here, each coefficient (or weight) associated with independent variable (  $x_i$  ) is represented by (  $\theta_i$  ), while  $b$  stands for intercept term; furthermore,  $X$  denotes feature vector.

Such mapping assigns values from linear combination into range [0,1]. These values correspond to probabilities because they indicate positive class membership (class 1) according to dependent variable which follows such distribution represented by sigmoidal curve shape. In other words, it tells us how likely it is for  $Y=1$  given  $X$  has been observed or measured when we say “probability” hereabouts what actually meant would be “chance”. Conversely if there were need talk about negative classes then those could also be done via  $P(Y=0|X)=1-P(Y=1|X)$ .

The output describes what proportion of times function should say yes given an input feature set  $X$ . If greater than or equal to 0.5 then observation predicted belong positive class otherwise negative class.

To estimate the parameters (theta) of a logistic regression model, it is common to maximize the likelihood function. The likelihood function represents the probability of observing the given set of data under the logistic regression model. Maximizing the likelihood function is equivalent to minimizing the logistic loss function, which measures the difference between the predicted probabilities and the actual class labels.

It may also be possible for us to regularize our models using L1 or L2 penalty terms in order prevent over fitting these approaches will allow higher generalization

performance by reducing large parameter values while maintaining fewer complex models if required.

In brief, logistic regression uses a logarithmic transformation called sigmoid which maps inputs onto probabilities which can be interpreted as belonging either positive or negative classes depending upon threshold level set. Model parameters are estimated through optimization methods such as maximum likelihood estimation (MLE) where regularization techniques like ridge (L2) or lasso (L1) could be implemented during estimation process thereby improving robustness against overfitting without sacrificing interpretability.

### ***Model Training:***

The most crucial part of building a logistic regression classifier is model training. It uses a dataset with known outcomes to determine the coefficients associated with each independent variable in the model. These parameters are estimated using a cost function that measures how far off predicted probabilities are from actual outcomes.

The data set need to be chosen which will be used for training and validation purposes. The chosen data should represent what is being studied and have all relevant features needed for classification tasks. Usually, this involves splitting it into two subsets: one used as train while another acts as validation or testing sets.

Before training models on datasets, one must ensure they are properly formatted by preprocessing them first. This may include dealing with missing values; encoding categorical variables; scaling or standardizing numerical ones so that their magnitudes do not differ too much from each other's ranges.

Once these steps have been taken care of, logistic regression can be trained using any optimization algorithm (e.g., Newton's method). In an iterative process called

gradient descent, the parameters are updated to find those values which minimize the cost functions i.e. difference between predicted probabilities and actual outcomes.

At every step during optimization algorithms execution phase where predictions based on current coefficient values are made against train data set observations then compared with corresponding real labels; thereafter errors obtained feed into computing cost functions used for updating corresponding coefficients involved until such time when either convergence achieved (minimum reached) or stopping criteria met.

Different metrics like accuracy rate; precision rate; recall rate; f1-score etc should be employed for monitoring performance throughout this stage more especially on training set. Such indicators help analysts understand how well our models capture underlying patterns within datasets hence detect potential overfitting/underfitting issues early enough before proceeding further with analysis procedures.

After finishing off above mentioned parts its high time we move further by evaluating trained models' generalization power using validation sets.

Here predictions produced by applying learnt theories onto unseen samples are compared against real labels provided alongside. Measures obtained from performance assessment tests done over these previously undisclosed records serve as approximate indicators concerning expected behavior patterns when faced with fresh inputs.

In conclusion, model training acts as an intermediate step towards building logistic regression classifiers. It involves data preprocessing; selection of optimization algorithms and incremental adjustment of parameter estimates so as to minimize the cost function. Analysts can gauge accuracy and reliability of trained models by monitoring performance on training sets while checking how well they generalize through validation sets.

### ***Model Interpretation:***

Understanding a model's interpretation in logistic regression helps us to know how the model makes predictions and determine what factors affect the outcome. This process is achieved by looking at each independent variable's coefficients, their magnitudes, and signs, as well as evaluating their significance.

The log-odds of the dependent variable changes by a factor equal to that of corresponding independent variable when its coefficient is changed by one unit other things held constant in logistic regression. Estimation of these coefficients is based on maximum likelihood estimates which seek to identify parameter values that make given data most probable according to the specified model.

The interpretation involves both size and directionality considerations; increased positive associations with increase in log odds while negative ones mean decrease respectively. So it should be noted that large values alone do not indicate practical importance but instead demonstrate stronger relationship between predictor variable(s) and outcome measure.

Statisticians typically rely on p-values associated with estimated  $\beta$ s when testing hypothesis about whether one thing affects another at all. In this case we compare observed significance levels against chosen alpha level (usually 0.05) for drawing conclusions concerning null hypotheses' rejection/acceptance states regarding any effect(s) under investigation. Analysts may choose different levels of significance depending upon context but usually use five percent level throughout unless they have strong reasons otherwise.

Analysts also evaluate global fit by checking if all individual fits are acceptable or not therefore overall goodness-of-fit tests can be employed just like in other models, for example penalized likelihood ratio test (PLRT). Generalized linear mixed models provide

AICc statistics for comparing two or more competing models where smaller numbers indicate better fit among those alternatives considered plausible explanations about what generated data based on theoretical framework adopted here.

Another way used frequently involves examining overall adequacy where predicted probabilities match actual outcomes across various levels expected outcome probability suggested so called non-significant results from such as those derived using Hosmer–Lemeshow goodness-of-fit statistic would imply good fit while significant deviations between predicted probabilities and observed frequencies indicate otherwise but this might be due to over-dispersion only.

Therefore, understanding logistic regression models requires a close look at the coefficients, their magnitudes and signs as well as evaluating significance levels associated with each one. Additionally overall goodness of fit tests should not be overlooked either during model interpretation or diagnostics phase since they offer valuable insights into how exactly such a model works when making predictions about certain outcomes being influenced by specific factors.

### ***Assumptions***

Logistic regression is a powerful method of statistics used in many binary classification tasks; however it is still a statistical technique with some assumptions and limitations that must be met or considered when applied practically.

Linearity of the Logit: Logistic regression assumes that the relationship between independent variables and the log-odds for the outcome are linear. In other words, this means that the logit function of probability an event falls under some category should be expressed as a sum over predictors with each weighted by corresponding regression



coefficient. Estimates will not only be biased but also predictions inaccurate if we violate this assumption.

**Independence of Observations:** A logistic regression model assumes error terms (or residuals) from separate observations to be uncorrelated; mathematically speaking  $Cov(e_i, e_j) = 0$  for all  $i \neq j$ . The probability of one observation belonging to a certain category should not depend on another's value(s). Thus, violating independence assumption may result into misleading standard errors especially when dealing with time-series and spatial data where they can inflate type I error rates.

**No Multicollinearity:** Another condition needed by logistic regressions has no presence multicollinearity among its predictor variables—the situation in which two or more independent variables are highly correlated such that their individual effects cannot be estimated. Hence if there is indeed any such correlation then parameter estimates become unstable and difficult to interpret because each variable does not provide a unique contribution towards predicting the response variable.

**Large Sample Size:** Logistic Regression requires enough samples so as to produce stable estimates for parameters which can further give reliable inference. Simply put, small sample sizes lead to overfitting where models fit noise rather than real patterns while underfitting implies failure to capture important relationships.

### ***Limitations***

**Binary Outcome Limitedness:** This technique was designed specifically for situations involving two groups only; hence it cannot directly deal with more than two categories. However, extensions like multi-nominal logistic regression exist enabling the handling of multi-nominal or ordinal outcomes depending on context.

**Linear Decision Boundary:** Logistic regression assumes that the boundaries between classes are linearly separable based on predictors. This means it can only capture those relationships which are linear with respect to predictor variables and log-odds of outcome event happening. Nonetheless when such relationship does not exist logistic regression fails in representing data correctly.

**Outliers Sensitivity:** The presence of outliers may distort, or bias model estimates especially if sample size is small. In this case, predictions will be heavily affected as certain observations might disproportionately affect parameter ranges used by algorithm for computing probabilities attributable to corresponding outcome categories. Thus, robust methods like using robust standard errors might be required to account for their influence during analysis stage.

**Missing Data Handling:** Logistic regression model does not account for how missing values should be treated in predictors' datasets i.e., independent variables' dataset. If there is anticipated any form(s) of missingness within these type(s) then they must first imputed upon before fitting into logistic regressions otherwise wrong conclusions could arise due improper use of available information.

Despite its assumptions and limitations, logistic regression is still widely used across different industries including but not limited to healthcare industry, financial institutions (such as banks), marketing companies etc. By considering these assumptions analysts are able decide when, where and how best apply them.

### **3.3.2. Decision tree**

Decision trees are a basic and broadly utilized algorithm in machine learning known for their simplicity, interpretability, and capacity to solve classification as well as regression problems effectively. In other words, decision trees mimic the thought process

of humans by dividing the feature space into different levels of decision nodes, where each node represents a decision based on input features' values. It is like a tree structure with branches showing decisions taken and leaf nodes showing outcome or prediction made.

The main reason why people love working with decision trees is because they are easy to understand which means that even someone without much knowledge about machine learning can still use them. They do not hide anything like neural networks hence stakeholders can know how predictions were arrived at since this model offers transparency and interpret ability unlike most black box models. This kind of explanation is important especially in areas where people need to be told why certain choices were made e.g., healthcare systems, financial institutions etc.

In addition to being able to handle numerical data types of even categorical ones without requiring much preparation beforehand such as dealing with mixed-type datasets where some variables are continuous while others may have discrete levels – these methods can also work with missing values automatically during splitting process unlike many other algorithms available which cannot do this at all hence decision trees remain one of most flexible methods across various fields and industries.

Decision trees are powerful models because they strike a balance between being simple enough for people understand yet strong enough make accurate predictions about complex phenomena. They're so intuitive transparent that everyone loves using them customer churn prediction through medical diagnosis may benefit greatly from these techniques. It's only by knowing how these trees are built what they can do as well as their limitations that practitioners will be able apply such knowledge effectively achieve goals involving difficult problems where lots information is available.

### ***Mathematic Formulation***

The formulation of mathematical decision trees involves recursive partitioning and split evaluation. It constructs the decision trees in an iterative way such that each node represents a feature or attribute of data while each edge indicates a decision rule depending on the values of that feature. At every node, among all possible splits the algorithm chooses the best one according to some criterion usually information gain maximization or impurity minimization.

For every tree node, potential splits are evaluated using different criteria based on features and a splitting criterion. Some common criteria include entropy, Gini impurity or information gain etc. For binary classification problems, Gini impurity ( $G(t)$ ) at node  $t$  can be given as:

$$G(t) = 1 - \sum_{i=1}^K p(i|t)^2$$

Where  $K$  is number of classes and  $p(i|t)$  refers to probability of class  $i$  at  $t$ . Similarly entropy and information gains can be calculated with respect to distribution over subsets obtained by all potential splits using these two measures.

Now the chosen criterion is maximized by selecting split which creates two child nodes representing corresponding subsets in data. Each of these children becomes parent during next iteration until some stopping condition is satisfied like maximum depth limitation reached or minimum samples per leaf condition met.

After creation of decision tree there may occur overfitting where model learns noise present within training examples rather than underlying pattern it should represent this can be resolved by employing pruning methods which simplify overly complex trees by eliminating nodes that don't significantly contribute toward better generalization on

validation sets. Pruning helps improve generalization ability and reduces risk for over fitting.

When trained, decision tree can make predictions about unseen instances by following rules from root down to leaves designed during learning phase where each internal uses value associated with its split feature determine which branch follow until it arrives at terminal leaf containing predicted response.

Mathematically, what drives formation of these trees is an optimization of split criteria so as partition space among features in manner that optimizes predictability while minimizing complexity and overfitting. In other words, decision trees are able to capture complex decision boundaries within data by evaluating potential splits repeatedly until no further improvement can be made through pruning thus making them ideal learners for environments with non-linear relationships between inputs and outputs.

### ***Model training***

Model training in decision trees requires building a tree-like structure that repeatedly divides the feature space according to the values of the predictor variables. The process starts with a root node representing the entire dataset, and at each node, an algorithm chooses one split from all possible ones by some criterion. This split partitions the dataset into two or more subsets, which become child nodes of this node. The procedure proceeds recursively for each child node until some stopping criterion is satisfied.

When training, amongst other criteria, decision tree algorithms compare different splitting properties to find splits as good as possible. Most frequently used ones are information gain and Gini impurity. Information gain measures reduction of entropy in target variable after division; Gini impurity measures probability of misclassifying

randomly chosen element if it were randomly labeled. Here we want splits that generate purest or most homogenous subsets.

To do this, features are checked against impurity or information gain by checking each potential dividing place on data set for them in turn. Feature-impurity pairs where divisions cause maximum decreases in either indicator measure will be preferred over others within same stage of division process (node). When choosing such pair at any stage among others – currently those available – constitute optimal partitioning point for that stage only according to its merits relative others considered together so far here called “stage top”.

Afterwards model selection bias might creep into final structure created by our algorithm due too much flexibility given during its training phase leading us into overfitting which means capturing noise instead underlying pattern represented there within trainings sets used. For instance, pruning comes handy when trying to fix such situations because it simplifies complexity hiding behind decision trees until only those parts responsible greatly improving predication power remain unpruned while removing some less useful sub-trees whose inclusion significantly impacts overall performance estimation ability during validation stages thereby reducing risk associated with over fittings.

When lacking information about any attribute value during evaluation procedure they can either take no account of this lack or treat it as another possibility which should be treated like any other category by creating surrogate splits on related variables whose behavior closely resembles original splitting characteristics demonstrated when missing values encountered thus enabling better predictions about those instances having measurements unavailable. Surrogate-splits are created based on alternative splits that mimic primary ones' behavior when missing values come into play.

Another advantage is interpretability and visualizable of decision trees because tree-like structures gotten after finishing construction process can easily be presented in graphical form so that anybody can understand them even without going through much explanation making it possible for stakeholders get clear insight into how model arrived at its decisions. This is exactly why decision trees remain one among most popular algorithms used for such cases across different sectors since they offer good simplicity along with clarity needed within black box systems where lots more may not always apply but here everything should work out fine indeed.

To sum up, predictive models built using decision trees involve breaking down the feature space into parts to predict what happens next. The algorithm considers several ways of dividing data at each stage and avoids overfitting by pruning underperforming sections of the tree. Decision Trees provide visibility throughout an organization because they allow people see where things have been done right or wrong based on certain conditions being met or not met.

### ***Model Interpretation***

The interpretation of models based on decision trees implies the comprehension of how the tree structure itself and the decision-making process contribute to predictions and insights. Decision trees enable users to extract valuable information about relationship, feature importance as well as decision paths although they are simple.

At the heart of model interpretation is its tree structure. In a decision tree, nodes represent features or attributes; branches denote decision rules depending on feature values while leaf nodes stand for predicted outcomes or classes. By viewing this graphical representation (tree structure), one can tell what makes up a forecast according to such models and which input variables are most influential in making decisions.

Feature importance ranks very high among other things that need to be considered when interpreting prediction models using decision trees. Features' significance is measured by their ability to decrease disorderliness within an object being studied i.e., target variable uncertainty reduction method; therefore, those appearing near the root node often used repeatedly during splits have more weightage since they affect greatly on predictions made by our model. It's crucial for users understand this concept because it helps them know what matters most in driving such decisions made by these systems and prioritize for further investigation/action those inputs found responsible.

Decision paths reveal much about classification made for individual observations via any given model created with decision trees. The path from one specific leaf node back through all intermediate points or nodes towards root illustrates which features were considered at each step along such journey as well as values considered leading up to outcome being predicted. We may also further aggregate over many examples if we want to find common pattern among them all.

### ***Assumptions***

Despite their comprehensiveness and adaptability, decision trees work on several assumptions that determine performance and applicability in different settings. These assumptions must be understood for effective employment of decision trees in predictive modeling.

For one, it is assumed that the association between predictor variables and the target is not linear and can be represented by simple rules. In other words, this assumption enables decision trees to represent complex nonlinear relationships between features and outcomes without using explicit mathematical equations. However, they may



fail to catch some very intricate interactions between variables especially when non-linearity is high or when dealing with data with many dimensions as well.

Another assumption made by decision trees is that data are free from noise with no missing values or outliers. Despite having mechanisms for handling outliers and missing values, these aspects can still affect interpretability plus performance of models produced by them. Outliers could skew splitting criteria leading to suboptimal splits while imputation or special treatment during tree construction might be necessary if there are missing values. In addition, noisy data which usually result into overfitting thereby making poor generalization about unseen data could also influence results obtained through decision trees even though they are known to deal poorly with such cases.

Additionally, another aspect taken into account here states that features selected for branching should have some information content about predicting what we want (target variable). This algorithm depends on feature's predictor process when it decides how should we split our sample space among branches. If an uninformative feature is used, then few splits will be made hence producing shallow trees or those containing redundant ones thus making performance less optimal. Selection as well engineering methods might help solve this problem by identifying only relevant attributes for prediction purposes only.

Fourthly, it assumes the dependent variable (also known as response) being categorical or ordinal while all else remain constant. While both classification regression tasks could be handled by them discretization intervals may need to be done so continuous predictions can approximate surrogate splits. Moreover, sometimes nonlinearity exists between continuous predictors and outcomes for such cases decision trees may not capture well the relationship between them.

Lastly, there is an assumption that data are independent identically distributed. This means each observation in our dataset comes from the same probability distribution and moreover no single draw depends on another. While it's true that dependencies among observations are not explicitly modeled by decision trees violation of this can lead to poor performance especially when dealing with temporal or spatially dependent data. Preprocessing techniques like decomposition clustering based on time series or space could be applied before using decision trees to handle such situations.

What should be noted however is that despite being flexible, interpretable, and easy to use; decision trees have their own weaknesses too thus knowing about assumptions behind them would help choose appropriate modeling techniques as well as preprocess data adequately which will eventually enhance interpretation of findings within real life settings. By recognizing these assumptions and implications thereof, practitioners can maximize utilization of decision trees in various predictive modeling tasks.

### ***Limitations***

Although they are powerful and widely used machine learning algorithms, decision trees have several limitations that practitioners must consider when applying them to real-world problems.

Overfitting the training data is one of the primary drawbacks of decision trees. The noise in the data may be captured by the decision tree and thus create models that are too complex but perform well on the training data and poorly on unseen data. This can happen if either or both the depth of a tree is allowed to grow beyond certain limits or proper pruning is not performed. Overfitting can reduce model interpretability as well as lead to poor performance on new samples.

Another limitation of decision trees is their instability. Small changes in training data may result in significantly different structures for a tree shifting predictions from one class to another. Therefore, when compared with other machine learning methods, especially those dealing with noisy or high-dimensional cases, this makes them less robust and reliable.

Besides these limitations concerning overfitting and stability, another disadvantage associated with decision trees relates to their inability to capture complex relationships among features within a dataset effectively. While it can handle simple decision boundaries and feature interactions, more intricate patterns or dependencies may not be represented correctly by this method due to its simplicity in terms of modeling capabilities.

Furthermore, choice sensitivity applies also here since it depends on what criteria are used for splitting as well as hyperparameters selection vis-a-vis performance sensitivity which could come about through various ways such as maximum tree depth setting, minimum samples per leaf point size specification or minimum impurity threshold level determination where finding an optimal set might require extensive experimentation coupled with tuning efforts.

Additionally imbalanced datasets prediction bias susceptibility: When there is imbalance between classes observed in target variable distribution i.e., majority versus minority class representation being skewed then decisions trees tend favoring more common groups during classification leading into lower predictive ability against less frequent ones even though cost misclassification for latter could be higher than former during such tasks.

### **3.3.3. Random Forest**

Random Forest is an extremely flexible and universally used machine learning ensemble algorithm in the field of artificial intelligence; it can handle both classification and regression tasks effectively. The algorithm is unique because of its strength, scalability, and extraordinary predictive precision among other algorithms. Random Forest is basically a group of decision tree algorithms which are known for their easy-to-understand representation of how decisions are made. However, what sets Random Forest apart from conventional decision trees is that it combines several prediction capabilities by using many trees thus making predictions more accurate and reliable.

The phrase “ensemble learning” means combining different models to enhance the performance of overall learning algorithm. In random forest these models comprise a collection or combination of decision trees that are trained independently on various subsets of training data. These trees are built with random selection for features at each node leading to the name “Random Forest”. By introducing some element of chance into model building process; this technique greatly reduces overfitting risks while increasing variations between individual trees thereby strengthening generalization abilities within final models.

Scalability has always been one key feature associated with Random Forest which makes it suitable not only for small but also large-scale applications. This method can be easily parallelized hence taking advantage of modern hardware architectures compute power to efficiently handle big volumes of information. Such extensibility becomes very useful when speediness matters most such as real time prediction systems or even big data analytics platforms.

Random forest remains a powerful toolbox algorithm within machine learning that provides several benefits over traditional approaches based on decision trees alone. It's competent enough to deal with complex datasets, avoid overfitting problems as well

as scaling appropriately therefore becoming widely applicable in different areas involving classification and regression analysis. Further sections will discuss math formulas behind RFs, how they should be trained, interpreted vis-a-vis results obtained; assumptions made during their use along with limitations associated with them.

### ***Mathematical Formulation***

The Random Forest algorithm is a machine learning method that uses a collection of decision trees to increase predictive accuracy. Now let's look at the mathematical notation of Random Forests:

- **Decision Trees:** Decision trees are hierarchical structures made up of nodes and branches. At each node, a decision tree splits the data according to a chosen feature and its threshold value. This process continues until some stopping criterion is met (e.g., maximum depth or minimum number of samples per leaf).
- **Ensemble Learning:** To create an ensemble of decision trees, Random Forest trains many trees on different subsets of the training data. It does this by bootstrap sampling, where each tree is trained on a random sample from the training set with replacement. Also, only a random subset of features is considered for splitting at each node in each tree, which adds more diversity among them.
- **Voting Mechanism:** For any input sample during prediction time every single one of the Random Forest's trees predicts an outcome independently from others. If its classification task then final prediction comes down to majority vote – class predicted by most votes among all trees wins as last prediction;

for regression tasks final prediction will be average over predictions made by all trees.

- **Aggregation:** In Random Forest idea is to aggregate multiple weak learners (individual decision trees) into one strong learner (the ensemble). By blending different predictions together, we hope that this way overfitting can be reduced as well as generalization performance improved because now base classifiers have been trained over diverse subsets taken from original dataset.
- **Feature Importance:** Another useful aspect offered by RFs concerns feature importance measures which show how much given variable contributes towards model's ability to make good predictions about unseen instances' target labels' values - usually such scores are calculated based upon Gini impurity or information gain computed after splitting data using attribute.
- **Hyperparameters:** To get better results with Random Forest it is necessary to adjust some of its hyperparameters. These include: the number of trees in the forest, the maximum depth of each tree and how many features should be considered at each split point. Tuning these settings plays an important role in achieving highest achievable prediction accuracy as well as making sure that model generalizes well beyond seen examples.

In a nutshell, Random Forests are powerful ensemble learners that make predictions by aggregating over many decision trees. Their power comes from being able to flexibly fit complex data patterns through bootstrapping multiple diverse subsets while training variously randomized base classifiers on them before combining their different outputs via voting procedures. This way, RF can handle very intricate relationships between inputs and output(s) variables thereby producing solid forecasts for unseen cases even if such knowledge was not present during training phase.

### ***Model training***

Random forest is implemented using a technique of ensemble learning, in which multiple decision trees are built and their predictions are combined for making accurate as well as robust classifications or regressions. The process of training involves various steps:

- **Bagging:** Known as bootstrap sampling, random forest does this where subsets from training data are taken randomly with replacement to create several bootstraps samples. In the forest each decision tree is trained on one such sample. This approach introduces diversity among trees which helps in reducing overfitting and improving generalization performance of the model.
- **Feature Bagging:** At every node of every decision tree in the forest, only a few features are considered for splitting. This is called feature randomization or feature bagging. By choosing different subset of features at each split randomly, the algorithm de-correlates trees and decreases chances to overfit any feature(s) or its subset(s). The number of features that should be considered at each split can be set as hyperparameter and chosen during model training phase.
- **Tree Construction:** Once we have decided bootstrap samples and feature subsets, then next step is constructing each individual decision tree recursively within Random Forests framework. Among available features at current node in the tree, algorithm selects best split based on some criterion (such Gini impurity or entropy for classification tasks; mean squared error for regression tasks) – it tries to find that division which will separate records into different classes maximally or minimize variance reduction (regression).

- **Ensemble Aggregation:** After building all decision trees composing our random forest — final prediction for an unseen observation is made by considering predictions made by every single one among those many trees altogether simultaneously. For classification task mode can be used (most frequently occurring class) while average (mean) should suffice when dealing with regression problem sets.

### *Assumptions*

Random forest is one of the most versatile and robust machine learning algorithms. However, like any other algorithm, it has its own set of assumptions that may affect the way it performs or should be interpreted. Here are some of them:

- **Independent and Identically Distributed Data:** This assumption means that every sample in the training dataset must have been selected independently and has equal probability to any other observation within this set. It also implies that no autocorrelation should be present between observations or no temporal dependencies should exist in time series data because if violated so, models estimate can become biased while predictions wrong.
- **Random Sampling:** To create multiple decision trees forming an ensemble, the algorithm uses random selection along with replacement (bootstrap sampling). According to this assumption bootstrap samples are supposed to represent well true underlying data distribution. Thus, if during sampling process there is introduced any kind of bias or if real variabilities inherent in data were not captured then less accurate model will be obtained.
- **Feature Independence:** Random forest assumes independence between features (or weak correlation at least). This helps each tree in the forest to



explore different aspects of data which contributes unique insights from all possible angles necessary for making correct decisions about target variables. If any two given attributes were highly correlated with each other, some trees might appear redundant leading therefore decreased predictive performance by whole ensemble.

- **Homogeneous Decision Trees:** The individual decision trees forming random forests are expected not overfit on training dataset but rather have ability generalize new unseen test examples. This requires Algorithm setting limits how deep they grow down certain branches based on their size relative those others growing simultaneously alongside them. However when such conditions aren't met single biased overfitted classifiers end up being combined into ensemble thus violating homogeneity assumption
- **Balanced Classes (for Classification):** Random Forests assume equal number of each class labels being used during constructing trees. In case one has imbalanced datasets where some classes are represented much more frequently than others, it may be necessary to apply techniques like class weighting or resampling in order achieve better generalization across all categories represented within given dataset.
- **Noisy Data:** Random forest is known to be robust against noise but assumes that level of noise is not too high. Outliers or erroneous observations can add unnecessary complexity which affects decision boundaries learned by algorithm. Performance improvements might require removing outliers through preprocessing steps such as data cleaning and also, It's worth noting that even if doesn't strictly adhere to these assumptions still works quite well most times except when there's too much noise

## *Limitations*

Random Forest is known as a powerful and widely used machine learning algorithm. Although it has several advantages, there are still some limitations. So, knowing these limitations well is important to apply Random Forest models properly in real-world applications.

- **Computational Complexity:** The foremost limitation of Random Forest is its computational complexity mainly during the training phase. It can be computationally expensive when building many decision trees on bootstrap samples and evaluating multiple features at each node especially when dealing with large datasets with many features. Consequently, much computation time and resources may be required to train a Random Forest model particularly during hyperparameters tuning or cross-validation.
- **Memory Consumption:** Apart from computational complexity, memory consumption could also be considerable in Random Forest models particularly for large datasets or ensembles containing many trees. Each decision tree in the ensemble has to keep track of its structure such as split points and feature importance scores which leads to high memory usage specifically for deep trees or ensembles having numerous trees.
- **Model Interpretability:** On one hand where random forest models achieve high predictive accuracy, on the other hand they are often referred as black box models because they do not reveal much about how decisions were made within them. Unlike simple interpretable models like decision trees which can be easily understood, random forests lack transparency in terms of understanding individual feature contributions towards predictions thereby

making it difficult for people to explain model predictions to others or understand why did model make certain decisions based on what factors?

- **Overfitting:** Random forest is less prone to overfitting compared with single decision tree algorithms but that does not mean it cannot be overfit at all especially when there are too many trees in the forest or their depth is not controlled properly. Overfitting happens if noise or irrelevant patterns within training data are captured by our model resulting into poor generalization performance on unseen data thus, we need to prevent this by setting maximum tree depths among other things.
- **Imbalanced Data:** Random forests may struggle when dealing with imbalanced datasets where one class is much more dominant than others since such cases can cause bias towards majority class which in turn leads to suboptimal performance for minority classes. To address this issue, we could use techniques like class weighting, resampling, or alternative evaluation measures such as AUC-ROC.
- **Scalability:** Though random forest can handle datasets with large number of features, but it may not scale well on extremely high dimensional data or those having millions of instances because as the number of features or instances increases so does the computational and memory requirements needed to build evaluate multiple decision trees in the forest thereby limiting its scalability for big data.
- **Parameter Sensitivity:** There are several hyperparameters within random forest models that need to be tuned to optimize their performance but sometimes choosing wrong values might result into poor performing model hence it is important conduct comprehensive hyperparameter search using

grid/random search etc., if we want best possible outcomes from our random forest model.

### 3.3.4. XG Boost

XGBoost is a leading machine learning algorithm that has gained popularity because of its record-breaking performance over a wide range of predictive modelling tasks. This algorithm rose to stardom due to its speed, scalability, and accuracy; thus, making it the number one choice for data scientists across industries. XGBoost is based on gradient boosting which has been celebrated for being able to produce highly accurate models while still being computationally efficient; hence becoming one of the foundations of modern artificial intelligence.

In essence, XGBoost works by adding weak learners (usually decision trees) together sequentially into an ensemble model. The method used here allows the system to correct itself as it learns from previous mistakes made by the current ensembles. Through successive approximations that minimize some measure of errors between actual and predicted observations using gradient descent optimization on model parameters, this algorithm updates predictions iteratively by optimizing an objective function in relation to closeness with regards targets. These updates are done until no further improvement can be made or a certain maximum number have been reached. At each step during training many base models may be fitted but only some contribute much towards final output so pruning techniques such as regularization or early stopping are employed to prevent overfitting.

At every iteration  $n$  sample pairs  $(x_n, y_n)$  with  $x_n \in \mathbb{R}^m$ ,  $y_n \in \mathbb{R}$  are randomly drawn from the dataset  $D = \{(x_i, y_i)\}_{i=1}^n$  according bins rules where  $x_i \in \mathbb{R}^m$ . Then these samples acts as input for each tree  $T_n(x) = \sum_{j=1}^m f_j(x)$ . For classification

problems we have  $y_n \in \{-1,+1\}$  and regression this is continuous variable so it can take any value within certain range. In order obtain good quality predictions we must construct accurate estimators which map  $x \rightarrow y$  that would work well on unseen examples too thus avoiding underestimation or overestimation when applied new data points outside training set bounds.

### ***Mathematical Formulation***

There are several factors involved in the mathematical setup of XGBoost that control its learning and optimization process. At its heart is a gradient boosting algorithm, which means it sequentially builds an ensemble of weak learners – often decision trees – to minimize a pre-specified loss function. Now let's go into the math behind XGBoost:

- **Objective Function:** The objective function measures how far predicted values are from actual values for some target variable in the model being trained by XGBoost. In regression tasks, this usually takes the form of squared error loss while logistic loss tends to be used for binary classification tasks; these functions compare estimated probability distributions against true ones.
- **Gradient Boosting:** Gradient boosting is employed by xgb, a method which involves adding weak learners iteratively to minimize an objective function. During each round (iteration) new trees are fit to correct mistakes made by previous models i.e., they learn residual errors relative to current predictions; this idea forms foundation for all variants using gradients.
- **Regularization Terms:** Regularization terms are added into the objective function of xgboost to reduce overfitting and enhance generalization performance. These additions penalize complex models by imposing extra costs based on parameter complexity; common examples include L1 (Lasso)

or L2 (Ridge) regularization where magnitudes of coefficients can be controlled.

- **Optimization Algorithm:** An optimization algorithm such as gradient descent which works with gradients is used by XGBoost to minimize its objective function. Descent iteratively adjusts parameters towards negative direction where rate drops value down until convergence point reached at minimum value; traditional GDHRO offers more speed through various tweaks like tree pruning but this package uses MSE as default criterion.
- **Tree Construction:** Decision trees within XGBoost grow one level at time following split criterion that makes best distinction between records on different sides. Trees are formed by selecting splits which maximize reduction in loss function achieved through each division thus created; additional features consist of maximum depth limitation for trees and minimum weight needed by child nodes during creation.
- **Model Prediction:** Final prediction made by an xgb model after training is completed can be calculated as sum of the predictions from all trees weighted according to their overall contribution. For regression tasks, this amounts to adding up predicted values produced per tree while in classification tasks it involves taking softmax distribution applied across raw sums.

Overall, the math behind XGBoost combines elements from gradient boosting with those relating to regularization and optimization; this creates a highly effective algorithm for ensemble learning that is both powerful and efficient. It iteratively optimizes the objective function thereby able to discover non-linear patterns within data hence making accurate predictions across many ML problems.

### ***Model Training:***

XGBoost, or eXtreme Gradient Boosting, is a powerful machine learning algorithm that combines multiple weak learners to make predictions. In this case, the weak learners are typically decision trees. The training process of XGBoost entails adding one tree after another into an ensemble until it stops improving the model's parameters with respect to some pre-specified objective function.

- **Gradient Boosting:** XGBoost follows a gradient boosting framework which is a technique for building predictive models with high accuracy by ensembling many weak models such as decision trees. As its name suggests, gradient boosting involves sequentially adding trees to the ensemble where each new tree is trained to correct errors made by existing ones.
- **Objective Function:** To start with training in XGBoost, an objective function has to be defined which measures differences between predicted and actual values on the training set. Examples of common objective functions include logistic loss for binary classification problems, squared loss for regression tasks and SoftMax loss when dealing with multi-class classification challenges. In essence these functions guide optimization by telling us how far we are from making better predictions.
- **Regularization:** In order not to overfit, XGBoost introduces regularization techniques during training including a regularization term in its objective function so that it avoids fitting noise present within training data samples observed during modeling process; some popular ones being L1 (Lasso) and L2 (Ridge) regularization methods which control complexity of learnt models by penalizing magnitudes of model parameters.

- **Tree Construction:** Each XGBoost tree within an ensemble is constructed using greedy algorithm that recursively partitions feature space to minimize loss function at every step i.e., node split point selection based on maximum reduction in value of given splitting criterion derived from chosen objective function. Creation halts when maximum depth or minimum number of samples per leaf reached according to stopping criterion.
- **Learning Rate:** Learning rate also called shrinkage parameter controls influence each individual tree has over final ensemble; smaller values slow down learning but may improve generalization by reducing sensitivity of model to noise while bigger steps accelerate convergence speed albeit with higher risk facing overfitting. Therefore, finding appropriate learning rates is critical for successful implementation of XGBoost algorithm in practice.
- **Model Evaluation:** XGBoost performs model evaluation during training using validation dataset separate from training set so as to prevent overfitting and guide further iterations on what trees should be added or pruned from current ensemble during training process until performance starts deteriorating indicating that more trees will only worsen situation due to excessive complexity brought about by them.
- **Parallelization:** XGBoost supports parallelizable training on multicore CPUs or distributed computing frameworks such as Apache Spark thus allowing for faster processing times mainly aimed at dealing with big data applications where large datasets are involved which makes it highly scalable too.

In summary, the XGBoost model training procedure involves adding trees to an ensemble iteratively, optimizing parameters of a model to minimize an already defined objective function and introducing regularization methods in order not to overfit the data.



Also, hyperparameter tuning along with monitoring performance throughout the process can lead practitioners into building highly accurate predictive models with XGBoost.

### ***Model Interpretation***

The interpretation of the model in XGBoost involves understanding the importance of different features in the dataset and their contribution to the predictions made by this model, which is necessary for discovering hidden patterns in data as well as selecting features more wisely or refining models better or even comprehending what drives predicted outcomes. There are a number of ways through which one can interpret a model using XGBoost; these include but not limited to feature importance, SHAP values and visualization techniques.

- **Importance of Feature:** In XGBoost, feature importance is calculated according to how many times does a feature split data across all trees within an ensemble on average. If frequently used for splitting and yielding higher gains in information, then such features are considered more significant than others. For each feature, XGBoost gives an inbuilt score for its importance thus enabling faster identification of influential ones among them by users who may wish to apply further analysis based on that.
- **SHAP Values:** SHAP (which stands for Shapley Additive Explanations) values provide us with a much richer picture about the relevance or otherwise of various attributes towards any prediction made by our model. It does that by telling us how much each input variable contributes towards moving from some baseline average prediction output over all possible inputs up-to reaching final prediction value associated with specific observation being considered currently. A positive sign attached here indicates those factors

which tend to raise our forecast while negative signs reveal opposite behavior i.e., things that make it lower instead. By looking at individual observations' SHAPs or else aggregating them across entire dataset we could learn much more regarding what affects forecasts made by this.

- **Visualization Techniques:** We can use visualization techniques to make XGBoost models easier to understand. One way is through feature importance plots which provide graphical representations showing scores attributed against different features – thus making it simple for anyone interested identify most important ones easily at glance. Partial Dependence Plots (PDPs) on the other hand demonstrate changes in predicted outcome as a single feature varies while all others are fixed throughout. This helps us understand relationship between individual input variables and our target variable. Interaction plots extend PDPs so that we can see how two or more features' combined values affect predictions made by our model.
- **Applications:** The practical applications of XGBoost model interpretation are vast across many fields. For instance, within finance sector; knowing which factors contribute most towards credit risk would be very helpful when trying to manage such risks effectively. Similarly, in healthcare domain it might enable us to identify what drives poor outcomes among patients thereby leading appropriate measures being taken towards improving them. In marketing industry too these techniques could assist marketers gain deeper understanding as well as insights into customer behaviour patterns thereby enabling them come up with better strategies for product design or service delivery. By employing these methods practitioners will be able to demystify

black box nature of XGBoost models and thus make more informed decisions within their respective areas of expertise.

### *Assumptions*

XGBoost, much like many other machine learning algorithms, is based on several assumptions that ensure its reliability and efficiency in making predictions. XGBoost is robust and flexible but it depends on some assumptions for accurate output. Here are the major assumptions of XGBoost:

- **Independence of Observations:** One assumption made by XGBoost is that each observation in the data set is independent from the others. This implies that the presence or absence of one observation should not influence another's presence or absence. If violated, this may cause biased model estimates and prediction instabilities.
- **Linearity:** Although it can account for non-linear relations between predictors and response variables, XGBoost still assumes some level of linearity within data sets. It suggests that there should be reasonably good approximation by linear functions between features and target variables. If this relationship happens to be extremely nonlinear, then XGBoost might struggle to accurately model such kind of data.
- **Homoscedasticity:** Another assumption implied by XGBoost states that the errors (residuals) variance remains constant across all levels of predictors. In simpler terms, it means residuals should have equal spread along predicted values range i.e., across different groups defined by predictor variable(s). Heteroscedasticity occurs when there are significant differences in error

variances at various levels of predictors thus leading to biased estimates as well as unreliable forecasts.

- **No Multicollinearity:** There should be no multicollinearity among predictor variables according to XGBooster's expectations about their interrelationships with response variable(s). When two or more explanatory factors are highly correlated with each other (i.e., exhibit high degree correlation), this can render coefficients unstable during estimation process due convergence problems associated with singularity caused by perfect collinearity between these predictors leading poor interpretability towards individual feature importance estimation capability of XGBoost model.
- **Normality of Residuals:** XGBoost doesn't assume that residuals follow normal distribution, but it performs better when they are approximately normally distributed. If deviations from normality are detected, then some relevant information in the data has not been captured by the model thus resulting in biased predictions.
- **Stable Feature Importance:** For different datasets and iterations of models, features importance should remain constant as per XGBoost assumption about their relevance across various contexts. Despite being robust to noisy data or irrelevant predictors, if there is considerable variation in feature importance between different datasets or even within same dataset with respect to two separate runs then it could suggest instability within the algorithm itself so care must be taken while interpreting such measures.

XGBoost is a powerful and flexible algorithm; however, one should bear in mind its assumptions which may not hold true for all datasets. Failure to satisfy these conditions can lead to wrong estimates, less accurate predictions, and lower performance

of models by XGBooster. Hence, it is important that users critically analyse their situation against this backdrop before applying XG Boosting in practice.

### *Limitations*

Despite being so popular and effective, XGBoost does have its limitations. It is important for users to understand them if they want to make informed decisions about when and where it should be used. Here are some of the key ones:

- **Hyperparameter sensitivity:** There are many different hyperparameters which control how XGBoost works such as learning rate, tree depth or regularization parameters – frequently achieving top performance requires careful selection but this can take long time because there are many possible combinations, and they need to be tested.
- **Computational complexity:** Although known for its speed and scalability even on big datasets, training large ensembles with many trees may still consume a lot of computation especially if data is vast too therefore limiting the usefulness in terms of computational resources required by an application.
- **Memory consumption:** During training all these trees are stored in memory hence this algorithm often leads to high memory usage particularly when dealing with either large datasets or ensembles containing numerous trees which poses challenge on systems having limited RAM size available for use at any given moment.
- **Limited interpretability:** Like most ensemble methods, XGBoost generates models that are difficult to comprehend once built; nevertheless feature importance metrics can shed light on what features matter most relative to others however understanding exactly how decisions were made by our model

becomes hard mainly due non linearity exhibited among variables within dataset under consideration thus making it impossible establish clear cut rules driven from these models alone which describe decision boundaries between classes being predicted

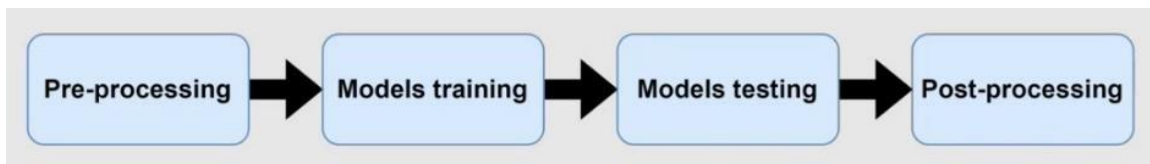
- Difficulty with imbalanced datasets: When one class is far more frequent than other classes – such scenario represents imbalance; now sometimes called positive event rare situation – then tree boosting does not perform well unless some corrections are done either via changing objective functions or re-sampling techniques like SMOTE etcetera although achieving good results remains challenging because of inherent nature imbalance problem itself coupled with other factors.
- Overfitting: XGBoost tries to avoid overfitting by using different regularization techniques but still this is possible especially when training noisy or high dimensional data, so model validation must be done carefully along with hyperparameter tuning in order not only prevent it from happening but also ensure that our models generalize better over unseen examples.
- Limited handling of outliers: Outliers can greatly affect decision boundaries production within machine learning model which usually leads to poor performance unless steps were taken during pre-processing stage where such extreme values were addressed properly before feeding them into ensemble method like XGBoost even though it may not handle them well too much especially those associated with features having many levels or skewed distributions.

As a summary, XGBoost is a very powerful and flexible algorithm. However, we should always keep in mind these limitations as well as other potential challenges that

might come up during analysis so that we can make the most out of it for our projects in machine learning.

### 3.4 Research Design

The research design includes the methodical strategy used to successfully answer the study questions and goals. This study employed a thorough research methodology to examine the effectiveness of the predictive process mining technique in improving the performance of order lifecycle management (OLCM) (Figure 3). The components constitute the research design are pre-processing, Model Training, Model Testing and finally, post-processing. Each of these steps are explained in detail below:



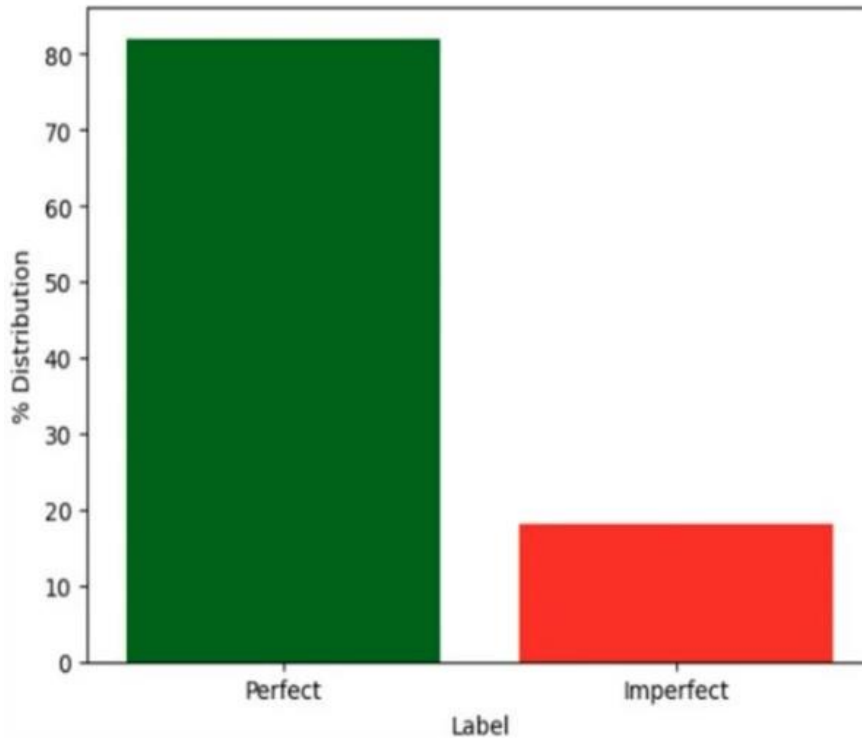
*Figure 3: General overview of research design.*

#### 3.4.1 Pre-Processing

The first step of data analysis is to load the dataset into the computational environment. The dataset contains 1.5 million records from a lifecycle event log, must be carefully examined to guarantee that every piece of information is captured correctly and transferred for later use in analysis. Once the dataset is successfully loaded, attention shifts to data cleaning – a key aspect of preprocessing. In this phase, different tasks like managing missing values and eliminating redundancies are performed to rectify any mistakes or inconsistencies.

Managing missing values ensures that the dataset remains whole and ready for analysis. Techniques such as imputation can be used to estimate or swap out missing

values based on the available data. By doing so, we minimize the impact of absence of data on subsequent analyses. Similarly, duplicate records are removed from the dataset to reduce redundancy and make sure each individual event log entry is unique and contributes to the analysis.

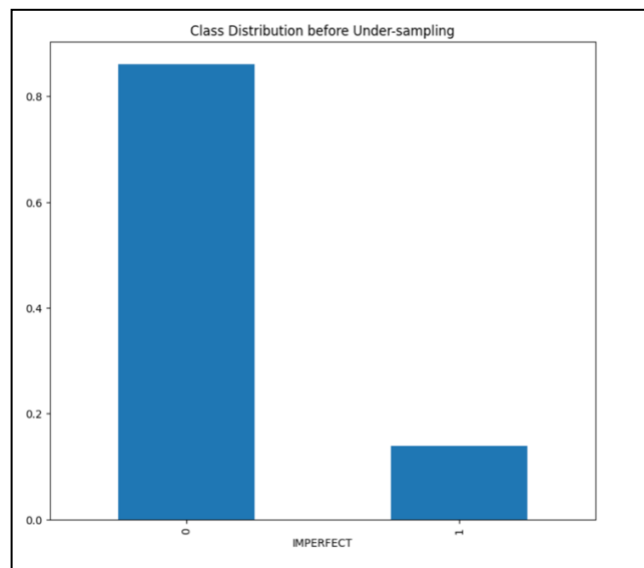


*Figure 4: EDA: Percentage distribution of perfect and imperfect.*

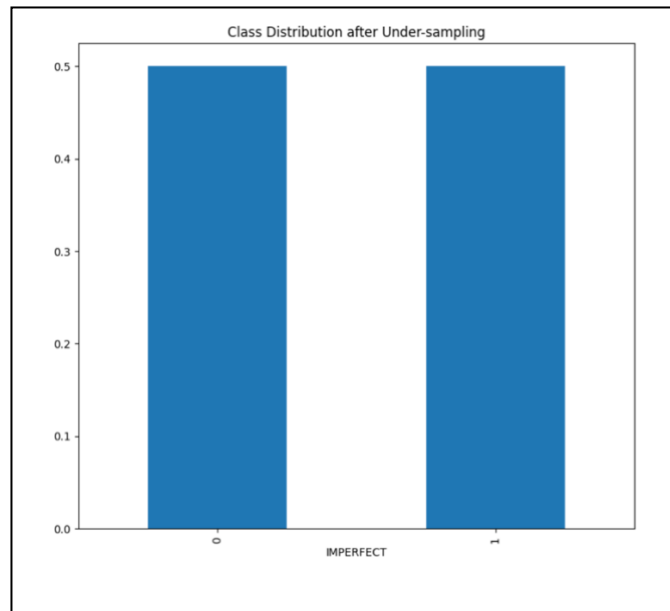
After cleaning the data, the study starts with the Exploratory data analysis phase, Exploratory Data Analysis (EDA) is a vital first step in data analysis. It is the process of scrutinizing the dataset visually and statistically, and finding patterns, trends, and anomalies. The study start by summarizing basic information about the dataset like its dimensions and statistics. Then move to individual variables and analyze them one at a time (univariate analysis). Followed by taking a look at relationships between pairs of variables (bivariate analysis), followed by interactions among multiple variables



(multivariate analysis). Several visualization methods come into play here including histograms, scatter plots, and heatmaps — all which help understand data distribution and relationships. EDA also includes outlier detection, missing values analysis, and identifying data quality issues. Overall, EDA helps to get an intuitive sense of the data's structure. Additionally, it informs subsequent analyses and decision-making processes. Some of the EDA as shown in the Figure 4, 5, 6, 7, 8 which offer informative visualizations of the distribution of perfect and imperfect orders, the first occurrence position of an imperfect activity, and the most frequent imperfect activity during the order lifecycle.



*Figure 5: Class Distribution before Under-sampling*



*Figure 6: Class Distribution after Under-sampling*

Figure 4 indicates the percentage distribution between perfect and imperfect orders. It provides an initial look at how prevalent imperfections are in relation to completed orders. Here the perfect order percentage is higher than Imperfect order percentage, which indicates the imbalance dataset and if the model is trained on the imbalanced dataset, the model prediction will be biased toward the Perfect Order, and hence there is a need to impute the dataset of imperfect orders to create a balanced event log.

Figure 5 demonstrates the original composition of the dataset before subsampling. In practice, there is a significant number of perfect orders than imperfect ones in real world scenarios, such as order management systems. Consequently, this can result in an imbalanced data set where the class imbalance is skewed towards the imperfection classes. Figure 5 visually represents this class inequality with more perfect orders than imperfections.

The implementation of under-sampling techniques has resulted in a transformation of the class distribution as shown in Figure 6. Under-sampling reduces class imbalance through randomly removing some instances from majority class, which are perfect orders to make both classes have equal representation. For instance, Figure 6 shows how under-sampling impacts on the distribution between non-conforming and conforming examples. This process of rebalancing is critical for improving machine learning models' accuracy since it ensures that classifiers do not become biased towards predicting most cases.

A comparison between Class distributions prior to and after Under-Sampling offers valuable insights into how samples were obtained and controlled for data imbalance issues. This EDA step prepares data for predictive modeling ensuring accurate and reliable machine learning approaches employed in order management life cycle prediction.

Figure 7 plots unique imperfect orders on the x-axis with their respective counts on the y-axis for each position of occurrence of their first imperfect activity. This figure shows the most imperfect orders activity starts at the second position in the order life cycle.

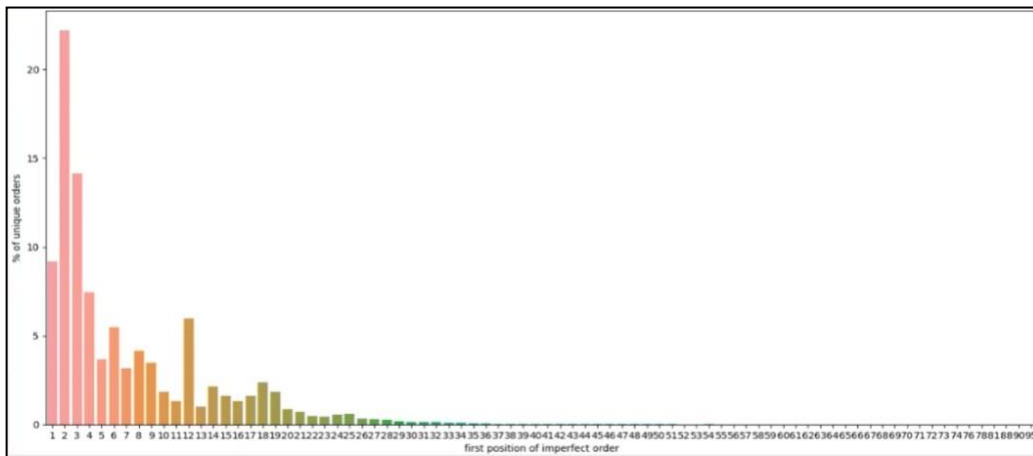
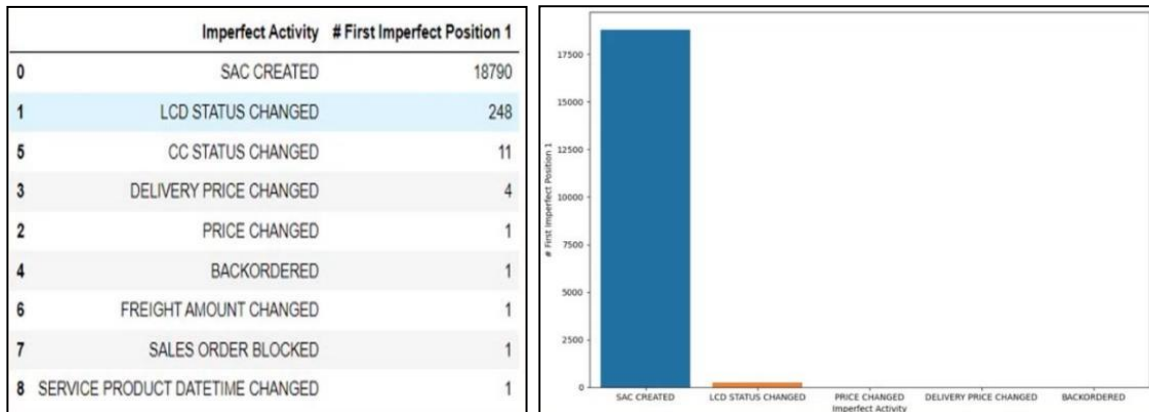


Figure 7: Percentage count of unique imperfect orders with a position of occurrence of the first imperfect order activity



*Figure 8: Most frequent imperfect activity that occur in OLC*

Figure 8 sums up all activities into a bar chart with their respective counts. It shares which specific activity are often associated with errors during our order lifecycle process. This information gives us a more precise idea as to what steps need improvement. From Figure 8, it is evident that the activity “SAC Created” is the most frequent activity which leads to order imperfection.

These three figures give us our first taste of insights into imperfections from EDA, namely distribution between perfect and imperfect orders; at which point along our order lifecycle they start going downhill; and what exact activities are causing these problems. As early as it is in this phase, it paves way for further analysis and model development that will help address these issues effectively.

After data cleaning and EDA, the dataset is split into distinct subsets to facilitate model training and evaluation. The partitioning process normally involves separating it into training and testing sets. Training sets help build predictive models while testing ones are reserved for checking how well these models perform. To provide an objective evaluation of the model's generalisation and prediction accuracy, data splitting aims to guarantee that the model is trained on one subset of the data and evaluated on another.

One-hot encoding is applied to categorical variables in the dataset and transforms into a format suitable for machine learning. Through this procedure, categorical variables are transformed into integer columns, where each column denotes a distinct category within the variable. This kind of encoding allows researchers to fully capture the range of categorical data in the dataset and efficiently integrate categorical variables into predictive models.

To guarantee that predictive models are trained on representative data and can effectively generalise to unseen cases, it is imperative to address class imbalance. To produce a balanced dataset with equal amounts of activity and non-activity occurrences, under sampling techniques are used. By using this method, biases resulting from unequal class distributions are mitigated and a representative and varied sample of data is used to train the prediction model.

Finally, a “Bag of Activities” feature is created to aggregate all activities that occur before an order imperfection. It’s entirely likely that there are patterns in this sequence of events – and by providing a deeper understanding of these processes, we can figure out which factors most significantly influence the outcome of our orders. Overall, these preprocessing steps lay the foundation for subsequent analyses and model development.

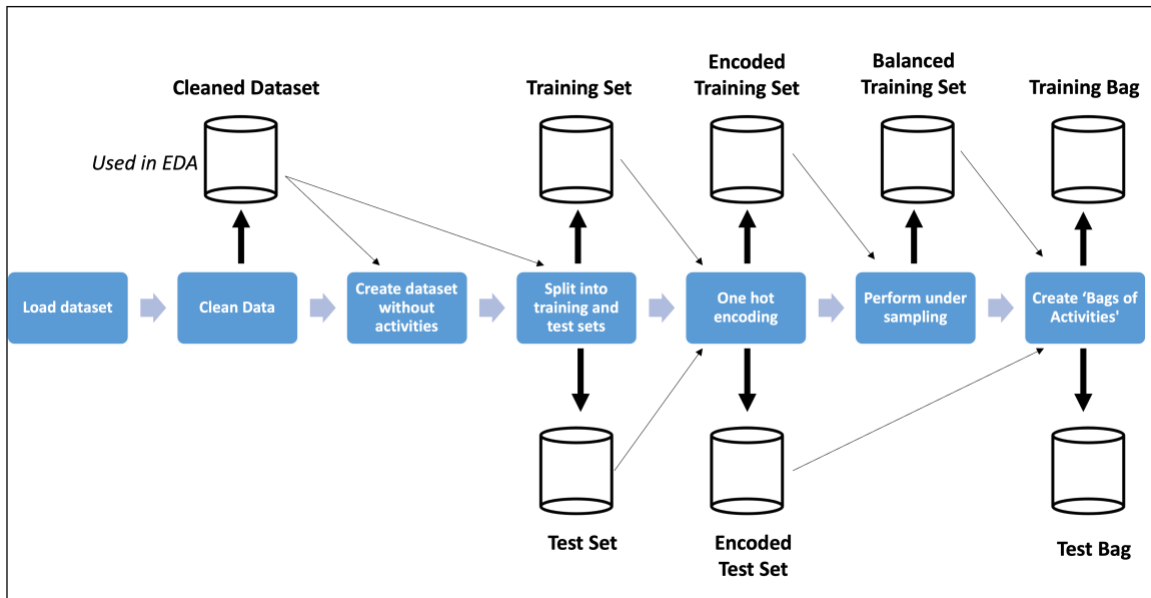
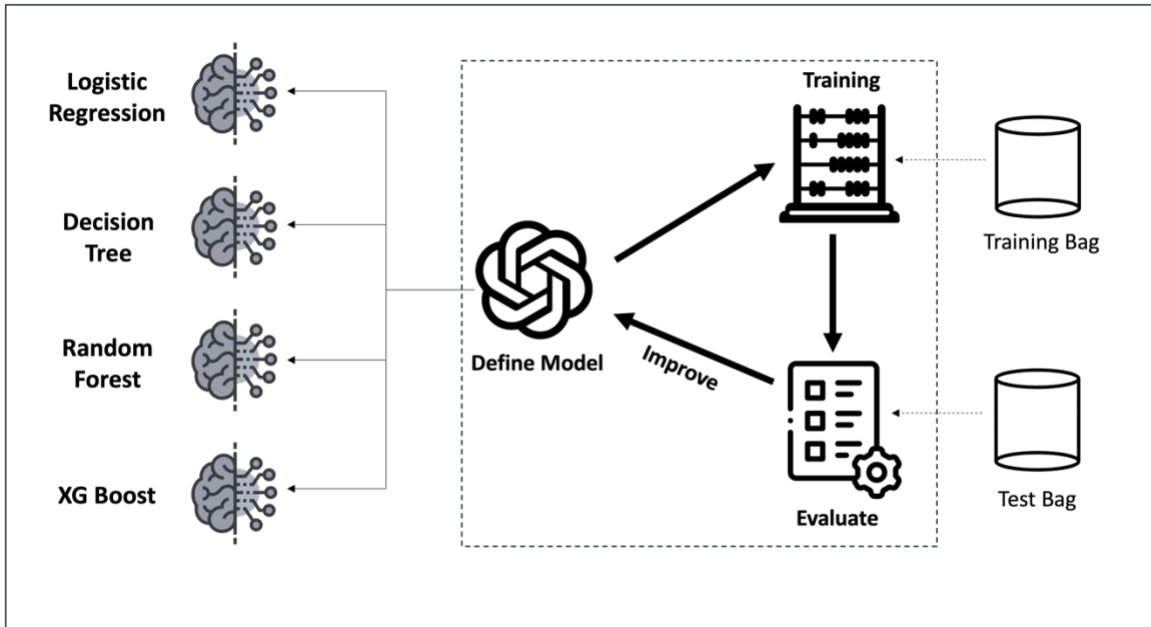


Figure 9: Data Pre-processing flow

### 3.4.2 Model Training

Model training is a crucial process in predictive process mining, where machine learning algorithms are trained to create models that can identify order imperfections. In model training, the activity bag is a significant factor when creating these models. It contains aggregated activities leading up to order imperfections. Balanced training sets come into play as well, which are created by under-sampling techniques to address class imbalance, ensuring the models train on representative data.

To capture different patterns and relationships within the data set, different machine learning algorithms are used. Some examples include logistic regression, which is known for its performance in binary classification tasks. Decision trees algorithm is a non-linear model and partition the feature space into hierarchical decision nodes. Random forest is an ensemble learning method that aggregates predictions from multiple decision trees for improved accuracy. Lastly, XGBoost is an efficient gradient boosting algorithm that handles structured data effectively.



*Figure 10: Model Training flow*

Throughout model training, encoded activity data — alongside other relevant features — are fed into these machine learning algorithms to learn patterns and relationships hidden within order imperfections. They adjust their parameters based on iterative optimization using training data so that errors in predicting imperfections can be minimized.

The study evaluates each machine learning algorithm's performance with metrics such as accuracy, precision, recall and F1-score. This allows to gauge the models' predictive capabilities and identify which ones work best for the specific scenario.

In conclusion, Model training plays a massive role in identifying order imperfections through predictive process mining. By using balanced training sets and various machine learning algorithms researchers have been able to create robust predictive models that enhance efficiency in managing order lifecycle processes

### **3.4.3 Model Testing**

Model testing is an important part of the predictive process mining workflow. This step ensures that the machine learning models have been properly trained and are able to perform well. During this phase, the models go up against a separate testing set to see how they hold up. The testing set typically contains instances from the activity bag that weren't used to train them originally, making it so the models aren't just memorizing data. This way ability of a model to accurately predict order imperfections on unseen data can be tested. Once input instances are inserted into the trained models, it allows to be able to assess how well model work when given new information and see if the model can correctly classify future instances as either imperfect or non-imperfect orders.

To measure model performance, a few evaluation metrics can be used including accuracy, F1 score, and recall. Accuracy will determine how many instances were correctly classified out of the total number in that dataset—essentially giving researchers a general overview of overall predictive performance. To get an optimal measure of prediction values, F1 score uses precision and recall as its harmonic mean—a more specific approach depending on imbalanced class distributions situations. Recall also tests true positive proportions but instead compares them against all actual positives in dataset.

By calculating these metrics for each model, strengths and weaknesses can be identified—ultimately helping with practical deployment in real-world applications. Models with higher accuracy, F1 score, and recall values have proven to identify imperfections more effectively than others—making them ideal for order lifecycle management systems.



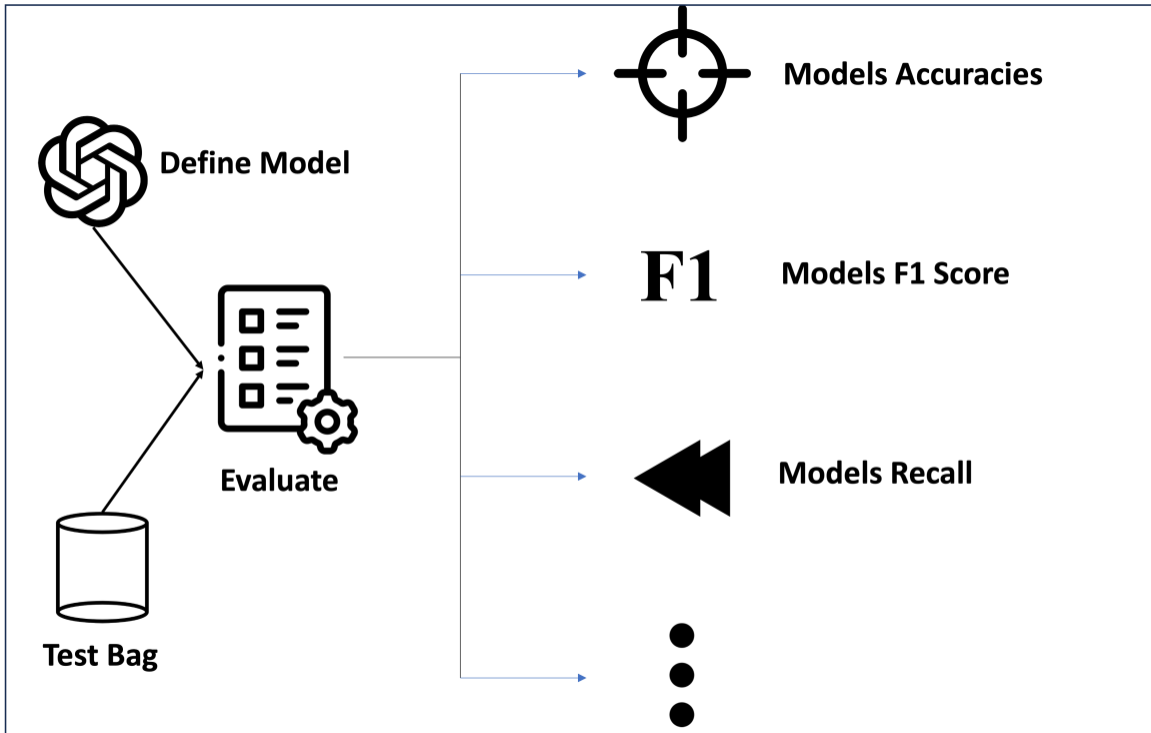


Figure 11: Model Testing Flow

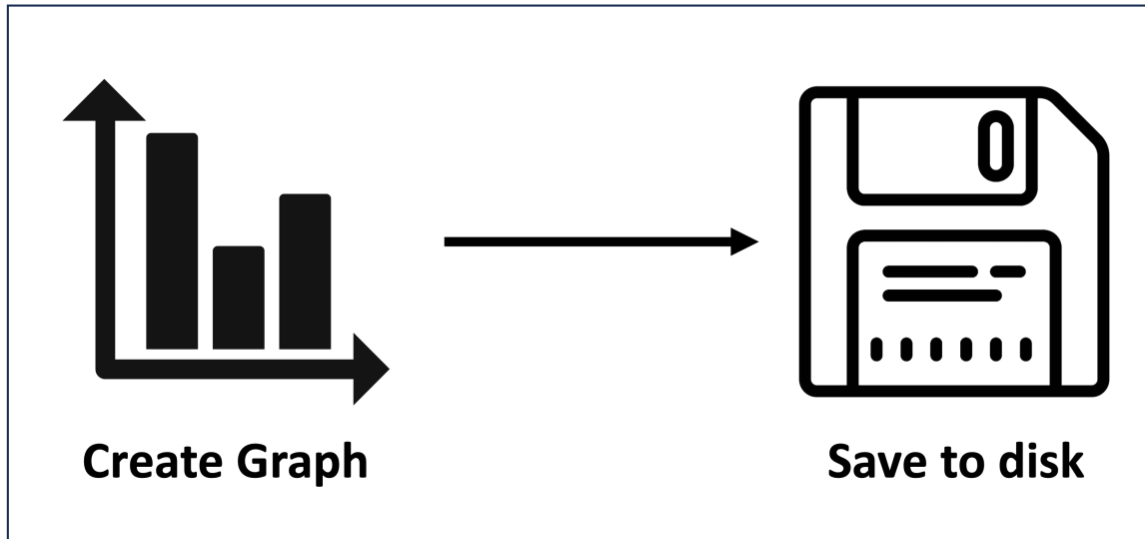
Overall, it's obvious that model testing plays an important role in determining if trained machine learning models are suitable enough for predicting order imperfections effectively or not at all. By incorporating reliable evaluation methods throughout the workflow researchers may find themselves leaning towards better choices for their system.

**3.4.4 Post Processing**

The actions performed after the model makes predictions are known as post-processing.

1. Compute the SHAP values: SHAP values indicate how each feature affects the model's prediction accuracy. By determining which characteristics have the most predictive power for predicting imperfect orders, it is possible

identify the root causes of order lifecycle events and inform choices that will enhance order handling.



*Figure 12: Post Processing flow*

2. Save the ML Output to Disc: The model maintains data, including predicted probabilities and classifications, after making predictions so that it will be accessed later for additional study or integration with other systems. This stage guarantees that predictions are preserved successfully can be utilise effectively in subsequent iterations of the study.

### **3.5 Population and Sample**

In this study, we go into details of order lifecycle processes in each country by using event logs and activity data to provide insights. These event logs and data repositories are sources of information that contain the full spectrum of activities involved in managing and fulfilling orders across various business domains. Every stage of an order's life cycle from initiation to completion is strictly documented and analyzed.

### **3.5.1 Overview**

Order management and customer satisfaction are facilitated through a series of interrelated activities within the order lifecycle process. This starts with capturing customers' demands through their creation orders and extends through subsequent stages such as processing, fulfillment, post-deliveries and so on. Many tasks or operations involving different departments, systems, or stakeholders are done at each stage to work towards meeting the expectations and needs of clients.

These order lifecycle processes can be better understood by studying them using event logs and activity data which unravel all sorts of intricacies when it comes to managing orders from their birth up to delivery. By analyzing timestamps, sequences, outcomes, or results among other factors associated with these activities we can uncover patterns, bottlenecks, or inefficiencies capable of affecting overall performance as well as efficacy of an order management system.

Moreover, concentrating on one country's Order Lifecycle Processes adds contextual relevance which enables us to mold our analysis and discoveries around some attributes as well as difficulties imposed by that market. We need localized contexts for capturing such things as industry practices, regulations or even nuances that shape how orders are managed thereby increasing applicability and relevance for our findings.

To effectively perform our analysis, we developed a representative sample drawn from the population described above consisting of event logs and activity data. Thus, this sample represents real-world scenarios on order management showing diverse activities noticed during its lifespan.

In developing this sample however, it was imperative that we synthesize data that can mimic actual complexities plus dynamism found in order management processes. We

aimed at replicating these real-world intricacies while ensuring dataset integrity hence authenticity through carefully curating timestamps and activities associated with each order.

Each entry in the sample reflects stages such as creating, processing, fulfilling and making a delivery of an order with some key milestones. Time stamping each activity comes in handy for providing temporal context which would enable us to reconstruct the sequence of events chronologically while looking into time spent on different order tasks.

Additionally, this sample is designed to mirror the heterogeneity and diversity inherent in order management processes by capturing various types of orders as well as transaction volumes and operational contexts. This ensures that the analysis covers a wide range of scenarios so that we can make sound conclusions about real-world order management environments.

### **3.5.2 Sample Creation**

The creation of this sample is a crucial procedure in our research methodology because it yields a representative dataset which captures many features seen during actual Order lifecycle processes. We can thus analyze using this synthesized data towards deriving necessary insights so that meaningful recommendations can be made to improve efficiency and effectiveness of these systems.

To ensure that our data set is relevant and representative, we used the purposive sampling technique. This approach enabled us to choose the most vivid examples to match a picture of variety in the country's lifecycle of goods and service orders.

Real world event logs served as the basis for generating synthesized data for our sample. It helped us capture the subtleties and complexities inherent in real order lifecycle processes since this way it was possible to realize order life cycle systems of

any kind. Thus, it made sure that our dataset was authentic and valid hence forming a solid foundation for rigorous analysis and meaningful inference.

Our sample size decision was guided by several considerations such as statistical significance and capturing different types of lifecycles with variation. We used 1.5 million records as this number is sufficient for obtaining statistically significant results while keeping computational load under control which is important given the high computational intensity associated with predictive process mining techniques, resource requirements for robust model training.

Our choice of this sample size aimed at achieving both statistical soundness and computational feasibility so that our findings remain robust and reliable. Also, we could generate insights that are both actionable and relevant to business entities operating within chosen jurisdiction since we concentrated on datasets reflecting complexity of order life cycle processes.

### **3.5.3 Inclusion Criteria**

The inclusion criteria were carefully laid down so that only useful event logs and activity details would be selected into our sample thus ensuring its quality and relevance. An essential factor for each case in question had been meeting defined criteria of accuracy, completeness, consistency. This allowed us to sustain veracity even when some deviations or inconsistencies have occurred during data collection thereby preserving our claims' validity.

Besides, if every event log or datum about certain activities concerned order management did not give valuable insights into various timestamps from which these transactions were enacted then it could not be included in the list of those logs needed to be incorporated in research materials as they lacked representation at multiple time points

during an item's existence through creation until it reached the client. We focused on datasets that offered useful information on order management thereby constructing a sample full of intricacies and complexities observed in real life orders' lifecycles.

Moreover, the inclusion criteria encompassed the relevance of each event log or activity datum to our study objectives. For this reason, only those data sets that came close to our aims and targets were included in the list. Hence the dataset had all events and activities used for analysis directly related to it thereby making it possible for us to do some finding and concluding remarks about this issue within a certain country's situation.

In conclusion, the inclusion criteria acted as a frame of reference for choosing event logs and activity data in order to ensure that only good quality, pertinent and informative datasets were included in our sample. In line with these standards, we wanted to construct a dataset that was strong, extensive, and fit for conducting thorough analyses leading to useful findings on order management processes.

#### **3.5.4 Exclusion Criteria**

For the sample group we developed, we utilized strict exclusion criteria to separate data that had considerable differences from actual real-world order management practices. Among other things, this criterion ensured that only genuine and representative order lifecycle datasets were considered thereby making our findings more reliable.

Meanwhile, event logs or activity data that strayed significantly from the accepted industry norms or best practices could not be included in our sample and were therefore rejected. This criterion helps in reducing bias and inaccuracies as datasets which were detached too much from reality do not give any insights that can be applied to order management context.

Moreover, any dataset including abnormal or outlier points of data that would badly skew results or lead to misrepresentation was also removed from the sample set. The purpose of such action is to ensure reliability and robustness of the analysis because only accurate decisions which are arrived at using reliable and representative information can be made.

Synthetic data that did not mirror real-life order management processes closely enough or those without relevance to our research goals were likewise excluded from the sample set. By doing so, it helped us narrow down on what was essential for our dataset hence focusing on meaningful analyses with respect to actionable insights on order life cycle process.

Overall, stringent exclusion criteria improved our sample by ensuring inclusion of only those sets closely related to real world order management practices and relevant to this research. It aimed at upholding these measures purposely for validity, reliability and application i.e., to strengthen the quality of our study altogether.

### **3.5.5 Event Log Dataset of Order Life Cycle**

Figure 13 shows the order life cycle dataset used in the process, and detailed explanation of all the features are as follows.

1. **createDateTime:** createDateTime is a timestamp representing when an order was created within the system. This element provides time-related details about each placed deal assisting in monitoring how long it takes processing orders as well as evaluating workflow standards throughout the organization.
2. **SalesSetId:** This attribute is likely a unique identifier or reference code for a particular sales transaction or order set in the system. It is useful in

differentiating individual orders or sales sets as well as organizing and managing all sales-related data.

3. Case key: On another note, Case key is another unique identifier used to distinguish and reference individual cases or instances within the dataset. The use of this element can be seen as a primary key or identifier of records linking these cases with transactions across various systems and processes.
4. Order No: Order No on the other hand represents each order number assigned to a certain order made through sales within the system. Every Order NO serves as an exclusive identity for tracking all single orders while it ensures easy references and retrieval of specific information about them.
5. Quantity: The Quantity attribute shows how many items or products are included in each order. It is the indicator of how much bigger or smaller a transaction processed in the system has; it is important for inventory management, demand forecasting and order fulfillment planning.
6. Imperfect: The Perfect/Imperfect attribute is a categorical variable indicating the status or condition of each order in terms of its completeness or accuracy. Orders termed as “Perfect” may be defined as those that went through processing efficient enough to ensure no discrepancies occurred while “Imperfect” orders point out to any faults detected during their processing like errors, defects and abnormalities among others.

In general, this screenshot gives a holistic view on data related to orders with timestamps, unique identifiers (UUID), order numbers, quantities and status indicators which are crucial in managing and analyzing the effectiveness of order lifecycle processes.



### 3.5.6 Bag of Activity (BoA) Feature

The dataset mentioned in the previous section was not able to generate the correct output when machine learning model were trained on the dataset. The study came up with the additional powerful attributes, which with the support of other attributes has a better prediction model. The bag of activity feature is a collection of all the activity for a case key in a sequence right before the imperfect activity occur.

Figure 14 and Figure 15 represent screenshots containing activity-related data from the order management lifecycle with several attributes that are essential for analysis and modeling. Below is a description of what attributes they contain:

**Case Key:** This attribute serves as an identifier of every individual case within the dataset. This case key distinguishes between different orders thereby aiding in monitoring and analysis of activities related to orders throughout their lifecycle. Each case key represents an individual order which may encompass multiple activities linked to this order.

**Activity\_en:** The attribute represents sequential actions performed during ordering process. It captures sequence of events leading from initiation through completion/imperfection of an order. The activity\_en field allows one to understand how activities flow during an Order's lifecycle thus providing grounds for process flow analysis and performance measurement.

**Perfect/Imperfect:** This Boolean attribute shows whether an order meets certain predefined criteria or business rules regarding perfection hence used as target variable for machine learning models built around predicting imperfections in orders. Usually, true values are meant for perfects whereas false indicates imperfection. This attribute is significant in training and evaluating predictive models for forecasting order defects that might arise in the future in the order management process.

The screenshots provide a snapshot of the order-related data, showcasing the key attributes necessary for predictive modeling and analysis. By utilizing these attributes, organizations can gain insight into their order management processes, identify patterns or anomalies, and create predictive models to anticipate and address potential imperfections or issues in order fulfillment.

	createDateTime	salesSetId	_CASE_KEY	orderNo	Quantity	IMPERFECT
0	2022-11-20 11:00:00	359728788.0	622386968A01208	6.223870e+08	4.00	0.0
1	2022-12-03 14:15:32	359728795.0	622386975A01208	6.223870e+08	1.00	0.0
2	2022-12-12 13:17:53	900230388.0	640551878A01288	6.405519e+08	702.00	1.0
9	2021-11-05 10:05:56	799576685.0	1230910706A01422	1.230911e+09	2.00	0.0
10	2021-11-05 08:57:39	817003598.0	1228815388A01301	1.228815e+09	3.00	0.0
11	2021-11-16 10:06:34	819880982.0	1229259956A01422	1.229260e+09	60.21	0.0
13	2021-11-05 11:52:20	803208071.0	1231088488A01422	1.231088e+09	10.00	0.0
14	2021-11-14 11:16:20	819398772.0	1230854185A01301	1.230854e+09	37.00	0.0
15	2021-11-03 11:33:56	816522981.0	1230552519A01301	1.230553e+09	2.00	0.0
16	2021-11-12 11:02:26	818287617.0	1230602019A01288	1.230602e+09	53.00	0.0
18	2022-01-14 11:47:47	834344149.0	1232246340A01008	1.232246e+09	1.00	1.0
19	2021-11-08 10:47:17	817802443.0	1231717518A01422	1.231718e+09	8.00	0.0
20	2021-11-08 09:51:16	817792001.0	1231821719A01288	1.231822e+09	7.00	0.0
21	2021-11-08 11:03:15	817819452.0	1231919173A01202	1.231919e+09	1.00	0.0
23	2021-11-15 11:31:31	819611552.0	1233450939A01422	1.233451e+09	81.00	0.0
24	2021-11-11 11:58:40	818595405.0	1232568765A01202	1.232569e+09	2.00	0.0
25	2021-11-02 10:11:36	816217071.0	1232603620A01288	1.232604e+09	1.00	0.0
26	2021-11-02 09:56:24	816182095.0	1232613821A01422	1.232614e+09	3.00	0.0
27	2021-11-06 08:55:34	817257985.0	1233403899A01422	1.233404e+09	32.00	0.0
29	2021-11-15 14:24:22	819646338.0	1233480786A01208	1.233481e+09	1.00	0.0
31	2021-11-14 13:19:15	819443388.0	1233211779A01202	1.233212e+09	2.00	0.0
32	2021-11-16 11:22:42	819893827.0	1233841685A01208	1.233842e+09	23.00	0.0
33	2021-11-17 09:55:26	820135003.0	1233940594A01422	1.233941e+09	7.00	0.0
35	2021-11-19 11:15:59	820639158.0	1234489670A01301	1.234490e+09	41.00	0.0
36	2021-11-01 12:24:23	815989586.0	1234497421A01208	1.234497e+09	2.00	0.0
37	2021-11-09 09:40:10	807632132.0	1234324448A01202	1.234324e+09	2.00	0.0
39	2021-11-02 09:29:48	816184694.0	1234748524A01208	1.234749e+09	21.00	0.0
40	2022-01-10 06:30:14	833168592.0	1238481825A01008	1.238482e+09	1.00	1.0
41	2021-11-03 11:42:40	815943919.0	1234899122A01288	1.234899e+09	4.00	0.0
42	2021-11-21 12:25:36	821205723.0	1234901147A01288	1.234901e+09	12.00	0.0
43	2021-11-03 12:52:04	816474393.0	1234916135A01208	1.234916e+09	37.00	0.0

Figure 13: Sample event log dataset of an Order Life Cycle

	<u>_CASE_KEY</u>	<u>ACTIVITY_EN</u>	<u>IMPERFECT</u>
0	1058948332A01288	CREATED, CREATED, SENT FOR FULFILLMENT, SALES ORDER CREATED, CUSTOMER NOTIFIED, SALES ORDER CONVERTED, LM_WORK ORDER CREATED, RELEASED FOR PICKING, PICKED, READY FOR DISPATCH, LM_DISPATCH COMPLETED, HANDED OVER TO TSP, RELEASED FOR PICKING, PICKED, READY FOR DISPATCH, LM_DISPATCH COMPLETED, HANDED OVER TO TSP, LM_RECEIVED AT HUB, RECEIVED AT LSC, LM_RECEIVED AT HUB, LM_RECEIVED AT HUB, LOADED ON DELIVERY TRUCK, DELIVERED, CUSTOMER DELIVERY COMPLETE, LOADED ON DELIVERY TRUCK, DELIVERED, CUSTOMER DELIVERY COMPLETE, PAYMENT EXECUTED, SALES ORDER COMPLETED	0
1	1083822576A01208	SALES ORDER CREATED, SALES ORDER CONVERTED	0
2	1233772043A01422	SALES ORDER CREATED, SALES ORDER CONVERTED	0
3	1236246036A01208	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SERVICE PRODUCT STATUS CHANGED, SERVICE PRODUCT STATUS CHANGED, SERVICE PRODUCT STATUS CHANGED, NOTIFICATION RECEIVED, SERVICE PRODUCT STATUS CHANGED, SERVICE PRODUCT STATUS CHANGED, SALES ORDER COMPLETED, CUSTOMER NOTIFIED	0
4	1236332895A01208	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	0
5	1236479406A01208	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	0
6	1236513346A01288	SALES ORDER CREATED, SERVICE PRODUCT STATUS CHANGED, SALES ORDER CANCELED, SERVICE PRODUCT STATUS CHANGED	0
7	1238022675A01422	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	0
8	1238459535A01008	SALES ORDER CREATED, SALES ORDER CONVERTED	0
9	1238545129A01422	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	0
10	1238578682A01202	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	0
11	1238722886A01008	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	0
12	1238977982A01301	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	0
13	1239053004A01202	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	0
14	1239206769A01422	SALES ORDER CREATED, SALES ORDER MODIFIED, ORDER AUTHORIZED, SERVICE PRODUCT STATUS CHANGED, SERVICE PRODUCT STATUS CHANGED, SERVICE PRODUCT STATUS CHANGED, SALES ORDER COMPLETED	0
15	1239243193A01208	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	0
16	1239590623A01202	SALES ORDER CREATED, SALES ORDER CONVERTED, LM_WORK ORDER CREATED, CREATED, PAYMENT EXECUTED, SENT FOR FULFILLMENT, LM_WORK ORDER UPDATED, RELEASED FOR PICKING, LM_DISPATCH COMPLETED, READY FOR PICKUP FROM STORE, PICKED UP BY CUSTOMER, CUSTOMER DELIVERY COMPLETE, SALES ORDER COMPLETED	0
17	1239732976A01288	SALES ORDER CREATED, SERVICE PRODUCT STATUS CHANGED, SALES ORDER CANCELED, SERVICE PRODUCT STATUS CHANGED	0
18	1239981628A01422	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	0
19	1240021269A01301	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	0
20	1240362009A01202	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	0

Figure 14: Sample bag of activity dataset with all perfect activities in sequence

	<u>_CASE_KEY</u>	<u>ACTIVITY_EN</u>	<u>IMPERFECT</u>
0	1015424538A01202	SALES ORDER CREATED, SALES ORDER CONVERTED	1
1	1019799229A01202	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SERVICE PRODUCT STATUS CHANGED, SERVICE PRODUCT STATUS CHANGED, SERVICE PRODUCT STATUS CHANGED, SERVICE PRODUCT STATUS CHANGED, SALES ORDER COMPLETED	1
2	1036844282A01288		1
3	1058948004A01208	SALES ORDER CREATED, SALES ORDER CONVERTED	1
4	1058948159A01422	SALES ORDER CREATED	1
5	1059277450A01208		1
6	1059441025A01422	SALES ORDER CREATED, SALES ORDER CONVERTED, LM_WORK ORDER CREATED, CREATED, CUSTOMER NOTIFIED, CUSTOMER NOTIFIED, PAYMENT EXECUTED, LM_WORK ORDER UPDATED, SENT FOR FULFILLMENT, CPS_PLANNED, CPS_ASSIGNED, CPS_PICKING, RELEASED FOR PICKING, CPS_PICKED, CPS_CHECKING, LM_DISPATCH COMPLETED, PICKED, CPS_CHECKED, CPS_COMPLETED, READY FOR DISPATCH, LM_RECEIVED AT HUB, RECEIVED AT LSC, LM_RECEIVED AT HUB, RECEIVED AT LSC, LM_RECEIVED AT HUB, RECEIVED AT HUB, LM_RECEIVED AT HUB, LM_RECEIVED AT HUB, LOADED ON DELIVERY TRUCK, DELIVERED, CUSTOMER DELIVERY COMPLETE, SALES ORDER COMPLETED	1
7	1077592044A01008	SALES ORDER CREATED	1
8	1078768218A01008	SALES ORDER CREATED, SALES ORDER CONVERTED, SALES ORDER MODIFIED	1
9	1078768524A01208	SALES ORDER CREATED, SALES ORDER CONVERTED, LM_WORK ORDER CREATED, CUSTOMER NOTIFIED, CREATED, CUSTOMER NOTIFIED, CUSTOMER NOTIFIED, PAYMENT EXECUTED, CPS_PLANNED, LM_WORK ORDER UPDATED, SENT FOR FULFILLMENT, CPS_ASSIGNED, RELEASED FOR PICKING, CPS_PICKING, CPS_PICKED, CPS_CHECKING, LM_DISPATCH COMPLETED, CPS_CHECKED, READY FOR PICKUP FROM STORE	1
10	1227619023A01422		1
11	1232246340A01008	SALES ORDER CREATED	1
12	1234240776A01422	SALES ORDER CREATED, SALES ORDER MODIFIED, SALES ORDER CONVERTED	1
13	1234338861A01422	SALES ORDER CREATED	1
14	1234900698A01202	SALES ORDER CREATED, SALES ORDER CONVERTED	1
15	1235342155A01208	SALES ORDER CREATED	1
16	1235554954A01008	SALES ORDER CREATED	1
17	1235986494A01202	SALES ORDER CREATED, SALES ORDER CONVERTED	1
18	1237180854A01422	SALES ORDER CREATED, SALES ORDER CONVERTED, PAYMENT EXECUTED, SALES ORDER COMPLETED	1
19	1237592424A01008	SALES ORDER CREATED	1
20	1237624543A01422	SALES ORDER CREATED	1

Figure 15: Sample bag of activity dataset with all imperfect activities in sequence

By using a large dataset generated by synthesized information relating to actual events, this study aims to see how effective predictive tools are for enhancing performance throughout each phase of an order's life cycle within our chosen country.

### **3.6 Data Collection Procedures**

In the first phase of data collection, many potential sources were identified to find event logs and activity data that relate to order lifecycle processes. There are a lot of issues with accessing real-world event logs like privacy concerns and proprietary limitations, so a synthetic data generation approach was necessary. Through synthetic data generation a bunch of simulated event logs and activity data were created that looked very similar to actual order lifecycle processes. We wanted to make sure there was as much diversity as possible for the purpose of analysis and model development.

When the synthesis was complete, there had to be further efforts put into making the synthesized data more refined. Those efforts included cleaning any missing values or duplicates in the dataset, standardizing formats through encoding categorical variables and scaling numerical features, which is called transformation. Another step with feature engineering means extracting or creating additional features that may help yield insights or enhance models. The study had to take these extra steps to ensure the best sample selected from all those other datasets.

As part of this sample method, process instances representing a wide variety of real-world order lifecycle management situations were carefully chosen from the synthesised dataset. The study's goal was to capture the subtleties and complexity present in order lifecycle processes by purposive sampling to enable thorough analysis and well-informed decision-making. The study established a strong basis for carrying out thorough analyses and obtaining significant insights into predictive process monitoring in order

lifecycle management by carefully adhering to these phases in the data gathering procedure.

### **3.7 Data Analysis**

In the data analysis phase, several analytical techniques were applied to the combined event logs and activity data in order to derive insights and create prediction models. To obtain a thorough grasp of the dataset's structure, distributions, and variable interactions, the investigation started with exploratory data analysis (EDA). Important patterns, trends, and anomalies were found using EDA techniques including data visualization and summary statistics, which gave important insights into the underlying order lifecycle processes.

Initially, exploratory data analysis (EDA) is done to get a good understanding at the structure, distributions, and relationships between variables within the order life cycle. By using techniques such as data visualization and summary statistics, the study can identify key patterns, trends, and anomalies that gave the insight into the underlying order lifecycle processes.

Subsequently, the predictive models were done to forecast order lifecycle outcomes while also being able to know what factors contributed most towards process performance. The machine learning algorithms were trained on these synthesized datasets before being evaluated based on their predictive performance. Once everything was refined enough, a collection of predictive models is obtained that boasted accuracy, precision, recall rates and F1-scores.

Moreover, feature importance analysis was conducted, it helped determine which variables carried the most weight when it came time to predicting things within this

system. Through SHAP values we can quantify how much individual features matter in terms of predictions - which can help us better understand the significance of the features.

The Bag of Activity (BoA) feature was introduced, and it enhanced both overall accuracy and capture of temporal dependencies within this specific lifecycle. BoA feature is generated by aggregating sequences of activities preceding order imperfections, which provided additional context and information to the predictive models, thereby improving their predictive capabilities.

In summary, the data analysis step involved a thorough review of the combined information, utilising sophisticated analytical methods to create precise prediction models and get understanding of the variables impacting order lifecycle management.

### **3.8 Methodological Insights: Unveiling How Selected Approaches Address Research Questions**

This section offers a detailed overview of the research methodologies in use and how they correspond to the stated research questions. In this section specifically, we delve into the approaches that were chosen to study predictive process mining as applied to order management systems and evaluate their effectiveness in addressing these goals. Consequently, this part scrutinizes methodologies used hence shedding light on how each method can be used to answer specific areas of study or meet desired objectives of a given investigation.

#### **3.8.1 Research Question 1: How efficient are methods for pattern recognition and outcome prediction in order lifecycle management using predictive process mining?**

Robust for answering the thesis's research questions about which methods are efficient in pattern recognition and outcome prediction within order lifecycle management were done by this predictive process mining methodology. The said method is a mix of process mining, predictive analytics, and machine learning hence it allows comprehensive examination of event logs as well as activity data from different stages in an order's life cycle. Predictive process mining uses sophisticated analytic tools to discover concealed information, foretell future results and refine organizational workflows. This part discusses how this approach can best tackle each component of the research question; starting from recognizing patterns efficiently upto predicting the right outcomes accurately while considering relevant KPIs associated with OLM.

### ***Recognition Efficiency for Patterns:***

Pattern recognition efficiency is concerned with how well predictive process mining methods can recognize and analyze complex patterns in organizational processes. These techniques use event logs and activity data to automatically detect frequently occurring sequences of activities, deviations from expected workflow paths and execution inefficiencies.

One advantage of predictive process mining being able to handle large quantities of event data is that it extracts useful patterns from them. Information systems generate event logs that can reveal much about the processing, fulfillment, and delivery of orders such as common activity sequences or steps taken; out-of-the-way anomalies during execution; potential bottlenecks indicated by certain patterns etc.

Additionally, algorithms behind these techniques are more advanced than simple rule-based models hence they improve on pattern recognition efficiency greatly. Machine learning models employed learn over time through various historical events thereby

becoming capable of identifying even subtle trends/patterns missed by human beings during manual analysis.

In summary; therefore, through automating the identification of both normal/abnormal activities along order flow lines, pattern recognition efficiency increases knowledge regarding underlying processes involved in realizing an order Outcome Prediction Accuracy – Another important capability provided by Predictive Process Mining relates with accurate forecasting specific outcomes related to an organization's management system for customer orders (OLM). With advanced analytics together with ML algorithms this method can accurately predict different types of future events like; when will we finish processing these orders or how satisfied are our clients likely to be after receiving them? Indeed, traditional approaches have always failed in capturing those hidden correlations which exist between various process steps leading towards outcomes but now thanks to this approach they are uncovered.

Additionally, predictive process mining allows for lively and adaptable prognosis models that can account for changing conditions and context. If new information is constantly being monitored and used to update the forecasted models; then the predictions will be accurate and relevant over time for any organization. This flexibility is especially useful in dynamic business environments with rapid changes in factors affecting results.

### ***Outcome Predictive Power***

Forecasting is an important part of predictive process mining. It helps businesses predict results in terms of their order lifecycle management. With this approach, it becomes possible to predict several outcomes using machine learning algorithms and



advanced analytics. Such predictions can range from when an order will be done to the level of customer satisfaction that will be achieved.

One of the major strengths of predictive process mining as far as outcome prediction accuracy is concerned lies in its ability to capture complicated relationships and dependencies within the order lifecycle. Traditional methods for predicting outcomes are usually based on simple models that do not consider all these factors. However, by analyzing event data predictive models can discover hidden patterns and correlations thus leading to better predictions about future results.

Predictive process mining also allows setting up dynamic adaptive prediction models which can respond to varying conditions or environmental influences surrounding them at any given time point during their lifetime span. As new information continuously feeds into these systems so does their capability shift according to current needs thus enabling organizations make timely decisions based on accurate knowledge about what might happen next around them. This feature becomes more useful especially when dealing with dynamic business environments where many things could change rapidly hence affecting different aspects related with anticipated end states' attainment.

Furthermore, another advantage associated with this technique pertains its ability handle large amounts of complex data sets generated through various operational activities such as those carried out along the supply chain management continuum starting from procurement planning stage up until goods delivery phase among others. Traditional analysis approaches fail here because most often they cannot cope with huge volumes or different types of events occurring within heterogeneous contexts like those typical for multi-national corporations with branches worldwide engaged in diverse lines of production/services provision concurrently across multiple regions/countries.

Outcome prediction accuracy serves as a foundation stone upon which rests all other aspects constituting successful completion any given project under taken by organization involved in process oriented activities within specific industry sector while utilizing modern technologies such as artificial intelligence and machine learning algorithms for data analytics purposes.

### ***Impact on Key Performance Indicators***

In order management, predictive process mining greatly affects key performance indicators (KPIs). The bottom line is that companies can realize tangible benefits in different KPIs by detecting potential problems before they occur and optimizing processes designed to prevent them. This should ultimately result in improved operational effectiveness as well as higher levels of customer satisfaction.

Efficiency in fulfilling orders is one of the main KPIs that can be influenced through predictive process mining. Organizations can predict where congestion points are likely to occur within their supply chain systems using historical event data together with forecasts derived from analytics models. Considering this knowledge, it becomes possible for them to streamline activities among departments or even sites to save time taken to serve clients' needs. Consequently, measures like period required processing orders and lead times for delivery could see a significant increase which would mean faster deliveries too.

Predictive process mining also impacts customer satisfaction rates significantly. By forecasting outcomes associated with an order such as when it will arrive at its destination or whether there is enough stock available, businesses ensure smooth service provision throughout all stages involved in meeting client requirements. What's more is that this technique allows firms foreseeing potential bottlenecks they may face while trying to meet demand hence preventing dissatisfaction among consumers.

Resource utilization is another area greatly affected by predictive process mining. When organizations use predictions about future events related to orders received; management personnel should take note on how best they could maximize available resources especially labour force, machinery etcetera so that every single penny counts towards achieving business objectives within specified timelines thereby ensuring

profitability while keeping costs low if need be. Predictive process mining therefore enables enterprises detect areas where staff are idling around due lack of proper workflow planning or where certain equipment is being underutilized leading to savings through optimal staffing levels among other things.

### ***Best Practices and Recommendations***

Predictive process mining provides insights into implementing predictive process mining in real-world situations. These suggestions come from empirical studies, data analysis, as well as practical experience with predictive process mining methods.

One of the most important best practices is data quality control and preprocessing. Accurate prediction modeling requires high-quality data, so enterprises must ensure that their event logs and activity data are clean, consistent, and representative of the processes underneath them. Data cleaning, transformation, normalization – all these things play an essential role in getting ready for analysis or modeling by preprocessing methods.

Secondly, organizations should choose suitable machine learning algorithms and modeling techniques according to different tasks of predictive process mining. Decision trees could be used sometimes while logistic regression may work better than random forests under some circumstances etc., this is why it's necessary for an organization to evaluate various algorithms carefully before adopting any one algorithm depending on what they want to achieve with their objectives.

Moreover, interpretability and explainability matter a lot when it comes down to predictive models' implementation within predictive process mining solutions. Although complex machine learning algorithms usually have higher prediction accuracy rates but may lack interpretability thus making it difficult for stakeholders such as managers or business owners who rely on those predictions either because they don't understand how

these models work themselves or simply aren't able trust them due too much complexity being involved somewhere along lines between input variables up until output variable has been reached etc.; therefore striking balance between model complexity alongside simplicity where possible remains critical point here.

Also, continuous improvement should be part of every strategy employed while implementing any form of predictive analytics including predictive process mining. Predictive models are not set in stone; they need frequent checks against reality so that over time businesses can continue reaping benefits out of them without realizing losses overtime through becoming irrelevant which was never intended at first place when coming up with idea behind building such models. Thus there should always exist ways through which model validation could be done on regular basis while at same time monitoring performance plus updating models as new data becomes available thus ensuring all predictions stay accurate in face changing conditions thereby remaining valuable tools for decision-making process within any given organization.

To make sure that predictive process mining works best for optimizing order lifecycle management processes, it is recommended that these guidelines and other related ones are followed. The above-mentioned ideas can be used by companies looking forward to driving operational excellence through predictive analytics or those interested in improving their organizational outcomes using this type of approach.

### **3.8.2 Research Question 2: What are the main elements affecting the predictability and accuracy of predictive models in terms of identifying imperfection or inefficiencies in order processing?**

To address the research questions of this thesis effectively, we have used a wide-ranging method which combines advanced machine learning algorithms with predictive

process mining techniques. This allowed us to determine the main drivers that make predictive models identify where there are defects or inefficiencies in order processing more accurately and reliably. The various stages of our approach were:

### ***Data Collection and Preprocessing***

Data collection and preprocessing are basic steps in any data-driven study that must be taken to ensure quality, reliability, and fitness for purpose of datasets. In our case, we created a very detailed dataset on different order processing activities during the order lifecycle management study. Some sources we used include transactional databases; customer relationship management (CRM) systems; enterprise resource planning (ERP) systems among others relevant business applications. These sources aimed at giving us an all-inclusive perspective about the entire process from when an individual places their order until it is fulfilled, and other services offered after delivery.

After this raw data was collected, it had to undergo some preprocessing before being ready for analysis. This involved several tasks whose objective was to improve data quality as well as its uniformity across different dimensions or attributes under consideration. First off, we performed what is commonly known as “data cleaning” by identifying dealing with anomalies such as inconsistencies within records or missing values within them too. Methods like outlier detection, imputation of missing values and dropping duplicates were employed during this phase so that only clean data gets used hence reducing chances for biasness or error during subsequent analysis.

Moreover, standardization together normalizing features contained within the dataset formed another part of data preprocessing step which could not be overlooked here because without doing so one feature may dominate over others simply due to its scale magnitude thereby causing unfair comparisons across variables. Various techniques

such z-score normalization or min-max scaling were applied on numerical features transforming them into standardized ranges thus making sure that all variables are measured in similar units while still retaining their original meaning.

Furthermore, among steps taken prior to modelling included exploratory data analysis (EDA) aimed at gaining insights about dataset's characteristics such as distribution patterns or relationships between different variables which might have helped us choose which ones were more relevant during subsequent predictive modeling exercises. For example, we could visualize how data points are distributed, find out if there is any correlation between two given features and even detect some underlying trends through this early analysis. Therefore, EDA played crucial part in guiding feature selection as well engineering decisions whereby only those attributes judged most informative for our predictive tasks would be selected.

The initial processes of collecting preparing information formed the basis upon which further investigations could be conducted thus providing a clean standardized well understood dataset ready for more detailed exploration and eventual modeling. Without them it would have been impossible to validate results obtained from subsequent stages hence making such findings meaningless since they lack credibility while shedding light on various aspects concerning order lifecycle management and predictive process monitoring.

### ***Feature engineering***

Creating characteristics is a major stage in predicting an ordered process's outcome, especially in the setting of predictive approach mining for order lifecycle management. It consists of picking and developing relevant features from the raw data that would best reflect the underlying structures and associations with respect to results of

processing orders. Feature engineering tries to extract useful knowledge from different sources like timestamps, activity logs, customer demographics as well as historical order information in relation to order life cycle management so that models can predict better.

One part of feature engineering involves converting raw data into a form suitable for use by machine learning algorithms. This might entail transforming categorical variables into numerical ones through methods such as one-hot encoding or label encoding. For instance, types of purchase could be transformed into binary features by recording whether certain methods were used while paying like cash on delivery (COD) = 0 otherwise 1; this enables them being included easily within prediction models alongside other input variables. Also time-related attributes derived from timestamps can be utilized to capture temporal patterns during order processing.

Another crucial aspect of feature engineering is about generating derived attributes that summarize higher-level knowledge hidden within source datasets without losing their interpretability. Here, we may aggregate across different levels or windows in time granularity which gives us insight into how things change over time at various levels within an organization or system around processing orders; this helps identify trends concerning events occurring frequently during business hours only vs those happening throughout day irrespective of whether it falls on weekend/holiday etc. One example could be calculating average processing time per customer (derived attribute) using historical data where each row represents an individual order made by someone unique.

### ***Model Selection and Training***

The selection and training of models is a critical part of the methodology for predictive process mining. During this stage we tried out different machine learning algorithms through careful evaluation to find the ones that were most appropriate for our



needs. We based our choice on their ability to handle complex, high-dimensional data, interpretability, scalability, and generalization capability too.

Among the top models considered was logistic regression, which is a popular linear classification algorithm known for its simplicity and interpretability. Logistic regression models the probability of a binary outcome given one or more predictor variables thus making it suitable for tasks like classifying if an order is perfect or not. Nevertheless, even though being simple in nature, it can still capture linear relationships between input features and target variable thereby providing us with some idea about what may be causing certain things happen during order processing.

Besides logistic regression we also investigated decision trees as well as ensemble methods such as random forest and XGBoost. Decision trees are non-linear models that split up feature space into smaller regions using simple decision rules so they're easy to understand intuitively too. On the other hand, random forest together with XGBoost are both examples of ensemble learning techniques where multiple weak learners (decision trees) are combined to improve prediction performance either individually or collectively depending upon circumstances involved here again these being robustness, scalability, capturing interactions among input features.

Each selected model was trained over preprocessed dataset during training phase which comprised features extracted from order processing activities along with corresponding outcome labels indicating whether there were any imperfections within orders or not at all. For example we used gradient descent optimization coupled up with backpropagation during training neural network-based models while employing standard training procedures such as these two named above only but this time around separately now because one deals exclusively with another type of model altogether whereas for decision trees, random forest and XGBoost ensemble learning techniques were employed

simultaneously to train several weak learners at once so that their predictions could be aggregated resulting into more accurate as well robustness achieved through different ways like bagging where multiple trees are built independently or boosting which involves training them sequentially etcetera.

To ensure that the trained models were reliable enough in terms of generalization ability we used k-fold cross-validation as a measure during this phase. The reason why is because we needed to know how they would perform when exposed to other unseen data sets outside those used during training. Therefore, what happens under such conditions is that dataset gets divided into many subsets also called folds after which each fold becomes test set for some number (k) times with remaining ones being used as training sets respectively too thereby enabling us to get better estimates about true performance across various scenarios considering possibility for overfitting detection.

### ***Evaluation metrics***

To accurately identify which areas, need improvement or development, ensuring that we choose the best evaluation metrics is crucial. Inconsistent with other predictive process mining for order processing systems, this helps us know how much do models talk about faults and inefficiencies in a life cycle of an order. Precision, recall and F1-score are some commonly used measures of performance on predicting models.

A metric which measures the correctness of positive predictions made by a model is called precision. It shows the ratio between true positives (correctly predicted positives) to all instances predicted as positive. As applied in order processing, precision denotes whether a given system can correctly identify those orders that are indeed imperfect or inefficient. In other words, it tells us about false alarms (positive errors). The higher the value for this index, then such alerts should be taken very seriously.

Recall (also known as sensitivity) evaluates if a model can find all relevant cases out of positive ones present in dataset under study by computing its ability to detect them correctly from total number available. For instance, let's assume there are 100 records containing different types of imperfections occurring during different stages within an entire lifecycle where only 70 were identified correctly while remaining 30 went unnoticed. If you were asked what is recall rate? You'll say 70%. This represents how well does our classifier capture every single anomaly happening along supply chain management process without leaving any behind undetected even once throughout analysis period until now?

F1-score takes into consideration both precision and recall by balancing them using harmonic mean formula so that no side gets favoritism over another when assessing efficiency levels achieved in terms of making accurate positive predictions plus capturing all corresponding instances respectively either way around too much emphasis could be placed on either one alone leading to biased outcomes, Therefore, it provides holistic view regarding performance measurement criteria employed during this study among other things too. It should be noted that F1 is an abbreviation for 'F measure' which can sometimes also be referred to as balanced F-score.

There are several evaluation metrics that can be used to gauge the performance of predictive models for order processing such as accuracy, specificity and area under receiver operating characteristic curve (AUC-ROC). These additional measures provide different perspectives on various aspects of a model's ability to address the research questions posed by this study. In general terms, it is important to carefully select appropriate evaluation criteria to get reliable estimates about how well does machine learning perform when applied within supply chain management systems especially ones dealing with procurement activities?

Throughout the machine learning pipeline, cross-validation and hyperparameter tuning are vital steps, especially in building predictive models for complex tasks such as predictive process mining in order processing. These methods are key to ensuring trained models are robust, reliable and can generalize well.

### ***Cross-Validation and Hyperparameter Tuning***

Cross-validation is a resampling technique used to assess a model's performance on a limited dataset. It works by dividing the dataset into multiple subsets or folds where one-fold is used as the testing set while others are used for training. This is done several times having each fold used once as the testing set. Cross-validation gives a more accurate estimate of how well the model will perform compared to a single train-test split by averaging the performance metrics across multiple iterations.

In our case we adopted k-fold cross-validation where k equally sized folds were created from the dataset. Each fold was tested once while k-1 remaining folds were used for training. This created k different estimates of how well the model could perform and by averaging them we got better understanding about its predictive power and ability to generalize on unseen data.

On the other hand, Hyperparameter tuning involves choosing best values for hyperparameters in an algorithm so that it performs optimally during learning phase but not when fed with new examples from test set. Unlike parameters which are learned from data directly, hyperparameters govern behavior of learning algorithm hence must be set before training begins. Examples include gradient descent learning rate, decision tree depth or random forest number of estimators.

During our project we applied grid search and random search techniques in tuning hyperparameters of predictive process mining models. Grid search involves going

through all possible combinations of hyperparameter values within a predefined grid and selecting combination yielding highest performance while random search samples hyperparameter values randomly from pre-defined distribution then evaluates their performance though being less exhaustive than grid search it can still give comparable results in practice since it's often more efficient.

Through cross-validation coupled with hyper parameter tuning we were able optimize on performance of our predictive process mining models so as to ensure their reliability and generalization ability. These methods enabled us to identify best performing model configurations as well as mitigating overfitting which eventually led to more accurate predictions about order imperfections and inefficiencies in the order processing lifecycle.

### ***Model Interpretability and Explanation***

Among all machine learning tasks, model interpretability and explanation are very important, especially in areas where the models' decisions can have significant consequences. In our studies on predictive process mining for order processing; however, this means that it is vital to make certain that predictions made by these models are understandable and explainable since doing so helps build trust with stakeholders as well as easing decision-making processes. Simply put, what this means is being able to know how a certain model comes up with its predictions and justifying these predictions in terms of influencing factors or features which could have brought them about — that's what we refer to as model interpretability and explanation respectively.

Feature importance analysis is one way of improving the interpretability of a model. Here, influential features or variables contributing towards making predictions by the models are identified. Stakeholders can understand which parts of their data on order

processing are more likely to cause errors or inefficiencies if they look at each feature's relevance ranking vis-à-vis others'. For example, it might be found out that types of orders; time taken during processing; and customer demographics account for most anomalies detected in this workflow signifying underlying drivers for inefficiency within this area.

Another method used for improving model interpretability involves SHAP (SHapley Additive exPlanations) values — they provide a universal framework through which any output produced by a machine learning system may be explained comprehensively. Each feature gets assigned its own 'contribution score' according to SHAP value approach thereby indicating its effect on what should be predicted by such models. By visualizing these SHAP values, therefore stakeholders can comprehend relationships between inputs and outputs as well appreciate complex nature of influence exerted by individual input variables upon various outcomes predicted by the model. In other words, there may exist some combinations of features whose joint impact cannot be described independently thereby showing why interaction terms must never miss out when conducting variable selection exercises during exploratory data analysis stage.

Additionally, model explanation implies giving straightforward and intuitive rationales behind predictions made by models particularly to non-technical decision makers who might not have background knowledge in machine learning. This can be achieved by generating explanations that are easy for humans to understand or creating visual representations of the decision-making process adopted by the model. Through presenting results obtained from these systems in a way which is transparent and comprehensible; hence stakeholders being able to trust them as well make informed choices based on insights offered by these algorithms.

In general, trust needs to be established through making models interpretable and explaining their decisions since this will facilitate decision making while at the same time increasing adoption rates for predictive process mining in order processing. Researchers should use methods like feature importance analysis together with SHAP values so as to provide useful information regarding determinants of imperfections within orders received for shipment; thus, leading into more accurate prediction models being used during such processes which eventually result into improved efficiency levels within organizations dealing with large volumes of transactions daily.

### **3.8.3. Research Question 3: How do various machine learning algorithms stack up in terms of how well they predict order lifecycle management outcomes?**

To maximize order lifecycle management, it is important to evaluate machine learning algorithms since they help in forecasting different results of the process. A systematic method for evaluating their efficiency should therefore be used when comparing performances among these various methods. This investigation aims at determining how well different machine learning models can predict order life cycle management outcomes given that we are within such an environment. To do so, we adopt a comprehensive methodology with several steps such as selecting data sets and algorithm types; collecting, cleaning and pre-processing data; selecting features or variables of interest; training models using provided examples or cases as well as evaluating them against each other followed by interpreting outcomes produced thereby. Our intention is that through this structured analysis will provide insights into which algorithms work best in predicting order completion times but more broadly help push forward predictive analytics within this area.

### *Selection of Machine Learning Algorithms*

In selecting machine learning algorithms, a variety of models were chosen to cover the question about order lifecycle management prediction. This process started by identifying predictive analytics and process mining literature which popularly use certain types of algorithms more than others. Logistic Regression, Decision Trees, Random Forests and XGBoost were selected because they are frequently used in different areas due to their versatility and demonstrated success.

Logistic regression serves as a basic model against which other models are measured since it is simple and understandable especially for binary classification tasks. Decision trees on the other hand can capture non-linear relationships between features thus providing insights into how decisions are made during an event. Random forests combine multiple decision trees together (an ensemble learning technique) to make better predictions than any single one can do alone – this was done here just in case there might have been some weaknesses if we only used one decision tree model. Finally, XGBoost algorithm was chosen over others because among all gradient boosting methods currently available for use; XGBoost has proved itself most robust when applied on large datasets where there exist complex relationships between variables being studied.

These selections were made according to what each algorithm is good at doing such as handling both categorical & numerical data types; scalability vis-à-vis big data sets; interpretability w.r.t output produced etc., thus allowing researchers involved in this study compare different methods well while still giving themselves chance of picking out best among them all that can be relied upon for predicting order lifecycle management outcomes.

Moreover, domain experts were consulted with during selection process besides reviewing past works to ensure that the chosen algorithms align nicely with research



goals even though dealing with intricacies surrounding order lifecycle management prediction may prove difficult sometimes. This methodological iteration helped lay strong foundations for selecting appropriate machine learning approaches considering our knowledge about them based on previous usage scenarios followed by rigorous evaluation concerning their respective predictive capabilities under such like settings.

Data collection and preprocessing are vital stages in every machine learning study. This is even more important if the study involves intricate real-life data sets like those of order lifecycle management. In this phase, researchers try to get needed information that represents different parts of an order's life such as creation, processing, delivery among others. Often this means going through various organizational systems for instance transactional databases; customer relationship management (CRM) systems and supply chain management (SCM) platforms just to name but a few.

The next thing after identifying these sources is extracting necessary details and putting them together into one dataset for analysis purposes — it can involve cleaning up too. To ensure integrity during analysis stages where some records might not have been recorded at all or contain wrong entries which need to be either corrected or dropped altogether so as not mess up with our findings later on while doing calculations, we may want impute them appropriately using various methods like mean/median/mode etc. but before that we must check their validity first hence maintaining data correctness always.

The most used method is to make sure that all variables are on a similar scale by converting them into z-scores is standardizing. However, there are situations where this wouldn't work well especially where there are features with different units or ranges because sometimes certain attributes might dominate over others during model training process thus leading us into trouble later when trying to interpret results obtained from fitted algorithms based on such input representations so what do we do about it then? We

normalize instead, Which means rescaling everything between zero (0) & one (1). Here categorical variables could also come in handy here since they need numerical values otherwise machines won't understand anything apart from numbers.

In summary, model training & evaluation depend on what has been done earlier at stage one (data collection/preparation). Through meticulous curation and preparation of the dataset, researchers may augment the accuracy and dependability of the predictive models created for order lifecycle management by guaranteeing that the machine learning algorithms are fed high-quality inputs.

### ***Data Collection and Preprocessing***

Any machine learning study, especially those that involve order lifecycle management systems, heavily relies on data collection and preparation. During this stage, the researchers strive to get useful information which represents all elements of an orders' lifespan like creation, processing, fulfillment, and post-delivery activities. Sometimes it requires one to investigate many places within the organization for different data sources like transactional databases; customer relationship management system (CRM) among others.

The next thing after identifying where these records are coming from involves pulling out relevant bits from each source before putting them together into a single dataset for analysis purposes. This might need some cleaning steps like but not limited to filling in missing values or dealing with inconsistent entries; outliers can also be handled at this point by either correcting them or removing altogether thus leaving only valid points behind so that they do not affect further analysis negatively.

Normalization often comes into play when trying to make sure everything is measured using similar scales or units while standardization helps achieve uniform

distribution throughout the range being considered. These become important especially if there are features having disparate measurement ranges because failure to do so might result in certain variables overwhelming others during model training phase. Moreover, categorical variables could be transformed into numerical representations through techniques such as one-hot encoding or label encoding so as to enable their usage by various algorithms employed in machine learning for analysis purposes.

Another criticality in pre-processing data has got something to do with feature engineering where additional attributes can be generated based on existing ones which will capture more patterns about relationships present within observations constituting some given dataset even though recorded separately from each other till now. For example, raw figures could easily reveal order processing time in reference to delivery lead times alongside customers' satisfaction scores thus providing deeper insights into how orders go through different stages until completion is reached. But then again, this activity calls for domain knowledge plus deep understanding of underlying operational flows so that engineered characteristics end up making sense apart from adding value towards final predictive model.

Data collection together with its preprocessing sets pace for what comes next like model training coupled with evaluation. By carefully selecting and cleaning all records that should form part of a given analysis, one can be sure about supplying MLAs (machine learning algorithms) with good enough inputs which in turn enhances accuracy as well reliability levels when it comes down to order management predictions.

### ***Feature Selection and Engineering***

To improve the forecasting capability of machine learning models, feature selection and engineering are very important, especially in complicated fields like order

lifecycle management. In doing this, several methods are used to recognize variables that matter most (features) and generate new features that represent essential trends or correlations within data.

At one level, we can look at feature selection as a process where we determine if there are any redundant or highly correlated variables by examining how different features correlate with each other. Removing these variables reduces dimensionality in the feature space which in turn prevents overfitting and enhances generalization of the model. Apart from that, another approach is through recursive feature elimination (RFE) technique or lasso regularization which identifies subset of features among others that contribute significantly towards predicting power for a given model.

Feature engineering on its part involves changing raw input data into formats suitable for modelling while extracting useful clues out of available predictors. For instance, categorical attributes may be represented numerically so that machine learning algorithms can understand them e.g., one-hot encoding or ordinal encoding. Similarly continuous predictors might undergo some form of scaling such as min-max scaling or standardization whereby all features will lie on same scale hence having comparable impacts on the model.

Other than the preprocessing steps mentioned above additional knowledge about domain may be used together with expertise in coming up with new features capable of capturing specific aspects within problem context being considered when building a given predictive system. For example, concerning order lifecycle management; it would be possible creating some engineered characteristics around customer demographics/orders volumes/frequency/history patterns etc., which can shed more light into underlying processes involved during this stage. These created predictors enable our

models understand complex relationships hidden deep down within datasets thereby making them much smarter than using only initial set of attributes.

All-inclusive, what should be done is keeping refining until you get it right because both feature selection as well engineering are iterative processes requiring repeated trials before settling down on those most informative or relevant features which enhance performance in machine learning models. By selecting carefully and engineering well, developers can make their algorithms more accurate while also ensuring that they remain interpretable thereby generating actionable insights for order lifecycle management systems.

### ***Model Training and Evaluation***

In the development and evaluation process of machine learning algorithms for predicting the order life cycle management, model training and model evaluation are very important steps. At this stage, selected algorithms learn patterns and relationships between input features and target outcomes by being trained on part of the dataset. Models adjust their parameters iteratively during training so as to minimize error between predicted and actual results. During this phase, various machine learning methods are used including decision trees; logistic regression; random forests among others like XGBoost which is applied to different modelling techniques to know what works best for this task.

Afterwards when we have finished training our models then they need to be evaluated using another portion from the dataset that has not been used in training. Evaluating models with unseen data helps us understand how generalizable these predictions can be or rather it serves as an indicator to show their ability in predicting order life cycle management on new observations. This is done quantitatively by

calculating such performance metrics like precision, accuracy, recall and F1-score for each algorithm. Additionally cross-validation technique may also be used at this point so that we get robustness in our evaluation results by avoiding over-reliance on peculiarities caused by splits of datasets.

There are several things which must be considered while evaluating a model since there could exist factors that can influence the algorithm's performance alone depending on quality representativeness complexity hyperparameters choice etcetera but not limited too. Therefore, one should compare different performances systematically analyze why certain approaches failed where others succeeded thus allowing us understand capabilities shown by various models towards making predictions.

Furthermore, interpretation of evaluation results goes further than simple performance metrics as it tries to bring out hidden forces behind predictive abilities possessed by these artificial intelligences. Feature importance analysis is one way through which we can achieve this because it tells us about those input variables that greatly affect expected values thus shedding more light onto what drives order lifecycle management processes. Additionally, visualization tools like ROC curves or confusion matrices may help reveal trade-offs between different evaluation measures while indicating areas for potential improvements.

In summary, model training and evaluation are an integral part of machine learning in order lifecycle management prediction. Researchers can therefore identify the best predictive models by testing them against each other under very strict conditions which will also enable to make good decisions about what works where when why how, but this should be done within a given period.

### ***Interpretation of Results***

Interpreting the results of the study's evaluation of machine learning algorithms in predicting order lifecycle management outcomes is such a big deal because it helps us understand more about the findings' implications and applicability. There can be many key observations made by having a deep look into performance metrics as well as those gained from insights during model evaluation.

Initially, these results give us valuable input as to how good at prediction various parts of an order life cycle each algorithm is. Quantitative measures like accuracy, precision, recall and F1-score can help assess the predictive power of any given model. For example, logistic regression might have shown high accuracy in predicting some outputs whereas decision trees or XGBoost may have had better precision or recall for certain classes within the dataset – knowing this difference will be important when picking out what works best with which situations.

For another thing, interpreting outcomes requires investigating what makes machine learning models perform well or not so much. This involves looking at feature importance; understanding decision boundaries created by each algorithm; spotting patterns/trends that could affect predictive accuracy in data etc.. e.g. there could be features tied up with time taken on processing orders or customer demographics which greatly impact their ability to predict while others don't do anything at all.

Additionally, weighing up between complexity & simplicity vis-a-vis accuracy becomes necessary during result interpretation phase because one needs to choose an appropriate method considering trade-offs involved. Random forests/XGBoost are known for higher predictive power but they are also computationally expensive since they require lots of parameter tuning whereas logistic regression being less complex may fail capturing intricate relationships within information.

Furthermore, we must interpret our findings from a practical standpoint i.e., how does this work outside? Therefore, organizations should know which machine learning algorithms work best in predicting different outcomes along the order's lifespan. For instance, optimizing resource allocation improving customer service processes; detecting areas for process optimization automation etcetera may form part of decision-making processes within organizations based on this knowledge.

Ultimately, through interpretation can reveal all these things about various approaches towards forecasting order management results using machine learning systems within an enterprise setting.

Overall, the research has given an intense and organized examination of machine learning algorithms during the order lifecycle management prediction by following this approach. Therefore, it considered all possible advantages as well as disadvantages inherent in every method which in turn helped gain insights into where they can be used most effectively or have greatest influence on real life situations.

#### **3.8.4. Research Question 4: How will putting predictive process mining technologies into practice affect customer satisfaction and organizational performance?**

To address the study topic about the influence on customer satisfaction and organizational performance, predictive process mining technologies were selected as the approach. This helped in numerous ways.

##### ***Data-driven insights***

Data-driven insights are valuable knowledge and understanding that come from analyzing large amounts of data with advanced analytical methods. Within predictive



process mining, these insights are achieved by examining event logs and activity data from order lifecycle processes. These records consist of details about what happens when an order is made starting with the first step to the very last.

Such datasets can also be studied with machine learning algorithms which will then expose patterns, trends, or relationships inherent in them. Predictive process mining uses these algorithms for instance to detect usual sequences of events that result into successful completing orders as well as where they deviate from this norm pointing inefficiencies or mistakes within a system. Besides correlating activity information against customer feedback or satisfaction scores helps companies understand which factors affect consumer perception towards their brands and loyalty therefore making it possible for them to know what matters most in business success today.

In other words, data driven insights are derived through predictive process mining thus enabling organizations make informed choices so that they could better their performance levels regarding customers' needs. For example, if analysis shows that one stage in an order life cycle tends being delayed or having many errors occur during its execution; efforts should be directed towards such points of weakness while streamlining things elsewhere along the chain Similarly when some activities have shown significant impacts on client satisfaction then resources must be channeled towards ensuring those particular tasks are handled first before any other part is optimized for maximum output.

To sum up, with predictive process mining we now have more detailed information about our procedures, clients as well as performance indicators than ever before which can help us realize continuous improvement opportunities while at same time reducing risks involved hence promoting innovation-led growth within all sectors of economy.

### ***Proactive Decision-Making***

Proactive decision-making is the capacity of organizations to anticipate and address potential problems before they arise or disrupt business activities adversely. In the predictive process mining setting, such a proactive choice is supported by making use of findings that are made from analyzing extensive sets of event logs and activity data related to different business processes like order lifecycle management.

Using predictive process mining, establishments can predict future results concerning order fulfillment customer satisfaction levels, and organizational performance indicators. Historical information may be examined through employment machine learning algorithms to detect patterns within it which would then help an organization identify potential pitfalls ahead of time while they are still easy to deal with.

For instance, with the aid of predictive process mining; one can predict when an order will be completed, where there might be bottlenecks in fulfilling orders as well as estimating how satisfied customers would be based on these predictions. Considering this knowledge predicted by prediction driven procedures for handling tasks organizations can distribute their resources more effectively optimizing systems and taking corrective measures necessary for ensuring timely delivery of orders thus improving customer experiences too.

Furthermore, strategic planning also falls under proactive decision making that has been enabled by means such as predictive process mining. Organizations should identify trends in customer behavior & preferences so that they can know what products or services will be demanded most in future markets hence being ready for any changes which may occur within them while remaining competitive against other firms operating within same industries during different periods when demands shift due to various

reasons like technological advancements among others thereby leading towards overall success over extended periods.

Proactive decision-making enabled through predictive process mining goes beyond just increasing operational efficiency but rather long-range planning. Customer habits are not consistent therefore companies must always stay prepared by adjusting themselves according to new market requirements whenever they arise earlier than rivals do since this approach ensures continuity even if needs change frequently over time because consumers keep evolving their tastes continuously.

### ***Continuous improvement***

Continuous improvement is the process of incrementally upgrading organizational systems, products or services over a period of time. Continuous improvement in predictive process mining involves analyzing historical data again and again, refining predictive models and making changes to maximize operational efficiency.

Predictive process mining also enhances one part of constant advancement which is order lifecycle management process. They do this by checking event logs and activity records which helps in detection of places where the performance can be improved due to slow down caused by too much work or even wrong task assignment that may lead to dissatisfied customers or poor organizational performance. The company should use this information as an opportunity for targeted intervention that will streamline operations, hence increasing general productivity.

Additionally, what makes perpetual enhancement possible are iterations which take place during predictive process mining. The more information organizations get and the better their models become in predictions; the more they can test how well these models work against different situations thus adapting where necessary. Such kind of an

iterative approach enables enterprises to respond quickly to market shifts brought about by changing needs among clients as well as emerging trends within various industries.

Another thing about continuous improvement has got something to do with creating a culture of learning new things all the time coupled with innovation within any given company setting itself apart from others through its ability to use technology such as predictive analytics based on processes involved while doing business like mining data sets for patterns so as identify areas requiring attention towards achieving customer satisfaction levels never experienced before. Organizations need not fear failure but rather embrace it when using methods like these because failure leads success eventually if taken positively.

In conclusion, what we have learnt is that continuous improvements are very essential if any organization wants to remain relevant amidst today's fast changing business environment as brought about by predictive analytics based on processes involved while doing business like mining data sets for patterns so as identify areas requiring attention towards achieving customer satisfaction levels never experienced before.

### ***Evaluation of Performance***

Any predictive modeling process requires the evaluation of performance. In this thesis, it served as a method for assessing the appropriateness of predictive process mining methods in customer satisfaction and organizational performance prediction. Accuracy and reliability were ensured through intensive testing and validation of predictive models with different measures.

To evaluate, the dataset was divided into two sets which are training set and test set. The training set was used to develop the predictive models while the test set was kept

separate so that model's ability to perform on unseen data can be evaluated. This splitting avoids overfitting thereby giving a more realistic indication of how well these models should generalize to new data.

Numerous performance metrics were employed during the assessment phase such as precision, recall or F1-score among others. Precision refers to true positives divided by all positive predictions made by a model whereas recall is proportionate with true positives recognized out of all actual positive instances by that same model while F1-score gives an evenhanded measure for evaluating any given model's effectiveness.

Other than these there may have been some usage like cross-validation or ROC curves depending on what exactly was required by the study concerned. Cross-validation includes breaking up data into several subsets then training a subset each time before using it to estimate how well would such a model perform with different samples from those available.

Moreover, we use ROC (Receiver Operating Characteristic) curves which plot true positive rate against false positive rate at various thresholds thus enabling them to select an optimal balance between sensitivity and specificity based on where their curve lies relative to other points along its path.

Evaluation of performance in predictive process mining therefore involves many different measurements as well as methods since if one wants their predictions accurate, they must also be reliable plus robustness can't be underestimated either. So, this kind rigorous approach should be taken when evaluating any prediction model so that people can trust what it says and act accordingly too.

In general, the method for predictive process mining that was chosen allowed us to understand how technology implementation affects customer satisfaction and organizational performance systematically. This approach involved using actual data and

advanced analytics to enable evidence-based decision making as well as strategic planning aimed at achieving favorable results for both the customers and organizations.

**3.8.5 Research Question 5: To increase productivity and lower mistake rates, how can companies successfully incorporate predictive process mining into their current order management systems?**

Companies are seeking ways to improve productivity and decrease the number of mistakes they make. They do this by using predictive process mining as an opportunity for enhancing their order management systems. Predictive process mining can expose inefficiencies, forecast results of processes and facilitate proactive decision making through its use of complex analytics and machine learning technologies. The purpose of this part is to explain why did we choose such methodological strategy while answering a question about how can predictive process mining be successfully integrated into current order management systems?

***Identification of Inefficiencies***

Predictive process mining focuses on finding inefficiencies in existing order management systems, among other things. This method uses many different strategies for studying historical event data and figuring out the processes that lie beneath them. One of these methods is process discovery; this entails automatically generating process models from event logs which show how orders move through a system by indicating what activities happen in what sequence and depend on each other.

By applying algorithms for process discovery to event logs, organizations may find hidden patterns or bottlenecks where they deviate from expected flow. For example, it might become apparent that some types of orders get stuck at certain stages causing

fulfillment delays. Also, outliers can be identified through process mining techniques which are cases differing greatly from average and thus represent either potential inefficiency or error.

Another technique employed by predictive process mining is conformance checking whereby actual observed behavior during a business operation is matched against pre-specified models or sets of rules. Through this comparison an organization can establish how close its processes come to meeting desired standards or benchmarks. In addition to pinpointing areas with poor compliance as well as those where expected behaviors differ notably from what was seen conformance checking also reveals places where improvements could be made due to inefficiencies being detected.

Furthermore, temporal aspects about executing processes such as waiting times between different stages in production cycle can also be analyzed alongside resource utilization levels within each stage etcetera while undertaking predictive process mining exercises. Such metrics should be quantifiable across various versions of a given procedure thereby enabling enterprises identify operational efficiencies which would result into faster throughput times reduction of idle periods among others so that materials do not wait too long for processing when there is little work done on them before moving forward.

In general terms, therefore, predictive-process-mining must enable businesses discover their weak points regarding systems used for managing orders thereby helping them make their processes better and faster too.

### ***Prediction of Process Outcomes***

Being able to predict the result of a process is perceived as an important part of predictive process mining since it gives organizations insights into future performance

and behavior. This approach involves the use of event logs from the past as a training set for machine learning models that can predict different aspects of order management process. Predictive process mining detects patterns or trends among data thus allowing organizations to forecast major outcomes of their processes with great precision.

One advantage is being able to know when orders will be finished using management systems by predicting process results. What this means is that predictive models estimate how much time it will take to complete any given order based on historical processing times together with other factors like resource availability and complexity levels among others. With such information at hand, companies can do proper scheduling, allocate resources better and manage customer expectations in terms of delivery timelines.

Predictive methods in process mining also help in estimating other critical results including fulfillment rates for orders placed within certain periods as well as resource needs required during such periods. For instance, a predictive model would make use of data on inventory levels over time together with workforce capacity utilization rates so that it predicts whether there are enough resources available to fulfill all orders within specific deadlines. Hence, through this approach bottlenecks can be identified early enough before they occur thus ensuring smooth running of operations throughout order fulfilment stages.

Furthermore, by telling you how likely an accurate prediction about completing orders accurately and on-time can be achieved with respect to customer satisfaction levels; predictive analytics could help improve insight into what might make customers happy beyond just getting things right every now and then. The model may consider such aspects like punctuality in delivery; feedback received from clients concerning their experience during ordering products/services etc., then come up with probability



estimates showing chances for meeting different client needs relating to these areas. In light of this knowledge, it becomes possible for companies to invest more efforts towards meeting those requirements which are most likely going enhance overall satisfaction levels among various segments served by them.

In summary, the ability to predict process outcomes through predictive process mining enables enterprises to optimize their order management systems, make better decisions and improve customer experiences. Organizations can anticipate behaviors or results with high accuracy rates during different phases of an order's life cycle across the supply chain by using historical records along with advanced analytics methods such as machine learning algorithms based on decision trees that allow one estimate what might happen given certain conditions. They are thus able to achieve efficiency gains along this entire process through accurate forecasting coupled with appropriate response planning whenever necessary.

### ***Integration with Existing Systems***

One of the most important things in predictive process mining is fitting it into current order management systems. In other words, this means that the predictive process mining methodology must work well with an organization's existing infrastructure, data sources and operational processes. Such should be done without any noticeable changes to routine activities as well as making sure that all possible outputs are achieved.

To ensure flexibility, tools used in predictive process mining should be able to handle different formats of data from various sources. For instance, transactional databases within an enterprise can provide necessary information for events logs which are also part of this system while customer relationship management (CRM) platforms might have some useful inputs too. Therefore, failure such as failing to integrate these

components may lead us not being able optimize our order processing methods because we would lack enough information.

Still on integration with current systems but now focusing more on decision-making frameworks employed by order management systems; another approach involves matching mined predictions against decisions made during different stages of processing customer orders. Normally insights got from forecasts should always be accurate enough hence coming up with models or findings that work well under specific conditions could greatly affect how users interact with them while using software interfaces like those found behind user friendly programs meant for managing sales activities among others may help realize this goal too either through reports displayed on screens after logging into such applications or just creating alerts whenever necessary thus enabling staff members take required action based on what they anticipate happening next.

Integration with existing systems also requires considering interoperability issues between various analytical tools and technologies deployed across different departments involved in managing sales transactions; departments dealing with supply chain management may require sharing certain sets of ordered goods' information needed when analyzing patterns associated with demand forecasting so as create holistic approaches towards optimization processes. An example here is where one needs use data cleansing procedures before carrying out any form calculations related quantities purchased at each stage along value chain starting from point purchase downstream side until reaching end user level.

Besides above, it should be noted that successful integration with existing systems will always be faced by some technical challenges which include but not limited to data preprocessing, cleaning as well transforming them into required format suitable for carrying out predictive modeling. Therefore, such tasks must take care of different types

of inconsistencies that may arise during this stage besides equipping pipelines used during integration process handle various forms of input data effectively, so they are ready use with methods employed within predictive process mining.

To summarize, incorporating predictive process mining in order management systems can only succeed if everything is aligned properly starting from infrastructure up decision-making levels.

### ***Continuous Improvement***

In business process management, continuous improvement is crucial, and predictive process mining is an essential part of this ongoing development within companies. Through incremental changes over time, organizations can make their systems better so that they can achieve more efficiency levels as well as higher quality outputs. Continuous improvement in order management systems involves continuously identifying bottlenecks inefficiencies or errors with a view of optimizing the process of fulfilling orders while at the same time improving customer satisfaction.

Continuous improvement relies on historical event data analysis through predictive process mining which gives insights to businesses. These insights help companies know where exactly they need to improve by finding patterns trends and deviations from normality within order management processes using machine learning algorithms combined with advanced analytics techniques. Such findings form basis for making informed decisions aimed at addressing those areas identified through appropriate interventions.

Another thing that can be done under continuous improvement using predictive process mining is monitoring real-time performance of a process. Organizations should compare incoming event data against set standards or benchmarks plus performance

measures continuously and frequently so as not only see when things have gone wrong but also act immediately towards correcting them if necessary For instance; sudden alerts may be raised when processing times take too long or there occurs rapid rise in errors during order placement thereby prompting stakeholders' attention to investigate what went amiss before taking corrective actions.

Moreover, it supports decision making based on facts rather than intuition thus enabling data-driven prioritization imitative selection by organizations Quantification allows organizations evaluate their efforts effectiveness in terms of fulfillment accuracy cycle time etc., hence leading into more refined approaches being adopted while still fostering evidence-based learning cultures across all levels within an organization where actors are encouraged experiment with different options depending on actual experience gained from various sources including past events depicted statistically

Additionally anticipatory capability provided herein can be regarded valuable since it allows businesses anticipate future problems before they occur Predictive models predict where blockages might happen next hence giving firms opportunity come up with strategies prevent such situations from happening or even prepare for worst case scenarios in order not disrupt smooth running of fulfillment activities The above approach not only strengthens operational flexibility but also builds trust among clients and partners towards consistent delivery excellence reliability by an entity.

In summary, predictive process mining is about continuous improvement and never about being perfect at any given time. Companies should therefore capitalize on data-driven insights through advanced analytics capabilities so as to keep refining their order management systems while striving for operational supremacy within today's ever changing business landscape.

Ultimately, the selected strategy for predictive process mining provides a strong foundation to answer the research question on how businesses can effectively incorporate predictive process mining into their current order management systems. This is capable of revolutionizing order management practices and fostering organizational success if it identifies inefficiencies, predicts process outcomes, integrates with existing systems as well as promotes continuous improvement culture. In conclusion, the preferred methodological approach of forecast-oriented procedural excavation offers a sturdy structure for tackling the study issue of how organizations can efficiently fit this technique into their present system of processing orders. The potentiality to transform practice in managing orders and bringing success at an enterprise level lies in its ability to identify where things are not working correctly; what may happen next steps concerning processes based on historical data; connecting various methods together so that they become seamless whole while still being continuously improved upon through habits engrained within people's mindsets toward work always done right first time every time without fail or fear but instead with love towards learning from mistakes made along our way.

### **3.9 Research Design Limitations**

When determining the research design, it's important to understand the limitations that come with the approach. The first limitation is that of synthesized data. To simulate real-world order lifecycle processes, this research created a sample dataset. However, this dataset might not fully duplicate the complexities and nuances of actual operations environments. This means that whatever findings we have may only be relevant for our study.

Another limitation comes from scope constraints on both depth and breadth of analysis. Focusing so much on the order lifecycle management, certain aspects of broader business context like external market dynamics or organizational factors were put aside. While this wouldn't affect the insights gotten for order lifecycle processes, other contextual factors that could influence operational performance weren't considered.

Overall, while our research offers valuable insights into predictive process mining for order lifecycle management, it's necessary to address and recognize these limitations to ensure the validity and reliability of the findings.

### **3.10 Conclusion**

In this section, we outlined research design which consists of Data Collection Procedures, Preprocessing Steps, Machine learning models and Study Limitations.

The design involves a synthetic dataset made up of various global contexts through an order lifecycle process. Leveraging this dataset, we conducted Cleaning and transformation & Feature selection which helped us prepare it for analysis. We have created a “Bag of Activity” features from the activities present in the dataset We then used machine learning models on the structured event logs dataset to compare their performances. Additionally, Feature Importance and SHAP values identifies key contributors within Order Imperfections at different stages in its lifecycle; giving us insight as to what causes problems within each stage.

The sample was made up of event logs from certain countries that included timestamps and order lifecycle activities, whereas the population of interest included order lifecycle processes with event logs and activity data globally. It's important to remember, too, that the dataset's synthesized character could make it less able to accurately capture the intricacies of the actual world. Moreover, the study's breadth may

limit the analysis's depth, especially when it comes to external contextual elements that affect operational effectiveness.

But even with these limitations, the methodology employed still offers valuable insights into predictive process mining for order lifecycle management. Recognizing and addressing them will allow researchers and organizations to get clear interpretations of their findings which in turn ensures a better understanding. Future studies should take a more thorough approach to addressing these constraints to further our understanding of order lifecycle management methods and how they affect organizational performance.

## CHAPTER IV

### RESULTS

#### **4.1 Research Question 1: How do various machine learning algorithms stack up in terms of how well they predict order lifecycle management outcomes?**

In addressing the research question, the research was done. A lot of real work data was analyzed with elaborate machine learning algorithms to come up with more profound comprehension about processing, completing, and delivering orders. The emphasis of this study was on finding out the correctness, exactness, exhaustiveness as well as usefulness (accuracy, precision, recall and F1-score) of each model using both train and test sets. It is hoped that through such an investigation into predictive process mining; light will be shed on its applicability in practice not only for efficiency but also decision making as far as operational activities during different stages of an order's life cycle are concerned.

In the discipline of order lifecycle management, the predictive power of spotting flaws ahead in time is highly valuable for businesses. It enables proactive intervention, streamlined operations and finally, better customer satisfaction. In this regard, machine learning models provide a way to analyze large datasets that can help to predict such issues which humans might have missed. Nevertheless, how effective these models are will depend on factors like algorithm choice, feature selection and dataset quality.

The Logistic Regression, Decision Tree, Random Forest and XGBoost algorithms were chosen as representatives for this evaluation since they depict different machine learning approaches with their own strengths and weaknesses. For instance, Logistic Regression is popularly known because it has simplicity and interpretability making it applicable where model transparency is important. However, decision trees present an intuitive approach to decision making but may fail when there are sophisticated relationships within data. On the other hand, random forests use ensemble learning to



increase predictive accuracy while XGBoost stands out due to its scalability and performance with big data.

We want therefore that subjecting these models through classification reports should inform how they performed in different metrics. Precision, recall and F1-score metrics shed light into how well the models classify both defective orders and perfect ones as well. A look into these metrics across training set versus test set enables us to assess generalization ability of the models as well as potential overfitting problems.

Below are the descriptions of each model along with its performance report to understand the potential of the model in Order Life Cycle Management and helps in understanding the contribution of various machine learning algorithms in prediction of imperfection of orders in OLCM:

#### **4.1.1 Logistic Regression**

Logistic regression performed well compared with other ML models especially regarding precision and recall rates. It might not have had highest accuracy among all tested options but being simple-to-understand & interpretable can be good enough reason why many organizations use it when involving any transparency or explainability features into PPMS process.

However, there are several advantages that come with logistic regression making it an effective tool in this type of predictive analytics. Linearity allows easy interpretation of the feature coefficients thus enabling stakeholders to comprehend what brings about imperfections in orders. Logistic Regression, therefore, becomes an attractive modeling option for organizations handling large datasets used in their order management system as it is computationally efficient and less prone to overfitting.

On the other hand, there are some drawbacks with respect to Logistic Regression. The linearity assumption of this model can reduce its performance when dealing with complex and non-linear relationships among variables. Moreover, Logistic regression assumes independence amongst features which is not true in real-world order management process that involves interconnections between activities.

Figures 16 and 17 present training and testing classification reports of logistic regression. The training classification report (Figure 16) shows that for category 0 (imperfect orders), precision is equal to 0.79, recall is equal to 0.94, while F1-score is 0.86; while for class 1 (perfect orders), these quantities are: precision =0.96, recall=0.83 and F1 score =0.89. The weighted average F1 score is about 0.88 which suggests a balanced performance across all classes. Similarly, in the testing classification report (Figure 17), the precision, recall, and F1-score for class 0 are given as follows: precision =0.97, recall=0.90 and f-1 score=94; for class one these are respectively defined as: precision =0.63, recall=0.87&f-1 score=73 with a weighted average F-1 score of approximately.

In conclusion, Logistic Regression as a predictive modeling approach is valuable in life cycle tasks for order management. Besides, its simplicity, interpretability and good performance make it favorable for organizations that wish to employ PPMTs for enhancing operational efficiency and customer satisfaction. Nevertheless, transparency and efficiency could still make logistic regression remain an important tool in predictive analytics toolkit even if it does not always outperform more sophisticated models.

Training Classification Report				
	precision	recall	f1-score	support
0.0	0.79	0.94	0.86	133915
1.0	0.96	0.83	0.89	196395
accuracy			0.88	330310
macro avg	0.87	0.89	0.87	330310
weighted avg	0.89	0.88	0.88	330310

*Figure 16: Logistic Regression: Training Classification Report*

Test Classification Report				
	precision	recall	f1-score	support
0.0	0.97	0.90	0.94	53502
1.0	0.63	0.87	0.73	10457
accuracy			0.90	63959
macro avg	0.80	0.89	0.83	63959
weighted avg	0.92	0.90	0.90	63959

*Figure 17: Logistic Regression: Test Classification Report*

#### **4.1.2 Decision Tree**

The Decision Tree model, renowned for its simplicity and interpretability, was one of the main machine learning techniques assessed in this study to predict imperfect orders within the order management lifecycle. This model is based on recursive partitioning of the feature space into smaller subspaces using different attributes value then finally creates a tree-like structure where each internal node makes decisions based

on one specific feature. The Decision Tree model has several advantages such as being able to handle both types of data (numeric and categorical), dealing with missing values, while having an easily interpretable rules for decision making purposes.

On evaluating performance of Decision Tree Model, we noted promising results across various classification metrics. In terms of precision, recall and F1-score, the Decision Tree model posted good scores implying that it is adept at accurately classifying imperfect orders. It had high precision in identifying imperfect orders such that most of the predicted imperfect orders were infact correct. Additionally, the recall rate reflected by the Decision Tree model was ok indicating that it can capture quite a considerable proportion of actual imperfect orders in our dataset. The trade-off between these two quantities illustrates how robustly it identifies imperfect orders minimizing false positives and false negatives.

Comparatively, amongst other machine learning models investigated in this research like Logistic Regression, Random Forest and XGBoost, Decision Tree model performed well. Each came with its own strengths and limitations but when it comes to simplicity and interpretability; decision tree stood out thus making it more suitable especially when transparency matters most since its explanation is needed behind any action taken by an organization's management team. However simple in structure compared to others like random forest or xgboost ensemble methods decision tree demonstrated similar predictive accuracy during prediction process.

Figures 18 and 19 representing a consistent performance achieved by both sets in the training and testing classification reports done by using decision tree model illustrates this idea very well. In the training classification report (Figure 18), the values for accuracy rate, sensitivity rate or true positive rate (TPR) and F-Measure equal to precision are all greater than those reported in table two except TPR which drops from

.91 to .84. For instance, in figure three if we consider instances predicted as class label zero then out of thirty misclassified instances eight were actually labeled zero. The f-measure averaged across categories was .89 i.e., almost similar with f-measure value obtained when considering only class zero. Setting a threshold at which positives would be selected resulted into an optimal alternative with respect to maximizing sensitiveness.

Similarly, in the testing classification report (Figure 19) the following precision, recall and F1-SCORE are obtained for class 0: precision =0.97, recall = 0.89 and F1-Score=93; whereas, for class one, the results are: precision =0.61, recall=0.88 and F1-score=72 thus the average weighted F-1 score equals .9.

However, note should be taken that Complex datasets containing many features with complex decision boundaries pose challenges overfitting problem particularly for the decision trees even though they are relatively flexible models. In order to minimize overfitting risk under such situations optimal performance needs to be guaranteed by careful hyperparameter tuning and pruning techniques. There are also cases where the Decision Tree model may not generalize well to unseen data, especially when the dataset is highly variable or contains noise or irrelevant features. Despite these limitations, decision tree remains a useful tool for predictive modeling in different contexts, balancing between interpretable models and accurate predictions.

Training Classification Report				
	precision	recall	f1-score	support
0.0	0.80	0.97	0.88	133915
1.0	0.97	0.84	0.90	196395
accuracy			0.89	330310
macro avg	0.89	0.90	0.89	330310
weighted avg	0.90	0.89	0.89	330310

*Figure 18: Decision Tree: Training Classification Report*

Test Classification Report				
	precision	recall	f1-score	support
0.0	0.97	0.89	0.93	53502
1.0	0.61	0.88	0.72	10457
accuracy			0.89	63959
macro avg	0.79	0.88	0.82	63959
weighted avg	0.91	0.89	0.89	63959

*Figure 19: Decision Tree: Test Classification Report*

### 4.1.3 Random Forest

One of the machine learning models used in this study to predict imperfect orders within order management lifecycle was Random Forest, a powerful ensemble learning method. During training phase, random forest algorithm builds several decision trees and combines their predictions to produce the final output. Such an approach allows Random

Forest to handle complex datasets and alleviate overfitting making it particularly suitable for our task of prediction process mining.

During the evaluation phase, Random Forest showed a robust performance in various classification metrics shown in Figure 20 and 21 such as recall and F1-score. During training, for both imperfect and perfect orders, it indicated a high level of precision with precision scores of 0.94 and 0.78 respectively. Similarly, Random Forest had strong recall scores for both perfect orders (0.93) implying its capacity to correctly match relevant instances from the dataset and imperfect orders (0.82). When both precision and recall are considered, the F1-score was computed at 0.88 for imperfect orders and perfect orders attained 0.85 hence this shows that the model fully captures the nuances in order management.

In comparison to other learning algorithms investigated in this paper; Random Forest gave a competitive performance. Logistic Regression and Decision Tree models also came out well although they were inferior to Random Forest on F1 scores for both training as well as test datasets. In particular, in comparison with other models, Random Forest achieved better recall rates for imperfect orders; thus it can predict potential flaws more accurately than those other methods. However, one should bear in mind that performance of RF may differ when we vary some properties of input data set or when we examine different types of order management process complexity.

Results imply that Random Forest is a promising predictive process mining technique for order management lifecycle analysis. With all these above features including ability to deal with complex data sets and giving reliable predictions make it an important model which helps businesses to optimize their order management processes leading to a reduced number of irrelevant ones nevertheless. More studies are still needed along such lines.

Training Classification Report				
	precision	recall	f1-score	support
0.0	0.78	0.93	0.85	133915
1.0	0.94	0.82	0.88	196395
accuracy			0.86	330310
macro avg	0.86	0.87	0.86	330310
weighted avg	0.88	0.86	0.87	330310

*Figure 20: Random Forest: Training Classification Report*

Test Classification Report				
	precision	recall	f1-score	support
0.0	0.97	0.87	0.92	53502
1.0	0.56	0.86	0.68	10457
accuracy			0.87	63959
macro avg	0.77	0.87	0.80	63959
weighted avg	0.90	0.87	0.88	63959

*Figure 21: Random Forest: Test Classification Report*

#### **4.1.4 XG Boost**

In our study, we tested the robustness of gradient boosting algorithm as a prediction tool for imperfect orders throughout the order management lifecycle. XGBoost was therefore discovered to be a strong predictive model in our study that produced promising results for forecasting imperfect orders within the order management lifecycle. Specifically, through extensive evaluation we observed that XGBoost exhibited good



precision, recall and F1-score metrics when compared with other machine learning models like Logistic Regression, Decision Tree, and Random Forest.

XGBoost attained a commendable F1-score of 0.7447 during training which reflected its capability to balance between precision and recall in identifying imperfect orders. This means that XGBoost learns effectively from training data by capturing the fundamental patterns and relationships inherent in order management process. Moreover, on operation level accuracy test, it had 90% accuracy rate to discern perfect from imperfect contracts.

Logistic Regression was followed by Decision Tree and Random Forest which were less effective than XGBoost on different evaluation parameters. Another important finding is that though Logistic Regression or Decision Tree possessed good performance metrics; XGBoost persistently outperformed them both in terms of precision, recall and F1 score over trainees and testees. The suitability of XGBoost is further highlighted by its superior performance since it captures complex relationships and non-linear interactions occurring during the entire order management lifecycle.

Moreover, being able to handle imbalances in datasets well is another strength demonstrated by XGBoost commonly encountered predicative modeling tasks. This makes predictions accurate since it has learned from both negative and positive samples hence reducing bias errors occurrence in predicting those instances which are wrongly classified as positives.

Additionally, one outstanding feature about this algorithm is its scalability which makes it an efficient choice when dealing with real-world business applications where large amounts of information need to be processed within extremely tight timeframes covering business hours only. It can also make fast predictions so that timely decisions

may be made based on company's large-scale data sets and their parallelized computations.

However, it is important to recognize that despite its excellent performance, XGBoost has limitations. In the same vein, XGBoost's effectiveness as a predictive model depends on the quality and representation of the training data used. Finally, interpreting XGBoost models might be problematic especially when they are being used to explain intricate decision-making processes to stakeholders.

Figures 22 and 23 exhibit the training and testing classification reports done using XGBoost model. In the training classification report (Figure 22), precision for category zero is equal to 0.79; moreover its recall rate is equal to 0.96 while F1-score is measured as being equal to .87; while for class one these quantities are respectively defined as: precision =0.97,recall=0.83&f-1 score=.89 with a weighted average F-1 score of about .88.Similarly,in the testing classification report(Figure 23), the following precision, recall and F1-SCORE are obtained for class 0: precision =.97 recall=.91 &F1-Score=.94; whereas, for class one, the results are: precision=.65 recall=0.88 and F1-score=.74 consequently the weighted average F-1 score is about .91.

In conclusion, the paper shows that XGBoost is an effective prediction technique for imperfect orders in the order management lifecycle. Its competitive superiority over other models as well as scalability and efficiency make it a strong choice for companies wishing to optimize its order management process and improve customer satisfaction. Nonetheless, further research work would be necessary to fully understand its potential and constraints.

Training Classification Report				
	precision	recall	f1-score	support
0.0	0.79	0.96	0.87	133915
1.0	0.97	0.83	0.89	196395
accuracy			0.88	330310
macro avg	0.88	0.90	0.88	330310
weighted avg	0.90	0.88	0.88	330310

*Figure 22: XG Boost: Training Classification Report*

Test Classification Report				
	precision	recall	f1-score	support
0.0	0.97	0.91	0.94	53502
1.0	0.65	0.88	0.74	10457
accuracy			0.90	63959
macro avg	0.81	0.89	0.84	63959
weighted avg	0.92	0.90	0.91	63959

*Figure 23: XG Boost: Test Classification Report*

Finally, from performance evaluation of machine learning models demonstrated that within such order management lifecycle XGBoost was found out by us as top-performing model in predicting imperfect orders. Its F1-score of 0.7447 was impressive showing higher precision, recall and overall accuracy than other models including Logistic Regression or Decision Tree. In contrast with these findings during testing phase; Logistic Regression scored 0.7330 while Decision Tree had 0.7175 F1-scores

respectively although Random Forest doesn't have predictive accuracy equivalent to that of XGBoost but exhibits good performance with an F1-score of 0.6818.

The models' comparison helps to underline the efficacy of XGBoost in capturing intricate patterns and dependencies within order management. Besides, it addresses imbalanced datasets, scalability, and performance making it a better choice for predictive modeling than other machine-learning algorithms in this setting. Nonetheless, we should not forget that each model has its own advantages and disadvantages and choosing the best one depends on several factors such as dataset characteristics, computational resources availability, interpretability requirements.

When it comes to real-time prediction of order imperfection or perfection in an order life cycle, XG boost gets the preference. Its excellent performance can be put into test when dealing with sophisticated business environment properties like scalability and speed of running systems. With advanced algorithms embedded in XGBoost organisations can predict various situations accurately; helping them take proactive actions early enough to mitigate risks involved throughout different stages of supply chain process.

XGBoost stands out as the optimal choice for predicting imperfect orders within the order management lifecycle, offering superior performance and accuracy compared to other machine learning models. Its robustness, scalability, and efficiency make it well-suited for real-world applications, enabling businesses to optimize their processes and enhance customer satisfaction effectively. Nevertheless, further research is needed on predictive modeling techniques to enhance their effectiveness in addressing emerging issues related to order management.

The benefits of using XGBoost are its ability to handle huge imbalanced data sets efficiently extracting complicated patterns & relationships which are easily read by users.

These techniques also ensure that the model's robustness is enhanced by achieving generalization thus increasing both stability and accuracy during forecasting. Additionally with regards to signal processing needs their flexibility regarding parameter tuning allowing for fine-tuning so as optimize model performance based on specific business requirements or objectives.

Additionally quick training time facilitated by computational efficiency ensures faster training times hence making it easier for organizations to carry out real-time monitoring as well as respond to changes in the order management process. Its scalability allows it to be easily integrated into existing systems and workflows thus enabling organizations to leverage fully the potential of predictive analytics towards enhancing operational efficiency and customer satisfaction.

In conclusion, XGBoost's versatility, performance and efficiency makes it ideal for real-time prediction of order imperfection or perfection in order management. This helps businesspeople to know what is happening throughout their processes by just using XGBoost. As such, this is a significant development in predictive analytics that has resulted into informed decision making hence optimized business results within the contemporary dynamic environment of businesses today.

#### **4.2 Research Question 2: How efficient are methods for pattern recognition and outcome prediction in order lifecycle management using predictive process mining?**

A thorough overview develops when we look at the results of a research question about how effective methods are for pattern recognition and outcome prediction in order lifecycle management through predictive process mining. There are three separate case studies that were analyzed to provide different perspectives on applying predictive

process mining techniques, which help us understand what can be predicted and done more effectively in operations. These case studies cover various industries such as e-commerce, manufacturing, and logistics so they serve as real-life examples where predictive process mining can be used to optimize order lifecycle processes. We can analyze these cases to find out if they show how much does predictive process mining contribute towards pattern recognition and outcome prediction in order lifecycle management thus showing its potential in driving operational excellence and improving customer satisfaction. Having said that let us now take a closer look at every single one of them: methodologies used; predictive abilities observed; implications for future research or practical application within OLM.

Each case study, characterized by distinct case keys, provides insights into the effectiveness of predictive process mining techniques in different operational contexts.

#### **4.2.1 Case Study One**

In case study 1, we are going to focus on the complex order path depicted by the case key 1358405196A01422 with a view of discovering whether the model can predict imperfect order within the order's life cycle management process. This case study takes an in-depth look at how well the model can detect failure points and give operational ideas for optimizing processes.

The essentiality of Case Study 1 resides in its ability to predict high with a value of Predicted Imperfection Probability = 0.9984907, which means that there is a huge probability of imperfection along the order journey. The significance of this prediction stresses that it is effective in identifying all possible hindrances and deviations from an ideal system on order fulfillment.



Figure 24: Figure 24 SHAP Explainer for case key 1358405196A01422

The SHAP value plot Figure 24 reveals important activities that contribute significantly to predicting reflectivity coefficient. “SALES ORDER COMPLETE” and “SALES ORDER MODIFIED” emerge as top activities, explaining their enormous impact on what happens during this journey. By narrowing down to these actions alone, one can know exactly where there are weak points during product lifecycle enabling businesses to anticipate problems before they occur thus enhancing efficiency.

	FeatureValue	Contribution
9	RETURN MONEY SENT = 1	5.625735
68	RECEIVED AT PUP = 1	0.235867
6	ISOM = 1	0.127187
67	CUSTOMER DELIVERY COMPLETE = 1	0.109331
123	CREATED = 1	0.066864

Figure 25: Feature Importance Heat map for case key 1358405196A01422

Furthermore, the heat map Figure 25 that shows the activity contributing factors provides a visual depiction of the activities that are most closely linked to order flaws. High contribution factor activities, such "SALES ORDER COMPLETE," highlight how important they are in creating order journey flaws. Businesses are provided with

actionable insights into areas of attention by this granular analysis, which enables focused interventions to reduce risks and improve overall process performance.

In conclusion, predictive modeling can uncover potential failures during procurement operations management based on Case Study 1 findings. Modern business organizations utilize these findings to improve upon productivity through reduced risk taking by optimizing process efficiencies and reducing delivery time lag thereby raising customer satisfaction levels considerably. This case study reveals how predictive process mining may change current practices of order management and contribute to operational excellence in highly volatile business sphere.

#### 4.2.2 Case Study Two

The case study 2 covers an analysis on another specific order journey within the order life cycle management process and gives a more detailed description of the predictability of this model. The core objective here is the review of case key 1345897825A01288 which involves series actions or activities undertaken throughout the process of fulfilling an order.



Figure 26: SHAP Explainer for case key 1345897825A01288

The order journey begins when the purchase process is initiated and consists of a series of actions designed to make the order fulfilment process easier for the consumer. Crucial events that shape the order's trajectory as it moves through different stages include the fulfilment of sales orders and the receiving of items at the last mile (LM\_RECEIVED AT HUB). One extremely important step in the fulfilment process is



the last mile receipt, which indicates that the purchase will soon be delivered to its intended location.

	FeatureValue	Contribution
114	SALES ORDER COMPLETED = 1	1.750307
127	LM_RECEIVED AT HUB = 1	0.247706
129	READY FOR PICKUP FROM STORE = 0	0.189547
102	SALES ORDER CONVERTED = 1	0.124617
2	288 = 1	0.112690

Figure 27: Feature Importance Heat map for case key 1345897825A01288

The constructed predictive model is utilised to evaluate the probability of order imperfections and predict possible deviations from the predicted order flow. The model carefully examines the order of activities related to the order to determine the likelihood of imperfection by utilising the bag of activity feature and predictive process mining approaches.

With a Predicted Imperfection Probability of 0.16466218, the predictive model estimates the possibility of order imperfection and produces meaningful findings. This probability correctly indicates that the order will not be imperfect.

The SHAP value plot Figure 26 further enables deep analysis of the model’s predictions, giving insights into how individual activities contribute to the predicted imperfection probability. Listed among some major contributors are “SALES ORDER COMPLETED” and “LM\_RECEIVED AT HUB” having contribution values of 1.750307 and 0.247706 respectively. Additionally, another key driver is identified as the

activity” READY FOR PICKUP FROM STORE”, this time showing a contribution value of 0.189547 (Figure 27)

Case Study 2 brings out how well this predictive model can detect imperfections in managing processes within the lifecycle of an order. Advanced machine learning algorithms and predictive process mining techniques help firms understand their dynamics in relation to orders, optimize their operational efficiencies, resulting in maximizing overall customer satisfaction levels.

### **4.2.3 Case Study Three**

The third case study, we examine an unusual situation in the order life cycle management process where an order seemed perfect initially but later exhibited imperfections. This case is with reference to case key number 1360331646A01288. We observed a series of events unfold in the order life cycle Figure 28 by examining the activity flow with timestamps. Key milestones such as payment execution and sales order completion were recorded, indicating the progression of the order towards fulfillment. However, this sequence took a different direction when "SAC CREATED" emerged thereby signifying a critical point of consumer dissatisfaction and initiation of imperfection in ordering.

	<u>_CASE_KEY</u>	<u>ACTIVITY_EN</u>	<u>SYSTEM</u>	<u>EVENTTIME</u>
17593512	1360331646A01288	PAYMENT EXECUTED	ISELL	2023-05-24 18:45:00.000
17593513	1360331646A01288	SALES ORDER CREATED	ISELL	2023-05-24 18:45:02.000
17593514	1360331646A01288	LM_WORK ORDER CREATED	CENTIRO-LM	2023-05-24 18:45:08.000
17593515	1360331646A01288	CREATED	ISOM	2023-05-24 18:45:21.000
17593516	1360331646A01288	SENT FOR FULFILLMENT	ISOM	2023-05-24 18:45:22.000
17593517	1360331646A01288	CPS_PLANNED	CPS	2023-05-24 18:45:26.000
17593518	1360331646A01288	CPS_ASSIGNED	CPS	2023-05-25 04:05:08.000
17593519	1360331646A01288	CPS_PICKING	CPS	2023-05-25 04:09:39.000
17593520	1360331646A01288	RELEASED FOR PICKING	ISOM	2023-05-25 04:09:40.000
17593521	1360331646A01288	CPS_PICKED	CPS	2023-05-25 04:30:58.000
17593522	1360331646A01288	CPS_CHECKING	CPS	2023-05-25 04:43:13.000
17593523	1360331646A01288	LM_DISPATCH COMPLETED	CENTIRO-LM	2023-05-25 05:42:37.000
17593524	1360331646A01288	PICKED	ISOM	2023-05-25 05:42:38.000
17593525	1360331646A01288	CPS_CHECKED	CPS	2023-05-25 05:42:38.000
17593526	1360331646A01288	READY FOR DISPATCH	ISOM	2023-05-25 05:42:40.000
17593527	1360331646A01288	CPS_COMPLETED	CPS	2023-05-25 05:42:40.000
17593528	1360331646A01288	HANDED OVER TO TSP	ISOM	2023-05-25 07:04:56.000
17593529	1360331646A01288	LM_RECEIVED AT HUB	CENTIRO-LM	2023-05-25 12:25:00.000
17593530	1360331646A01288	RECEIVED AT LSC	ISOM	2023-05-25 13:11:45.000
17593531	1360331646A01288	LM_RECEIVED AT HUB	CENTIRO-LM	2023-05-26 04:37:00.000
17593532	1360331646A01288	RECEIVED AT PUP	ISOM	2023-05-26 10:26:48.000
17593533	1360331646A01288	PICKED UP BY CUSTOMER	ISOM	2023-05-26 16:02:52.000
17593534	1360331646A01288	CUSTOMER DELIVERY COMPLETE	ISELL	2023-05-26 16:02:53.000
17593535	1360331646A01288	SALES ORDER COMPLETED	ISELL	2023-05-26 16:02:53.000
17593536	1360331646A01288	RETURN MONEY SENT	SAMS	2023-05-27 07:27:36.000
17593537	1360331646A01288	SAC CREATED	SAMS	2023-05-27 08:26:02.558

Figure 28: Order Life Cycle of Case Key:1360331646A01288

We cut the order exactly before the occurrence of the imperfect activity. Here, we segmented the activity sequence exactly at “SAC CREATED.” Then, this segmented activity sequence was given to the XGBoost model in form of bag of activity feature, enabling a detailed examination of the factors contributing to the imperfection.

The predictive model, leveraging machine learning techniques, showed that there is likelihood of an imperfect orders with Predicted Imperfection Probability coming out

as 0.99677306 Figure 29. Importantly, high probability underscored how significant it was for identified imperfection within order life cycle.

Valuable insight about activities' contribution towards predicted imperfection probability was also obtained through SHAP (SHapley Additive exPlanations) value plots analysis Figure 29. Specifically regarding prediction this included top three influential activities:



Figure 29: SHAP Explainer : 1360331646A01288

1. RETURN MONEY SENT = 1: The most important contributor to predicted imperfection probability was this action which initiated refunding process with its significantly high contribution value at 5.625735 (Figure 30).

2. RECEIVED AT PUP = 1(Received at Pick-Up Point): Its contribution value was lower than previous but still stood out at 0.235867 showing that it means when an order comes to the designated pick-up place (Figure 30).

	FeatureValue	Contribution
9	RETURN MONEY SENT = 1	5.625735
68	RECEIVED AT PUP = 1	0.235867
6	ISOM = 1	0.127187
67	CUSTOMER DELIVERY COMPLETE = 1	0.109331
123	CREATED = 1	0.066864

Figure 30: Feature Importance Heatmap for case key 1360331646A01288

This case study underscores just one example of how the model can detect imperfections in apparently perfect orders. Despite being complex, the Activity-based Predictive Model was remarkably accurate at pointing out and highlighting these shortcomings; hence, it enabled proactive interventions and process optimizations.

To sum up, Case Study 3 highlights how robust this predictive model is to be able to detect discrepancies in intricate order scenarios thereby enabling businesses respond quickly to customer dissatisfaction as well as improving overall process efficiency post haste.

The three case studies ultimately demonstrate, convincingly, how effectively the model predicts imperfections in the Order Life Cycle Management process. By means of predictive process mining techniques and advanced machine learning algorithms, organizations are now able to gain insights into their orders' movements as well as optimize process efficiency while also enhancing total customer satisfaction. These findings validate how predictive analytics can potentially revolutionize current practices of managing orders for quality services which drive operational excellence across today's dynamic business setting.

The methods of pattern recognition and outcome prediction in the context of predictive process mining for order lifecycle management have different efficiencies when applied in various operational settings. Some scenarios may be highly efficient and promising in terms of predictability while others are only slightly efficient but with potential to improve. Accordingly, more investigation should be done on these techniques so that they can be made better thereby increasing efficiency during order life cycle management.

### **4.3 Research Question 3: What are the main elements affecting the predictability and accuracy of predictive models in terms of identifying imperfection or inefficiencies in order processing?**

Our study on ‘what are the main elements affecting the predictability and accuracy of predictive models in terms of identifying imperfection or inefficiencies in order processing?’ has revealed several key findings:

#### ***Analysis of Feature Importance***

This is an important step in any predictive modeling exercise as it helps to understand how much each input variable contributes towards predicting the target variable. Fig 25, Fig 27 and Fig 30 shows the feature importance heatmap for different casekeys. In our case where we were trying to find out what causes order processing imperfections this analysis played a critical role in determining what factors should be considered when modeling for prediction accuracy.

The primary aim under this approach was to identify those characteristics about an order that greatly increase its chances of getting damaged during processing through various stages such as production, packaging, shipping among others. Such factors can

range from customer demographics/history; type or complexity level of goods being handled; time taken by different processes involved in fulfilling one single purchase request.

We then proceeded with data analysis which involved ranking features according to their relevance using different statistical methods till we came up with most significant ones. In doing so we utilized techniques like gini impurity, permutation importance or shap values depending on which machine learning algorithm was employed at that moment.

Feature Importance Analysis gave us a lot of insight into what could cause errors during order fulfillment but there were still more things left unexplained. For example, there seems to exist strong association between certain attributes like complexity & variability which makes them good predictors for errors detection thus leading to delays within supply chain management systems (SCMS).

Another thing worth mentioning is that feature selection process tends not only improve interpretability but also computation time required when building predictive models hence reducing dimensionality while ensuring same forecasting power remains intact. Therefore, through prioritizing influential variables I think we were able come up with simpler yet highly accurate algorithms for predicting imperfections within OPLs.

Moreover, this method enabled us to unearth some hidden patterns behind mistakes made during customer's requests handling process thus revealing complex relationships among different variables affecting order processing outcomes in general (SCM).

In summary it can be said that feature importance analysis played a key role in determining the accuracy and predictability of predictive models with regards to order processing imperfections. This was achieved by identifying the most influential variables

which were used as inputs into various algorithms for forecasting errors at different stages along supply chains management systems or any other business environment where goods/services are being provided on demand basis through E-commerce platforms.

### ***Model performance evaluation***

Predictive analytic model evaluation is very crucial in data science because it shows how useful and reliable machine learning models are for solving real world problems. This research engaged in a wide-ranging assessment of many different machine learning models to establish their competence in forecasting order imperfections within the life cycle of an order. In this assessment, we considered several metrics for evaluation such as precision, recall, and F1-score which help us understand predictive accuracy and robustness of these models.

One major aim behind evaluating various model performances was to contrast between them using training and testing datasets with different algorithms. We used logistic regression, decision tree; random forest etc., as they are known to be some of the best performing algorithms when predicting order imperfection. These methods were trained on one set of data (training dataset) but evaluated on another independent set (test dataset) so that there is no bias during performance estimation.

The evaluation process started by training each model on a subset of the total data then making predictions about unseen instances using those learned knowledge i.e., features extracted from previous examples. After that we compared predicted outcomes against true labels or actual observations to compute different performance measures like f1 score etc. What proportion among all positive prediction made by our system are actually correct? It is called Precision while Recall refers to what percentage among all



actual positive instance present in our database has been predicted correctly as such by our classifier?

We also went further by looking at performance across classes/categories which enabled us to see if there were any disparities or places where improvements could be done better. This means evaluating each algorithm's ability not only across general level but also within specific types or groups based on certain characteristics/features associated with them. For example: How good does method X work under condition A compared against B and C?

From this study findings were generated which helped in understanding strengths and weaknesses of different machine learning algorithms about identifying order imperfections. It was observed that logistic regression has high precision for identifying imperfection, but decision trees show better recall rates at the same time also random forest achieves good balance between these two measures (precision and recall) hence making it suitable for various kinds of orders with different scenarios.

In summary, what we did is basically testing several models against our dataset so as to find out which one among them fits best for predicting certain types of errors during processing orders. Hence this research gave a full assessment on prediction power associated with models used in addressing questions about detecting faults within an order process. Such insights guided us towards selection of most appropriate model that can be employed while forecasting defects at different levels as well influenced suggestions meant for improving management practices throughout an enterprise's supply chain system.

### ***Identification of Important Factors***

Our study looked at the many different things that can make it easier or harder to figure out when order processing isn't working well. We identified several key factors which significantly affect the performance of predictive models in this area.

First among these is order complexity. Some orders are much more complicated than others, such as those for customized products or with multiple items in them. These complexities add extra steps into the order processing workflow, which makes predicting what will happen next more difficult. By considering how complex an order is, businesses can improve the accuracy of their predictions about what will be needed to fill it and when.

The second factor we found was variability in processing times across orders. Time taken to complete any request can vary depending on factors like seasonal demand, availability of resources etc. This means that if you train a model using lots of data where this varies greatly from one case to another then its ability to accurately forecast completion time for new requests may be limited. Therefore, knowing about such variation allows organizations to take measures like allocating more staff during peak periods or implementing scheduling systems that adapt dynamically with changing workloads so as reduce detrimental effects caused by these variations.

Furthermore, customer satisfaction levels were also noted down as crucial predictors' performances because happy clients have predictable ordering habits while those who are not satisfied may keep on shifting goalposts hence making it hard for any system relying solely upon past events alone being able to know what they want next time. Hence looking at customer satisfaction levels helps businesses align their processes towards customer-centricity and adjust predictive models accordingly based on different consumer moods.

Moreover, historicity appeared important too; here we considered historical dataset's suitability vis-à-vis building robust forecasting tools given current technological advances; therefore, outdated records could cause false alarms besides missing out critical insights necessary for accurate predictions. On the other hand, comprehensive up-to-date information enhances reliability thus enabling firms detect anomalies early which might compromise service delivery; therefore, it is advisable that firms should ensure they have access to high quality historical data sets so as increase reliability levels associated with these predictive models.

In conclusion, understanding what factors affect predictability and accuracy of predictive models in order processing is necessary if one wants to come up with good strategies for managing order lifecycles. This can be improved by better understanding them and then addressing each point raised during our analysis thus leading into more efficient systems which can satisfy customers.

### ***Insights for Order Lifecycle Management***

Our study has given us insights which have far reaching implications on different industries' order lifecycle management. We came up with recommendations and strategies that can be used practically to optimize order lifecycle management by looking into predictive models' ability to recognize flaws in or inefficient aspects of order processing.

First, we need more advanced analytics and machine learning techniques if we want to know everything happening in process of an order. Organizations can predict where there might be blockages or congestions and smoothen their operational systems using some prediction models designed for this purpose. This observation makes decision

making about how orders should live easier since managers will now see them before they come hence reducing customer service disruptions caused by such occurrences.

Second, our investigation also shows that data-based decision making is key towards continuous improvement in the way orders are handled within a firm. When organizations analyze records containing history of previous orders; they can easily see patterns showing where things failed to go according plan thus helping them make necessary corrections so as not only improve efficiencies but also save resources like time and money for other areas of operation too.

Third, another thing learnt from undertaking research is that processes involved in managing life cycle need to focus more on customers than ever before. What businesses should do now are designing workflows which treat every client uniquely leading seamless transaction throughout their supply chains. It would therefore be important if predictive model considered customer feedback together with sentiments expressed during behavior analysis because it will enable companies anticipate people's needs while at same time ensuring services rendered remain uninterrupted thereby enhancing loyalty over longer periods among consumers.

Fourthly, agility at organization level coupled with willingness embrace change as well striving towards perfection cannot be ignored when dealing with complexities brought about by different stages through which goods pass prior delivery or consumption by end users. Prediction becomes so powerful especially during transformational processes hence allowing establishments transform faster. Otherwise without these qualities companies would find themselves being caught off-guard each time new challenges emerge along their life spans thus becoming rigidified even further than before thereby losing competitiveness within today's fast changing business world.

In summary, the insights gained through our study have provided implementable guidelines for enhancing order lifecycle management. Through predictive analytics there is possibility of streamlining efficiency levels during processing orders hence reaching high levels of effectiveness while at same time meeting needs demands from various customers. These findings act as catalysts that trigger change which empowers organizations thrive within dynamic competitive environments prevalent in current day enterprise settings.

In general, the findings from the study will contribute to more knowledge about what determines the success or failure of predictive models in recognizing bottlenecks and redundancies within transactional flow. As such, its implications are academic as well as practical- it should provide some direction on how best to go about managing order life cycles for companies that want them improved.

#### **4.4. How will putting predictive process mining technologies into practice affect customer satisfaction and organisational performance?**

The investigation focused on various order life cycle management processes in response to the research question about the effect of predictive process mining technology on customer satisfaction and organizational performance. We used predictive process mining to understand how different technological interventions can affect customer-focused results as well as efficiency within organizations. Our study therefore sought to identify where predictive analytics tools change traditional approaches towards order management while considering shifting consumer expectations, operational intricacies, and advancements in technology among other factors. We were also aiming at finding actionable insights that can be used for strategic decision making through looking into relations between key performance indicators (KPIs) with predictive process mining

technology so that enterprises may institute continuous improvement programs based on these findings. Based on empirical evidence coupled with theoretical frameworks, this research shows that there is great potential for transformation inherent within predictive process mining, but it must be noted that such kind of analysis has significant impact not only on customer experience but also organizational achievement.

### ***Enhanced Customer Satisfaction***

Predictive process mining technologies have been able to make customers happier in various ways. The first thing is that these methods were capable of predicting problems at every stage of the order lifecycle and organizations dealt with them before they could affect clients negatively. For example, this implies that such models could predict delays in processing or fulfilling orders thus enabling establishments to take corrective measures by speeding up those processes in order to meet customer expectations. This was an active approach towards solving problems because it ensured timely delivery of goods reducing chances for dissatisfaction caused by late arrival or wrong deliveries.

Secondly, predictive process mining systems made it possible for enterprises to increase accuracy and reliability levels during order fulfilment. This was achieved through studying past records and finding out common errors made while entering details then trying to avoid repeating them again. For instance, if there were high chances of leaving some fields blank or filling wrong information, the model would raise an alarm so that employees can recheck such areas before completing the rest parts. Paying close attention to each step taken during this phase played a vital role in improving satisfaction among buyers since goods were delivered correctly without necessitating any returns or exchanges.

Moreover, organizations became more satisfied when they discovered their ability to anticipate what customers want based on previous interactions thereby personalizing how those needs are met when handling orders. In other words, predictive analytics used purchase history as well as browsing habits among others so as to determine product preferences which can be reflected through tailor-made recommendations or promotions being sent during shopping experiences. Customer delight was seen through getting things right from the onset hence creating higher levels of engagement and enjoyment for shoppers.

In conclusion, implementation of predictive process mining technologies led to proactive problem solving, better order accuracy and personalized customer experience which contributed greatly towards improved levels satisfaction within different firms. By employing complex data analysis techniques coupled with foresightful measures companies managed not only meet but also exceed client demands thus gaining increased loyalty from them forever.

### ***Improved Organizational Performance***

Different dimensions of organizational performance have seen significant improvements through the implementation of predictive process mining technologies. One area where there has been a lot of growth is operational efficiency during the order lifecycle management process. Companies can predict possible bottlenecks and smoothen operations by making workflows more efficient using predictive analytics to anticipate these problems. The consequence for this optimization is shorter lead times, quicker order processing and overall better handling of customer orders.

Also, it is true that predictive process mining has helped organizations make better use of their resources. By knowing how things will turn out about fulfilling orders

or meeting customer demand in advance organizations can allocate their resources wisely. They can do this by maintaining optimal levels of inventory as well as staffing numbers required at different times depending on what they expect volume wise from customers' orders hence saving costs while increasing agility due to improved utilization rates realized through such optimizations.

Another way that predictive process mining improves organizational performance is by enabling data-based decision-making processes. Organizations can use machine learning algorithms together with advanced analytics to come up with actionable insights from large amounts of historical data which were not previously utilized fully for decision making purposes within an organization. These actionable insights informed decisions made about pricing strategies used during different phases along the order lifecycle management cycle including product offerings available as well as engagement initiatives aimed towards specific types of clients among others thereby enhancing effectiveness but also fostering continuous improvement plus innovation within an enterprise.

Furthermore, Predictive Process Mining has helped increase satisfaction levels among customers leading them prefer one company over another thus affecting positively on its performance level. Organizations should detect solve challenges associated with fulfilling orders earlier than expected so that they provide excellent services towards their clients who are expecting products delivered timely without any error being made throughout this entire period until when everything gets completed successfully according plan agreed upon between both parties involved in transactional exchange activities; because if there's dissatisfaction caused by errors made while processing requests then relationship may be ruined forever. This means that firms must create systems which will ensure that they identify problems within the supply chain early



enough thereby making it possible for them to offer unmatched customer experience. In addition, reducing delays in delivery also plays a major role when it comes customer satisfaction, as this may lead poor reputation within market segment targeted by organization thus affecting negatively its performance levels too; therefore, predictive process mining should be implemented carefully since every step counts towards achieving success or failure.

The overall effect of adopting Predictive Process Mining Technologies has been transformative on organizational performance. Among other things such technology optimizes operational efficiency resource utilization data driven decisions making and customer satisfaction which are key drivers of success in order life cycle management. It allows companies understand how best they can utilize their various resources depending on what is expected volume wise from customers' orders at different times hence saving costs while increasing agility due to improved utilization rates realized through such optimizations. As organizations continue using advanced approaches towards achieving competitive advantage through continual improvement based on dynamic business environment today may see significant improvements in revenue growth and profitability.

### ***Streamlined Processes***

Introducing anticipatory process mining technologies have streamlined the way organizations operate in several ways.

For one, it has allowed them to spot inefficiencies faster by identifying bottlenecks throughout the order lifecycle. Organizations can do this by reviewing historical event data and uncovering patterns of delays or disruptions during certain stages. This enables them to take proactive steps towards improving those processes; for instance, if there is always a delay in processing orders at payment stage they could

assign more resources or make changes that will speed up payment processing hence reducing lead time.

Secondly, predictive process mining technology can be used to optimize workflows by singling out redundant or unnecessary steps within an order's life cycle. Through looking at event logs and process models companies may find activities which contribute very little value to overall performance and therefore remove them thus making everything run smoothly while reducing processing time. Optimization ensures smooth flow of work through various stages leading to quick realization of customer orders which increases satisfaction among clients.

Besides that, these systems also bring about automation of repetitive tasks during different phases involved in taking care of customer demands thereby promoting efficiency further. By using machine learning algorithms alongside event data analysis capabilities provided by such tool's businesses are able automate routine jobs like verifying orders, managing inventory as well as routing orders themselves through supply chain networks etcetera. Apart from reducing manual involvement required for performing these tasks it also reduces chances occurrence errors which slows down operations causing jams in one area followed by another resulting into poor service delivery.

Again, predictive process mining technologies support continuous improvement programs based on real-time insights into performance levels achieved so far vis-à-vis desired targets set forth by management team members etcetera. Organizations keep track of key performance indicators (KPIs) including but not limited to average time taken from when an order is placed till when it gets delivered; accuracy rate at which correct items are packed together with customers' satisfaction scores among others. This allows enterprises detect areas that need adjustments immediately they arise rather than waiting

for later reviews or audits thus making necessary corrections right away. Continuous approach keeps business processes flexible enough accommodate changes required during different stages improvement ultimately leading to better efficiency gains over time.

In conclusion, predictive process mining technologies have streamlined organizational processes through identifying bottlenecks, optimizing workflows, automating repetitive tasks, and enabling continuous process improvement. Advanced analytics and machine learning algorithms should be deployed by firms in order ensure that their order lifecycle management processes are efficient, agile as well as responsive towards meeting customer requirements which leads to enhanced operational performance coupled with higher levels of satisfaction among customers.

### ***Data-Driven Decision-Making***

Modern business strategies are based on data-driven decision-making. Enterprises statistically process and use information to achieve their objectives like never before. However, organizations have always searched for knowledge hidden in raw data sets and acted upon it to be successful. How they do that has changed a lot over the years.

This approach becomes even more important when predictive process mining technologies are used for order lifecycle management optimization in terms of improving organizational performance.

First off, customer behavior and preferences can be better understood through the application of data-driven decision-making. Analyzing large amounts of historical records about order processing and customer interactions during fulfillment helps identify useful things like patterns, trends or correlations between events which indicate what customers need or expect most frequently from businesses they deal with. These findings

allow firms to customize products, services offered as well as methods used in providing them so as to satisfy clients' demands hence increasing their satisfaction rates and loyalty levels too.

Secondly, operational processes can be optimized by allocating resources accordingly using this same principle; secondarily only after my first point because it relates directly with it (order lifecycle management). This implies that organizations must understand where they go wrong along this cycle thereby finding ways on how best they can change those areas into more productive ones while still saving time plus money at once but not separately since these two things go hand-in-hand when dealing with any form of workflow systems such as these ones being analyzed here today.

Additionally, accurate predictions about future outcomes become possible due to basing decisions on facts rather than guessing games played by gut feelings alone which often led people astray especially during times when there is no sufficient evidence available yet concerning. In other words, adding another point after the second argument but still related closely enough with all others mentioned earlier before regarding various aspects surrounding utilization predictive process mining technologies within an entity's supply chain operation like this one under review today.

Moreover, it is important to note that data-driven decision-making does not only apply to supply chain operations but can be used in other areas as well. One such area is demand forecasting where past records are used together with current trends so as to come up with accurate predictions about what might happen going forward given different scenarios considered within one's business environment.

#### **4.5 In order to increase productivity and lower mistake rates, how can companies successfully incorporate predictive process mining into their current order management systems?**

The results of the thesis indicate that incorporating predictive process mining into current order management systems can significantly enhance productivity and reduce mistake rates. Through the analysis of historical event data and the application of machine learning techniques, companies can effectively leverage predictive process mining to achieve these objectives.

##### ***Identification of Inefficiencies***

The detection of imperfections within forecasting process management is one of the main points that are covered by predictive process mining. It enables firms to discover different bottlenecks, delays and deviations from best practice processes which can limit effectiveness and productivity through analyzing historical event data. Predictive process mining techniques can identify patterns of behavior that indicate inefficiencies by looking at event logs which give a blow-by-blow account of each phase in the life cycle of an order.

For instance, the study might expose such things as a series of recurring delays during order processing like when there is too much time taken between placing an order and it being fulfilled or many instances where mistakes were made during data entry resulting into wrong details being recorded against an order number. Again, this method can also highlight redundant or extraneous steps within the workflow for managing requests thereby suggesting points that could be simplified through automation while still achieving similar efficiency gains.

In addition, this approach helps in measuring how much inefficiency affects performance by examining certain KPIs (Key Performance Indicators) including but not limited to order cycle times; throughput rates; error rates etcetera. Companies may compare these measures with industry benchmarks or their own standards so as to evaluate seriousness of inefficiencies and areas needing more attention.

To sum it up, predictive process mining shows companies where their weak points lie as far as managing orders is concerned thus giving them insights on what parts need improvement most urgently in terms enhancing efficiency levels while cutting down costs at large scale across entire organization performance improvement efforts should be concentrated towards these areas.

### ***Prediction of Process Outcomes***

Predictive process mining empowers organizations to forecast future process outcomes with a high degree of accuracy, thereby enabling proactive decision-making and strategic planning. By leveraging historical event data and machine learning algorithms, companies can gain valuable insights into the potential trajectory of their order management processes.

One significant aspect of predicting process outcomes is the ability to anticipate completion times for various stages of the order lifecycle. Through the analysis of past event logs and the identification of patterns and trends, predictive models can estimate the time required for order processing, fulfillment, and delivery. This forecasting capability allows organizations to better manage resource allocation, optimize scheduling, and meet customer expectations by ensuring timely order completion.

Moreover, predictive process mining facilitates the prediction of order fulfillment rates, which is crucial for ensuring efficient and reliable service to customers. By

analyzing historical data on order processing times, inventory levels, and other relevant factors, organizations can predict the likelihood of successfully fulfilling orders within specified timeframes. This insight enables proactive measures to be taken, such as reallocating resources or adjusting workflows, to improve fulfillment rates and minimize delays.

Another important aspect of predicting process outcomes is the ability to forecast resource requirements and capacity utilization within the order management system. By analyzing historical event data and identifying patterns in resource usage, predictive models can estimate future demand for manpower, equipment, and other resources. This foresight allows organizations to optimize resource allocation, prevent bottlenecks, and ensure smooth operation of the order management process.

Furthermore, predictive process mining enables organizations to anticipate potential errors, deviations, or inefficiencies within the order management system. By detecting anomalies in historical event logs and identifying patterns associated with past mistakes, predictive models can flag potential issues before they occur. This early warning system allows companies to implement corrective actions, refine processes, and prevent costly errors or delays in order fulfillment.

In summary, the prediction of process outcomes through predictive process mining provides organizations with valuable insights into the future performance of their order management systems. By accurately forecasting completion times, fulfillment rates, resource requirements, and potential errors, companies can make informed decisions, optimize operations, and enhance customer satisfaction. This proactive approach to process management enables organizations to stay ahead of the curve, minimize risks, and achieve better outcomes in today's dynamic business environment.

### ***Integration with Existing Systems***

Integrating predictive process mining into current order management systems requires that the organization be able to tie it in with other systems. Among the most significant problems faced by businesses is making sure their already existing IT infrastructure, software applications and data sources are compatible with the predictive process mining framework. It is mandatory for this integration to happen since predictions models need access to necessary information for them to work correctly within the context of order management.

To solve this issue, companies should review what they have got and find out where predictive process mining can be incorporated. This means that they should know which transactional databases, ERP systems or CRM platforms among others contain useful information about orders already placed within an organization. In addition these repositories should be mapped so that links between them can be created thereby forming a single data environment which supports activities related to predictive process mining.

Additionally, there may be requirement for investments into other tools/technologies so that integration becomes possible. For instance; middleware solutions, API connectors or data integration platforms may need to be put in place by an entity seeking smooth communication between different systems. Such technologies are responsible for collecting data from various sources as well as standardizing their formats for purposes of predictive modelling through ensuring consistency and quality of such records.

Moreover, scalability together with flexibility needs of the company must also not escape attention during integration steps taken towards completion of this particular task. The future growth rates as well changes in business processes; data sources used for



analytics etc., require a good design thinking approach towards choosing right architectural structures that will accommodate any arising situation.

Data security compliance is another important factor when integrating various components into one system during its development phase . Organizations have to make sure sensitive information is handled appropriately especially when connecting with external systems or accessing third party data sources. This might entail putting measures like encryption controls over such kind of content so as only authorized personnel can get hold it which also helps meet GDPR, HIPAA & PCI DSS requirements.

Success in integrating with existing systems calls for strategic thinking on the part of an organization by taking into consideration various aspects like current IT landscape; scalability needs among others as highlighted above. Therefore through careful planning coupled with proper execution during integration stages, enterprises can leverage predictive process mining capabilities towards improving their order management systems thereby delivering tangible business results.

### ***Continuous Improvement***

Continuous improvement is a key concept in business process management. Essentially, it means that you should always be trying to make your processes, products, and services better bit by bit over time. In the age of order management systems with predictive process mining capabilities, this continuous improvement factor is what drives organizational excellence and competitive advantage in today's rapidly changing business environment.

One way predictive process mining enables continuous improvement is through iterative refinement of order management processes. In this approach, an organization keeps looking at historical event logs while also monitoring how well the system is

performing so as to find areas where things can be optimized then apply interventions aimed at making such improvements happen more efficiently. Through such iterations companies are able to deal with points of congestion; simplify work flows and get rid of duplicate tasks thus leading to smoother running operations coupled with increased productivity.

In addition to that, predictive process mining allows for real-time monitoring of performance during processing which enables early detection on deviations from expected results followed by immediate corrective measures being taken by the responsible personnel within an organisation. By utilizing machine learning algorithms trained using previous experiences businesses can discover outliers as they occur thus preventing bigger challenges from happening because predictions about possible future challenges were made before their occurrence. Such a pre-emptive problem-solving method ensures that disruptions do not affect any other part of the system hence maintaining high levels of customer satisfaction across all touchpoints.

Furthermore, continuous improvement fostered through predictive process mining promotes data-driven decision making among employees at different levels within organizations. Predictive process mining provides actionable insights based on facts thereby giving people involved in various aspects of value creation an opportunity to scrutinize their activities critically towards understanding what needs adjustment for maximum efficiency gains realization within limited resource envelope available at their disposal. This evidence-based reasoning encourages experimentation around best practices related to order management since there will be clear guidelines derived from proven methodologies.

Also worth noting is that continuous improvement supported by predictive process mining takes into consideration market dynamics as well as customer preferences

which keep on changing from time to time. Companies can leverage these two forces by establishing mechanisms for monitoring feedback given by clients with regard to their satisfaction levels vis-à-vis service delivery expectations then use such information together with ongoing performance evaluation findings related to different stages involved in processing orders – this will help them identify emerging trends, forecast future shifts in demand patterns and re-align strategic plans accordingly. It is crucial for any business operating within the current fast-paced environment characterized by frequent disruptions caused by technological advancements or policy changes.

Ultimately, continuous improvement remains crucial if one wants to achieve operational excellence through optimizing order management systems using predictive process mining technologies. The only way that organizations can do this is by creating an atmosphere of never-ending learning coupled with adaptability backed up by innovation at every level where people work so as not only bring out the best possible value from available data assets but also enhance process efficiency towards realizing higher customer satisfaction levels within shorter periods of time.

The study's results imply that firms can adopt predictive process mining in their existing systems of order management so as to enhance efficiency and reduce error rates. They can do this by maximizing on predictive analysis and artificial intelligence knowledge which will enable them to streamline their activities; make better decisions thus achieving more positive end results about order completion and customer happiness too.

#### **4.6. Summary of Findings**

The results of this study highlight the effectiveness of employing predictive process mining techniques to improve order lifecycle management. Through the

developed activity-based predictive model, we successfully predicted the likelihood of order imperfections within the order lifecycle management process. By integrating activities into the modeling approach, we observed a significant enhancement in the reliability of imperfection likelihood estimation. This underscores the importance of considering the sequence of activities in the order lifecycle when predicting imperfections.

Furthermore, the analysis revealed the three scenarios, which vary depending on the specific order and its lifecycle. This insight provides valuable information for businesses to identify and address key areas of concern within their order management processes. By understanding the underlying causes of imperfections, organizations can implement targeted strategies to mitigate risks and improve overall process efficiency.

Overall, the findings of this study underscore the potential of predictive process mining in enhancing order lifecycle management. By leveraging machine learning techniques and analyzing event logs, businesses can gain valuable insights into their processes, identify potential issues proactively, and optimize operations to ensure smoother order fulfillment and customer satisfaction. This research lays the foundation for further exploration and implementation of predictive process mining approaches in real-world business settings, offering promising opportunities for improving process efficiency and performance.

#### **4.7 Conclusion**

In conclusion, the results of this study highlight the effectiveness of employing predictive process mining techniques to improve order lifecycle management. We built a model with our activity-based features, and It successfully detected imperfections in the order lifecycle management process. We found that the modelling approach significantly

improved the accuracy of imperfection likelihood estimation when activities were included. This emphasises how crucial it is to take the activities of the order lifecycle's operations into account while predicting imperfections.

Additionally, the investigation identified the causes of order flaws, which differ based on the order and its lifecycle. Businesses can use this data to pinpoint and resolve major areas of concern in their order management procedures. Organisations can reduce risks and boost overall process efficiency by implementing targeted measures based on an understanding of the root causes of flaws.

The study's overall conclusions highlight the potential of predictive process mining to improve order lifecycle management. Businesses may ensure smoother order fulfilment and customer happiness by optimising operations, proactively identifying possible issues, and gaining useful insights into their processes by utilising machine learning techniques and event log analysis. With this research, there is potential to improve process performance and efficiency through additional investigation and application of predictive process mining techniques in practical business environments.

## CHAPTER V

### DISCUSSION

#### **5.1 Discussion of Results**

In the discussion section, we'll meticulously analyze and interpret of the research findings to get a better understanding of their significance and broader implications within the context of predictive process mining and order lifecycle management. Our research is an attempt to carefully traverse the complexities inherent in modern business operations, especially in the domain of order fulfilment and customer satisfaction. The ultimate objective was to drive operational excellence and improve organizational results by using predictive process mining techniques to large datasets of order lifecycle events and uncovering hidden insights.

#### **5.2 Discussion of Research Question**

The case studies provide valuable insights into the practical application and effectiveness of the developed predictive process mining model within the context of order lifecycle management. Each case study offers a unique perspective the predictive capabilities of the model and its ability to identify potential imperfections within the order lifecycle process.

Case Study 1 illustrates the effectiveness of the model is in predicting imperfect order when the order is imperfect for the case key: 1358405196A01422. Strong predictive power was indicated by the model's estimation of a high likelihood of order imperfection (0.9984907). With certain activities like "Sales Order Complete" and "Sales Order Modified" playing significant roles, the SHAP value plot related to this case key provided important insights into the variables leading to the imperfection prediction. Further emphasizing the significance of these actions in the order lifecycle process, the

accompanying heatmap offered a visual depiction of their contributions to the imperfection forecast.

In addition, Case Study 2 analyses a different case key (1345897825A01288) to further highlight the prediction power of the model. The model predicted a lower order imperfection likelihood in this instance (0.16466218), pointing to a better possibility of a successful order completion. The corresponding heatmap and SHAP value plot provided insights into the activities affecting the imperfection predicted; "Sales Order Completed" and "LM\_Received at Hub" were found to be key contributors. These results highlight how well the model predicts order outcomes based on activity-level data.

A unique situation is shown in Case Study 3, where an order that is initially flawless eventually becomes imperfect at the end of the order flow. The model used the case key (1360331646A01288) to correctly identify the imperfection, despite the intricacy of the scenario. The estimated likelihood of order imperfection (0.99677306) suggests a strong prediction accuracy. The SHAP value plot and heatmap connected with this case key gave insights into the actions leading to the imperfection predicted, with activities such as "Return Money Sent" and "Received at PUP" having major roles. This example demonstrates how well the model predicts flaws and can adjust to intricate order lifecycle circumstances.

Overall, the case studies offer strong proof of the predictive process mining model's potential in identifying potential imperfections within the order lifecycle process. The model provides useful insights that can inform decision-making, optimise process workflows, and improve overall operational efficiency within organizations by utilising activity-level data and modern machine learning algorithms. These findings underscore the significance of predictive process mining in improving order management practices and ensuring customer satisfaction.

### **5.3 Effectiveness of Predictive Process Mining**

Predictive process mining is highly effective in the context of order lifecycle management. It offers a new way to understand the complex processes that are involved in managing orders from inception to fulfillment. This enables organizations to use machine learning algorithms for actionable insights extraction from large event log data sets.

One of the main advantages of predictive process mining is that it allows one to predict future inefficiencies before they occur in the ordering cycle. Predictive models based on historical data can detect patterns that indicate potential problems and therefore flag up orders at risk during their lifecycles. Consequently, with this strategy, it becomes possible for a business enterprise to deal with problems even before such issues reach an alarming stage thereby reducing interruptions and improving overall operational efficiency.

Additionally, root causes of imperfections related to order can be identified through predictive process mining. Delays, errors, or deviations from standards may be examined as contributing factors to these imperfections hence giving helpful information about problem areas within an organization's processes. Therefore, by understanding this information, firms are able to introduce intervention measures specifically targeting these root causes thus enhancing their order management workflows.

Moreover, predictive process mining indirectly helps organizations optimize resource allocation and streamline operations. In other words, businesses can more efficiently allocate resources based on likely disruptions and configure tasks according so as to optimize time-based workflows for improved service delivery. This means it helps enhance operational efficiencies while also improving customer satisfaction by minimizing lead times and reducing risks of mistakes or delays in orders.



The above analysis shows how predictive process mining turns raw event log data into useful knowledge leading into continuous improvements regarding order life-cycle management. By leveraging advanced analytical techniques, organizations can gain a deeper understanding of their processes, anticipate potential challenges, and proactively mitigate risks, ultimately leading to enhanced operational performance and customer satisfaction.

#### **5.4 Integration of Activity-Based Modeling**

Integration of activity-based modelling is essential to improving predictive process mining's efficiency in order lifecycle management. By adding in-depth depictions of each step in the order lifecycle, this method offers precise insights into the order of events and how they affect the results of the operation.

Organizations can capture the subtleties of each activity in the order lifecycle, such as their dependencies, timing, and interconnections, by incorporating activity-based modelling. This makes it possible to comprehend how various actions affect the overall performance and results of the process in a more thorough manner. For instance, the success or failure of an order is largely dependent on operations like order formation, payment processing, inventory management, and shipment tracking.

In addition, activity-based modelling makes it possible to pinpoint the order lifecycle's crucial pathways and bottlenecks. Organisations can identify inefficiencies or delays that could impede order fulfilment by examining the sequence of tasks and their dependencies. This insight makes it possible to implement focused interventions that improve overall process efficiency by streamlining workflows, allocating resources optimally, and reducing potential bottlenecks.

Moreover, by capturing the intricate relationships between activities and their influence on process outcomes, the inclusion of activity-based modelling improves the precision and dependability of predictive models. Conventional process mining techniques could miss these subtleties, resulting in less precise predictions and suboptimal decision-making. By incorporating detailed activity-level data, predictive models can more accurately forecast the likelihood of order imperfections and anticipate potential issues before they arise.

Overall, the incorporation of activity-based modelling is a sophisticated method of predictive process mining that makes use of specific activity-level data to obtain more profound understanding of the order lifecycle. Organizations can customer satisfaction, operational efficiency, and order management processes by capturing the nuances of process execution and interactions across activities.

## CHAPTER VI

### SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

#### **6.1 Summary**

Pondering the application of predictive process mining in order lifecycle management (OLCM) within organizations is what this thesis is all about. There's a lot to unpack when it comes to managing the different phases of meeting customer orders, so it makes sense to lean on data-driven methods that can streamline these processes and keep customers happy. Predictive process mining is one such method – armed with event logs, this approach has shown potential in predicting inefficiencies or mistakes.

The problem at hand and thus the crux of our research lies in the vast quantity of data generated by event logs. Manually wading through it all is a huge pain and takes up too much precious time that could be spent improving business operations. Machine learning algorithms are at our disposal though, so why not use them? They're great for finding patterns, identifying root causes of problems and ultimately streamlining order lifecycle management. We want to bridge the gap between real-world order complexities and efficient decision-making driven by data.

Two objectives will drive our research: first, we need to develop predictive models that can let us know when something's bound to go wrong in the order lifecycle; secondly, we must study which factors contribute most to these imperfections while also looking for ways to improve them. With any luck, combining predictive process mining techniques with advanced machine learning algorithms will give us enough actionable insights for organizations to take control of their OLCM practices.

We'll have to collect and preprocess event data from as many orders as possible – cleaning the data before using it for analysis. Once that's done we'll train machine learning models with this new batch of processed data so they know what kind of

imperfections they should expect within an order lifecycle. By determining feature importance and SHAP values – both used for quantifying individual factors' impact on predictions – we'll get a better idea of what makes these imperfections happen.

The results of our research showcase the accuracy of predictive process mining in identifying potential bottlenecks and inefficiencies throughout an order lifecycle. Case studies and analysis revealed that these models are very good at predicting imperfections, allowing organizations to solve their problems before they even occur. We also managed to improve the reliability and interpretability of models by integrating activity-based modeling. Doing so gave us a deeper understanding of where these imperfections come from.

The scalability and generalizability of our findings are quite promising as well, making them perfect for implementation across different geographical regions and organizational contexts. With an evaluation of real-time predictive capabilities, we hope to emphasize how important it is that organizations address their issues as soon as possible instead of waiting until it's too late. After all, operational excellence and customer satisfaction are two things that can drive success when properly attended to.

The review of existing research and scholarly work on the application of predictive process mining in order lifecycle management covers essential aspects of relevant topics. We start by explaining the theory behind process mining, predictive analytics, and business process management to ensure we lay a solid foundation for our research. This section also includes key concepts, methodologies, and theoretical frameworks that inform these fields.

Next, we will discuss how process mining has turned into a field from traditional process modeling and analysis techniques. With input from well-known researchers such as van der Aalst and Dumas, we understand exactly how these techniques have evolved

over time. The goal is to help readers understand the theoretical underpinnings and practical applications of process mining techniques.

The study also talk about how machine learning algorithms can be used to predict future process behavior by analyzing event data. We find examples of different predictive analytics models such as monitoring, forecasting, and anomaly detection as well as their strengths, limitations, and potential applications in order lifecycle management.

Case studies where predictive process mining methods were applied in real-world situations. In these studies we noticed an improvement in efficiency for processes; bottlenecks were quickly identified making it easier to improve them which led to increased customer satisfaction rates. We break down everything we learnt from these studies with actionable insights for future research.

Moreover,The methodology section explains the way in which thesis was researched, the data collection process and the techniques used for analysis. The research approach was discussed first, highlighting the fact that both qualitative and quantitative methods were utilized to create a mixed-methods approach.

Then, it goes on to talk about the data collection process in detail. Detailing the problems of obtaining order lifecycle event logs from multiple sources, it also talks about how they dealt with data preprocessing by cleaning it up, transforming it and engineering featured. Machine learning algorithms were then used to make predictive models with activity-based modeling techniques integrated to improve accuracy.

Afterwards, we discover that performance metrics such as accuracy and precision were used to evaluate these models as well as cross-validation tech and model tuning procedures.

The limitations faced during this process are also mentioned such as quality issues with data or complexity of models or generalizability concerns.

The study presents the findings in the data collected from order lifecycle event logs during their study. They go into detail about the analysis and interpretation of that information.

It is concerned with the analysis and interpretation of data obtained from order lifecycle event logs. It opens by summing up how well the predictive models were able to forecast imperfections within an order in a process that ends with its delivery.

This section also provides details about case studies that were carried out to verify these predictive models as well as giving descriptions about those cases identified and their predicted probabilities of having errors. It entails SHAP (Shapley Additive explanation) values and heat maps for detecting major factors causing mistakes in orders thus making it clear what drives failure of orders.

Additionally, activity-based modeling approaches are shown to positively improve the model prediction accuracy. The incorporation process regarding the actions characteristics which have implications on performances of these models where they capture detailed information about processes hence increasing predictability.

Furthermore, it offers insight into various aspects such as performance metrics including; accuracy, precision, recall and F1-score among others during model evaluation. The verification explains how robustly they estimated order defects when different circumstances were used or datasets prepared by them.

Similarly, it outlines how the findings may impact order lifecycle management and business operations. This indicates that there are potential benefits related to predictive process mining in optimizing OMPs as a means of reducing IR, increasing efficiency in operational activities and enhancing customer satisfaction.

## **6.2 Implications**

The study has implications that go beyond academic boundaries and touch on business strategy, operations management and technological innovation. This research offers insights into how machine learning models are able to predict customer behavior in order management lifecycle processes which poses several significant organizational issues. It is applicable to managerial decision making, operational optimization, and strategic positioning among other areas whose results can bring positive changes in performance of businesses and satisfaction of customers.

### **6.2.1 Practical Implications**

The practical implications of the thesis are numerous, giving vital ideas and advice to businesses that want to better their order lifecycle management processes. This will enable organizations to extract actionable intelligence from their order data that can be used in the improvement of operational efficiency, customer satisfaction, and overall business performance.

Additionally, predictive process mining is essential for identifying bottlenecks and potential problems early during order lifecycle. Thus, using these models developed as part of this study, enterprises may predict possible flaws or disruptions within the order management process before they manifest themselves in delays or poor customer experience. Companies can optimize resource allocation by utilizing advanced analytics to forecast process behavior and outcomes, streamline workflows and ensure smooth servicing from order placement through delivery.

More likely than not this thesis also points out how crucial data-driven decision making is in relation to life cycle management of orders. On the other hand, through

incorporating predictive process mining models into their decision-making processes firms have embraced a strategy that enables them to make better decisions more quickly taking into consideration the real-time information gained from the analysis of orders made so far. As a result, companies can adjust efficiently under shifting market conditions, consumer preferences basis as well as internal dynamics thereby enhancing their agility and adaptability over cycles.

Moreover, improvement and optimization remain very important aspects in dealing with an organization's order lifecycle management according to this research paper. Still there are other potential uses for such kind of model which focuses on predicting next event based on historical data however it is mainly useful in refining processes associated with O2C cycle itself since all parameters affecting prediction results are chosen from its context. This approach calls for continuous learning and innovation cultures within an organization whereby life cycle management practices continue being aligned with changing corporate objectives as well as meeting customer demands.

Further still practical implications go beyond individual organizations to incorporate broader industry trends and best practices. In this sharing knowledge acquired during investigation with industrial peers plus stakeholders' firms may support other businesses' study efforts while enhancing general level of order lifecycle management. Moreover, industry-wide collaboration and sharing of knowledge among the industry players may promote standardization, innovation and adoption of best practices in order management leading to establishment of a market leader.

To sum up, the practical implications of this thesis show how predictive process mining can be used to optimize order lifecycle management processes. Therefore, by using data driven techniques companies will be able to improve their operational



efficiency, customer satisfaction, and maintain competitiveness in today's fast changing business environment.

### **6.2.2 Theoretical Implications**

The theoretical significance of this thesis goes far beyond its findings and contributes to the general knowledge of predictive process mining in order lifecycle management. The study reveals that business process management is changing and data driven decision making is becoming central. One of the major theoretical contributions is to show how predictive process mining can enhance traditional process analysis techniques by predicting forthcoming behavior of processes and outcomes on the foundation of historical event data. This contradicts conventional approaches to process management that mainly depend on retrospective examination and manual intervention.

Additionally, this dissertation adds to theoretical arguments about linking machine learning algorithms with process mining techniques. By building and evaluating predictive models for order lifecycle management, it highlights how machine learning may improve prediction accuracy as well as reveal patterns for complex structure processes. Besides, there are other exploratory studies on algorithmic methods of predictive modelling for Process Mining such as feature engineering practices and model selection scenarios.

This thesis also extends our theoretical understanding of data-driven decision-making in organizations. It explains how it can be practically used in optimizing order lifecycles through predictive process mining. Moreover, it justifies the need for organizations to embrace data analytics so as to achieve operational excellence and strategic decision making by emphasizing the practical relevance of Predictive Process Mining (PPM) in the optimization of order life cycle processes. These trends are in line

with those witnessed in business intelligence and analytics that now underscore the importance of forecast based insights towards enhancing organizational performance and competitiveness.

In conclusion, implications from this theory are not only limited within case(s) but have a basis upon which various theoretical tenets about process mining, predictive analytics, and managerial decision-making can be built from. Also, this research work enhances our understanding regarding how Predictive Process Mining can be implemented practically; thus contributing to current debates on aspects like Data Analytics in Management Processes or Organizational Performance Improvement through Predictive Models

### **6.2.3 Methodological Implications**

This research has methodological implications far beyond the specific research context and presents valuable insights into the wider field of predictive process mining. A key methodological implication is that activity-based modeling in predictive process mining has been proven to be effective. Better predictions about how likely an imperfection will occur can be made by using refined process-level data in predictive models. Therefore, it would appear useful for future studies aimed at improving methodologies in business process management.

The thesis also shows a way for further refinement and extension through its methodological approach. Predictive models can be built on XGBoost which is among the machine learning algorithms used in handling large-scale event log data with flexibility and scalability. In addition to making accurate predictions, this choice of methodology allows for the exploration of complex dynamics and relationships within processes. If they want to increase predictive accuracy even more, future researchers

should consider hybrid modeling approaches and alternative machine learning techniques.

Apart from anything else, one important aspect of this thesis is its ability to incorporate industry-specific knowledge into the overall framework of prediction making. The researchers collaborated with stakeholders who are experts in their respective fields leading to capturing minor details unique to order lifecycle management requirements within each domain as well as customizing prediction models themselves directly based on these details. This methodology indicates a need for interdisciplinary cooperation in process mining where subject matter expertise must go hand in hand with sophisticated analytical techniques if we want any meaningful results.

Moreover, this thesis developed a comprehensive methodological framework that outlines all stages associated with conducting a successful PPM project starting from data pre-processing activities such as feature extraction through model training before validation stage. The thoroughness of this framework makes it an invaluable tool for those academics or practitioners who intend applying PPM techniques across different kinds of organizations. It is designed so that other researchers can replicate or extend this study by giving detailed procedures and best practices; hence contributing towards more developments in predictive analytics as an evolving field.

In conclusion, what these methodological implications teach us is that a rigorous and systemic approach must be followed in conducting research on the topic of predictive process mining. Hence, with carefully selected advanced machine learning techniques and inclusion of industrial domain knowledge as well as following a structured methodology framework, it then becomes possible for researchers to create reliable and accurate predictive models which have great significance to complex business processes.

These methodological insights allow future studies aimed at addressing critical issues in process mining while advancing state-of-the-art in predictive analytics.

#### **6.2.4 Managerial Implications**

This thesis has several managerial implications which can significantly improve the quality of organizational decision making. Managers leveraged predictive process mining techniques to gain a better understanding of their order lifecycle management system for more informed and strategic decisions. Among the key managerial implications are identification and elimination of bottlenecks and inefficiencies in order management process. By analyzing historical data, predicting possible flaws in order lifecycle, managers will come up with areas requiring improvement hence target-oriented remedies may be applied to facilitate easier work.

Furthermore, insights from predictive process mining can also help in better resource allocation as well as optimization. Predictive models developed based on this research can help managers forecast where orders could go wrong and thus allocate resources more effectively to help mitigate such risks. In this regard, additional resources would be allocated to high-risk orders or certain activities given priority so that imperfections are reduced. This proactive resource management approach leads to cost savings, improved efficiency and customer satisfaction.

Additionally, another important implication is an opportunity for improving customer service and satisfaction. By identifying flaws within the order lifecycle, organizations will be able to smoothen the order processing systems leading timely delivery of goods and accurate fulfillment of orders. As a result, this increases customer satisfaction levels leading increased loyalty among customers served by them (Kurtz & Boone 2011). Managers can use information obtained from predictive process mining

towards developing strategies that put customers at the center of their operations thereby adapting to changing demands in the market.

Moreover, when predictive process mining techniques are implemented by organizations' managers; they become aware of how they can enable continuous improvement initiatives through action plans by acting on them upon discovery. Managers should look for trends in order lifecycle data they continuously monitor and analyze them for optimization opportunities that may exist. The findings could hence be used to improve existing processes through iteration or even test new ones to effect ongoing enhancements in operations (Goldratt 2004). Therefore, it creates a culture of learning and agility within organizations by assisting them to adapt and thrive in a turbulent business environment.

In short, the managerial implications of this study highlight the transformative power of predictive process mining in optimizing order lifecycle management systems. By use of advanced analytics techniques, managers can gain deeper insights into their operations thereby improving decision making processes and facilitating sustainable business growth. Nevertheless, organizations must invest in resources such as ability and infrastructure that will support successful implementation and utilization of predictive process mining techniques if these gains are to be realized (Agarwal et al., 2010).

### **6.3 Recommendations for Future Research**

With businesses evolving and embracing data-driven approaches, there is a need to enhance predictive model's capabilities so as to match industry's changing demands and challenges. In this segment, we will highlight various areas on which future research can concentrate with regards to improving the effectiveness and applicability of

predictive process mining techniques for optimizing order lifecycle processes and driving operational excellence.

### **6.3.1 Customized Training and Prediction**

Customized training and prediction in predictive process mining refers to the ability to customize a training process or model parameters so that they suit specific organizational requirements. The objective is to create tools and methodologies for users to use. This allows users to customize any aspect of the training process, such as how they select their data or what performance metrics they're using.

Furthermore, feature engineering techniques can be provided which further enhances the process. Organisations may have certain variables, performance metrics, or process features that are pertinent to their order lifecycle management procedures. Predictive models may record a wider range of parameters impacting order outcomes by allowing users to create and include domain-specific data, such as customer segmentation attributes, product attributes, or operational KPIs.

Customised training also includes fine-tuning model parameters and setups to maximise performance for particular use cases and objectives, in addition to feature engineering and data pre-treatment. Choosing the right machine learning algorithms, hyperparameters, and optimisation techniques may fall under this category, depending on the order lifecycle management procedure's specifics and the data at hand. Giving users authority over these parameters enables them to modify the model in response to shifting limitations, preferences, and business needs.

Furthermore, users can evaluate the applicability and efficacy of prediction models in accordance with their own objectives and goals thanks to the customisable performance indicators and assessment criteria. Depending on their operational goals and

limitations, organisations may give various performance metrics—like accuracy, precision, recall, or F1 score—priority. Predictive process mining solutions enable users to specify and rank performance criteria, which can yield useful insights and suggestions based on the specific requirements of the organisation.

All things considered, predictive process mining's ability to provide tailored training and prediction skills enables businesses to fully use data-driven insights and decision-making. Organisations may create more accurate, dependable, and useful predictive models for optimising operations by customising the training procedure, feature engineering approaches, model parameters, and performance measures to organisational contexts and goals.

### **6.3.2 Feature Enrichment for Improved Accuracy**

Feature enrichment is a process where one can add extra features to modeling processes that help with things like accuracy and robustness. There are two ways to go about this: either by incorporating variables that have a known impact on processes or by having complex relationships between different variables.

In terms of incorporating variables with a known impact, this could look like considering customer behavior, product characteristics, market dynamics, operational factors (etc.). Imagine being able to predict something like shipping problems based on inventory levels or pricing information.

For creating complex relationships between different variables, organizations can use advanced data analysis and machine learning techniques. This will allow us to create new features that represent higher-level patterns or trends in the data by combining multiple other variables together. For example, we combine order frequency with customer loyalty status to create a new feature indicating the value of repeat customers.

To automate the whole thing, there are dimensionality reduction and feature extraction algorithms. These will help identify which variables provide meaningful insights and are worth incorporating into the predictive models.

Lastly, there is also an option where we can load up on external data sources such as market trends, weather conditions, economic indicators or social media sentiment for example. Although it might sound overwhelming at first, when combined with internal process data these sources can lead towards a very comprehensive view of how exactly all these things impact order lifecycle management.

So overall while feature enrichment may seem like just another step in the model-building process it plays a critical role in enhancing accuracy and effectiveness for order lifecycle management models - which ultimately leads to better decision-making capabilities for businesses.

### **6.3.3 Cross-Country Model Adoption**

Essentially, cross-country model adaptation is when predictive process mining models are taken that are developed for one country's order lifecycle management (OLM) processes and adapt it to another. Due to differing order managements practice, customer behaviors, regulatory frameworks, and market dynamics across the globe this has become a necessity. A model that performs well in one country simply won't generalize well in another.

The best way to tackle this method is by fully understanding the domain analysis and unique factors of the target country. This includes cultural preferences, regulatory requirements, supply chain dynamics, market competition and customer expectations specific to the target market. These insights will help organizations better find ways on



how they can tailor their predictive process mining models so that it aligns with the local context.

Another aspect is retraining or fine-tuning existing predictive models using data collected from the target country. This allows organizations to incorporate country-specific patterns, trends, and variations into the model which helps ensure it capturing all relevant factors driving order outcomes in a new market. Additionally, organizations may need adjust feature weights or algorithmic settings to optimize performance for specific countries OLM environments.

Transfer learning techniques could also be very helpful if properly utilized during this process as well. It's essentially leveraging knowledge gained from training models on one dataset so that one can apply it to improve performance of models on another related but different database. By doing this companies can accelerate adaption time all while reducing amount of labelled data required for training a new model.

Collaborating with local stakeholders can go along way too being efficient at cross-country model adaptation as well. By having experts in said countries collaborate with the researcher, not only does it help increase the chances of tailoring the predictive model effectively but helps address any county-specific challenges or considerations too.

Overall cross-country model adaptation is crucial for success in global markets. It's all about making sure that the predictive process mining models are customized to the specific needs of each country's OLM environment. By doing this, companies can make better decisions. Improve order management processes and ultimately drive business success on a more global scale.

### **6.3.4 Integration with Order Systems**

Integration with order systems means including predictive process mining models in the existing order management system that organizations use. The goal of this is to make the functionality and decision-making abilities of these systems better. Providing real-time insights, predictive analytics, and actionable recommendations for stakeholders who are involved in the lifecycle of orders.

One crucial part of integrating order systems with predictive process mining models is making sure they work well together. They need to be able to communicate and exchange data seamlessly, which might mean having APIs or connectors. Establishing standardized protocols will help with this so that different software components can interact without any problems.

Once everything is integrated, there's a few ways that predictive process mining models can improve order systems. One way is by offering real-time predictions and forecasts about how an order will play out based on historical data and other algorithms. This allows businesses to address issues before they occur or optimize their operations beforehand.

Another thing these types of models can do is offer prescriptive insights to users which guides them towards optimal decisions and action plans. By letting someone know what adjustments they should make to a current process if it predicts a high likelihood of error in an upcoming order, businesses can fix things before it becomes too late. This helps increase efficiency.

Order management is an ongoing task so one important aspect of integrating these models into existing systems involves continuous monitoring and performance evaluation from within the production environment in which organizations operate in. Being able to track down where errors are coming from always allows for quick fixes.

Predictive analyzation tools are not always the easiest things for people to understand but when they're directly put into their familiar workflow or user interface like existing Order Management Applications, it makes things easier for everyone involved because no additional training must be done.

### **6.3.5 Scalability and Generalizability**

Scalability and generalizability are two of the top priorities in predictive process mining. If we want these models to be useful and practical in real-world business scenarios, then scalability and generalizability must be the focus. Scalability is if the model can effectively handle large amounts of data, as well as increasing computational demands that come with increasing dataset sizes.

Scalability isn't just about how quickly a model can compute. It's also about how it can scale horizontally across multiple systems or cloud infrastructures with ease. Implementing efficient data processing techniques will help achieve this scalability. Parallelization for example speeds up analyses by allowing the model to process large amounts of event log data simultaneously.

Data partitioning and caching mechanisms are also helpful when it comes to reducing data movement and improving overall processing speed. So, this all goes together with ensuring fast computations. However, even if a system that performs computations lightning quick - that doesn't mean it has scalability.

Generalizability on the other hand is when models perform well across different datasets and business contexts. This means they're able to extract meaningful insights accurately when applied to completely new sets of data from different domains or environments.

To create generalizable models, one must consider multiple factors including feature engineering, model selection, and validation strategies. Features should be carefully selected so they represent underlying processes accurately while minimizing noise levels as much as possible. L1 & L2 regularization are both techniques that prevent overfitting which improves a model's ability to generalize by penalizing overly complex systems.

Proper validation procedures are crucial for assessing generalization performance too. Cross-validation is an example technique where a model will train and evaluate itself on multiple subsets of data. Doing this provides some insight into how robust/stable the system is across differing datasets.

Lastly there's transfer learning, which allows to take knowledge gained from one domain and transfer it to another. This technique can prove useful when adapting models to new contexts.

If scalability and generalizability is a concern within the organization, then organization would be able to deploy predictive process mining solutions that are capable of handling large-scale data analysis.

These systems should also provide actionable insights and predictions that are both relevant and applicable across various domains in a business setting.

### **6.3.6 Evaluation of Real Time Predictive Capabilities**

Judging real-time predictive capabilities of a predictive process mining model means observing its success in fast predictions. As the new event data flows in, it's important to know how swiftly and accurately it responds with forecasts. Proactive decision making can be based on timely forecasting and that's what results in intervention for a change in dynamic business processes. This flexibility is what allows

organizations to take prompt action when either issues arise, or opportunities present themselves.

One approach to assessing real-time prediction capabilities is to measure the model's latency. This term refers to how long it takes for predictions to come through from the moment new event data has been received. Low latency is crucial as having quick access allows swift implementation based on the model's insights. Techniques such as stream processing and efficient algorithm implementations help minimize latency ensuring that predictions are delivered within acceptable timeframes.

Furthermore, evaluating the accuracy of said model in a real-time setting is necessary so we can better understand its effectiveness when practically applied. Pairing up the predictions made by our model against actual outcomes gives us a sense of how reliable our process is. From there, we can then measure performance metrics like precision, recall, and F1 score to give us even more insight into our efficacy. Continuous monitoring will also keep track and feedback any adjustments needed as well as identify if there has been any drift or degradation in performance.

Another side of evaluating real-time capabilities involves assessing just how much this model adapts to evolving data distributions and changing business contexts. Seasonality, market trends, or organizational changes are all factors that influence time dependent processes. A solid predictive model should be able handle these changes while continuing to maintain its accuracy.

Additionally, scalability needs assessment holds weight when we want the system handling high volumes of incoming event data effectively. By incorporating scalable architecture designs and distributed computing techniques one will be able to manage multiple streaming inputs at once without overloading the machine. Load balancing and

resource allocation optimization work in tangent to optimize performance and response time.

By meticulously evaluating the real-time predictive capabilities of our models, organizations can trust their abilities for timely and actionable insights. By doing so, operational efficiency can be enhanced, resource allocation can be optimized, and business performance overall can improve in fast-paced environments.

#### **6.4 Conclusion**

This thesis has investigated the utilization of predictive process mining approaches for improving order lifecycle management systems. A few important findings have been made from the making and checking of predictive models that show factors that cause order imperfections and areas for optimization. The results obtained from this study indicate that the use of PPM in organizations can help to increase efficiency, enhance decision-making as well as improve customer satisfaction.

Predictive process mining also assists managers to identify bottlenecks, inefficiencies and possible risks in their order lifecycle proactively. In other words, it is a proactive approach which helps organizations optimize resource allocation; minimize possibilities of order imperfections associated with operations; and streamline processes. Further, insights extracted from predictive process mining can guide development of customer-driven strategies, leading to greater levels of customer satisfaction and loyalty.

Moreover, this research underscores the significance of continuous improvement & innovation in the context of order lifecycle management. From an ongoing perspective though examining iterative information on lifecycle orders along with refining predictive models, firms can engender innovation and ongoing improvement. In so doing there

builds a culture for continuous improvement thereby enabling businesses to adapt themselves to prevailing circumstances in a relatively volatile business environment.

Nevertheless, it is imperative to note the limitations inherent in this research findings as well as ppp challenges when applying such techniques practically. In other words, whether activities are effective may depend upon organizational readiness besides data quality or model accuracy concerning predictive process mining initiatives too often overlook these island-like issues creating fewer effective outcomes among any others vitiating its positive impacts. Thus, it is essential that companies allocate adequate resources needed to derive full benefits from PPP while ensuring they possess requisite competencies and infrastructure thereof.

This dissertation adds on existing knowledge base about PPM applied for managing orders within their entire lifecycles. By giving useful recommendations based on scientific grounds' analysis carried out within this research project has put power into hands of managers who would use it in driving innovation, efficiency and client satisfaction. As more companies embrace digitization as well as data-driven decision-making, insights emanating from this study will be useful in directing future activities that guarantee sustainable business growth.

## REFERENCES

Alves de Medeiros, A. K., van Dongen, B. F., van der Aalst, W. M. P., and Weijters, A. J. M. M. (2004a). *Process Mining: Extending the  $\alpha$ -algorithm to Mine Short Loops*. Eindhoven University of Technology, Eindhoven.

Alves de Medeiros, A. K., van Dongen, B. F., van der Aalst, W. M. P., and Weijters, A. J. M. M. (2004b). *Process mining for ubiquitous mobile systems: an overview and a concrete algorithm*. In L. Baresi, S. Dustdar, H. Gall, and M. Materaseries (Eds.), *Ubiquitous Mobile Information and Collaboration Systems (UMICS 2004)*. Springer Verlag.

Burattin, A., Maggi, F. M., and Sperduti, A. (2019). *Predictive Process Monitoring Methods: Which One Suits Me Best?* In *Proceedings of the International Conference on Advanced Information Systems Engineering*.

Di Francescomarino, C., Maggi, F. M., Dumas, M., and Ghidini, C. (2014). *Predictive monitoring of business processes*. In M. Jarke et al. (Eds.), *CAiSe 2014 proceedings*. Thessaloniki.

Dumas, M., La Rosa, M., Mendling, J., and Reijers, H. A. (2018). *Fundamentals of business process management*. Springer.

Evermann, J., Rehse, J.-R., and Baesens, B. (2016). *Predictive Process Monitoring with LSTM Neural Networks*. In *Proceedings of the International Conference on Advanced Information Systems Engineering*.

Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal, M., and Shan, M.-C. (2004). *Business process intelligence*. *Computers in Industry*, 53(3), 321–343.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., and Ullah Khan, S. (2015). *The rise of "big data" on cloud computing: review and open research issues*. *Journal of Network and Computer Applications*.

Kang, B., Kim, D., and Kang, S.-H. (2012). *Real-time business process monitoring method for prediction of abnormal termination using KNNI-based LOF prediction*. *Expert Systems with Applications*, 39(5), 6061–6068.

Lund, S., Manyika, J., Nyquist, S., Mendonca, L., and Ramaswamy, S. (2013). *Game changers: five opportunities for US growth and renewal*.

Mans, R. R., Van Der Aalst, W. M. P., and Vanwersch, R. J. B. (2015). *Process Mining in Healthcare*. SpringerBriefs in business process management.



Medeiros, A. K., Weijters, A. J. M. M., and Van Der Aalst, W. M. P. (2007). *Genetic process mining: an experimental evaluation*. *Data Mining and Knowledge Discovery*, 14(2), 245–304.

Park, S., and Kang, Y. S. (2016). *A Study of Process Mining-based Business Process Innovation*. *Procedia Computer Science*, 91, 734–743.

Rozinat, A. A., and Van Der Aalst, W. M. P. (2008). *Conformance checking of processes based on monitoring real behavior*. *Information Systems*, 33(1), 64–95.

Schuh, G., et al. (2020). *Improving Production Systems Using Process Mining*. *Procedia CIRP*, 91, 734–743.

Tax, N., et al. (2017). *Predictive Business Process Monitoring with Structured and Unstructured Data*. *Information Systems*.

Tiwari, A., Turner, C. W., and Majeed, B. (2008). *A review of business process mining: state-of-the-art and future trends*. *Business Process Management Journal*, 14(1), 5–22.

Van der Aalst, W. M. P., et al. (2016). *Process Mining Manifesto Update: Towards Explainable Process Mining*. In *Proceedings of the International Conference on Business Process Management*.

Van der Aalst, W. M. P., et al. (2016). *Process Mining: Overview and Opportunities*. *ACM Transactions on Management Information Systems*.

Van der Werf, J. M. E., et al. (2016). *Process Mining for Inter-organizational Workflows: A Bag-of-Tasks Approach*. *Journal of Information Technology and Management*.

Van Geffen, F., and Niks, R. (2013). *Accelerate DMAIC using process mining*. *CEUR Workshop Proceedings*.

Van Zelst, S. J., et al. (2018). *Bag of Tasks: A Systematic Approach to Process Model Decomposition*. In *Proceedings of the International Conference on Business Process Management*.

Van der Aalst, W. M. P., et al. (2016). *Process Mining for Inter-organizational Workflows: A bag-of-tasks approach*. *Journal of Information Technology and Management*.

Theis, J., et al. (2022). *Improving the In-Hospital Mortality Prediction of Diabetes ICU Patients Using a Process Mining/Deep Learning Architecture*. IEEE Journal of Biomedical and Health Informatics, 26(1), 388–399.

Schuh, G., et al. (2020). *Improving Production Systems Using Process Mining*. Procedia CIRP, 69(1), 381–384.

Lorenz, R., et al. (2021). *Using process mining to improve productivity in make-to-stock manufacturing*. International Journal of Production Research, 59(16), 4869–4880.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., and Ullah Khan, S. (2015). *The rise of "big data" on cloud computing: review and open research issues*. Journal of Network and Computer Applications.

Lund, S., Manyika, J., Nyquist, S., Mendonca, L., and Ramaswamy, S. (2013). *Game changers: five opportunities for US growth and renewal*.

Shmueli, G., and Koppius, O. R. (2011). *Predictive analytics in information systems research*. Management Information Systems Quarterly, 35(3), 553–572.

Schuh, G., et al. (2020). *Improving Production Systems Using Process Mining*. Procedia CIRP, 91, 734–743.

Kratsch, W., Manderscheid, J., Reißner, D., and Roglinger, M. (2017). *Data driven process prioritization in process networks*. Decision Support Systems, 100, 27–40.