



**ENHANCING ABUSE / FRAUD DETECTION IN OPD INSURANCE THROUGH  
RULES- BASED CUSTOMER RISK SCORING APPROACHES**

by

**Mukul Jain**

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfillment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

AUGUST 2024

**ENHANCING FRAUD DETECTION IN OPD INSURANCE THROUGH RULES  
BASED CUSTOMER RISK SCORING APPROACHES**

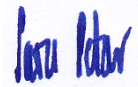
by

MUKUL JAIN

Supervised by

Hemant Palivela

APPROVED BY



---

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

RECEIVED/APPROVED BY:

---

Admissions Director

## **Dedication**

This work is dedicated to individuals tirelessly working in the insurance industry, risk and fraud prevention teams, data analytics, and healthcare professions. Your commitment to maintaining integrity and upholding best practices in your respective fields forms the bedrock of the systems that millions rely upon every day. May this study serve as a testament to your unwavering diligence and contribute to our shared mission of combating fraud and fostering a trustworthy insurance sector.

## **Acknowledgements**

Life's journey is profoundly shaped by those who support and inspire us. This dissertation is the result of such remarkable contributions.

I am deeply grateful to my family. To my wife, Siwani, and my daughter, Drisha — your unwavering support and encouragement have been my foundation. To my parents, your faith in me and the freedom to pursue my path have been instrumental.

A special thanks to Sagar Maheshwari for his invaluable assistance in realizing this research and applying it to detect fraudulent activities in OPD insurances.

I also extend my gratitude to Himank Jain for his meticulous review and proofreading, which have been crucial to refining this work.

Lastly, my deepest thanks to my mentor, Dr. Hemant Pallivela, whose guidance, and encouragement have been vital throughout this journey.

This achievement would not have been possible without each of you. Thank you.

## ABSTRACT

### ENHANCING ABUSE / FRAUD DETECTION IN OPD INSURANCE THROUGH RULES-BASED CUSTOMER RISK SCORING APPROACHES

Mukul Jain

2024

Dissertation Chair: <Chair's Name>

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

This thesis investigates the challenges of outpatient (OPD) health insurance fraud detection and proposes a solution that uses a rules-based approach to score every single customer based on their policy utilization patterns & behavior and assigns a risk score to enhance abuse / fraud detection in the field. The research is motivated by the growing prevalence of abuse / fraud in the OPD insurance industry and the need for more effective abuse / fraud detection methods to protect both insurers and policyholders. The study aims to identify the key characteristics of OPD insurance abuse/fraud and develop a comprehensive set of rules for customer risk scoring for abuse/fraud detection. The research questions focus on the effectiveness and efficiency of the proposed solution.

The study uses a large secondary dataset of OPD insurance claims and policyholder data and shows that the combination of rules leads to a more robust customer risk scoring which helps in raising alerts and signals to prevent abuse and/or fraud. The proposed solution is expected to have a significant impact on the OPD insurance industry, including the identification of high-risk customers, discovery of their syndicates and nexus, blocking their policies, recovering lost money, and reducing operational expenses.

We propose a novel approach utilizing a set of pre-defined rules to flag customer data points indicative of potential risk factors. These flags can be assigned weights based on their relative

importance in predicting claim behavior. A customer's risk score will be calculated as a normalized value between 0 and 100, with higher scores indicating a greater risk of claims.

We will be deriving the Customer Risk Score (RS) and it will be calculated using a weighted/non-weighted sum of the rules.

The formula can be expressed as follows:

$$RS = w_1R_1 + w_2R_2 + w_3R_3 + \dots + w_nR_n$$
$$RS \text{ (Normalized)} = RS / (\text{Total Rules})$$

$W_{i0-n}$  = weights

$R_{i0-n}$  = Rules

The masked and anonymized data on Outpatient Department (OPD) claims will be used to develop and validate the risk scoring model. Analyzing this data will allow us to evaluate the effectiveness of the proposed rule-based approach in identifying high-risk customers within the insurance population.

In summary, this thesis contributes to the development of more effective abuse / fraud detection methods in the OPD insurance industry and provides valuable insights for practitioners in the field. By using a rules-based approach, the proposed solution offers an effective and efficient solution to identify and prevent abuse / fraud in the industry. The findings of this study have important implications for the OPD insurance industry and pave the way for further research in this area.

## TABLE OF CONTENTS

LIST OF TABLES .....	ix
LIST OF FIGURES .....	x
CHAPTER I: INTRODUCTION .....	1
1.1 Introduction .....	1
1.2 Background and context of the research problem .....	3
1.3 Problem Statement .....	7
1.4 Research Questions .....	11
1.5 Objectives.....	11
1.6 Scope and limitations of the study .....	13
1.7 Significance and relevance of the research .....	14
1.8 Business Relevance.....	15
1.9 Thesis Structure.....	26
CHAPTER II: LITERATURE REVIEW.....	27
2.1 Overview of the insurance industry and fraud detection.....	27
2.2 Review of previous research on OPD insurance fraud detection.....	31
2.3 Discussion of machine learning techniques vs rule-based model in fraud detection.....	40
2.4 Review of relevant regulations and policies related to insurance fraud.....	45
CHAPTER III: METHODOLOGY .....	51
3.1 Introduction .....	51
3.2 Operationalization of Theoretical Constructs.....	52
3.3 Research Design.....	55
3.4 Data Collection and Instrumentation.....	58
3.5 Data Cleaning and Pre-processing .....	74
3.6 Exploratory Data Analysis .....	82
3.7 Outlier Detection .....	96
3.8 Hypothesis Testing.....	103
3.9 Rule Generation.....	117
3.10 Summary .....	123
CHAPTER IV: RESULTS .....	128
4.1 Introduction .....	128
4.2 Evaluation of Rule-Based Risk Scoring Model .....	129
4.3 Findings related to each hypothesis and research question.....	137
4.4 Case Study.....	139

CHAPTER V: DISCUSSION, CONCLUSIONS, AND IMPLICATIONS ..... 146

5.1 Discussion ..... 146

5.2 Comparison with previous research and contributions to the field..... 153

5.3 Limitations and challenges faced during the research..... 155

5.4 Recommendations on Future Research ..... 157

5.5 Conclusions ..... 159

**REFERENCES..... 162**



## LIST OF TABLES

Table 1-1 Key differences between OPD & IPD.....	4
Table 1-2 Abuse Penetration - OPD vs IPD .....	5
Table 1-3 Quantity Metric .....	12
Table 1-4 Monetary Metrics .....	12
Table 1-5 TAT Metric.....	13
Table 2-1 Differentiation between Fraud and Abuse.....	31
Table 2-2 Examples of Fraud Schemes .....	38
Table 3-1 Selected fields across all tables/objects .....	69
Table 3-2 Primary Data Objects and their attributes.....	70
Table 3-3 Sample attribute values for all three data objects .....	71
Table 3-4 Sample data of Claim object .....	72
Table 3-5 Claim object fields values uniqueness & unavailability.....	72
Table 3-6 Sample data of Policy holder (Account) object.....	72
Table 3-7 Account object fields values uniqueness & unavailability .....	72
Table 3-8 Sample data of Policy object .....	73
Table 3-9 Policy object fields values uniqueness & unavailability .....	73
Table 3-10 Missing Values analysis from data objects.....	74
Table 3-11 Claim amount statistics.....	76
Table 3-12 Inconsistent invoice number values.....	78
Table 3-13 Data points and missing value count .....	81
Table 3-14 Claim Approved Amount range .....	96
Table 3-15 Customer-Policy Relationship .....	97
Table 3-16 Claims Amount Percentile Distribution-1 .....	99
Table 3-17 Claims Amount Percentile Distribution-2.....	99
Table 3-18 Customer Age correction using mean values .....	101
Table 3-19 Customers with Highest Severity .....	102
Table 3-20 Age characteristics.....	104
Table 3-21 Gender characteristics.....	105
Table 3-22 City level characteristics .....	106
Table 3-23 Pin code characteristics.....	106
Table 3-24 Customer Mobile characteristics .....	108
Table 3-25 Payment characteristics .....	109
Table 3-26 Product level characteristics .....	110
Table 3-27 Claims characteristics basis policy start date .....	111
Table 3-28 Claims characteristics basis policy end date.....	113
Table 3-29 Claims characteristics basis source and channel of purchase.....	114
Table 3-30 Customer Risk Score Rules.....	121
Table 3-31 Example - How customer risk score is generated.....	121
Table 4-1 Metric Summarization basis category .....	130
Table 4-2 Distribution of Incidence, Severity & population against risk score bins .....	133

## LIST OF FIGURES

Figure 1-1 Major Stages in OPD claims submission journey.....	10
Figure 1-2 Approved Claims trend - month wise .....	18
Figure 1-3 Forged/Fake documents submitted by syndicate entities.....	20
Figure 1-4 Duplicate case identification: CASE01357339, CASE01372976, CASE01382711.....	20
Figure 1-5 CASE02763863: Amount editing, CASE02872614: Date editing .....	22
Figure 1-6 CASE02875848 – Invoice & Prescription with different names .....	24
Figure 1-7 4 identical claims from same customer having sequential invoice .....	25
Figure 3-1 Trend of number of claims (Month over Month).....	61
Figure 3-2 Trend of number of total claims in INR (Month over Month).....	62
Figure 3-3 Trend of NOPs purchased .....	63
Figure 3-4 Claims Trend.....	63
Figure 3-5 Data Sources and Architecture .....	64
Figure 3-6 Final Dataset insights (Before Pre-Processing).....	74
Figure 3-7 Claim Amount Skewness .....	77
Figure 3-8 Case Number CASE01016781 .....	79
Figure 3-9 Final Dataset insights (After Preprocessing).....	80
Figure 3-10 Bill amount (amount asked by customer) histogram .....	83
Figure 3-11 Approved amount (Actual amount reimbursed to customer) histogram .....	84
Figure 3-12 Correlation between Bill Amount and Approved Amount .....	84
Figure 3-13 Pin code Distribution.....	85
Figure 3-14 Age Distribution.....	86
Figure 3-15 Age-Claims Correlation .....	87
Figure 3-16 Age-Claims-Gender Correlation .....	88
Figure 3-17 Distribution of insurance claims by gender .....	88
Figure 3-18 Box plot for Claim Status.....	90
Figure 3-19 Amount distribution at City level.....	91
Figure 3-20 Number of claims vs claimed amount.....	92
Figure 3-21 Claim Amount skewness.....	93
Figure 3-22 Claim Distribution at City level .....	93
Figure 3-23 Claim Amount Distribution at City level .....	94
Figure 3-24 Claims per Person .....	94
Figure 3-25 Relationship between Frequency and Severity .....	105
Figure 3-26 Frequency vs Severity at City level .....	107
Figure 3-27 Phone number series with claims severity .....	108
Figure 3-28 Frequency vs Severity at UPI level.....	109
Figure 3-29 Frequency vs Severity at Product level.....	111
Figure 3-30 Frequency vs Severity with Policy effective date .....	112
Figure 4-1 Fraud Detection Matrix .....	129
Figure 4-2 Confusion Matrix .....	133
Figure 4-3 Claims of user with risk score=100.....	141
Figure 4-4 Claims with risk score=100.....	142
Figure 4-5 High risk claims linked to same bank account.....	144
Figure 5-1 Mind Map.....	159

## CHAPTER I: INTRODUCTION

### 1.1 Introduction

The insurance industry is a critical component of the financial landscape, providing individuals and businesses with a diverse range of risk management products. These include health insurance, property insurance, life insurance, and more. Insurance fundamentally operates on the principle of spreading and mitigating risks among a pool of policyholders, offering financial protection against unforeseen events.

Within the realm of health insurance, various policies cater to different aspects of healthcare. One significant category is Outpatient (OPD) insurance, designed to cover medical expenses incurred outside of hospitalization, such as doctor visits, diagnostic tests, and prescribed medications. OPD insurance aims to provide policyholders with financial support for routine medical care and preventive services, contributing to overall well-being and reducing out-of-pocket expenses.

While insurance products are designed to bring financial security and peace of mind to policyholders, the industry faces challenges in maintaining profitability and ensuring fairness due to the increasing prevalence of fraudulent activities. Fraud can manifest in various forms, posing a particular challenge in OPD insurance. Instances of fraud in this domain may involve falsified claims, billing for services not actually rendered, or collusion between policyholders and healthcare providers. (Legotlo and Mutezo (2018))

#### **Fraud Committed by Medical Providers:**

- **Double billing:** Multiple claim intimation for the same service.
- **Phantom billing:** Charging for services or supplies that were never provided to the patient.
- **Unbundling:** Separately billing for components of a service that should be billed together.

- **Upcoding:** Billing for a more expensive service than what was actually provided to the patient.

#### **Fraud Committed by Patients and Other Individuals:**

- **Bogus marketing:** Deceiving individuals into sharing their health insurance information for fraudulent billing, identity theft, or enrolment in fictitious benefit plans.
- **Identity theft/identity swapping:** Illegally using another person's health insurance or allowing someone else to use your insurance.
- **Impersonating a healthcare professional:** Providing or billing for healthcare services or equipment without the required professional license.

#### **Fraud Involving Prescriptions:**

- **Forgery:** Creating or using forged prescriptions.
- **Diversion:** Using legally prescribed medications for illegal purposes, such as selling them.
- **Doctor shopping:** Visiting multiple healthcare providers to obtain multiple prescriptions for controlled substances or obtaining prescriptions from unethical medical practices.

This research is dedicated to investigating and addressing the challenge of abuse / fraud detection in OPD insurance, with a primary emphasis on introducing an innovative solution centered around customer risk-scoring using a rules-based approach. The pivotal objective of the research is to identify the key characteristics associated with abuse / fraud in OPD insurance, laying the foundation for the development of a comprehensive set of rules meticulously tailored to the intricacies of abuse / fraud detection. The unique feature of this proposed solution is its emphasis on customer risk scoring, which involves systematically assessing and scoring the risk associated with individual policyholders.

The research scope is deliberately confined to a rules-based approach, excluding the incorporation of machine learning algorithms. This deliberate limitation stems from a strategic focus on leveraging predefined rules to assess and score customer risks. The significance of this research lies in its contribution to the advancement of fraud detection methods within the OPD insurance sector. By crafting a nuanced approach, the study aims to curtail fraudulent

activities, thereby fortifying the protection of insurers and policyholders and fostering an environment conducive to improved industry profitability.

Acknowledging the inherent limitations, such as the constraints imposed by the availability and quality of data, and the potential constraints on the generalizability of results, the study lays the groundwork for subsequent research endeavors. These limitations, along with any emerging challenges, will be thoughtfully addressed and discussed in future research, ensuring a comprehensive understanding of the proposed rules-based customer risk scoring model for enhanced fraud detection in OPD insurance.

## **1.2 Background and context of the research problem**

Outpatient Department (OPD) insurance provides coverage for medical expenses incurred outside of a hospital stay. The rising demand for affordable healthcare and the increasing prevalence of non-hospital medical treatments have fueled the growth of OPD insurance. Approximately 62% of healthcare expenditure in India is allocated towards Outpatient Department (OPD) costs (Insurance Regulatory and Development Authority of India (2022)). This substantial portion of healthcare spending highlights the critical role of OPD services in the country's healthcare system. Patients seeking non-hospital-based medical care, consultations, diagnostic tests, and other essential healthcare services significantly contribute to the healthcare economy. However, this growth has also attracted fraudulent activities that jeopardize the integrity and sustainability of the insurance industry.

Insurance fraud is a pervasive problem that drains resources and leads to financial losses for both insurers and genuine policyholders. Fraudsters exploit various vulnerabilities within insurance processes, including filing false claims, exaggerating expenses, or seeking reimbursement for non-existent or non-covered treatments. Detecting and preventing fraud in the context of OPD insurance is essential to maintain the industry's viability and ensure that legitimate policyholders receive the benefits they deserve.

Fraud detection within the OPD insurance industry is a longstanding issue, with numerous methods having been employed to tackle it. Nevertheless, conventional approaches frequently fail to capture the unique aspects of OPD insurance fraud. The characteristics of outpatient medical treatments and the related claims complicate the detection of fraudulent activities with standard methods. Moreover, the ever-changing tactics used by fraudsters necessitate adaptive and innovative strategies to effectively combat their schemes.

In this research, we aim to enhance fraud detection in OPD insurance through the application of rules-based customer risk scoring approaches. These approaches involve the development and implementation of sophisticated rules and algorithms that evaluate individual policyholders' risk levels based on various parameters. By leveraging historical data, transaction patterns, and other relevant factors, these rules aim to identify suspicious claims and behavior indicative of potential fraud.

Before we deep dive into OPD insurance, in below table we highlight key differences between OPD and IPD insurance, and significant challenges faced in OPD.

*Table 1-1 Key differences between OPD & IPD*

Feature	IPD (Inpatient Department)	OPD (Outpatient Department)
Coverage	Hospitalization expenses for more than 24 hours	Doctor consultations, diagnostic tests, medications (without hospitalization)
Typical conditions	Surgeries, chronic illnesses, severe injuries	Minor illnesses, checkups, preventive care
Benefits	Room charges, doctor fees, surgery costs, nursing care, medication (during stay)	Consultation fees, lab tests, x-rays, some medications
Offered as	Standalone plan or part of comprehensive plan	Often as an add-on rider to an IPD plan

The above table presents nature of OPD and IPD insurance. Now in below table we present OPD vs IPD from an Abuse perspective.

Table 1-2 Abuse Penetration - OPD vs IPD

Parameter	IPD	OPD	Insights
Severity	Average 1,00,000 INR	Average 700 INR	Higher severity in IPD claims acts as a deterrent for fraudulent attempts.
Incidence Rate	Low	Very high	Higher frequency of OPD visits increases opportunities for fraudulent claims.
Ease of Claims	Complex documentation	Simple procedures	Simple claim procedures in OPD may facilitate easier submission of fraudulent claims.
Documents Required	Admission summaries, bills (~50 pages)	Bills, prescriptions (~2-3 pages)	Lesser documentation in OPD reduces scrutiny and facilitates fraudulent claims.
Provider Network	Network hospitals, standardized	Diverse clinics, less standardized	Less standardized network in OPD allows for easier inclusion of fake or fraudulent providers.

### Outpatient Department (OPD) Challenges:

- **Smaller Claim Amounts:** OPD claims typically involve smaller amounts compared to IPD claims. This makes it difficult to identify abuse patterns as anomalies may not stand out as significantly.
- **Service Verification:** Verifying if consultations and tests happened for OPD claims can be challenging. Insurance companies often rely on documentation and provider confirmation, which can be forged or manipulated. Also, investigation cost may become more expensive than the actual claim amount.
- **Treatment Necessity:** Unlike IPD claims with a clear hospitalization need, the necessity for some OPD services can be less clear-cut. This creates subjectivity in claim assessment, making it easier for providers or patients to justify unnecessary consultations or tests.
- **Higher Claim Volume:** OPD claims vastly outnumber IPD claims. This high volume makes it difficult to manually review each claim for potential fraud. Implementing automated fraud detection systems becomes crucial, but these require careful design to avoid flagging legitimate claims.

- **Provider Collusion:** Fraudulent OPD practices can involve collusion between patients and providers. Patients might be incentivized to receive unnecessary services for a kickback, or providers might inflate charges on claims submitted with cooperation from patients.

#### **Additional Considerations:**

- **Data Sharing:** Limited data sharing between healthcare providers and insurance companies can hinder abuse detection efforts. Sharing relevant data securely can help identify patterns of fraudulent activity across different facilities.
- **Regulatory Landscape:** The regulatory environment around healthcare abuse can vary. Insurance companies need to stay updated on relevant regulations to ensure their fraud detection methods comply with legal requirements.

Above listed challenges assure us that there is a need to develop a robust and agile abuse and fraud management system. Our rule-based customer risk scoring is one of the capabilities of the fraud management system.

This research seeks to employ a rules-based approach to develop and fine-tune rules capable of distinguishing between legitimate and fraudulent claims, thereby improving the efficiency of the claims processing system. The study aims to address the frequent problem of false positives in fraud detection by optimizing the rules and scoring mechanisms, reducing the unwarranted examination of valid claims.

The proposed solution is expected to have a significant impact on the OPD insurance industry by identifying high-risk customers, finding their syndicates and nexus, blocking their policies, recovering lost money, and reducing operational expenses.

By meeting its objectives, this research seeks to significantly enhance the effectiveness of fraud detection within the OPD insurance sector, thereby creating a more secure and trustworthy insurance environment for both insurers and policyholders.



### **1.3 Problem Statement**

Despite technological advancements and improved processes in the healthcare industry, fraudulent activities remain a significant concern for insurers and policyholders in the outpatient (OPD) insurance sector. OPD insurance fraud involves any intentional act or deception by a policyholder or healthcare provider that causes financial loss to the insurer. This issue presents substantial financial risks to the industry, resulting in higher premiums for policyholders and eroding trust in the insurance system. In recent years, the frequency and complexity of fraudulent activities have escalated, creating an urgent need for more effective fraud detection methods.

The existing fraud detection techniques in the OPD insurance industry have several limitations, primarily due to their reliance on manual and outdated processes. Insurers typically rely on a combination of manual review and basic business rules to detect fraudulent claims, which are time-consuming and not always effective. This approach also lacks scalability and agility, making it difficult to keep up with the evolving tactics used by fraudsters.

Hence, the primary issue addressed in this thesis revolves around the imperative necessity for a data driven and a more potent fraud detection methods specifically tailored for the OPD insurance industry. This research sets out to introduce a solution that not only elevates the fraud detection capabilities of insurers but also shields them from potential financial losses stemming from fraudulent activities. The proposed solution places a focus on a customer risk score-centric approach, aiming to provide insurers with a nuanced tool for risk assessment and fraud prevention.

In essence, the research endeavors to present a solution characterized by a rules-based methodology. This approach is designed to identify a diverse spectrum of fraudulent activities, with a specific emphasis on assessing and scoring the risk associated with individual policyholders. The comprehensive set of rules employed in this solution draws upon historical data and expert knowledge of fraudulent behavior, ensuring a thorough and informed evaluation.

While machine learning has shown promising results in various applications, there are certain reasons why a rule-based approach might be preferred over machine learning for fraud detection in some cases:

- **Interpretability:** Rule-based systems are more transparent and easier to interpret than complex machine learning models. Understanding the rules that trigger fraud alerts allows investigators to gain insights into the reasoning behind flagged transactions, which can be essential in fraud investigations.
- **Performance in Rare Events:** Fraudulent transactions are typically rare compared to legitimate ones, which can make it challenging for ML models to identify these infrequent events, often resulting in a higher number of false negatives (missed fraud cases). Rule-based systems, however, can be specifically crafted to identify distinct patterns and indicators of fraud more effectively.
- **Data Requirements:** Machine learning models need substantial amounts of labeled training data to perform effectively. In fraud detection, acquiring labeled data for fraudulent transactions is difficult due to their rarity and variability over time. Conversely, rule-based systems can be developed using expert domain knowledge and can function effectively with smaller datasets.
- **Adaptability:** Fraud patterns can change rapidly, and machine learning models might require continuous updates and retraining to stay effective. On the other hand, rule-based systems can be easily adapted and updated by domain experts to capture emerging fraud patterns.
- **Resource Constraints:** Training and deploying machine learning models can be computationally expensive and require significant computing resources. In scenarios with limited computational power or real-time processing requirements, rule-based systems can be more efficient.

## **Generic Claim Submission Journey:**

Submitting an OPD claim for insurance reimbursement involves several crucial steps to ensure that insured members receive timely and accurate compensation for medical expenses incurred.

Below are some general steps in Claims Submission Journey:

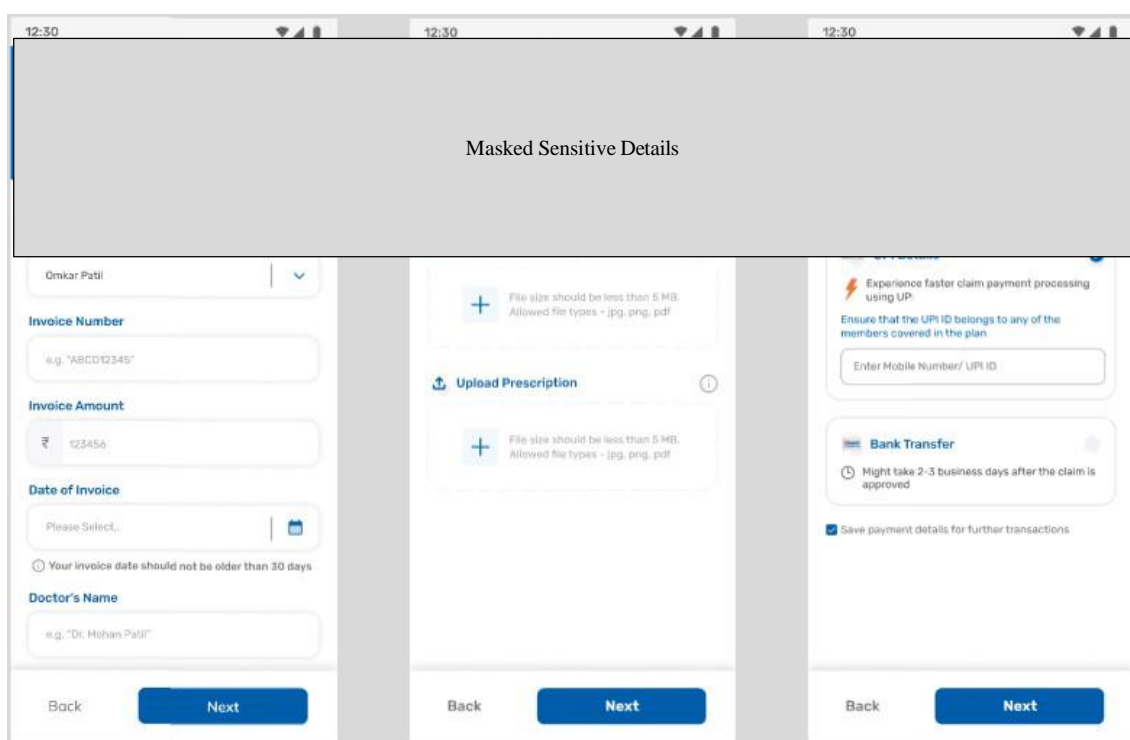
1. Basic Information:
  - a. Insured Person Details: Name, age, gender, contact information (phone number, email), policy number, and relationship to the policyholder (if applicable).
  - b. Policyholder Details: If different from the insured person, include name, contact information, and relationship to the insured.
  
2. Medical History (if applicable):
  - a. Pre-existing Conditions: Details of any pre-existing medical conditions relevant to the current claim. Usually not necessary in OPD claims.
  - b. Previous Treatments: Information about any previous medical treatments or surgeries that are related to the current claim. Usually not necessary in OPD claims.
  
3. Hospitalization Details:
  - a. Admission Date: Date when the insured was admitted to the hospital.
  - b. Discharge Date: Date when the insured was discharged from the hospital.
  - c. Treatment Details: Description of the medical treatment received during the hospital stay.
  - d. In most OPD claim submission we do not have above details, mostly we need invoice details like invoice number, amount and provider details like name, address.
  
4. Medical Records:
  - a. Invoice: Detailed breakdown of charges incurred during hospitalization.
  - b. Prescription: Copies of prescriptions issued by the treating physician.
  - c. Lab Reports: Results of any diagnostic tests conducted during the hospital stay.
  - d. Other Supporting Documents: Any additional documents that support the claim, such as referral letters, specialist reports, or discharge summary.

5. Billing Details:

- a. Breakdown of Hospital Bill: Itemized list of charges for services, procedures, medications, and supplies provided during the hospital stay.
- b. Explanation of Benefits (EOB): Summary of what the insurance company will cover and any out-of-pocket expenses.

6. Bank Account Information:

- a. Bank Name: Name of the insured's bank.
- b. Account Holder Name: Name of the person to whom the reimbursement payment should be made.
- c. Account Number: Bank account number for direct deposit.
- d. Can also be details for other modes of payment like UPI.



*Figure 1-1 Major Stages in OPD claims submission journey.*

Above steps and reference figure (1-1) highlight how easy it is to file an OPD claim. Just by providing mere details like name, invoice number, date, provider details, by submitting an invoice and prescription and submitting payment details, we can easily get reimbursement for healthcare services. All these details are very easy to fabricate and further strengthen the need for a strong abuse and fraud prevention framework.

## 1.4 Research Questions

This research seeks to address crucial aspects of outpatient insurance fraud detection. The proposed research questions explore the key characteristics of fraud, assess the effectiveness of a rules-based approach, and evaluate its efficiency in combating outpatient insurance fraud:

1. What key characteristics define outpatient insurance fraud, encompassing traits like falsified claims and collusion between policyholders and providers?
2. How well does a rules-based approach identify outpatient insurance fraud, utilizing historical data and expert insights?
3. How streamlined is a rules-based method for outpatient insurance fraud detection, balancing thoroughness, and efficiency as compared to traditional methods?

## 1.5 Objectives

Grounded within the research framework and oriented around PMBOK, PRINCE2, and AI in Agile settings, the study manifests the subsequent objectives:

1. **Objective 1:** To identify the key characteristics of outpatient insurance fraud and the challenges associated with detecting it.  
By understanding the key characteristics of outpatient insurance fraud, the study aims to provide insights into the common patterns, techniques, and indicators used by fraudsters. The study also aims to shed light on the challenges faced in detecting outpatient insurance fraud.
2. **Objective 2:** To develop a comprehensive set of rules for customer risk scoring to detect outpatient insurance fraud.  
To achieve this objective, the study will involve an in-depth analysis of historical data, patterns, and characteristics of fraudulent outpatient claims. The aim is to identify key indicators and red flags that can help distinguish fraudulent claims from legitimate ones. This analysis may involve examining various data sources, such as claim forms, medical records, billing codes, and transactional data.
3. **Objective 3:** To assess the effectiveness of the rules-based customer risk scoring approach in detecting outpatient insurance fraud by comparing it to traditional methods.

By doing so, the study aims to advance fraud detection techniques in the insurance industry.

Below tables represent sample KPIs that can be used to measure effectiveness & efficiency of proposed solution: -

\* *Entity heading means the category/description of subsequent values*

*Table 1-3 Quantity Metric*

ENTITY	Formula	Month 1	Month 2
Flagged	x	30% of Entity	20% of Entity
Investigated	y	-	-
Fraud	y1	-	-
Fraud %	y1/y	0.6	0.80
Suspicious	y2	-	-
Suspicious %	y2/y	0.2	0.20
Genuine	y3	-	-
Genuine %	y3/y	0.2	0.00

Table 1-3 focuses on the quantity of customers who were investigated and the percentage share between fraud, suspicious and genuine.

*Table 1-4 Monetary Metrics*

ENTITY	Formula	Month 1	Month 2
Total Amount/Services	x	-	-
Amount/Services Utilized	y	-	-
Amount Prevented	x-y	-	-
Amount Prevented%	(x-y)/x	0.4	0.70

Table 1-4 highlights cost related metrics - total amount, how much amount was utilized and

percentage of amount that was saved from being lost.

*Table 1-5 TAT Metric*

ENTITY	Formula	Month 1	Month 2
First Claim/Appointment	P date	-	-
Flagged Date	Q date	-	-
Investigation Date	R date	-	-
Block listed Date	S date	-	-
Flagging TAT Days	Q - P days	2 Days	0 Days
Active TAT Days	S – P days	-	-
Action TAT Days	S – Q days	-	-

Table 1-5, TAT metric is more towards efficiency - how are we able to reduce the time taken from highlighting suspicious customers to investigation followed by final status update.

## 1.6 Scope and limitations of the study

The scope of this study is focused on outpatient (OPD) insurance customer risk scoring model for fraud detection and proposes a rules-based approach to enhance fraud detection in this field. The study specifically investigates the fraud detection methods used by OPD insurers, identifies the key characteristics of OPD insurance fraud, and develops a comprehensive set of rules for customer risk scoring. The study also examines the effectiveness and efficiency of the proposed solution in identifying high-risk customers, detecting syndicates and nexus, blocking their policies, and reducing operational expenses.

The limitations of this study include the exclusive use of rule-based algorithms in the proposed solution, which may not capture all types of fraud. The study also relies on a limited dataset

for analysis, which may not fully represent the diversity of OPD insurance fraud cases. Additionally, the study does not consider the legal and regulatory frameworks in different countries, which may have implications for fraud detection methods and strategies.

Despite these limitations, the study aims to provide valuable insights and contribute to the development of more effective customer risk scoring models for fraud detection methods in the OPD insurance industry. The proposed solution and findings can be used as a foundation for future research to further enhance fraud detection and prevention in this field.

### **1.7 Significance and relevance of the research**

The proposed research on customer risk scoring model enhancing fraud detection in outpatient (OPD) insurance using a rule-based approach is significant and relevant for several reasons.

First, the OPD insurance industry has been facing more fraudulent activities in recent years. According to Deloitte's latest Insurance Fraud Survey conducted in 2023 (Deloitte (2023)), an alarming 60 percent of the surveyed respondents indicated a substantial increase in incidents of fraud within the insurance sector. Additionally, an additional 10 percent noted a marginal uptick in fraudulent activities. These findings underscore a growing concern within the industry regarding the surge in fraudulent practices, necessitating urgent and strategic measures to combat and mitigate such risks effectively.

These fraudulent activities not only cause financial losses for insurance companies but also adversely affect policyholders, who may encounter difficulties in having their legitimate claims approved. Therefore, the proposed solution aims to tackle this issue by developing a more effective rule-based customer risk scoring model for fraud detection, which can identify fraudulent activities and minimize false positives.

Secondly, the proposed solution has practical implications for the OPD insurance industry by helping insurers to better understand the characteristics of fraud and develop appropriate



strategies to prevent it. This can result in significant cost savings for the insurance companies, as well as a more streamlined claims processing and investigation process.

Thirdly, the research enhances the field of rule-based customer risk scoring models for fraud detection by offering insights into the effectiveness and efficiency of a rules-based approach in the context of OPD insurance fraud. These insights can guide the development of future fraud detection systems and strategies within the industry.

Finally, the proposed research is relevant to policymakers responsible for regulating the OPD insurance industry. The study's findings can inform policy decisions related to fraud prevention and detection, which can benefit consumers by improving the industry's overall integrity.

Overall, the proposed research has significant implications for the OPD insurance industry and the field of rule-based customer risk scoring model for fraud detection, making it a relevant and important area of study.

## **1.8 Business Relevance**

The outpatient (OPD) health insurance sector is increasingly vulnerable to abuse and fraud, posing significant challenges for insurers. Effective fraud detection and prevention mechanisms are essential for safeguarding the financial health of insurance providers and maintaining customer trust. This section explores the critical business perspectives related to enhancing abuse and fraud detection in OPD insurance through rules-based customer risk scoring approaches.

By implementing a sophisticated, rules-based risk scoring system, insurers can achieve numerous advantages, including improved operational efficiency, enhanced financial performance, and increased customer satisfaction. Conversely, without such enhancements, insurers will face growing challenges that can hinder their business operations and competitiveness.

The following subsections delve into the specific benefits of adopting a rules-based approach to fraud detection and highlight the escalating business challenges that insurers will encounter

without this advancement.

## **1. Operational Efficiency**

### 1.1 Challenges

- **Manual Fraud Detection:** Consumes time and resources, reducing efficiency.
- **Delayed Processing:** Affects customer satisfaction and increases administrative costs.

### 1.2 Gains

- **Streamlined Detection Process:** Automation reduces manual efforts and allocates resources more effectively.
- **Proactive Risk Management:** Early detection and intervention prevent fraudulent claims from progressing.

## **2. Financial Performance**

### 2.1 Challenges

- **Uncontrolled Fraudulent Claims:** Leads to substantial financial losses and higher payout ratios.

### 2.2 Gains

- **Cost Reduction:** Effective detection mitigates financial losses from fraudulent claims.
- **Recovery of Funds:** Identifying fraud syndicates enables recovery of lost funds.
- **Premium Pricing Accuracy:** Accurate fraud detection maintains actuarial integrity for competitive pricing.

## **3. Customer Trust and Satisfaction**

### 3.1 Challenges

- Eroding Trust: Frequent fraud incidents and processing delays erode customer trust.
- Negative Publicity: Damages the insurer's market reputation.

### 3.2 Gains

- Enhanced Trust: Reliable fraud prevention enhances customer loyalty.
- Fair Treatment: Swift and fair processing of genuine claims maintains positive customer relationships.

## 4. Regulatory Compliance

### 4.1 Challenges

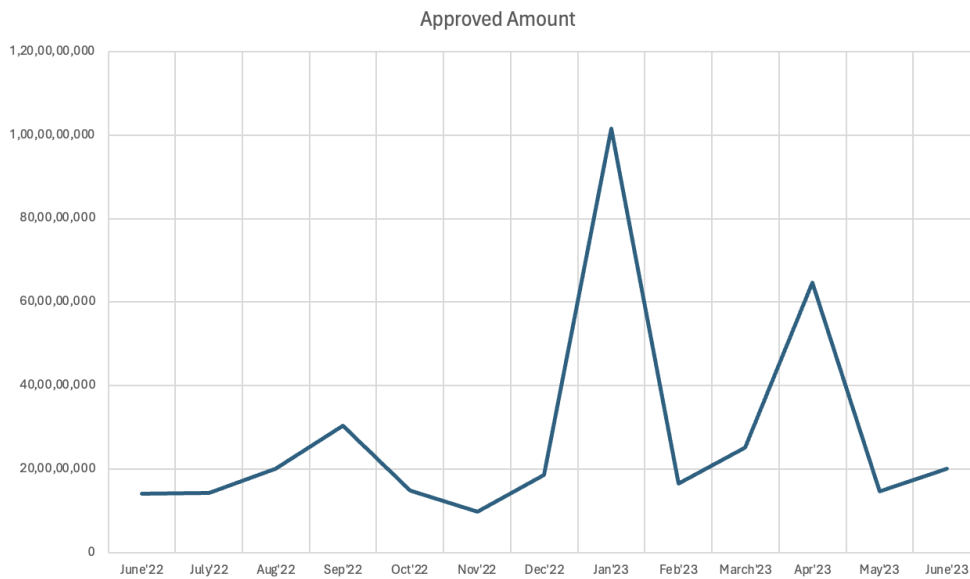
- Non-Compliance Penalties: Leads to legal penalties and increased scrutiny.

### 4.2 Gains

- Transparency and Reporting: Facilitates detailed and transparent reporting to regulators and stakeholders.

In the dynamic landscape of outpatient (OPD) health insurance, the growing incidence of fraud and abuse poses substantial challenges for insurers. Below analysis delves into the data provided to underscore the critical need for enhancing fraud detection mechanisms.

By examining the dataset of insurance claims, policies, and customer interactions over a series of months, this analysis highlights the fluctuations and anomalies that suggest potentially fraudulent activities. It also outlines the business advantages of implementing sophisticated fraud detection methods and the escalating challenges insurers face without such enhancements.



*Figure 1-2 Approved Claims trend - month wise.*

Based on the provided table, we can derive several insights that highlight the need for enhancing abuse/fraud detection in OPD insurance through a rules-based customer risk scoring approach:

- **Fluctuations in Claims and Approved Amounts:**
  - From Table 1-6 and Figure 1-2 we observe that there are significant fluctuations in the number of claims and the approved amounts over the observed period. For instance, the approved amount peaked dramatically in January 2023 (1,01,60,39,545) and April 2023 (64,76,33,387).
  - These sudden spikes could indicate potentially fraudulent activities or abuse of the insurance policies during these periods, necessitating a more rigorous detection method.
  
- **High Volume of Policies and Customers:**
  - The dataset shows a consistently high number of policies and customers each month.
  - In June 2023, the number of policies reached 554,664 and the number of customers was 545,945. Managing fraud detection manually in such a large

dataset is impractical, underlining the need for automated, rules-based approaches to effectively handle the volume and complexity.

- Correlation Between Claims and Lives Covered:
  - Months with higher lives covered, like January 2023 (329,604) and April 2023 (388,722), also show higher claims.
  - Investigating these correlations can help identify patterns that are indicative of fraud, reinforcing the need for sophisticated risk scoring systems.
- Sustainability of Claims and Payouts:
  - The sustainability of the insurance business model could be threatened by unchecked fraudulent claims.
  - Consistently high or increasing approved amounts without corresponding increases in policy numbers or lives covered could indicate inefficiencies or vulnerabilities to fraud.

Below we present several different types of abuse-based / fraud activities those were captured historically. The aim is to highlight the variation with which fraudulent activities are being committed while emphasizing the monetary loss an insurance company must incur due to such incidences.

### 1.8.1 Syndicate Identification

<div style="text-align: center; background-color: #f0f0f0; padding: 5px;">Masked Hospital Name</div> <div style="text-align: center; background-color: #f0f0f0; padding: 5px;">Masked Customer and Doctor name</div> <table border="1" style="width: 100%; border-collapse: collapse; font-size: 8px;"> <thead> <tr> <th>Phone No</th> <th>Email</th> </tr> </thead> <tbody> <tr> <td colspan="2"><b>Sr No Services Booked</b></td> </tr> <tr> <th>Department</th> <th>Price</th> </tr> <tr> <td>1. ESR (ERYTHROCYTE SEDIMENTATION RATE) PTHA</td> <td>110.00</td> </tr> <tr> <td>2. BUNAM* - FASTING PTBC</td> <td>70.00</td> </tr> <tr> <td>3. E.C.G. CARD</td> <td>200.00</td> </tr> <tr> <td>4. CBC WITH PLATELET (THROMBOCYTE) COUNT PTHA</td> <td>300.00</td> </tr> </tbody> </table>	Phone No	Email	<b>Sr No Services Booked</b>		Department	Price	1. ESR (ERYTHROCYTE SEDIMENTATION RATE) PTHA	110.00	2. BUNAM* - FASTING PTBC	70.00	3. E.C.G. CARD	200.00	4. CBC WITH PLATELET (THROMBOCYTE) COUNT PTHA	300.00	<div style="text-align: center; background-color: #f0f0f0; padding: 5px;">Masked Hospital Name</div> <div style="text-align: center; background-color: #f0f0f0; padding: 5px;">Masked Customer and Doctor name</div> <table border="1" style="width: 100%; border-collapse: collapse; font-size: 8px;"> <thead> <tr> <th>Phone No</th> <th>Email</th> </tr> </thead> <tbody> <tr> <td colspan="2"><b>Sr No Services Booked</b></td> </tr> <tr> <th>Department</th> <th>Price</th> </tr> <tr> <td>1. ESR (ERYTHROCYTE SEDIMENTATION RATE) PTHA</td> <td>110.00</td> </tr> <tr> <td>2. BUNAM* - FASTING PTBC</td> <td>570.00</td> </tr> <tr> <td>3. E.C.G. CARD</td> <td>200.00</td> </tr> <tr> <td>4. CBC WITH PLATELET (THROMBOCYTE) COUNT PTHA</td> <td>800.00</td> </tr> <tr> <td>5. CREATININE - BLOOD PTBC</td> <td>190.00</td> </tr> </tbody> </table>	Phone No	Email	<b>Sr No Services Booked</b>		Department	Price	1. ESR (ERYTHROCYTE SEDIMENTATION RATE) PTHA	110.00	2. BUNAM* - FASTING PTBC	570.00	3. E.C.G. CARD	200.00	4. CBC WITH PLATELET (THROMBOCYTE) COUNT PTHA	800.00	5. CREATININE - BLOOD PTBC	190.00	<div style="text-align: center; background-color: #f0f0f0; padding: 5px;">Masked Hospital Name</div> <div style="text-align: center; background-color: #f0f0f0; padding: 5px;">Masked Customer and Doctor name</div> <table border="1" style="width: 100%; border-collapse: collapse; font-size: 8px;"> <thead> <tr> <th>Sr#</th> <th>Service Particulars</th> <th>Unit</th> <th>Amount</th> </tr> </thead> <tbody> <tr> <td>1.</td> <td>UW HMO</td> <td>1</td> <td>2250</td> </tr> <tr> <td>2.</td> <td>2.5L DIALY</td> <td>1</td> <td>2000</td> </tr> <tr> <td>3.</td> <td>LPT</td> <td>1</td> <td>600</td> </tr> <tr> <td>4.</td> <td>HST</td> <td>1</td> <td>650</td> </tr> <tr> <td>5.</td> <td>COMPLETE BLOOD COUNT</td> <td>1</td> <td>300</td> </tr> </tbody> </table>	Sr#	Service Particulars	Unit	Amount	1.	UW HMO	1	2250	2.	2.5L DIALY	1	2000	3.	LPT	1	600	4.	HST	1	650	5.	COMPLETE BLOOD COUNT	1	300
Phone No	Email																																																							
<b>Sr No Services Booked</b>																																																								
Department	Price																																																							
1. ESR (ERYTHROCYTE SEDIMENTATION RATE) PTHA	110.00																																																							
2. BUNAM* - FASTING PTBC	70.00																																																							
3. E.C.G. CARD	200.00																																																							
4. CBC WITH PLATELET (THROMBOCYTE) COUNT PTHA	300.00																																																							
Phone No	Email																																																							
<b>Sr No Services Booked</b>																																																								
Department	Price																																																							
1. ESR (ERYTHROCYTE SEDIMENTATION RATE) PTHA	110.00																																																							
2. BUNAM* - FASTING PTBC	570.00																																																							
3. E.C.G. CARD	200.00																																																							
4. CBC WITH PLATELET (THROMBOCYTE) COUNT PTHA	800.00																																																							
5. CREATININE - BLOOD PTBC	190.00																																																							
Sr#	Service Particulars	Unit	Amount																																																					
1.	UW HMO	1	2250																																																					
2.	2.5L DIALY	1	2000																																																					
3.	LPT	1	600																																																					
4.	HST	1	650																																																					
5.	COMPLETE BLOOD COUNT	1	300																																																					
<div style="text-align: center; background-color: #f0f0f0; padding: 5px;">Masked Hospital Name</div> <div style="text-align: center; background-color: #f0f0f0; padding: 5px;">Masked Customer and Doctor name</div> <table border="1" style="width: 100%; border-collapse: collapse; font-size: 8px;"> <thead> <tr> <th>Test Name</th> <th>Result</th> <th>Unit</th> <th>Ref. Int.</th> <th>Interpret</th> </tr> </thead> <tbody> <tr> <td colspan="5"><b>PROTHROMBIN TIME (PT TIME) - SICULAR CITRATE PLASMA</b></td> </tr> <tr> <td>TECH</td> <td>10.00</td> <td>Seconds</td> <td>10-15 Sec</td> <td>Operational/Not Ok</td> </tr> <tr> <td>CONTROL PT</td> <td>11.00</td> <td>Seconds</td> <td></td> <td>Operational/Not Ok</td> </tr> <tr> <td>P INDEX</td> <td>115.00</td> <td>%</td> <td></td> <td>Operational/Not Ok</td> </tr> <tr> <td>PT INDEX</td> <td>0.87</td> <td></td> <td></td> <td>Normal</td> </tr> <tr> <td>HR</td> <td>0.87</td> <td></td> <td></td> <td>Normal</td> </tr> </tbody> </table>	Test Name	Result	Unit	Ref. Int.	Interpret	<b>PROTHROMBIN TIME (PT TIME) - SICULAR CITRATE PLASMA</b>					TECH	10.00	Seconds	10-15 Sec	Operational/Not Ok	CONTROL PT	11.00	Seconds		Operational/Not Ok	P INDEX	115.00	%		Operational/Not Ok	PT INDEX	0.87			Normal	HR	0.87			Normal	<div style="text-align: center; background-color: #f0f0f0; padding: 5px;">Masked Hospital Name</div> <div style="text-align: center; background-color: #f0f0f0; padding: 5px;">Masked Customer and Doctor name</div> <table border="1" style="width: 100%; border-collapse: collapse; font-size: 8px;"> <thead> <tr> <th>SrNo.</th> <th>Particulars</th> <th>Amount</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>CONSULTANT FEES</td> <td>3500.00</td> </tr> </tbody> </table>	SrNo.	Particulars	Amount	1	CONSULTANT FEES	3500.00	<div style="text-align: center; background-color: #f0f0f0; padding: 5px;">Masked Hospital Name</div> <div style="text-align: center; background-color: #f0f0f0; padding: 5px;">Masked Customer and Doctor name</div> <table border="1" style="width: 100%; border-collapse: collapse; font-size: 8px;"> <thead> <tr> <th>SrNo.</th> <th>Particulars</th> <th>Amount</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>CONSULTANT FEES</td> <td>3160.00</td> </tr> </tbody> </table>	SrNo.	Particulars	Amount	1	CONSULTANT FEES	3160.00							
Test Name	Result	Unit	Ref. Int.	Interpret																																																				
<b>PROTHROMBIN TIME (PT TIME) - SICULAR CITRATE PLASMA</b>																																																								
TECH	10.00	Seconds	10-15 Sec	Operational/Not Ok																																																				
CONTROL PT	11.00	Seconds		Operational/Not Ok																																																				
P INDEX	115.00	%		Operational/Not Ok																																																				
PT INDEX	0.87			Normal																																																				
HR	0.87			Normal																																																				
SrNo.	Particulars	Amount																																																						
1	CONSULTANT FEES	3500.00																																																						
SrNo.	Particulars	Amount																																																						
1	CONSULTANT FEES	3160.00																																																						

Figure 1-3 Forged/Fake documents submitted by syndicate entities.

In our continuous efforts to combat insurance fraud, we continuously analyze historical data to scrutinize suspicious cases based on certain hypotheses. These hypotheses focused on detecting unusual claim activities, such as abnormal claim amounts and repetitive usage of provider and bank details. The analysis flagged cases that fit these patterns, leading us to identify a fraud syndicate exploiting the system. By drilling down into the flagged cases, we observed commonalities in documents and case properties, particularly in claim amounts, provider details, and bank details.

Through comprehensive syndicate analysis, we mapped out all entities involved in the suspicious activities, linking them via UPI IDs, bank details, device IDs, and email IDs. The fraudulent syndicate involved 63 interconnected entities linked to 261 cases, of which 211 were paid. The suspects used 36 different bank accounts across 111 policies. Financially, the total wallet amount allocated was ₹12,00,000, out of which ₹8,00,000 was utilized by the syndicate.

Our findings revealed that syndicate members consistently purchased low-cost, high-benefit products, enabling them to maximize the wallet amount allocated for reimbursements. Most suspects utilized the entire wallet amount through single reimbursement claims per benefit, minimizing the number of transactions to avoid detection. A notable trend was the repeated use of the same bank details across multiple cases, with suspects editing the account holder names to create the illusion of different accounts.

### 1.8.2 Duplicate Claims Submission

The figure shows three side-by-side screenshots of insurance claim documents. Each document has a header with 'Masked Hospital Name' and 'Masked Customer and Doctor name'. Below the header is a table with columns: #, Service Code, Service Name, NAC Code, Rate, Discount, and Total. The data in the tables is identical across all three screenshots, indicating duplicate submissions. At the bottom of each document, there is a summary section with fields for Settlement, Payment, Receipt No, Mode, Amount, Total Reimburse, and Net Bill Amount.

Figure 1-4 Duplicate case identification: CASE01357339, CASE01372976, CASE01382711

The above findings clearly highlight the submission of duplicate claims, with three claims from

Apollo Hospital exhibiting identical invoices. The only difference was the patient's name, while all other details, such as bill number, bill date and time, amount, and doctor's name, remained consistent across the claims. Upon further investigation, we discovered 11 more historical instances with identical invoices. In some cases, even the lab test results were the same for different customers. The overall amount paid for these fraudulent cases totalled ₹66,000.

Identifying such cases necessitates real-time document duplicity detection, meaning that each document submitted (invoice or lab report) by a customer needs to be compared against millions of existing documents swiftly and accurately. This real-time comparison poses a significant challenge due to the need for fast response times while maintaining accuracy in detection. Duplicate document checks are crucial because they represent one of the simplest yet most damaging methods of committing fraud, potentially resulting in substantial financial losses for the insurer.

To address this, we must implement advanced algorithms capable of efficiently scanning and comparing new submissions against our extensive database of existing documents. Techniques such as optical character recognition (OCR), natural language processing (NLP), and machine learning can be employed to identify similarities and flag potential duplicates. Additionally, leveraging cloud computing and parallel processing can help manage the vast volume of data and ensure quick turnaround times.

Enhancing our document verification process will involve continuous improvement and adaptation of our algorithms to stay ahead of sophisticated fraudulent tactics. Integrating these advanced technologies will not only streamline our fraud detection capabilities but also ensure that legitimate claims are processed swiftly, maintaining customer satisfaction while safeguarding our financial resources.

1.8.3 Document Tampering

Masked Hospital Name

Masked Customer and Doctor name

RECEIPT NO: 252      DATE: 30/06/2023

NAME: .....      PAYMENT: 1500/-

AGE/SEX: .....      PHONE NO: .....

ADDRESS: .....

Check One:  Annual Examination  Follow-up Visit  Emergency  Other *Consultation*

*One thousand Five hundred*

PARTICULARS AMOUNT(RS.) 1500/-

Masked Hospital Name

Masked Customer and Doctor name

NT: 1500/-

Masked Hospital Name

Masked Customer and Doctor name

171 cm.  
1201100 mmHg  
86 min.

15/07/2023 10:39 AM

Sl	Particulars	SAC	Ref. Code	Date	Qty	Unit Rate	Gross Amount	Disc. Amt	Net Amt
1	Registration		999311	15/07/2023	1	1,000.00	1,000.00	0.00	1,000.00
Sub Total :							1,000.00	0.00	1,000.00
Sl No.	Date	Receipt No	Original Amt.		Adjusted Amt.				
1	11/05/2023 10:58	RCF-1718379	1,000.00		1,000.00				
Total								Payable Total :	1,000.00
								Discount Amount :	0.00
								Advance :	0.00
								Pending Payment :	
								Final Payment :	1,000.00

Rs. One Thousand Only.

Masked Hospital Name & Address

2023 10:39 AM

Figure 1-5 CASE02763863: Amount editing, CASE02872614: Date editing



Document tampering represents another significant form of fraudulent activity in reimbursement claims. This type of fraud involves altering invoices, lab reports, or other supporting documents to inflate claim amounts, change service dates, or manipulate other critical information to unjustly obtain higher reimbursements. Perpetrators might edit scanned documents using image editing software or even physically alter paper documents before scanning and submitting them. As seen in Figure above, in the first claim, amount has been edited from Rs. 500 to Rs. 1500 by writing “one thousand” in front of actual consultation fees and by adding “1” in front of invoice amount. Similarly in second image we can clearly see the date being stand out from rest of the document because of its font and size compared to the text on the remaining areas of the document.

Identifying such tampered documents by the naked eye is exceedingly challenging. Fraudsters often employ sophisticated techniques to ensure that the alterations are nearly indistinguishable from genuine entries. Minor changes, such as altering numbers or dates, can be especially difficult to detect without a detailed and time-consuming manual review. Furthermore, with the volume of claims that insurers process daily, relying solely on human inspection is impractical and inefficient.

Advanced fraud detection systems must be deployed to tackle this challenge effectively. These systems can use machine learning algorithms to detect anomalies and inconsistencies in the documents. Optical character recognition (OCR) can convert scanned documents into text that can be analysed for signs of tampering. Additionally, image analysis techniques can compare the pixel patterns in documents to identify alterations that might not be visible to the naked eye.

For instance, inconsistencies in font styles, sizes, or alignments can indicate tampering. Machine learning models can be trained on large datasets of genuine and tampered documents to recognize subtle patterns indicative of fraud. By implementing these technologies, insurers can enhance their ability to detect tampered documents, thereby reducing fraudulent claims and protecting their financial interests.

## 1.8.4 Customer Name Mismatch

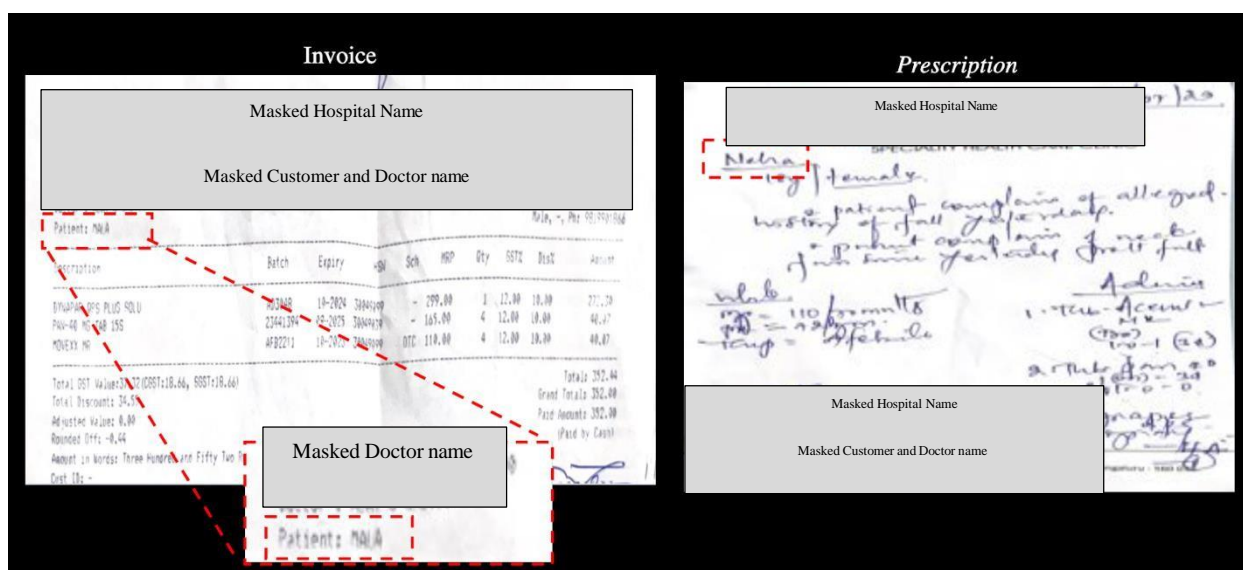


Figure 1-6 CASE02875848 – Invoice & Prescription with different names

In our ongoing efforts to detect and prevent fraud in OPD insurance claims, a critical aspect of our analysis involved the scrutiny of submitted documents. One such fraudulent pattern that emerged was the mismatch of customer names across different documents. This section highlights a specific case where the invoice and prescription submitted by a claimant exhibited inconsistencies in the patient's name, indicating potential fraud.

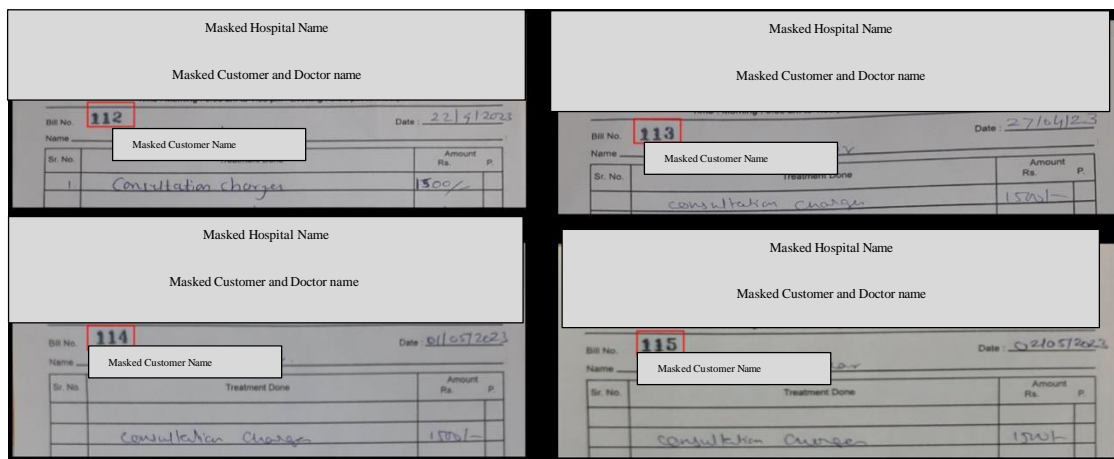
As seen in figure, the invoice submitted showed the patient's name as "Mala", whereas the prescription showed a different name, "Neha", which did not match the name on the invoice. Processing thousands of OPD insurance claims daily is a highly demanding task that requires considerable manual effort. The volume alone presents a significant challenge, and when combined with the need for meticulous attention to detail, the difficulty increases exponentially.

Manual processors face several hurdles in maintaining the necessary level of alertness to effectively identify fraudulent claims, particularly those involving handwritten documents. Handwritten invoices and prescriptions vary significantly in legibility and style. Deciphering these documents requires careful attention, which is difficult to sustain over long periods.

Manual processing of OPD insurance claims is fraught with challenges, particularly when it involves the scrutiny of handwritten documents. The combination of high claim volumes,

variability in document formats, and the need for continuous alertness makes it difficult for manual processors to effectively identify fraudulent activities. This underscores the necessity for advanced, automated fraud detection mechanisms that can efficiently handle large datasets, recognize patterns, and flag discrepancies, thereby enhancing the overall accuracy and reliability of the insurance claims process.

### 1.8.5 Identification of Sequential Invoices



*Figure 1-7 4 identical claims from same customer having sequential invoice*

Detecting fraudulent activity, such as the submission of sequential invoices by the same customer, is an intricate task that becomes significantly more challenging when handled manually. The provided example illustrates how a customer submitted multiple invoices with bill numbers differing by a constant number (in this case, by 1), indicating potential fraud. The invoices are numbered sequentially (112, 113, 114, 115), a subtle pattern that can easily go unnoticed during manual processing, especially when claims are reviewed by different processors. Such small variations between invoices can be difficult to detect without a system designed to recognize and flag these patterns automatically.

Different claims are often processed by different individuals or teams. This segmentation means that a processor handling a single claim is unlikely to have visibility into the sequence of invoices submitted by the same customer. With thousands of claims processed daily, identifying a sequential pattern among a vast number of invoices is nearly impossible without automated assistance.

The identification of sequential invoice fraud, where a customer submits invoices with numbers differing by a constant value, is a prime example of the limitations of manual processing.

## 1.9 Thesis Structure

This research thesis has been divided into five chapters. The structure of thesis is as follows:

**Chapter I - Introduction:** This chapter highlights the core aim of the thesis, outlines its objectives, and emphasizes the research's significance.

**Chapter II - Literature Review:** This chapter conducts a comprehensive review of the relevant literature, focusing on key variables aligned with the research objectives and providing precise operational definitions for terms.

**Chapter III - Methodology:** This chapter elucidates the chosen research methodology and details the strategies employed for data analysis, as well as the process leading to the ultimate conclusions.

**Chapter IV - Results:** This chapter provides a detailed analysis of the data collected during the research, accompanied by a thorough discussion of the results and their interpretation, leading to conclusive insights.

**Chapter V- Discussion, Conclusions, and Implications:** This chapter further illuminates the insights derived from data analysis and visualization, advancing the narrative towards a comprehensive conclusion.

## CHAPTER II: LITERATURE REVIEW

### 2.1 Overview of the insurance industry and fraud detection

India, with its population of 1.3 billion people, is a rapidly developing country. However, its healthcare system is complex and fragmented, comprising both public and private providers. The public sector primarily relies on government funding, while the private sector is largely supported by out-of-pocket payments. Unfortunately, a significant portion of the population, approximately 30% or 40 crore individuals, known as the "missing middle" (Kumar and Sarwal (2021)), lack any financial protection for healthcare expenses, including medicines, vaccinations, and diagnostics. Consequently, they must bear all medical costs themselves due to the absence of health insurance coverage.

For the remaining 70% of the population with health insurance plans (whether public, private, or government-sponsored), coverage is typically limited to incidents requiring in-patient care or hospitalization. Other healthcare expenses such as doctor consultations (both online and in-clinic), lab tests, medication, and diagnoses, commonly referred to as out-of-pocket expenditures (OOPE), are typically not covered by insurance plans in India. This disparity underscores the need for outpatient department (OPD) insurance coverage, which encompasses doctor consultations, diagnoses, and short treatments that do not require hospitalization. OPD insurance plans offer financial assistance for various health-related issues, alleviating the burden of medical expenses, especially as medical inflation continues to rise.

In India, approximately 62% of total healthcare spending constitutes OOPE, with 65% of this spending allocated to OPD treatments. This highlights the significance of OPD insurance coverage (Maiti (2021)). Unlike hospitalization incidents, OPD incidents are more frequent across all age groups. Acquiring OPD coverage from an early age enables insured individuals to take preventive measures for maintaining good health, reducing the need for frequent hospitalization.

Recognizing the importance of including OPD coverage in conventional health insurance plans, several Indian Health Tech startups have collaborated with insurance providers to offer comprehensive health insurance plans that cover various OPD expenses. These plans include benefits such as preventive health checkups, lab tests, in-clinic, and tele-consultations, all at a nominal cost. However, fraudsters often exploit these insurance benefits in various ways.

The Indian healthcare system is structured into three tiers: primary, secondary, and tertiary care. Primary care is delivered by general practitioners, community health workers, and primary health centers. Secondary care involves specialists and hospitals, while tertiary care is provided by specialized hospitals and medical colleges. To enhance the healthcare system, the Indian government has launched initiatives like the National Health Mission, National Rural Health Mission, and National Urban Health Mission (Ministry of Health and Family Welfare (2012)). Private hospitals, clinics, and health insurance companies also play vital roles in extending healthcare access, especially in urban areas, and offering financial coverage for individuals unable to afford out-of-pocket expenses. Despite these efforts, the Indian healthcare system continues to grapple with several challenges, particularly in the realm of detecting and preventing fraud (Chokshi et al., (2016)).

Some of the key challenges include a lack of standardization within the healthcare system, inadequate technology and infrastructure for effective fraud detection, a lack of awareness among consumers regarding OPD fraud, unscrupulous providers exploiting the system, limited legal and regulatory frameworks for addressing fraud, complex and time-consuming investigations, and insufficient data management systems for tracking fraudulent claims and monitoring potential abuses. (He et al. (2020))

Common types of insurance claims include medical, dental, vision, accident, and disability claims. Medical claims encompass expenses incurred for medical illness, injury, or procedures, such as doctor visits, hospitalizations, medical devices, and medications. Dental claims cover dental care expenses such as cleanings, fillings, and crowns. Vision claims include expenses for vision care services like exams, eyeglasses, and contact lenses. (Fraud.com, (2022))

By leveraging data techniques, companies can process substantial volumes of data to discover obscured relationships and irregularities that may denote deceptive conduct. These techniques involve using algorithms and statistical models to sift through vast datasets, identifying

deviations from normal transactional behavior that may indicate potential fraud. By detecting and analyzing these irregularities, organizations can take proactive measures to mitigate fraud risks and protect their assets. (Borah, Saleena, and Prakash (2020))

Organizations must continually update and refine their data strategies to effectively combat advancing fraud tactics. By embracing a data-driven approach and harnessing advanced analytical tools, organizations can strengthen their capacity to identify and thwart fraudulent activities. This proactive stance not only protects their financial assets but also safeguards their reputation within the industry.

Fraud detection in claim data presents a significant challenge, as the rule base used for detection may not cover all possible fraudulent activities. Hence, there is a need to minimize false positives while extending an existing rule-based detection system. The aim is to enhance the detection accuracy without compromising on identifying true fraud cases (false negatives). This objective can be incorporated into the optimization function to strike a balance between precision and recall in fraud detection. By refining the rule base and optimizing the detection process, it is possible to enhance the efficiency and effectiveness of fraud detection in claim data. (Sheikhalishahi et al., (2023))

In conclusion, while fraud remains a dynamic and elusive challenge, data mining techniques present a promising avenue for identifying and combating fraudulent transactions. By continuously refining and applying data methodologies, organizations can reinforce their abuse detection abilities and stay resilient to emerging threats.

Fraud detection has long been a challenging and costly task for insurance companies, often requiring manual investigation and extensive resources. However, with the advent of data analytics, there is now a more effective and proactive approach to combating fraud. By leveraging data analytics, insurance companies can uncover valuable insights from their vast datasets, enabling them to identify transactions that exhibit signs of fraudulent activity or pose a higher risk of fraud.

Data analytics is crucial in the battle against fraud, enabling insurers to utilize advanced algorithms and statistical methods to identify anomalies and detect patterns that suggest fraudulent behaviour. Through the analysis of historical data and real-time transactions,

insurance companies can build predictive models that promptly flag suspicious activities. This proactive approach enables insurers to take timely action to prevent or minimize losses (Bănărescu, (2015)).

One of the significant advantages of data analytics in fraud detection is its ability to process and analyze large volumes of data quickly and efficiently. This scalability ensures that insurance companies can handle massive datasets and identify fraudulent patterns across a wide range of transactions.

Furthermore, data analytics not only aids in detecting existing fraud but also helps in the proactive identification of potential risks and emerging fraud trends. By continuously monitoring and analyzing data, insurers can stay ahead of evolving fraud schemes and adapt their fraud detection strategies accordingly. (Ikhsan, W., Ednoer, E., Kridantika, W. & Firmansyah, A. (2022))

Overall, data analytics serves as a powerful ally in the fight against fraud for insurance companies. It enhances their ability to identify and combat fraudulent activities effectively, leading to reduced losses and improved operational efficiency. As data analytics continues to advance, its role in fraud detection will become even more crucial in safeguarding insurers' assets and ensuring the financial well-being of their policyholders. (Hargreaves and Singhanian (2016))



## 2.2 Review of previous research on OPD insurance fraud detection

Fraud involves deliberate deception or misrepresentation leading to unauthorized benefits or payments, whereas abuse encompasses actions that are improper, inappropriate, or do not meet professional standards or medical necessity. Rebecca S. Busch, in "Healthcare Fraud: Auditing and Detection Guide," defines healthcare fraud as intentionally executing a scheme to defraud a healthcare benefit program. According to the National Healthcare Anti-Fraud Association (NHCAA (n.d.)), healthcare fraud involves intentional deception or misrepresentation with the knowledge that it could result in unauthorized benefits.

The largest healthcare provider fraud takedown in US history, as reported by the U.S. Department of Justice, involved charges against 412 defendants across 41 federal districts, with schemes totalling \$1.3 billion. The Centres for Medicare and Medicaid Services (CMS) provides examples of common fraud and abuse within healthcare.

*Table 2-1 Differentiation between Fraud and Abuse*

Examples of Fraud	Examples of Abuse
Submitting claims for services not provided or used	A pattern of waiving cost or deductibles
Falsifying eligibility, claims or medical records	Failure to maintain adequate medical or financial records
Misrepresenting dates, frequency, duration, or description of services rendered	A pattern of claims for services not medically necessary
Billing for services at a higher level than provided or necessary	Improper billing practices
Failing to disclose coverage under other health insurance	Refusal to furnish or allow access to medical records

To make progress in improving the healthcare industry, it is crucial to understand the methods and techniques used by fraudsters. (Tricare (2022))

Detecting and preventing healthcare fraud can be complex due to fluctuating volumes of both fraud and legitimate cases, multiple styles of fraud occurring simultaneously, changing legal behavior over time, and the continuous evolution of new or modified fraud styles by professional fraudsters in response to detection systems. (Phua et al. (2010))

There are various types of insurance fraud involving different individuals. Healthcare fraud is commonly categorized into two types:

- **Financial fraud** involves deceptive activities with the primary motive of obtaining financial gain or causing financial loss to others. This category includes various schemes and manipulations targeting monetary assets (Hilal, Gadsden, and Yawney (2021)). The three main types of financial fraud are:
  - **Asset Misappropriation:** This type of abuse implies the theft or misuse of an business's resources or assets by individuals within the organization. Examples include stealing, theft of cash, or misuse of company funds.
  - **Corruption:** Corruption-related financial fraud involves acts of bribery, extortion, or collusion. Individuals may engage in corrupt practices to gain undue advantages, such as securing contracts or favorable business deals.
  - **Financial Statement Fraud:** This type of abuse entails deliberate misrepresentation of financial information, often with the aim of deceiving investors, creditors, or the public. It can include inflating revenues, understating expenses, or manipulating financial statements.
  
- **Non-financial fraud** encompasses deceptive activities that do not necessarily involve monetary transactions but can cause harm or loss to individuals, organizations, or systems. Some common types of non-financial fraud include:
  - **Cybercrime:** This involves criminal activities carried out through digital means, such as hacking, identity theft, phishing, or spreading malware. Cybercriminals often exploit vulnerabilities in computer systems for various malicious purposes. (Jain, Shrivastava, and Professor (2014))
  
  - **Identity theft** take place when somebody steals personal information, such as social security numbers or financial details, and uses it for fraudulent activities like opening unauthorized accounts or making purchases in the victim's name. (Koops and Leenes (2006))

- Insurance Fraud: This type of fraud involves false or exaggerated claims made to insurance companies to obtain benefits or compensation. Individuals may stage accidents, exaggerate injuries, or submit fake documentation to support their claims.

In summary, financial fraud revolves around monetary gain or loss, while non-financial fraud encompasses a broader range of deceptive activities that extend beyond financial transactions. Both types pose significant challenges to individuals, businesses, and regulatory authorities, requiring comprehensive strategies for prevention and detection.

Fraud can also be categorized based on entities involved:

- **Provider fraud:** This occurs when doctors, pharmacies, or hospitals are involved.

Examples:

- Lapping: Stealing premiums and disguising it by crediting a fake customer account with another customer's premium. Theft within the insurance sector involves illicitly appropriating premiums and concealing this fraudulent act by falsely attributing the credited amount to a fictitious customer account, which is camouflaged as another legitimate customer's premium (Morley, Ball, and Ormerod (2006)).
- Skimming: Stealing premiums before they are credited to customers' accounts by the insurer. Misappropriating premiums before they are rightfully credited to customers' accounts represents a deceptive practice within the insurance domain. This fraudulent act involves the illicit diversion of funds intended for customers, eroding the financial trust and security that policyholders place in their insurers (Cather (2018)).
- Fictitious policies: Creating policies by "investing" their own money as premiums, using funds from insurance companies' incentive programs to cover the investments, and letting the policies lapse afterward. This deceitful tactic involves leveraging funds obtained from insurance companies' incentive programs to cover the initial investments. Subsequently, these policies are

intentionally allowed to lapse, leading to financial losses for the insurance companies (Insurance Fraud Examples, n.d.)

- Forgery: Forging policyholders' signatures to steal premiums and cash values of insurance policies. Engaging in fraudulent activities within the insurance landscape involves a reprehensible tactic: the forgery of policyholders' signatures. This unscrupulous practice is undertaken with the malicious intent of pilfering both the premiums and cash values associated with insurance policies. By fortifying these protective measures, insurers can safeguard the interests of policyholders and maintain the credibility of the insurance sector. (Derrig (2002))
  - Churning: Persuading customers to terminate existing policies and buy new ones to earn extra commissions, often without informing them of the financial loss involved. Unethical practices within the insurance sector extend to duping customers into terminating existing policies under the guise of offering new ones, all for the purpose of accruing additional commissions. Through such measures, the insurance industry can uphold its commitment to fair practices and maintain the confidence of its clientele. (Federal Bureau of Investigation, n.d.)
  - Unscrupulous agents operating within the insurance realm may engage in fraudulent practices by selling policies to customers and deliberately withholding crucial documentation from carriers. This deceitful strategy allows these agents to siphon off payments without the carriers' knowledge, compromising the integrity of the insurance process. To counteract such fraudulent activities, it becomes imperative for insurers to implement robust documentation verification processes and employ advanced technologies that can detect anomalies in policy transactions. (content.naic.org, (2022))
- Doctor fraud: This type of fraud involves doctors (Chavali (2015)).

Examples include:

- Falsified claims schemes: Participating in insurance fraud may extend to the creation of fabricated medical personas and engaging in identity theft. This deceptive practice involves individuals assuming false identities to perpetrate fraudulent claims, leading to financial losses for insurers. To combat such fraudulent activities, insurance companies must employ advanced identity verification technologies and cross-reference information to ensure the legitimacy of the individuals involved in claims processes. These measures collectively contribute to enhancing the security and credibility of insurance operations.
- Engaging in insurance fraud can manifest in various forms, including the unethical practice of prescribing unnecessary medications to patients solely for financial gain. This fraudulent behavior compromises the integrity of the healthcare system and puts patients at risk. By actively combating fraudulent prescription practices, the healthcare industry can uphold ethical standards and prioritize the well-being of patients.
- Phantom billing: A prevalent form of insurance fraud involves the deceptive billing for services or supplies that were never provided to the patient. This unscrupulous practice not only undermines the financial integrity of insurance systems but also jeopardizes the trust between healthcare providers and insurers. Ultimately, safeguarding against false claims contributes to the sustainability of the healthcare system and ensures that insurance funds are allocated to genuine patient care. (Kumaraswamy et al. (2022))
- Upcoding: A deceptive practice observed in insurance fraud involves billing for a service of higher value than the actual service provided to the patient. This fraudulent tactic not only distorts the financial landscape of insurance but also poses a threat to the credibility of healthcare providers and the insurance industry at large. (Geruso and Layton (2020))
- Engaging in unnecessary medical procedures or treatments with the sole motive

of financial gain constitutes a severe form of healthcare fraud. Perpetrators of such fraudulent activities may subject individuals to surgeries, therapies, or medical interventions that are medically unwarranted but serve their financial interests. This unethical practice not only endangers the well-being of patients but also results in inflated healthcare costs and places an undue burden on insurance providers. (Liu, Wang, and Yu (2023))

One prevalent fraudulent practice within the realm of insurance pertains to the misrepresentation of dates, locations, or service providers. This deceptive maneuver, often employed by unscrupulous individuals seeking to exploit insurance coverage, introduces inaccuracies into the claims process. By providing false information regarding the timing, locations, or entities involved in services rendered, fraudsters attempt to manipulate the system for personal gain. Establishing robust mechanisms to cross-verify information ensures the accuracy and reliability of claims data, thereby fortifying the defenses against fraudulent activities within the insurance domain. (Villegas-Ortega, Bellido-Boza and Mauricio, 2021)

The involvement of criminal syndicates in orchestrating intricate and sophisticated fraudulent schemes poses a significant challenge to the insurance industry. These criminal gangs, often organized with precision, exploit vulnerabilities within insurance processes to execute large-scale fraud operations. Their activities may encompass various types of insurance fraud, such as staged accidents, falsified claims, and identity theft. Addressing this level of organized crime requires a multi-faceted approach involving advanced analytics, AI, and collaboration between insurers, law enforcement, and regulatory authorities. Effectively combating such criminal enterprises necessitates continuous vigilance, proactive detection measures, and the development of strategies that can stay ahead of evolving criminal tactics.

The insurance industry must remain resilient and adaptive to counter the threats posed by these criminal organizations, safeguarding the integrity of insurance processes, and maintaining trust in the industry. (Reuter and Paoli (2020))

- Customer fraud: This occurs when customers engage in fraudulent activities.  
Examples:
  - Unbundling: Presenting several bills for the same service. Unbundling fraud

refers to a deceptive practice where a comprehensive service or product is intentionally broken down or "unbundled" into its individual components, with each component then billed separately. This fraudulent tactic is commonly observed in various industries, including healthcare, insurance, and finance. (Nabrawi and Alanazi (2023))

- Double billing: Presenting several claims for the same service. Double billing insurance fraud occurs when a healthcare provider submits multiple claims for the same service, procedure, or treatment for a single patient, leading to the fraudulent collection of payments from the insurance company more than once. (Pitler and Bonomi (2006))
- Forgery/Prescription Medication Abuse: refers to deceptive practices where individuals manipulate or falsify medical prescriptions or related documents to illegitimately obtain benefits from an insurance provider. This type of fraud typically involves creating or altering prescriptions to acquire prescription medications for personal use, resale, or to submit fraudulent claims to insurance companies. Fraudsters may forge signatures of healthcare professionals, change prescription details, or use stolen prescription pads to create fraudulent documents. (Haddad Soleymani et al. (2018))
- Booking multiple appointments with the same doctor using OPD (Outpatient Department) insurance can be indicative of potential insurance fraud. In this scenario, individuals may exploit the insurance coverage by scheduling several appointments with the same healthcare provider within a short span. This behavior raises concerns about the legitimacy of these appointments and the necessity of the associated medical services. Fraudsters might attempt to exaggerate or fabricate medical conditions to maximize their insurance benefits, leading to financial losses for insurance providers.
- Non-disclosure of pre-existing medical conditions during the application stage is a form of insurance fraud that involves the intentional withholding of relevant health information by the policyholder when applying for an insurance policy. In this scenario, individuals fail to provide accurate details about existing

medical conditions they might have, such as chronic illnesses, previous surgeries, or ongoing treatments. (Villegas-Ortega, Bellido-Boza and Mauricio, 2021)

- Employee fraud: In this scheme, an insider, typically an employee, manipulates the system for personal gain. This can manifest in various forms, such as submitting fictitious claims or modifying banking information to reroute legitimate payments to unauthorized accounts. The submission of false claims involves an employee creating fictional scenarios or exaggerating legitimate incidents to claim insurance benefits improperly. (Videnović and Hanic (2021))

**Syndicate fraud:** Involves organized groups attempting various fraudulent activities, such as submitting false membership applications, changing bank details to redirect claims payments, admitting healthy members to hospitals for cash-back insurance, and colluding with healthcare funders' employees. (Pourhabibi et al. (2020)). Syndicates may also commit identity theft to defraud medical schemes and build fraudulent health profiles for other fraudulent activities in different member products. (Ogunbanjo and Bogaert (2014))

*Table 2-2 Examples of Fraud Schemes*

Fraud Schemes	Example
Phantom Billing	Submitting claims for services not provided.
Duplicate Billing	Submitting similar claims more than once.
Bill Padding	Submitting claims for unneeded ancillary services to Medicaid.
Upcoding	Billing for a service with a higher reimbursement rate than the service provided.
Unbundling	Submitting several claims for various services that should only be billed as one service.
Excessive or Unnecessary Services	Provides medically excessive or unnecessary services to a patient.
Kickbacks	A kickback is a form of negotiated bribery in which a commission is paid to the bribe-taker (provider or patient) as a quid pro quo for services

The review of previous research on outpatient department (OPD) insurance fraud detection using rule-based method for customer risk scoring reveals significant insights into the effectiveness and challenges of this approach. Rule-based systems have been extensively



explored in the domain of insurance fraud detection due to their simplicity and transparency in identifying fraudulent activities.

Here, we have summarized the key findings from the existing literature (Ahmed et al. (2021)):

- **Rule-Based Approach:** Researchers have emphasized the use of predefined rules and expert knowledge to detect fraudulent claims in health insurance. These rules are designed based on domain expertise and historical data, enabling the system to flag suspicious claims that deviate from expected patterns.
- **Advantages of Rule-Based Systems:** The primary advantage of rule-based systems is their interpretability. Insurance companies and investigators can easily understand the rules applied to identify fraud. Additionally, rule-based systems are relatively easy to implement and can deliver real-time results, making them practical for operational use.
- **Limitations:** Despite their simplicity, rule-based systems have certain limitations. They heavily rely on predefined rules, which may not cover all potential fraud scenarios. As fraudsters continually evolve their tactics, rule-based systems may struggle to keep up with emerging fraud patterns. Moreover, rule-based systems may generate false positives when rules are overly conservative, leading to additional investigation costs.
- **Rule Refinement:** Researchers have explored different techniques to improve rule-based systems. This includes rule optimization, where rule parameters are adjusted based on performance metrics and feedback from investigators. Rule learning algorithms have also been studied to automatically generate rules from historical data.
- **Hybrid Approaches:** Some studies have explored combining rule-based systems with machine learning techniques. By integrating the strengths of both approaches, researchers aim to enhance fraud detection accuracy and reduce false positives.
- **Case Studies:** Several research papers have presented case studies and evaluations of

rule-based fraud detection systems in real-world insurance datasets. These studies demonstrate the effectiveness of the approach in identifying fraudulent OPD insurance claims.

- **Challenges:** One of the significant challenges in using rule-based systems for OPD insurance fraud detection is maintaining the rule base as fraud patterns evolve. Ensuring the rules remain up- to-date and relevant is crucial to the success of the system.

In conclusion, the research on OPD insurance fraud detection using rule-based rules highlights the importance of interpretability and real-time processing. While rule-based systems offer transparency and simplicity, they need continuous refinement and adaptation to stay effective in detecting evolving fraud patterns. Combining rule-based approaches with advanced machine learning techniques may provide a promising direction for improving fraud detection accuracy and reducing false positives in the future. (Baumann (2021))

### **2.3 Discussion of machine learning techniques vs rule-based model in fraud detection**

In traditional methods of detecting healthcare fraud and abuse, auditors have limited time to review each claim, resulting in a narrow focus on specific claim characteristics rather than considering the provider's overall behavior. This approach is time-consuming and inefficient, particularly in low- income and middle-income countries. Fraudsters adapt their strategies to avoid detection when they become aware of detection methods. Therefore, it is necessary to update thresholds regularly for rudimentary methods and adjust parameters linearly for rule-based models. It is important to emphasize unsupervised learning/models over supervised learning/models. (Kazeem (2023))

Detecting fraud in financial institutions is a critical concern, and traditional approaches like rule- based systems and supervised learning models have their limitations in identifying complex fraud cases. As a result, there is a growing trend among financial institutions to adopt both rule based and machine learning techniques for fraud detection. Implementing a

comprehensive strategy that includes product and benefit design, consumer education, advanced claims handling, and direct action against offenders is essential. (Islam et al. (2024))

To bolster fraud detection efforts, software applications can be employed to flag suspicious claims or deviations from typical patterns for further scrutiny. Common red flags to monitor include consultations with inappropriate specialists, unusually high claim values, and durations or amounts exceeding industry norms. The specific attributes utilized for detecting each type of fraud typically remain consistent.

Management data commonly encompasses financial ratios derived from accounts receivable, provisions for doubtful debts, and net sales figures. Health insurance data incorporates ratios related to compensation, premiums, liabilities, customer behaviour, and financial standing. Medical insurance data may encompass patient demographics, treatment specifics, and policy and claim details.

Rule-based flagging is a method used for fraud detection that applies predefined rules or heuristics to identify suspicious transactions or behaviors. These rules are based on various criteria such as transaction amounts, frequency, and geographical locations. For instance, a rule could be set to flag transactions exceeding a specific amount or originating from high-risk areas. If a transaction meets one or more of these rules, it is marked as suspicious and subject to further investigation.

Rule-based flagging can be a simple and efficient approach to detect fraud, enabling quick decision-making. However, it may generate false positive alerts, leading to wasted time and resources investigating harmless transactions. Fraudsters can also adapt their tactics to evade rule-based systems, emphasizing the need for regular updates and fine-tuning of the rules. Customer risk scoring models employ predefined rules and algorithms to assign a risk score to each customer. These models analyze customer attributes and behaviors, such as transaction history, demographics, and IP address, to determine the likelihood of fraud.

Here are a few examples of rules used in a rule-based customer risk scoring model for fraud detection:

- **High-Risk Transactions:** The strategy of flagging transactions exceeding a specific

amount plays a critical role in fortifying the security apparatus of financial institutions and businesses. This proactive measure involves singling out transactions that surpass a predetermined threshold, providing a meticulous examination of high-value financial activities. This meticulous scrutiny extends to transactions originating from a new or infrequently used device, acknowledging the evolving landscape of fraud tactics where perpetrators exploit various devices to conceal their activities. (Cherif et al. (2022))

- **Suspicious Behavior:** The strategic flagging of transactions involving repeated attempts, rapid succession, or a high volume within a condensed time frame is a pivotal component in fortifying the security infrastructure of financial institutions and businesses. This proactive measure seeks to identify and scrutinize potentially suspicious behavior that deviates from regular transaction patterns. Repeated attempts, especially when occurring in rapid succession, can be indicative of unauthorized or fraudulent activities. This could include scenarios where an individual or entity is trying multiple times to authenticate or execute a transaction, signaling potential malicious intent. The high volume of transactions within a brief period is another red flag, as it may suggest an abnormal surge in financial activity that requires closer inspection.
- **Irregular Payment Patterns:** Identifying potential fraud, transactions triggering substantial deviations from a customer's established payment patterns are flagged for scrutiny. This includes abrupt shifts in payment methods or frequencies. (Mensah, Acquah, and Akpah (2019))
- **Blacklisted IP Addresses:** To fortify security measures, the system automatically blocks or flags transactions originating from IP addresses linked to prior instances of fraudulent activity. By block-listing IP addresses with a history of fraudulent behavior, the system helps mitigate risks and protect users from unauthorized transactions. This pre-emptive strategy bolsters the overall security infrastructure, fostering a more resilient environment that ensures the integrity of financial transactions and safeguards against recurrent fraudulent attempts.
- **Device and Identity Linking:** The system employs a sophisticated fraud detection mechanism that promptly flags transactions displaying patterns involving multiple

devices or identities. This strategic alert is triggered when there is an indication of potential evasion tactics, such as the use of various identities or devices to circumvent detection measures. This multifaceted approach underscores the importance of vigilance against orchestrated attempts to deceive security systems, contributing to a robust defense against evolving strategies employed by malicious actors in the digital landscape. (Mao et al. (2022))

Rule-based customer risk scoring models stand out for their simplicity in implementation and user-friendly nature. Their strength lies in adaptability; businesses can effortlessly customize these models by tweaking existing rules or introducing new ones to align with emerging fraud patterns. This inherent flexibility empowers organizations to stay ahead of evolving threats without requiring intricate adjustments or a deep understanding of complex algorithms. This simplicity in customization not only enhances the responsiveness of the risk scoring system but also reduces the need for constant algorithmic recalibration. As a result, rule-based models offer a practical and efficient solution, providing a dynamic defense against the ever-changing landscape of fraudulent activities in the business domain. (Ahmed et al. (2021))

However, these models have limitations. Fraudsters can evade rules by changing their tactics, and rule-based systems may lack flexibility to adapt to new fraud patterns or changing circumstances. Additionally, they may generate false positive alerts, resulting in wasted resources investigating benign transactions.

Advantages of human-driven approaches in fraud detection include:

- **Experience:** Human investigators have experience and expertise in identifying and investigating fraudulent activity, which can be beneficial in complex fraud cases.
- **Contextual Understanding:** Humans can consider the context of a transaction and understand the nuances of customer behavior, making it easier to identify suspicious activity.
- **Flexibility:** Human investigators can be flexible in their approach and can adapt to changing fraud patterns and circumstances.
- **Personal Touch:** In some cases, human investigators may be able to engage with

customers and gather additional information that could help to identify and prevent fraud.

- **Collaboration:** Human investigators can collaborate with other departments, such as legal and IT, to gather additional information and resources needed to detect and prevent fraud.
- **Better Customer Experience:** Human-driven fraud detection can result in a better customer experience, as customers may feel more valued and trusted when a person is handling their case.
- **Improved accuracy:** With the right training and resources, human investigators can be highly accurate in detecting fraud, which can result in fewer false positive and false negative alerts.

Advantages of rule-based flagging in fraud detection:

- **Ease of Implementation:** Rule-based flagging is relatively straightforward to implement and does not require specialized skills or expertise in machine learning.
- **Speed:** Rule-based flagging can quickly identify suspicious transactions, allowing organizations to respond quickly to potential fraud.
- **Customizability:** Rule-based flagging can be easily customized to meet the specific needs and requirements of an organization.
- **Transparent Decision Making:** Rule-based flagging provides a clear and transparent decision-making process, making it easy to understand why a transaction was flagged as suspicious.

In summary, this comprehensive literature review has examined various published studies on the detection of fraud in outpatient department (OPD) insurance. It has discussed the characteristics of fraud, the types of fraud, data considerations, performance metrics, and different methods and techniques. The review has also pointed out the limitations of current approaches and identified the potential for improvement in this field. Taking all these factors into account, it can be concluded that rule-based approaches show promise in detecting suspicious and fraudulent transactions in OPD healthcare claims. Particularly, for small startup companies seeking to address business challenges, rule-based approaches are well-suited.

Although machine learning and AI-based methods are widely used in fraud detection in sectors like banking, finance, and credit cards, the research on detecting OPD insurance fraud using these techniques is limited. The literature review has revealed a scarcity of machine learning and AI-based solutions in the OPD healthcare industry, which makes rule-based approaches an appealing alternative. (Baumann (2021).

This review emphasizes the need for more research in this area and suggests that rule-based approaches can contribute to the development of a comprehensive solution for detecting OPD insurance fraud.

## **2.4 Review of relevant regulations and policies related to insurance fraud.**

In India, insurance fraud is a growing problem that has significant financial and legal consequences. Here is a detailed review of some of the key regulations and policies related to insurance fraud in India:

**Insurance Regulatory and Development Authority (IRDA):** The IRDA is the regulatory body in India that oversees the insurance industry. It develops and enforces regulations aimed at preventing fraud, covering areas such as claims processing, underwriting, and risk management.

**Indian Penal Code (IPC):** The IPC is India's primary criminal code that criminalizes insurance fraud. It prohibits fraudulent activities related to insurance, including false claims, forgery, and misrepresentation.

**Prevention of Money Laundering Act (PMLA):** The PMLA is an Indian law that criminalizes money laundering. It mandates insurers and other financial institutions to maintain transaction records, report suspicious activities, and cooperate in investigations concerning money laundering.

**Securities and Exchange Board of India (SEBI):** SEBI regulates the securities markets in India. It develops and enforces regulations aimed at preventing fraud, such as insider trading, market manipulation, and other fraudulent practices in the securities industry.

**Insurance Fraud Investigation Unit (IFIU):** The IFIU is a specialized unit that is dedicated to investigating insurance fraud in India. The unit works with law enforcement agencies, insurers, and other stakeholders to investigate fraud and develop strategies to prevent it.

**General Insurance Public Sector Association (GIPSA):** The GIPSA is a group of four public sector general insurance companies in India. The association has developed guidelines and best practices related to fraud prevention and detection, including guidelines related to claims processing and underwriting.

The Ayushman Bharat – Pradhan Mantri Jan Arogya Yojana (PMJAY) is a health insurance initiative by the Government of India, designed to offer health coverage of Rs. 5,00,000 to over 10 crore beneficiary families, targeting more than 40% of the country's population. Recognizing the high risks of fraud in health insurance programs, the National Health Agency (NHA) has developed comprehensive Anti-Fraud Guidelines to assist state governments in preventing, detecting, and deterring fraud within PMJAY. The guidelines emphasize a zero-tolerance approach to fraud and are based on principles of transparency, accountability, responsibility, independence, and reasonability. They outline various forms of fraud, including beneficiary, payer, and provider fraud, and recommend robust mechanisms for fraud management.

The guidelines also highlight the responsibilities of both national and state health agencies in combating fraud. The NHA is tasked with developing anti-fraud frameworks, providing oversight, and offering technical assistance to states. It also emphasizes the importance of IT infrastructure for advanced data analytics and fraud detection. The State Health Agencies (SHA) are responsible for adapting and implementing these guidelines, developing state-specific IT platforms, and conducting awareness programs. The institutional arrangements include dedicated anti-fraud cells at both national and state levels, equipped with specialized personnel for legal, medical, and data analytics functions. The guidelines also detail procedures for managing fraud complaints and measuring the effectiveness of anti-fraud efforts, ensuring a holistic approach to maintaining the integrity of the PMJAY scheme.

In summary, there are several regulations and policies in place in India that are designed to prevent, detect, and prosecute insurance fraud. These include the IRDA regulations, the IPC, the PMLA, the SEBI regulations, the IFIU, and the GIPSA guidelines. It is important for



insurers and other stakeholders to be familiar with these regulations and policies to effectively prevent and detect insurance fraud in India.

Insurance fraud is a deliberate and deceptive act committed with the intention of obtaining financial advantages, either at the expense of an insurance company or facilitated by an insurance agent. This deceptive practice is not limited to a single point in the insurance process but can involve various parties, including applicants, policyholders, third-party claimants, and professionals providing services to claimants. Furthermore, insurance agents and company employees themselves may engage in fraudulent activities. Some common schemes include claim inflation, providing false information on insurance applications, making claims for fictitious injuries or damages, or orchestrating staged accidents.

Over the course of decades, assessments of the yearly costs associated with insurance fraud may have been underestimated, overlooking crucial updates for factors like inflation and other essential data components. A recent study conducted in 2022 by The Coalition Against Insurance Fraud (CAIF) sheds light on the staggering impact of insurance fraud, revealing that it could potentially cost U.S. consumers a substantial \$308.6 billion (about \$950 per person in the US) annually. This comprehensive figure encompasses estimated annual fraud costs spanning various liability areas, notably Life Insurance (\$74.7 billion (about \$230 per person in the US)), Property and Casualty (\$45 billion (about \$140 per person in the US)), Workers Compensation (\$34 billion (about \$100 per person in the US)), and Auto Theft (\$7.4 billion (about \$23 per person in the US) (about \$23 per person in the US)).

This eye-opening study emphasizes the pervasive nature of insurance fraud across diverse sectors, underscoring its significant financial ramifications for consumers and the insurance industry at large. The detailed breakdown of fraud costs by specific areas, such as life insurance and auto theft, provides a nuanced understanding of the breadth of the issue. By incorporating inflation-adjusted estimates and accounting for different facets of insurance, the study seeks to provide a more accurate reflection of the economic impact of fraud on both insurance companies and the individuals they serve.

As we navigate the complex landscape of insurance fraud, it becomes evident that addressing this issue requires a multifaceted approach. The substantial figures revealed by the CAIF study emphasize the urgency of implementing robust anti-fraud measures, fostering collaboration

across the industry, and leveraging advanced technologies to detect and prevent fraudulent activities.

Rule-based insurance fraud detection systems are an integral part of the broader regulatory framework designed to combat insurance fraud. Various regulatory authorities and international organizations have recognized the significance of such systems in protecting insurers and policyholders' interests. By complying with relevant regulations and policies, insurance companies can improve their fraud identification abilities and contribute to a more resilient and trustworthy insurance industry.

Healthcare insurance fraud is a primary obstacle that can have substantial economic and legal impact. As a result, there are many regulations and policies in place designed to prevent, detect, and prosecute insurance fraud. Here is a detailed review of some of the key regulations and policies related to insurance fraud over the globe:

**State insurance laws:** Each state has its own insurance laws that regulate the insurance industry and help prevent fraud. These laws typically require insurers to maintain certain standards of conduct, provide clear disclosures to policyholders, and cooperate with investigations into fraud.

**National Insurance Crime Bureau (NICB):** The NICB is a nonprofit organization dedicated to preventing and detecting insurance fraud. The organization works with law enforcement agencies, insurers, and other stakeholders to investigate fraud and develop strategies to prevent it.

**Federal Insurance Fraud Statutes:** There are several federal statutes that criminalize insurance fraud, including mail fraud and wire fraud statutes. These statutes make it a federal crime to use mail or electronic communications to commit fraud, including insurance fraud.

**National Health Care Anti-Fraud Association (NHCAA):** The NHCAA is a nonprofit organization dedicated to preventing and detecting health care fraud, including insurance fraud. The organization works with law enforcement agencies, insurers, and other stakeholders to investigate fraud and develop strategies to prevent it.

**False Claims Act:** The False Claims Act is a federal rule that permits individuals to file court case on behalf of the government against bodies or individuals alleged of duping federal programs. This law has been utilized to prosecute cases involving insurance fraud.

**Insurance Information and Privacy Protection Act (IIPPA):** The IIPPA is a federal law governing the handling of personal information by insurers. It mandates insurers to obtain consent before using personal information for certain purposes and grants individuals' specific rights concerning their personal data.

**Health Insurance Portability and Accountability Act (HIPAA):** HIPAA is a federal law that legalizes the privacy and security of protected health information (PHI). It imposes obligations on insurers to safeguard PHI and mandates written consent for the use or disclosure of PHI for specific purposes.

Fraud management encompasses a broad set of measures aimed at tackling fraudulent activities effectively. It can be categorized into three main components: (Štefan and Bajec (2008))

**Fraud Prevention:** This involves implementing stringent access controls and usage restrictions to create a secure environment that minimizes the possibility of fraudulent activities. By employing robust authentication mechanisms, encryption, and role-based access, organizations can deter potential fraudsters from gaining unauthorized access to sensitive data or services.

**Fraud Detection:** The process of fraud detection involves continuous monitoring and analysis of various indicators, such as service usage metrics and transaction patterns. Real-time or non-real-time observations are made to identify suspicious activities that may indicate fraud. When a potential fraud instance is detected, appropriate actions are triggered, such as blocking access to the service or generating alerts to notify relevant personnel.

**Fraud Reduction:** Acknowledging that absolute fraud prevention is challenging, fraud reduction focuses on minimizing the frequency and impact of fraudulent incidents. Real-time detection plays a vital role in promptly identifying fraudulent activities, allowing organizations to take immediate corrective actions and mitigate potential damages. (Burge et al. (1997))

To achieve effective fraud management, organizations must implement a multi-layered approach that combines prevention, detection, and reduction strategies. By employing advanced analytics, organizations can enhance their fraud management capabilities and stay ahead of evolving fraudulent tactics. (McGibney and Hearne (2003))

It is essential to continuously update and adapt fraud management strategies to address emerging threats and vulnerabilities in the digital landscape. By fostering a proactive and vigilant approach to fraud management, organizations can safeguard their assets, protect customer data, and maintain their reputation in an increasingly interconnected world.

This literature review explored the potential of customer risk scores, specifically rule-based approaches, in detecting fraud within the OPD insurance domain. The lack of readily available OPD-specific fraud research necessitated drawing insights from studies across diverse insurance sectors, granting a holistic perspective on fraud detection strategies.

However, recognizing the limitations of rule-based systems, further research should explore hybrid approaches integrating rule-based and machine learning techniques to capture both the interpretability and adaptability needed for a robust OPD fraud detection system.

While our study utilized insights from broader insurance fraud research, dedicated investigations into OPD fraud are paramount. Future research should delve deeper into OPD-specific claim patterns, fraud typologies, and risk factors to refine and optimize fraud detection strategies.

In conclusion, customer risk scores employing a rule-based approach present a promising avenue for combatting OPD insurance fraud. Their advantages in transparency, fairness, and ease of implementation make them well-suited for this unique domain. Nevertheless, continuous research and development, including exploring hybrid approaches and OPD-specific analysis, are crucial to refine these tools and ensure their effectiveness in safeguarding the integrity of the OPD insurance system.

This conclusion emphasizes the key takeaways of your research, highlighting the advantages of rule-based approaches for OPD fraud detection, acknowledging their limitations, and urging further research in this understudied area.

## CHAPTER III: METHODOLOGY

### 3.1 Introduction

The essence of this study lies in evaluating customers' risk propensity based on their historical transactions with the insurance company. Our goal is to establish a quantifiable risk score derived from various customer interactions, including claim history, policy acquisition habits, and the way customers file insurance claims. The objective is to implement an early warning system, termed the Customer Risk Score, which, once operational, would identify potentially fraudulent or high-risk claims. This would ensure that such claims are directed to a specialized investigation unit. Such proactive measures can significantly optimize resources, streamlining the processing of genuine claims while subjecting potentially deceptive ones to more intensive scrutiny.

In this chapter, we delineate the methodology employed in the research to address the core question: "How can a rule-based risk score enhance abuse or fraud detection instead of using Artificial Intelligence, and what are the key characteristics of a rule-based customer risk score?"

Our research adopts a Cross-Sectional Design. The objective is to observe and describe the situation as it exists currently. Given that our primary goal is to understand existing patterns and correlations that pertain to risk behaviors of policyholders, this design is apt. Cross Sectional design ensures a detailed portrayal of situations at a particular time frame, making it a fitting choice when the aim is to correlate variables without any external intervention.

The foundation of the research is an approach that combines qualitative and quantitative techniques along with cross-sectional research. This approach enables a thorough understanding of the characteristics, advantages, and drawbacks of using rule-based risk scoring and explores how it can enhance fraud identification processes. The subsequent sections of this chapter elaborate on the research design, data collection techniques and instruments, sampling approach, and data analysis methodologies. Ethical considerations related to the study are also discussed, along with the inherent limitations of the method.

The results are anticipated to have practical ramifications for organizations incorporating and integrating into the system.

## 3.2 Operationalization of Theoretical Constructs

### 3.2.1 Research Philosophy

**Interpretivism:** This investigation embraces interpretivism as its guiding philosophy. Given the inherent complexity and subjectivity of customer behavior and risk assessment, interpretivism allows for a deep and nuanced understanding of how rule-based customer scoring systems operate in real-world contexts.

#### Why Interpretivism?

**Multiple Realities:** Customer actions and motivations are multifaceted, influenced by individual characteristics and situational factors. Interpretivism acknowledges this inherent subjectivity, enabling us to delve into the diverse perspectives shaping customer behavior and risk profiles.

**Constructing Meaning:** Rule-based scoring systems rely on specific criteria and assumptions. Interpretivism empowers us to unpack these assumptions, explore the reasons behind them, and understand how individuals interact with and interpret these rules within unique contexts.

**Focus on lived experiences:** Claims processors insights gleaned through interviews, observations, and other qualitative methods form the backbone of robust rule-based systems. Interpretivism prioritizes these subjective experiences, allowing us to capture the rich tapestry of customer behavior beyond readily quantifiable data points.

#### Interpretivism in practice:

**Triangulation:** involves integrating both qualitative and quantitative methods, such as analyzing existing customer data alongside subjective insights. This approach enhances the depth and validity of our findings by cross-verifying different sources of information, revealing potential biases and uncovering blind spots that may arise from relying solely on data-driven approaches. By triangulating data and insights, we gain a more comprehensive understanding of the phenomena under study, ensuring a more nuanced and accurate analysis.

**Contextualizing results:** Interpretivism emphasizes the importance of understanding the specific context in which rule-based systems operate. By investigating the organizational culture, industry trends, and broader market dynamics, we can contextualize our findings and develop generalizable insights.

Embracing interpretivism in this research allows us to move beyond superficial data points and delve into the heart of customer behavior, enabling the development of nuanced and effective rule-based scoring systems that capture the complexities of real-world interactions.

### **3.2.2 Research Approach**

**Exploring Rule-Based Customer Scoring:** This study employs a dual approach, integrating qualitative and quantitative analyses along with cross sectional data analysis to delve into the complexities of rule-based customer risk scoring. This method is chosen to achieve a thorough understanding that encompasses both the subtleties of customer behavioral patterns and the factual insights derived from data analysis during a defined timeframe.

#### **Quantitative Lens:**

- **Statistical Analysis:** The comparative evaluation of different parameters for rule-based scoring models will involve statistical technique of percentile analysis on various attributes and features.
- **Data & Human-Driven Insights:** Adjudicator's (processor's) validation for all finalized rules and the results. This objective data provides a solid foundation for refining and optimizing the scoring system.

#### **Qualitative Depth:**

- **Stakeholders' reviews:** Conducting reviews with customers and risk management professionals allows for detailed exploration of their experiences and perspectives on the scoring system. This qualitative data is vital for understanding the subjective

realities that drive customer behavior and risk assessment.

- **Grounding the Rules:** Analyzing real-world customer cases can shed light on the underlying reasons behind specific scoring rules. This qualitative approach helps ensure the rules are rooted in practical experience and relevant to actual customer behavior.
- **Open-Ended Feedback:** Analyzing adjudicator's responses can uncover unforeseen concerns and opportunities for improvement in the scoring system. This qualitative perspective helps ensure the system remains flexible and adaptable to evolving customer behavior and risk patterns.

### **Balancing Objectivity and Subjectivity:**

The approach balances the objectivity of quantitative data with the richness of qualitative insights. This synergistic approach provides a comprehensive understanding of rule-based customer risk scoring, leading to:

- **Robust and Trustworthy Findings:** The triangulation of quantitative and qualitative data enhances the research's credibility and generalizability.
- **Data-Driven Optimization:** Objective metrics guide the refinement of scoring rules, while qualitative insights ensure the system remains user-friendly and relevant to customers.
- **A Nuanced Understanding:** By delving into both the quantitative and qualitative aspects of customer risk scoring, the research paints a complete picture of this complex subject, revealing not just how the system works, but also its impact on customer behavior and risk assessment.

By combining quantitative and qualitative methods, this study aims to offer a comprehensive and nuanced exploration of rule-based customer risk scoring, fostering a deeper understanding of its strengths, limitations, and potential for evolution.



### **3.3 Research Design**

Our research methodology utilizes a Cross-Sectional Research Design, selected to observe, and analyze data at a single point in time. This approach aims to provide a thorough understanding of the risk behaviors of policyholders based on data gathered from a specific moment.

The suitability of a Cross-Sectional design arises from its inherent ability to provide a detailed snapshot of a population or phenomenon at a particular moment in time. This design allows us to accurately capture and document the current state, offering valuable insights into prevailing trends and correlations without the need to track changes over time.

In the context of our research, where the primary focus is on unraveling patterns and correlations in policyholders' risk behaviors, the Cross-Sectional Design proves to be highly suitable. By opting for this approach, we aim to uncover relationships between variables based on data collected at a defined time frame, without introducing any external factors that might alter the organic state of the observed phenomena.

In conclusion, the Cross-Sectional research design is instrumental for our research goals, enabling a thorough investigation and documentation of the current risk behaviors among policyholders. This approach seamlessly supports our objective to comprehend the current patterns and relationships as they stand at a specific moment in time.

#### **3.3.1 Supporting Questions**

In crafting a robust rule-based customer risk score as part of our abuse or fraud detection framework for outpatient insurance, our journey involved extensive consultations with business and domain experts in the field. These interactions aimed to illuminate the gaps, challenges, and critical questions that demand answers in the realm of customer risk score. As a result, this section encapsulates the insights gained from these engagements, providing a comprehensive exploration of crucial considerations in the development of a rule-based risk score model. From understanding the breadth of available data sources to pinpointing fraud & abuse indicators and ensuring data reliability, each query plays a pivotal role in guiding the meticulous construction of the scoring model. Let us delve into these questions to illuminate

the path toward creating a robust framework that meets the unique challenges of outpatient insurance.

- What data sources will you use?
  - What types of data are available for analysis (e.g., claims data, customer profiles)?
  - How reliable and up to date is the data?
  - What is fraud & abuse indicators?
  
- What are the common indicators or red flags of fraudulent claims?
  - Are there specific behaviors or patterns associated with high-risk claims?
  - How will we define suspicious behavior?
  
- What behaviors or activities will trigger suspicion (e.g., frequent claims, changes in claim patterns)? (highlights feature importance and selection)
  - What thresholds or criteria will be used to identify suspicious behavior?
  
- Have there been previous instances of fraud claims that you can learn from?
  - Are there identifiable historical patterns or trends accessible for analysis in outpatient insurance?
  
- Do risk patterns vary by location (e.g., different regions, cities)?
  - Will we consider geographic factors in your scoring rules?
  
- How will we detect and handle anomalies that do not fit typical patterns?
  - What processes will be in place to investigate and verify anomalies?
  
- How to consider a customer's claims history and behavior over time?
  - How will we factor in the customer's overall history with the organization?
  - What role does customer history play?
  - If available, how can we use external data?
  
- How will we assign scores?
  - What scoring system will we use in rule-based mechanisms (e.g., points-

based, weighted factors)?

- What is the role of expert knowledge?
  - Will we involve domain experts (e.g., fraud investigators, actuaries, business stakeholders) in defining scoring rules?
  - If yes, how will we leverage their expertise to improve rule accuracy?
  
- How often will we update rules?
  - Will the scoring rules be static or regularly updated?
  - How will we incorporate new data and adjust rules over time?
  - Will the updation process be manual or via automation?
  
- What is the appeal and review process?
  - Is there a process for customers to provide additional information or appeal decisions? (Scrutiny queue & customer support)
  - How will we handle cases where customers dispute their scores / flagged claims?
  - What are the consequences of high scores or flagging claims as suspicious / fraud?
  
- What actions will be taken when a claim receives a high-risk score?
  - How to balance fraud prevention with maintaining a positive customer experience?
  
- What reporting and monitoring will be implemented?
  - How to track the effectiveness of scoring rules?
  - What reporting mechanisms will be in place to identify trends and anomalies?

In upcoming sections, we will provide a detailed summary of the research methods employed, the data collected, and the analysis conducted to answer the research questions posed at the beginning of the study. Next sections will cover data collection and instrumentation, summarizing the methods and tools used to gather data; data sampling strategies, detailing the approaches used to select representative samples; data cleaning and preprocessing, outlining steps taken to prepare the data, including handling missing values and ensuring data quality;

exploratory data analysis, reviewing the initial investigations to uncover patterns, anomalies, outlier detection, explaining techniques used to identify and manage outliers; and rule generation, describing the rules and patterns identified during analysis. Additionally, the section will outline any limitations encountered during the research process and provide recommendations for future research.

Overall, the methodology conclusion section will also highlight how the research design was implemented to address each specific research question and will discuss the findings and insights gained from the study. It will also serve as a synthesis of the entire research process, demonstrating how the study's objectives were achieved and contributing to the broader knowledge of the research topic.

### **3.4 Data Collection and Instrumentation**

Data collection is a crucial step in developing an efficient customer risk-scoring model using data analytics. There are various data collection methods that can be used, including EHR (Electronic Health Records), insurance claims data, surveys, and interviews. The sources of data include hospitals and clinics, health insurance companies, government agencies, and fraud detection organizations. By collecting and analyzing comprehensive and accurate data we can create a reliable rule-based customer risk score model using which healthcare providers can detect fraud and prevent abuse in the OPD transactions, leading to improved healthcare quality and reduced healthcare costs. (Martinez-Cruz, Blanco & Vila, 2012)

To develop a rule-based customer risk scoring model using data analytics, data collection is an important step. The data collected should be comprehensive, accurate, and representative of the OPD transactions. We have chosen a secondary data collection strategy to ensure that we have enough reliable and properly collected representative data for our cross-sectional analysis and research.

Considering the objectives and scope of this thesis, we have decided to pick a secondary data, as the source of data needed to present our proposed solution. Below are few of the benefits of using a secondary dataset collected from an external data source:

- **Relevance and Specificity:** the data is highly relevant to the specific problem or context being addressed in the thesis. This ensures that the data directly aligns with the research objectives and is more likely to yield meaningful insights.
- **Data Control:** When using secondary data, we have greater control over data quality and collection methods. This allows for data refinement and validation, ensuring that the data is reliable and accurate, which is crucial for drawing valid conclusions.
- **Holistic Understanding:** The comprehensive view enables us to understand the problem from multiple angles and develop solutions that account for interdepartmental or cross- functional intricacies.
- **Real-world Applicability:** The secondary data reflects real-world scenarios and challenges faced by the organization. This makes it universally applicable and pragmatic for developing solutions that can be implemented in a real business setting.
- **Confidentiality and Security:** The collected secondary data has been managed with greater confidentiality and security measures compared to external datasets, ensuring the protection of sensitive information.

### **Data Collection Sources –**

There are various sources from where we have collected the secondary data for OPD fraud detection:

- **Claims Data:** It is the information submitted to health insurance companies by customers to get reimbursement of their healthcare expenses. Claims data contains information such as patient information, provider information, healthcare treatment information and billing information etc.

- **Account & Policy Data:** It is the information submitted to health insurance companies by customers at the time of policy purchase and account creation.
- **Discussions & Inputs from stakeholders:** Discussions conducted with business stakeholders and SME to gather information about their experiences and perceptions of fraud and abuse in the healthcare industry.

Analyzing anonymized and masked secondary data for developing risk score model requires proper data preprocessing, cleaning, and validation to ensure data quality and accuracy. It is crucial to ensure that the data used for the rule-based model is consistent, free from errors, and representative of the actual scenarios the model will encounter. Secondary data which we will be using will include several types of information, such as customer transactions, historical data, hashed user demographics, and any other data that the organization collects and stores for its operations.

To ensure data security and proper extraction of relevant data from the appropriate database systems, formal approval was obtained to access anonymized/ masked data for any consecutive four-month period. This time duration was chosen to provide a sufficient sample size for model development and validation.

While collecting the secondary data from the data warehouse and consulting with the data experts, we realized that all the data required for our research are mostly stored in a few objects/tables in the warehouse. After consolidating and de-normalizing, the data model is finally divided into 3 data objects namely claims data, policy data and accounts data. Accordingly, the data was requested.

Emphasis was placed on data privacy during the data collection process. All data were masked and anonymized throughout the research to ensure that no individual customer could be identified from the analyzed dataset. We have mentioned several mobile numbers and other sensitive information in a masked format throughout this document for illustration and representational purposes only.

The approval allowed us to securely access entire claims, policy, and accounts for a specific

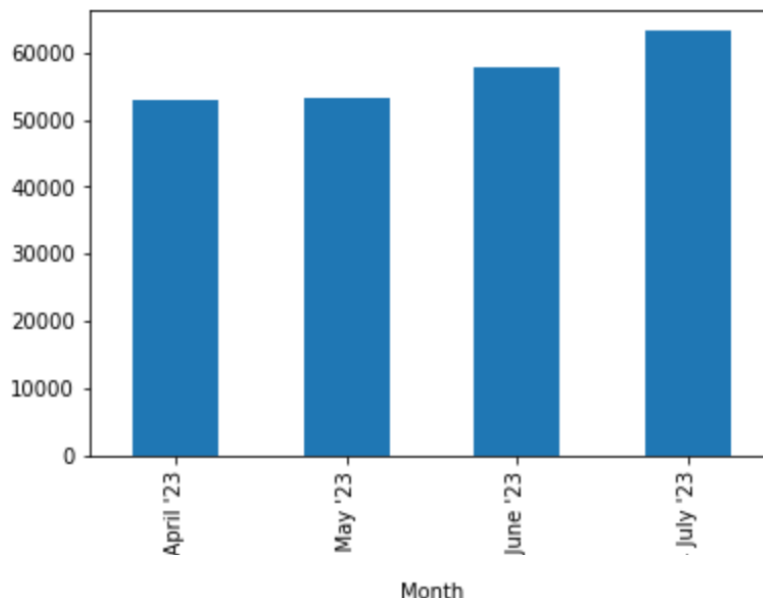
period of April 1<sup>st</sup>, 2023, till July 31<sup>st</sup>, 2023. Post approval we received a masked and anonymized data dump of requested data from the data team in the form of excel spreadsheets.

### Data Sampling Strategy:

Our study focuses on data collected over a four-month period for several key reasons:

- **Data Stability:** During this period, claims activity tends to stabilize and exhibit near-linear trends, enabling a clearer and more accurate analysis of risk assessment and customer behavior.
- **Sufficient Data Volume:** A four-month timeframe provides an ample volume of data for comprehensive analysis. This period strikes a balance between collecting enough data for robust insights and managing logistical constraints without compromising data quality.

The graph below displays the monthly volume of claims processed in the Outpatient Department (OPD) from April 2023 to July 2023.

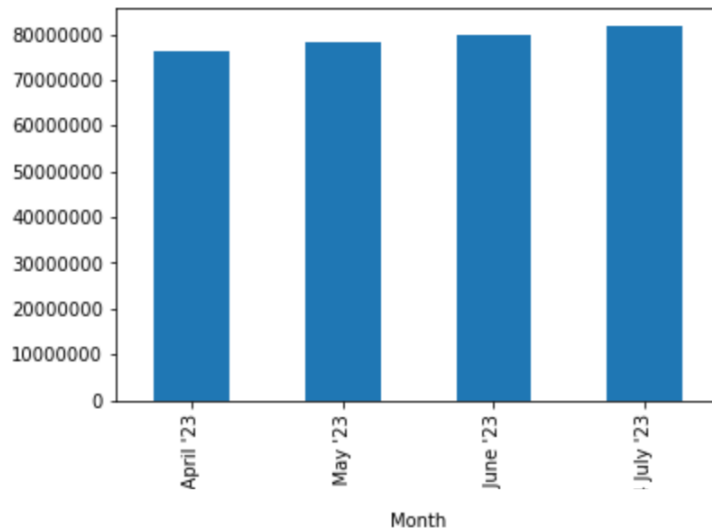


*Figure 3-1 Trend of number of claims (Month over Month)*

The consistent and increasing trend in claims volume, as seen in figure above, supports the

notion that a four-month period provides sufficient data volume for robust analysis, as it captures stable and predictable patterns in claims activity.

Now in the figure 3-4, the total monetary value of claims processed in the Outpatient Department (OPD) from April 2023 to July 2023.



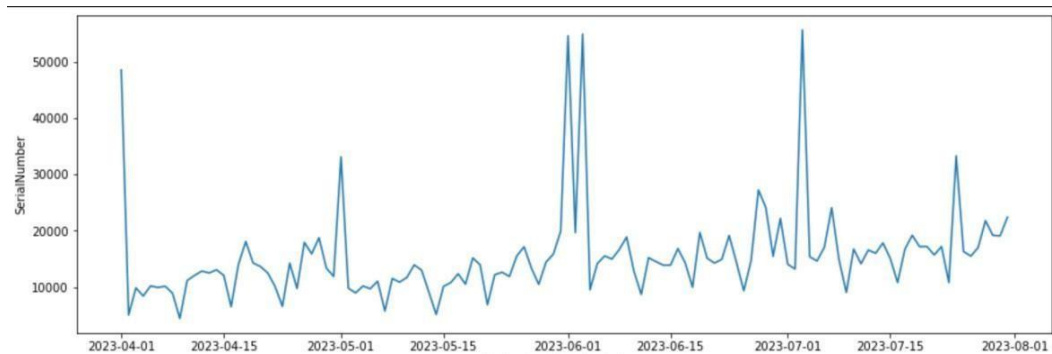
*Figure 3-2 Trend of number of total claims in INR (Month over Month)*

The stable trend in the claims' value (figure 3-4) indicates a predictable pattern in claim costs, which is useful for budgeting and financial forecasting. The consistent data over the four-month period suggests that this timeframe is sufficient for capturing stable financial patterns in OPD claims, providing a solid basis for further analysis and model validation.

Analyzing data from a substantial period allows us to assess the effectiveness of the initial scoring system and gather valuable insights for further refinement and optimization.

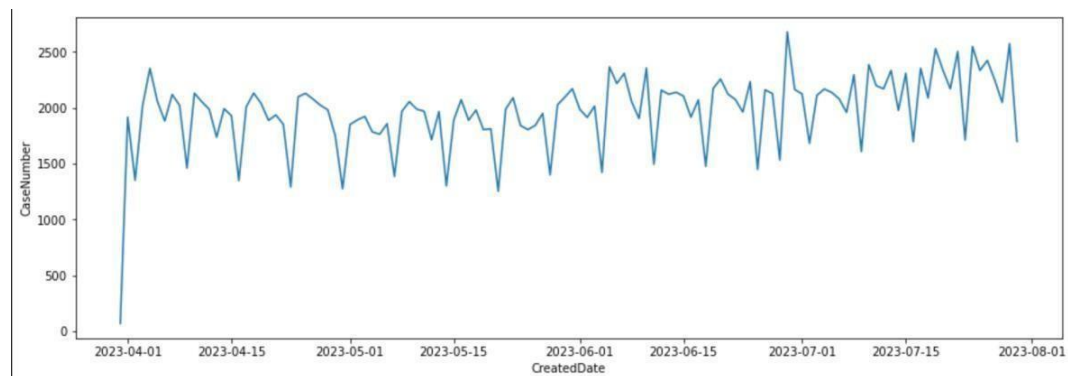
In the graphs below we have used examples of insurance policies purchased and how the number of claims is trending week on week to understand how time series analysis helps us to gather important insights.





*Figure 3-3 Trend of NOPs purchased*

From Figure 3-5 we got an indication that in the starting of the month there is always an evident spike with NOPs sold/purchased, after which we drilled down on the insight and built a hypothesis that insurance sales agents try to complete their monthly sales targets at the very starting of the month.



*Figure 3-4 Claims Trend*

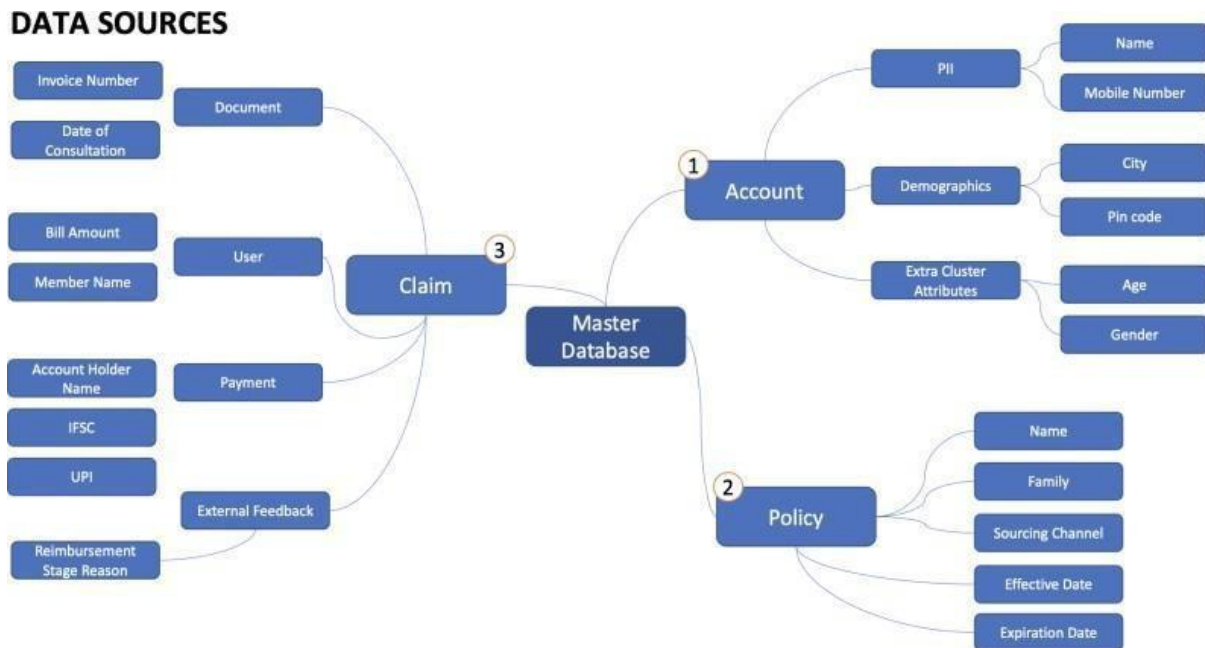
Figure 3-6 offers a compelling visual narrative of the evolving business landscape and the consequent impact stemming from a consistent upswing in claims. The upward trajectory of claim numbers is an insightful indicator of the business's expansion and activities. The ascending curve eloquently illustrates a pattern characterized by sustained and gradual growth in claims over time.

This growth can be attributed to a combination of factors, including heightened customer interactions, business expansion initiatives, or shifts in market dynamics. The importance of this visualization lies in its ability to effectively convey the augmented business volume propelled by the increasing number of claims. This evolving pattern serves as a pivotal context

for the rule-based customer risk scoring model currently in use. With an understanding of this progression, we are better poised to interpret and contextualize the outcomes of abuse detection efforts, ensuring that any deviations from the growth trend are subject to meticulous scrutiny, potentially revealing instances of fraudulent activity.

The collected data follows a data model, which is divided into 3 data objects –

- **Customer object:** It contains all the customer attributes, for ex. Name, Age, Demography.
- **Policy object:** It contains all the policy attributes linked with the account object, for ex. Policy Type, Purchase Date, Premium paid, Bounces.
- **Claim object:** It contains all transactional level data for reimbursement claims linked with the account and policy object, for ex. Claim Id, Claim Date, Claim Amount, Status of claim.



*Figure 3-5 Data Sources and Architecture*

### Data Dictionaries:

We have mentioned below an exhaustive list of all the fields/columns and their

definitions that were available in all the three objects discussed above. We will focus on the fields/variables relevant to our study in the subsequent sections.

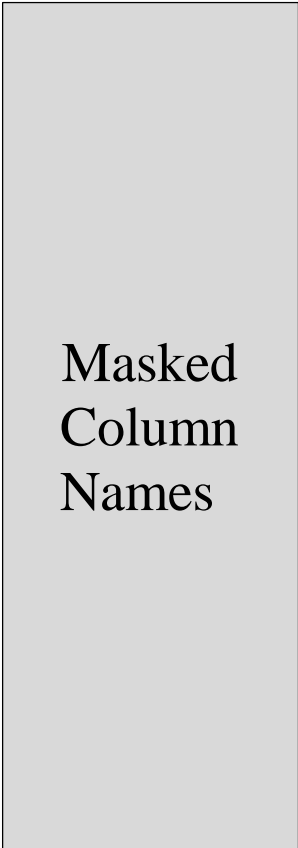
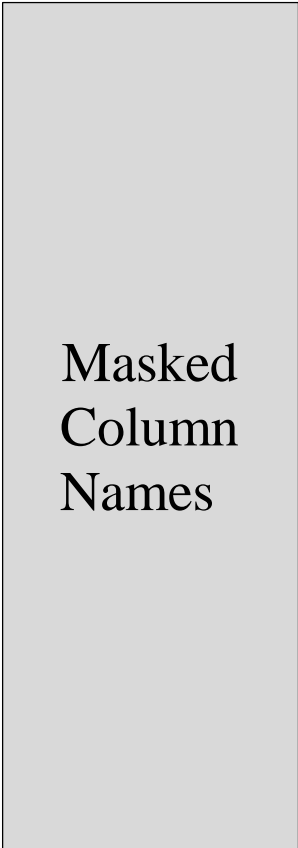
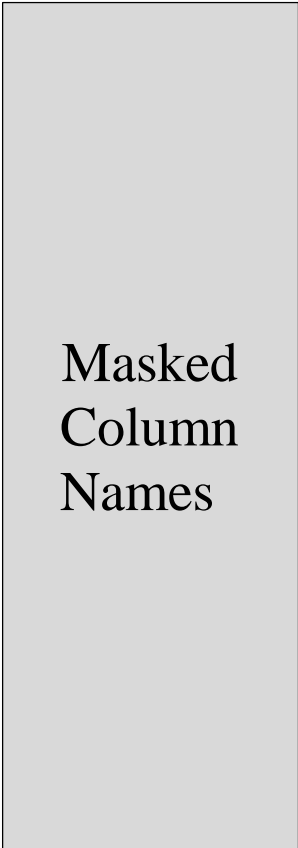
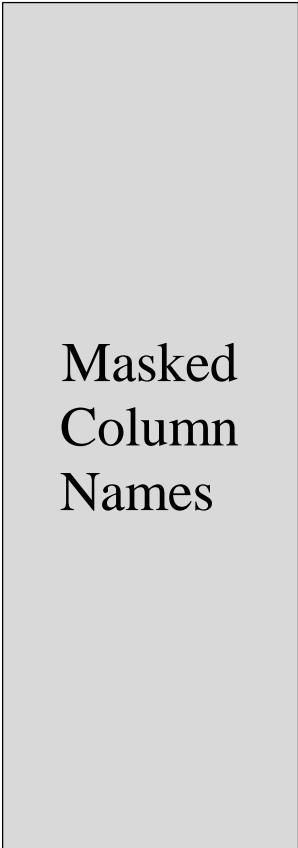
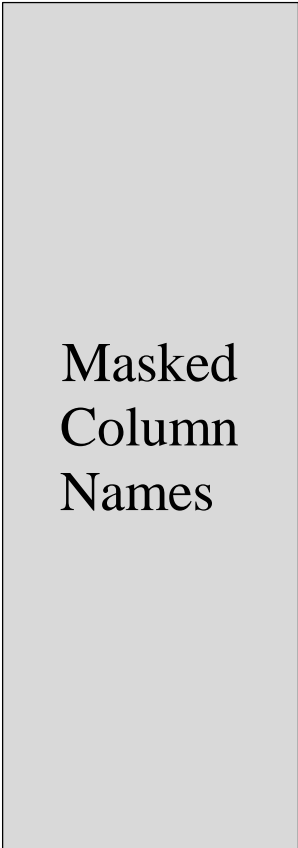
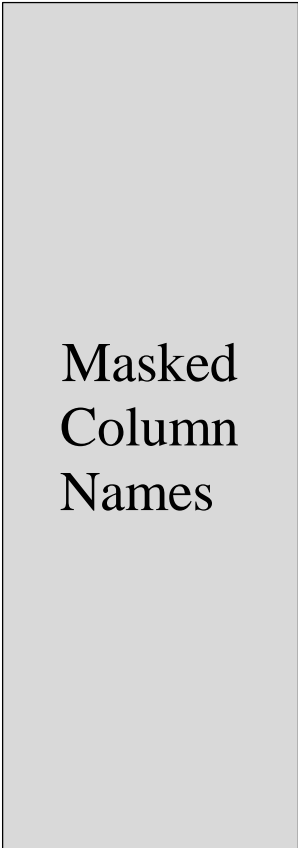
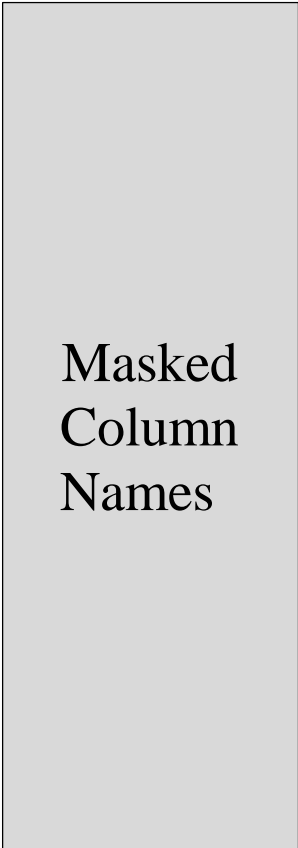
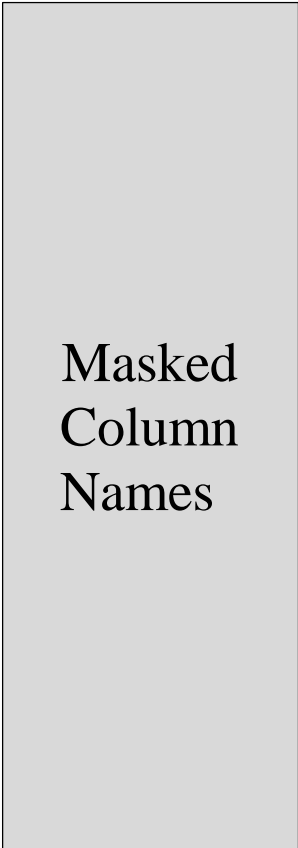
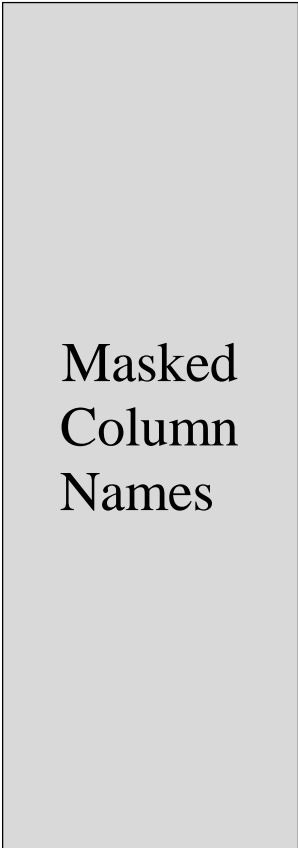
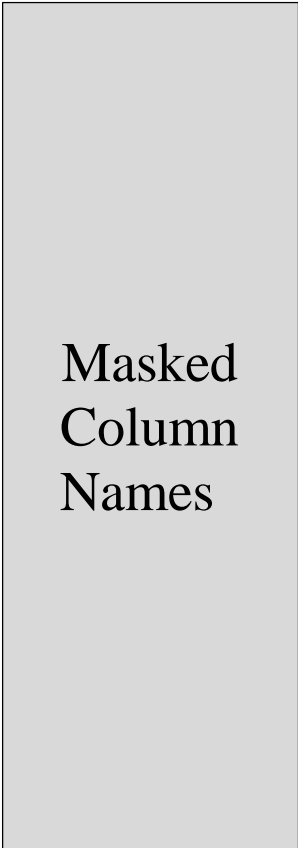
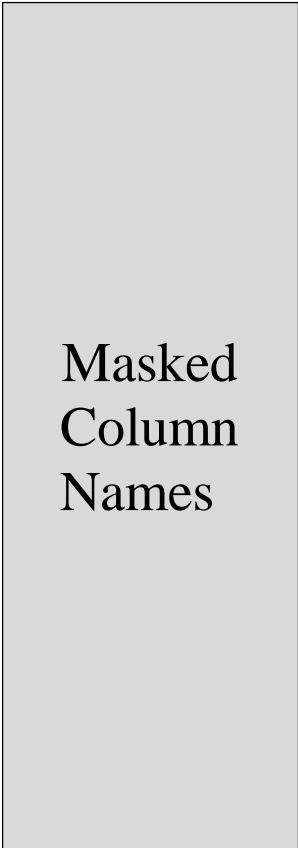
### Data Dictionary for Customer Account Fields:

Masked Column Names	•	Customer's full name.
	•	Identifier for the record type of the customer account.
	•	Date and time when the customer account was created.
	•	Date and time when the customer account was last modified.
	•	Date of birth of the customer.
	•	Gender of the customer.
	•	First line of the customer's address.
	•	Second line of the customer's address (if applicable).
Masked Column Names	•	Postal code or PIN code of the customer's address.
	•	Personal health record information (if applicable).
	•	Age of the customer calculated from the date of birth.
	•	City where the customer resides.
	•	State where the customer resides.
	•	Masked Mobile phone number of the customer.
	•	Email address of the customer.
	•	Alternate contact number provided by the customer.
	•	Contact number of the person who referred the customer.
	•	Name of the person who referred the customer.
	•	Name of the account holder.
	•	Account number associated with the customer.
	•	Name of the bank where the customer holds an account.
	•	IFSC associated with the customer's bank.

**Data Dictionary for Customer Insurance Policy Data Fields:**

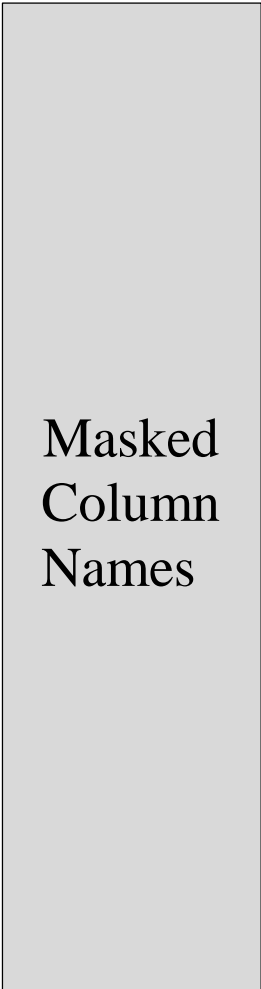
<p><b>Masked Column Names</b></p>	•	Unique identifier for the insurance policy record.
	•	Identifier for the customer account associated with the policy.
	•	Code representing the insurance product.
	•	Family to which the insurance product belongs.
	•	Description of the insurance product.
	•	Date and time when the insurance policy record was created.
	•	Datetime when the insurance policy record was last modified.
	•	Name of the insurance policy.
	•	Serial number or unique identifier of the insurance policy.

<p><b>Masked Column Names</b></p>	•	Date when the insurance policy was purchased.
	•	Status of the insurance policy (active, expired, etc.).
	•	Price or cost of the insurance policy.
	•	Number of policy (usually 1 for individual policies).
	•	Additional description related to the insurance policy.
	•	Annual premium amount for the insurance policy.
	•	Effective date of the insurance policy coverage.
	•	Effective term or duration of the insurance policy.
	•	Expiration date of the insurance policy coverage.
	•	Monthly premium amount for the insurance policy.
	•	Total amount paid for the insurance policy term.
	•	Total amount paid for the insurance policy.
	•	Total amount paid for insurance claims.
	•	Total sum insured for the insurance policy.

- Members\_Covered c: Information about the members covered.
- Product\_Name c: Name of the insurance product.
- SourceName c: Source or origin of the insurance policy.
- CancellationReason c: Reason for the cancellation of the insurance policy
- TotalCoverages c: Total number of coverages included in the policy.
- TotalMembers c: Total number of members covered under the policy.
- Agent\_Id c: Identifier of the agent associated with the policy.
- Base\_Policy\_Number c: Base policy number (if applicable for group policies).
- No\_Of\_Years c: Number of years the policy covers or the term duration.
- Premium\_Paid c: Amount of premium paid for the insurance policy.
- Product\_Type c: Type of insurance product.
-  : Date of the transaction related to the insurance policy.
-  : Code representing the distribution channel for the policy.
-  : Previous policy number (if applicable).
-  : Amount paid as a down payment for the insurance policy.
-  : Tenure EMI for premium payments.
-  : Number of down payments made for the policy.
-  : Number of EMIs for premium payments.
-  : Mode of payment (monthly, quarterly, annually).
-  : Code of the Insurance Marketing Distributor
-  : Name of the Insurance Marketing Distributor
-  : Code representing the sub-distributor of the insurance

**Data Dictionary for Customer Insurance Claims Fields:**



- Id: Unique identifier for the insurance claim record.
- Invoice\_Number c: Invoice number associated with the claim.
- CaseNumber: Case number associated with the claim.
- Date Consultation: Date of consultation for which the claim is made.
- CreatedDate: Date and time when the claim record was created.
- Paid\_Time c: Time when the insurance company paid reimbursement.
- Mobile c: Masked Mobile number associated with the claim.
- Status: Current status of the claim (e.g., Working, Paid, Rejected).
- OpenLoop\_Hospi Name of the hospital for claims.
- ClosedLoop\_Doc Name of the doctor for claims.
- Approved\_Amount Approved reimbursement amount for the claim.
-  Identifier of the asset/policy associated with the claim.
- Identifier of policy related to the claim.
- Name of the member making the claim.
- Name of the benefit for which the claim is made.
- Name of the account holder associated with the claim.
- Bank Account number associated with the claim.
- IFSC code of the bank associated with the claim.
- UPI ID related to the claim.
- Reason for the stage of reimbursement for the claim.
- Type of benefit claimed.
- Sub-type of the benefit claimed.
- Identifier of the customer account associated with the claim.
- Age of the member making the claim.
- Record type identifier associated with the claim record.

### **Variable Selection:**

In our thesis, we have chosen not to utilize all available fields due to several reasons.

Firstly, some fields may not be relevant to our research objectives or may not provide meaningful insights into the phenomenon under study. Certain variables may be redundant, leading to unnecessary duplication of information.

For example, Total Amount For Term and Total Amount is the same field and contains the same data.

Secondly, there may be practical constraints such as data availability, quality, or accessibility, which prevent us from incorporating certain fields into our analysis. Additionally, considering all available fields may lead to information overload, making it challenging to focus on the most critical aspects of the research. For example, Effective Term was not useful in our context, also the data available was not accurate and reliable. Instead, it can be easily derived from Effective Date and Expiration Date as and when required.

Therefore, we have carefully selected only a subset of fields mentioned below, those are most pertinent to our research questions and objectives, ensuring that our analysis remains focused and meaningful. By prioritizing these fields over others, we can streamline our research process and concentrate our efforts on investigating the most relevant factors influencing the phenomenon under investigation.

*Table 3-1 Selected fields across all tables/objects*





Data Object	Field Name	Description
Claims	Masked Column Names	Primary Key
		Invoice Number of claimed document
		Primary Key
		Appointment Date mentioned on claim document
		Claimed Date
		Reimbursement Payment Date
		Present Status of claim
		Provider Name
		Doctor Name
		Reimbursement Amount asked by customer
		Reimbursed Amount by us
		Policy member under which claim is raised
		Benefit under which claim was raised
		Bank Name of payment instrument to which amount is reimbursed
		Account Number of payment instrument to which amount is reimbursed
IFSC Code of payment instrument to which amount is reimbursed		
UPI Id of payment instrument to which amount is reimbursed		
Account	Masked Column Names	Remarks by Agent
		Name of Primary Policy Owner
		Mobile Number of Primary Policy Owner
		Age of Primary Policy Owner
		City of Primary Policy Owner
Policy/Asset	Masked Column Names	Pincode of Primary Policy Owner
		Gender of Policy Owner
		Present Status of policy
		Product Tag of policy
		Product Name of policy
		HAN of Policy
		Sourcing channel of policy
		Primary channel channel of policy
		IMD Agent name of policy
		Sub IMD Agent code of policy
Policy Effective Date		
Policy Expiry Date		

Table 3-3 Sample attribute values for all three data objects

Masked Column Names	5008p0000064QTVvAAM	5008p0000064U0HAAU	5008p0000064ZB5AAM
	MKDPH/2375	2324/234	NaN
	SR02335567	SR02355446	SR02343830
	01/04/23	05/04/23	02/03/23
	01/04/23	05/04/23	03/04/23
	02/04/23	NaT	NaT
	Paid	Rejected	Rejected
	Masked Sensitive Details		
	xxxx017529@api	xxxxnanpater64673@oknncbank	xxxx
	NaN	Illness & Injury is not mentioned on the presc...	Invoice does not contain Invoice number
	913601xxxx	957410xxxx	981832xxxx
	Masked Sensitive Details		
	NaN	390009	110095
	32	34	43
	Male	Male	Female
Masked Sensitive Details			
NaN	NaN	NaN	
07/08/23	28/08/23	27/10/23	
01/01/23	29/08/22	28/10/22	

From the above-mentioned datasets (Table 3-1, Table 3-2, Table 3-3), we got an idea about

what all important and relevant information is available in the company’s internal dataset which can be leveraged for better results of our algorithm. The variables that can be included in a dataset for customer risk scoring can vary depending on the available data and the specific problem being addressed.

Table 3-4 Sample data of Claim object

<b>Masked Column Names</b>	5008p0000064QTVAAAM	5008p0000064UDHAAU	5008p0000064ZBSAAM	5008p0000064amfAAA	5008p0000064bBvAAI
	MKDPH/2375	7324/234	NaN	1311	45014
	SR0233567	SR02335446	SR02343830	SR02348703	SR02350304
	01/04/23	05/04/23	02/03/23	31/03/23	04/04/23
	01/04/23	05/04/23	03/04/23	04/04/23	04/04/23
	02/04/23	NaN	NaN	04/04/23	NaN
	Paid	Rejected	Rejected	Paid	Rejected
	Masked Sensitive Details				
	xxxx017329@apl	xxxxhanpatel64673@okhdfcbank	xxxx	xxxxdawalokesh@oksbi	xxxxjumar.ambulkar@axl
	NaN	Illness & Injury is not mentioned on the presc...	Invoice does not contain Invoice number	NaN	Prescription / Invoice is edited; Illness & Inj...

Table 3-5 Claim object fields values uniqueness & unavailability

Column	Total Values	Unique Values	Null Values
<b>Masked Column Names</b>	240296	240296	0
	225505	124752	14791
	240296	240296	0
	239706	538	590
	240296	236080	0
	171901	156222	68395
	240296	19	0
	230538	128062	9758
	202903	117560	37393
	239718	7320	578
	227643	6562	12653
	228050	63644	12246
	239707	40	589
	205847	47496	34449
	74792	23837	165504
	74771	12045	165525
	147847	33139	92449
86556	1114	153740	

Table 3-6 Sample data of Policy holder (Account) object

<b>Masked Column Names</b>	913601xxxx	957410xxxx	981832xxxx	902143xxxx	702087xxxx
	Masked Sensitive Details				
	NaN	390009	110095	NaN	441901
	32	34	43	31	39
	Male	Male	Female	Male	Male

Table 3-7 Account object fields values uniqueness & unavailability

Column	Total Values	Unique Values	Null Values
Masked Column Names	240296	42933	0
	240296	44273	0
	144769	667	95527
	145093	5083	95203
	239767	89	529
	235209	2	5087

Table 3-8 Sample data of Policy object

Masked Column Names	GM01CSHP07263182	ML01HPWR05397663	ML01HPWR06178945	GM01CSHP07253793	01CHSU07227013
	Masked Sensitive Details				
	NaN	NaN	NaN	NaN	NaN
	07/08/23	28/08/23	27/10/23	31/12/23	22/01/24
	01/01/23	29/08/22	28/10/22	01/01/23	23/01/23

Table 3-9 Policy object fields values uniqueness & unavailability

Column	Total Values	Unique Values	Null Values
Masked Column Names	239876	51080	420
	239876	8	420
	239876	229	420
	239876	7	420
	239876	9	420
	239872	65	424
	123378	2813	116918
	58788	4525	181508
	239876	1176	420
	239876	584	420

In Table 3-4: 3-9, we present a comprehensive snapshot of the data points within the Claim, Policy, and Policy Holders (Account) object. This overview encapsulates the total values for each field, the count of distinct values, and the prevalence of null values in each field. As we progress further, we will delve into the potential ramifications of null values in our dataset and explore the pertinent measures to address and mitigate the impact of these null values.

```
Duration : 2023-03-31 to 2023-07-30
Claims : 240296
Claimed Amount : 316082580.0
Max Claimed Amount : 57670.0
Min Claimed Amount : 0.0
Avg Claimed Amount : 1388.5
Min Age : 0.0
Max Age : 124.0
Avg Age : 44.12
```

*Figure 3-6 Final Dataset insights (Before Pre-Processing)*

The above Table 3-8 gives us an idea about the data and a few key measures. Some important insights that we derive from the data are -

- Our dataset covers a time span from April '23 to July '23.
- It comprises a substantial 2,40,296 claims, reflecting the sheer volume of data.
- Notably, these claims cumulatively amount to an impressive 31.6 crores INR in value.
- The claim amounts span a wide range, from as low as 0 INR to approximately 57,670 INR, with an average ticket size of 1,388 INR.

### **3.5 Data Cleaning and Pre-processing**

After data collection, the preprocessing stage ensures its quality and reliability. During preprocessing, missing values are handled either through imputation techniques or removed if they are minimal. Outliers, which can skew the analysis, are also identified and appropriately managed.

*Table 3-10 Missing Values analysis from data objects*

Column	Missing Values	Missing Records %
Id	0	0.00
	14791	6.16
	0	0.00
	590	0.25
	0	0.00
	68395	28.46
	0	0.00
	9758	4.06
	37393	15.56
	578	0.24
	12653	5.27
	12246	5.10
	589	0.25
	34449	14.34
	0	0.00
	165525	68.88
	0	0.00
	153740	63.98
	0	0.00
	0	0.00
	95533	39.76
	95203	39.62
	529	0.22
	5087	2.12
	420	0.17
	420	0.17
	420	0.17
	420	0.17
	420	0.17
	424	0.18
	116918	48.66
	181508	75.54
	420	0.17
	420	0.17

Masked  
Column  
Names

Table 3-10 highlights the variations in missing rates across the three datasets, allowing for a quick assessment of the data completeness. These findings are instrumental in shaping the course of the research, as they inform the selection and formulation of rules based on the availability of data.

Missing values could occur due to several reasons, such as data not being recorded, data corruption, or data not being applicable to certain cases. These missing values can pose challenges during analysis and may lead to biased results. Data experts typically use imputation techniques to replace missing values with estimated values based on the existing data. Alternatively, if there are too many missing values in a particular feature, it might be prudent to remove the entire feature from the dataset to avoid any potential bias.

In this study, a thorough assessment of the data quality and completeness has been conducted by analysing the fill rates of the account, policy, and claim objects within the dataset. Fill rates, which indicate the proportion of non-missing values in a dataset, are a critical indicator of data completeness and reliability.

The calculated missing rates for each dataset were derived by evaluating the percentage of null / missing values for every field within the respective objects. The findings shed light on the extent to which the dataset's key attributes are populated with information.

Upon analysis, it is evident that the Account dataset demonstrates high fill rates across most of its fields, signifying a substantial level of data completeness. This suggests that the customer account data is rich with information, making it a robust foundation for rule-based customer risk scoring models.

The policy dataset exhibits similar trends in data completeness, with most fields displaying satisfactory fill rates. This implies that the policy-related information is well-represented, thereby enhancing the potential for effective rule-based risk identification.

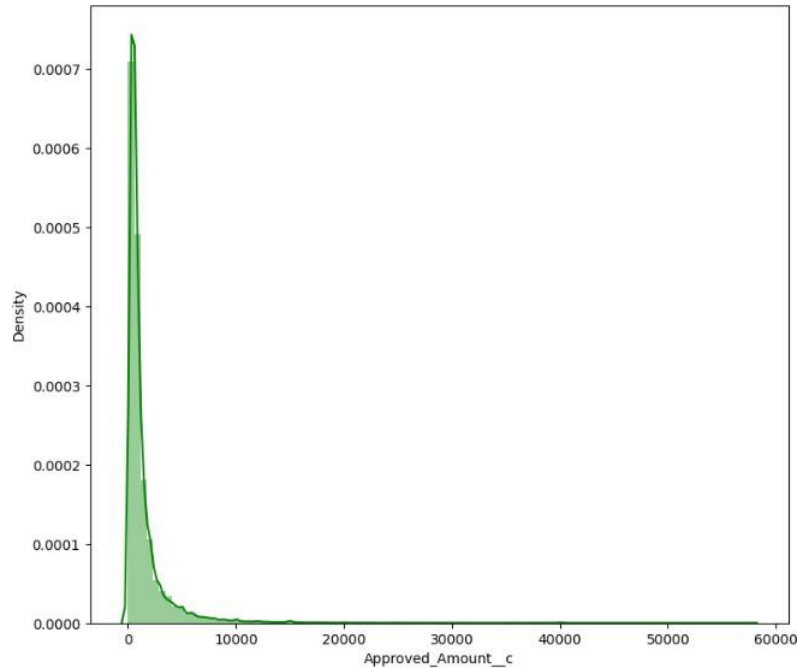
In the case of the claim dataset, the fill rates vary across different fields. While some fields exhibit high fill rates, indicating comprehensive data availability, others demonstrate lower fill rates, suggesting the presence of missing values in certain instances. Addressing these fields with lower fill rates could be a critical consideration for refining the rule-based customer risk scoring process.

The insights gained from this analysis guide the subsequent stages of rule formulation, ensuring that the developed rules are anchored in comprehensive and reliable data attributes. This, in turn, contributes to the robustness and effectiveness of the proposed rule-based customer risk scoring methodology for abuse detection.

Before addressing the null and missing values, it is essential to gain a deeper understanding of these data points. This will enable us to determine the most appropriate actions to take.

*Table 3-11 Claim amount statistics*

Description	Approved_Amount_c
count	227643
mean	1389
std	2236
min	0
0.25	400
0.5	700
0.75	1500
max	57670



*Figure 3-7 Claim Amount Skewness*

The distribution of reimbursed claim amounts, as depicted in the provided Table 3-11, exhibits a distinct right-skewed pattern (Figure 3-9). This skewness implies that most of the claim amounts tend to cluster towards the lower end of the range, while fewer claims extend towards higher values. This distribution characteristic is indicative of a scenario where a considerable number of claims involve relatively smaller amounts, while a limited subset of claims may involve significantly larger amounts.

It is important to acknowledge that the presence of higher claimed values in the distribution should not be disregarded or outrightly removed. Doing so could potentially lead to the loss of critical insights and information, especially concerning the differentiation between potentially fraudulent and genuine customer activities.

The higher claimed values in the distribution, despite being fewer in number, hold crucial significance in the context of risk scoring and abuse detection. These higher claim amounts

could potentially indicate instances of fraudulent behaviour that require specific attention. Removing these observations with higher claimed values would diminish the dataset's ability to capture and distinguish such anomalous activities, which are vital for accurate and effective customer risk scoring for abuse identification.

In essence, while the left-skewed distribution may present challenges in handling extreme values, it is imperative to approach the data with a nuanced perspective. Neglecting higher claimed values could lead to a skewed perception of the overall claim landscape, potentially undermining the risk score for fraud detection process.

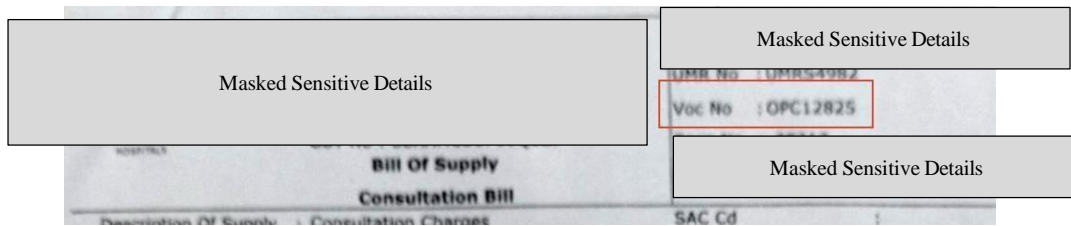
Data pre-processing involves correcting errors and inconsistencies in the data. Errors can arise from human errors during data entry or data extraction, which can lead to incorrect values or formatting issues. Inconsistencies may occur when data is collected from various sources or systems, resulting in discrepancies between similar data points. Addressing these errors and inconsistencies is crucial to ensure the data's accuracy and reliability for subsequent analyses.

In our dataset, the invoice number from the billing documents is one such data point that contains instances of human error.

*Table 3-12 Inconsistent invoice number values*

<b>Invoice_Number__c</b>	<b>CaseNumber</b>
15	01016777
540	MASKED 01016779
	01016780
	01016781
	01016783





*Figure 3-8 Case Number CASE01016781*

In the context of reimbursement claims, a noteworthy observation emerges from the data illustrated above. Specifically, a substantial number of reimbursements claims exhibit instances where certain critical fields, such as the invoice number, remain unpopulated or incompletely captured by the claim’s processors – Table 3-12 and Figure 3-10. This aspect introduces a noteworthy challenge in the process of establishing a reliable and accurate relationship between invoices and the corresponding claim dates, ultimately impacting the identification of outliers within the dataset.

The absence or incomplete nature of the invoice number field significantly hinders the ability to precisely link claims with their respective invoice records. This linkage is pivotal for a comprehensive understanding of the reimbursement landscape, as it aids in the identification of anomalies or outliers within the claims data. Anomalies, in this context, could refer to situations where claims deviate significantly from the norm, raising suspicion of potential irregularities, such as abusive activities.

Data preprocessing is a critical step as it ensures that the subsequent analysis and modeling are based on clean and accurate data. By gathering relevant data and meticulously pre-processing it, we set the foundation for effective customer risk scoring for abuse detection and build a robust framework for the subsequent stages of the process. (Al-Hashedi and Magalingam)

In the data preprocessing phase, one of the crucial steps is to remove duplicates and irrelevant data. Duplicates can occur in datasets because of several explanations such as data input mistakes, system glitches, or merging different sources. These duplicate records can lead to biased analysis and inaccurate results. Hence, identifying and eliminating duplicates is essential to maintain data integrity.

Irrelevant data refers to information that does not contribute to the analysis or does not align

with the research objectives. Such data can clutter the dataset and increase processing time unnecessarily. It is essential to carefully review the data and eliminate any irrelevant information to focus only on the most relevant data points.

- Date of Consultation and Reimbursement Paid Date: These were imputed with the "Claim Created Date" based on the assumption of same-day consultation and claim resolution.
- Open-Loop Hospital Name: Null values were filled with "Doctor Name," assuming a clinic under the doctor's name.
- Claimed Member and Bank Account Member: Replaced with "Primary Policy Holder." While this simplifies data handling, it disregards potential discrepancies between the policyholder and the actual claimant or beneficiary.
- Null Gender: Classified as "Others."
- Age: Filled with the "Average Age" of the data.
- Records with Null Claim Amount: Dropped based on the assumption of dummy/test claims.

```
Duration : 2023-03-31 to 2023-07-30
Claims : 227242
Claimed Amount : 315888648.0
Max Claimed Amount : 57670.0
Min Claimed Amount : 0.0
Avg Claimed Amount : 1390.1
Min Age : 0.0
Max Age : 124.0
Avg Age : 43.98
```

*Figure 3-9 Final Dataset insights (After Preprocessing)*

Initially, we had 2,40,296 claims as seen in figure 3-11. After dealing with missing values, null values, duplicate values, and irrelevant data points, we see the number of claims dropped to 2,27,242 as seen in figure 3-11. For any accurate rule-based model, data quality is of utmost importance and so, even though our data points are reduced, now we have a richer quality of data available at our disposal. Below we dive deeper into pre-processing of individual data points of our 3 main objects – Claims, Account & Policy.

Table 3-13 Data points and missing value count

Column	Missing Values	Missing Records %
Id	0	0.00
	10487	4.61
	0	0.00
	0	0.00
	0	0.00
	55343	24.35
	0	0.00
	3306	1.45
	30987	13.64
	0	0.00
	0	0.00
	0	0.00
	0	0.00
	0	0.00
	0	0.00
	156632	68.93
	0	0.00
	150154	66.08
	0	0.00
	0	0.00
	93046	40.95
	92747	40.81
	0	0.00
	0	0.00
	0	0.00
	0	0.00
	0	0.00
	0	0.00
	0	0.00
	0	0.00
	111549	49.09
	172239	75.80
	0	0.00
	0	0.00

Masked  
Column  
Names

It is essential to recognize that these specific attributes might lack direct relevance to the identification of customer abuse patterns. In the context of constructing robust and unbiased modelling, it becomes imperative to discern and address such attributes.

The strategic identification and subsequent exclusion of columns with a 100% fill rate but limited relevance hold significant importance. The presence of such attributes, while complete, may introduce potential biases during the modelling and rule generation phases of the analysis. There were few columns with low fill rates also, but we still must keep them as they play a significant role in risk scoring and abuse detection, for ex. IFSC code, it has lot of missing records, but we can handle it by creating a feature by aggregating Account Number, Account Holder Name and IFSC Code.

Biases can arise when the modelling process assigns undue importance to these less-relevant attributes, potentially overshadowing other critical variables that are truly indicative of abuse patterns.

Overall, data preprocessing is a fundamental step in data analysis projects. It lays the foundation for accurate and meaningful insights by ensuring that the data is clean, relevant, and consistent. Proper data preprocessing enhances the reliability of the analysis and aids in making informed decisions based on the data-driven outcomes.

After data pre-processing, our dataset is now free from absurd values. We have successfully created reliable input data without any discrepancies. Based on this clean dataset, we initiated our Exploratory Data Analysis (EDA), which will be covered in subsequent sections. The EDA will include summary statistics, data visualization, correlation analysis, and data distribution examination to identify patterns, trends, and outliers, providing valuable insights for further analysis.

### **3.6 Exploratory Data Analysis**

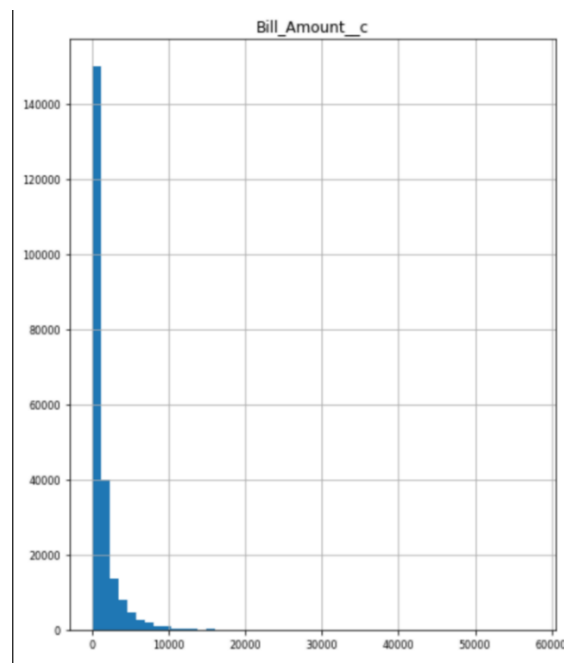
Exploring and visualizing the data are crucial steps in risk analysis to gain insights and identify patterns or trends that may indicate potential abuse or fraudulent activities.

**Data Exploration:** In this step, we closely examine the dataset to understand its structure, size, and distribution. Along with data understanding we looked for basic statistics such as mean, median, minimum, maximum, and standard deviation for numerical features. For categorical

features, we checked the frequency of various categories. Exploratory data analysis (EDA) techniques are used to uncover patterns, correlations, and initial insights.

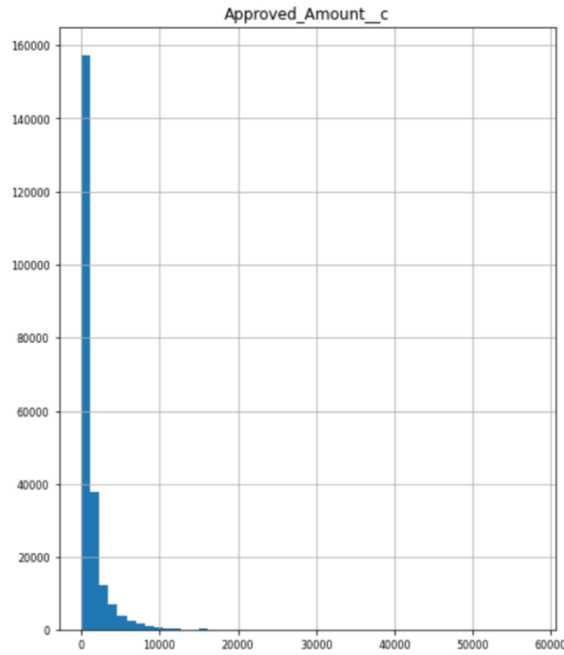
**Data Visualization:** Data visualization is a powerful tool to understand the data and detect patterns visually. It involves creating plots, charts, and graphs to represent the data's distribution and relationships between variables. Common visualization techniques include scatter plots (Hao, Ming & Dayal, (2010)), histograms, bar charts (Gorai, Pal, and Gupta, 2016), line plots, box plots (Stojanovic, (2021)), and heatmaps (Argyriou, (2013)).

Now we explore some data points we discovered in last section via various charts and graphs.



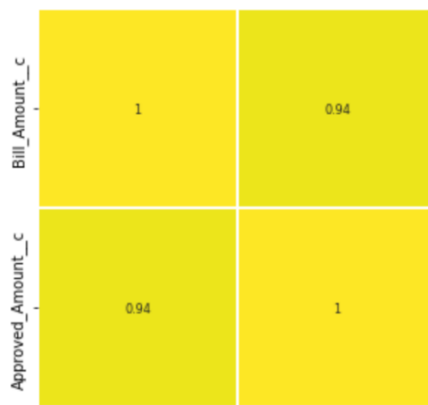
*Figure 3-10 Bill amount (amount asked by customer) histogram*

The histogram displayed above represents the distribution of the bill amounts requested by customers. The data is highly right skewed, with most bill amounts concentrated towards the lower end. As the bill amount increases, the frequency sharply declines, with very few occurrences of higher bill amounts.



*Figure 3-11 Approved amount (Actual amount reimbursed to customer) histogram*

The above graph reveals most customer approved amount falls into lower amount categories. This suggests two things: first, customers typically receive reimbursements for smaller expenses. Second, there might be an opportunity to streamline the process for frequent, lower-value refunds.

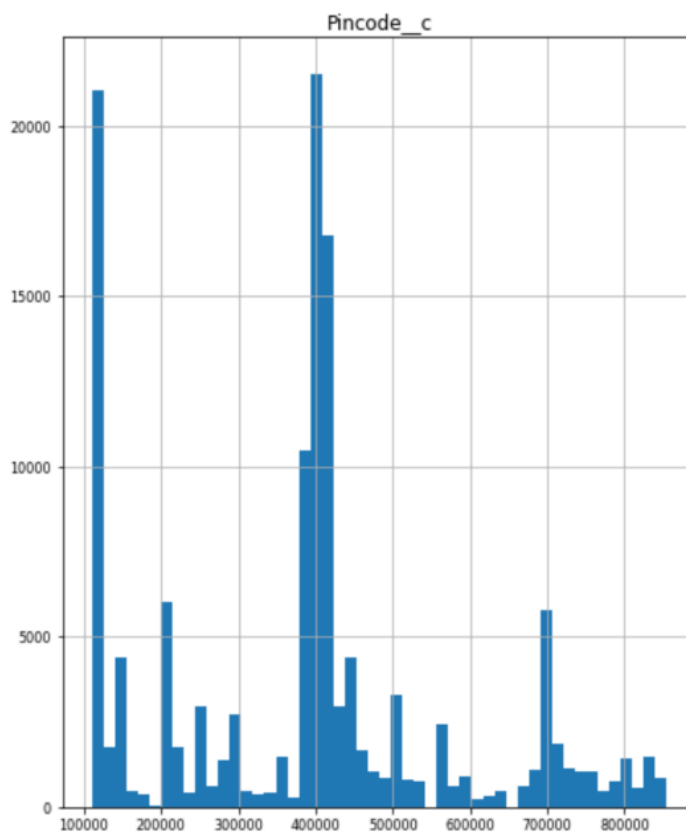


*Figure 3-12 Correlation between Bill Amount and Approved Amount*

In the analysis of bill amounts requested by customers and the corresponding approved amounts, both histograms exhibit a right-skewed distribution (Figure 3-12 and Figure 3-13), indicative of a prevalence of lower values with a long tail extending towards higher values. However, a nuanced distinction arises in the bin sizes of these histograms, reflecting a subtle

yet crucial aspect of the reimbursement process. This discrepancy is attributed to instances where partial payments occur in adherence to the standard operating procedures (SOPs) governing product transactions. Consequently, this partial payment dynamic introduces a variation in the distribution of approved amounts compared to the requested bill amounts. The exploration of these bin size disparities provides valuable insights into the intricacies of customer reimbursement behaviours and aligns with the overarching theme of rule-based customer risk scoring, shedding light on the nuanced financial interactions within the context of established operational guidelines.

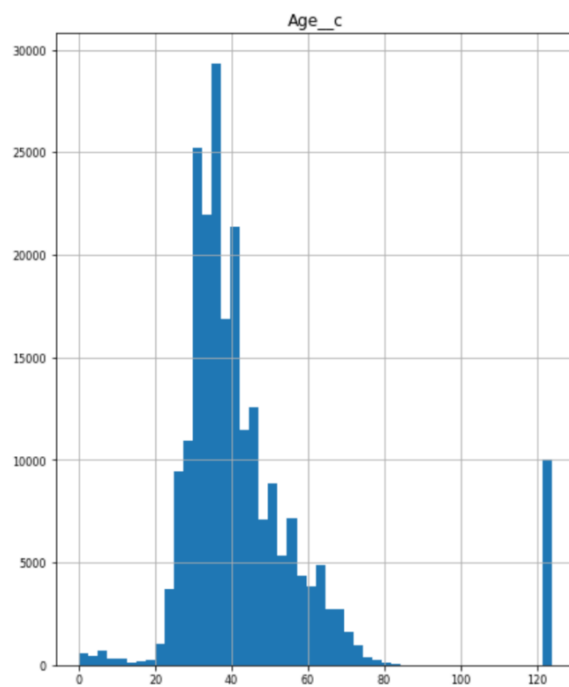
The above graphs helped us get some insights on bill amount and approved amount with their trends, let shift our focus towards other data points – pin code, age, and gender of customer and how it correlates with claims.



*Figure 3-13 Pin code Distribution*

The examination of the histogram depicting claims distribution across different pin codes, Figure 3-15, reveals notable trends that significantly contribute to the understanding of customer risk scoring dynamics. Particularly noteworthy are the peaks in claims originating

from pin codes within series 1 and series 4. This concentration of maximum claims in these specific pin code series underscores the importance of geographical patterns in influencing claim frequencies. The clustering of claims in certain pin codes suggests a potential correlation between regional factors and customer behaviours, contributing crucial insights for the development of a rule-based customer risk scoring model. Analysing the prevalence of claims in distinct pin code series enables the identification of geographical hotspots, thereby enhancing the predictive accuracy of the risk scoring algorithm by incorporating regional nuances. This observation aligns with the overarching theme of the research paper, emphasizing the significance of localized patterns and rules in optimizing customer risk assessment within the insurance domain.



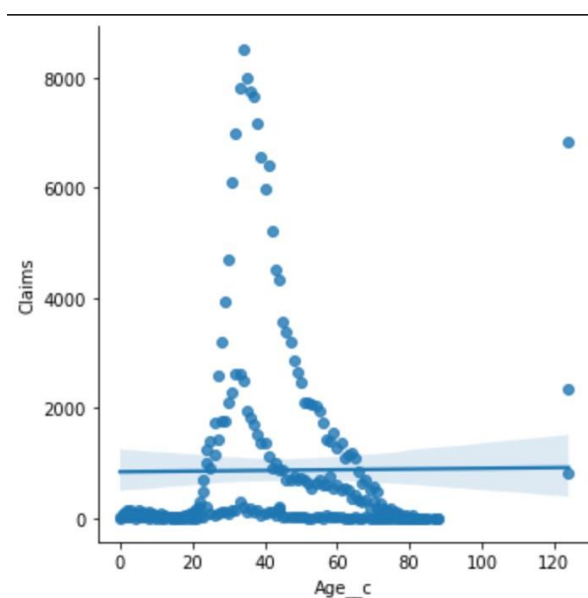
*Figure 3-14 Age Distribution*

Upon scrutinizing the histogram illustrating the correlation between insurance claims and the age of customers, Figure 3-16, discernible patterns emerge, holding profound implications for the formulation of a rule-based customer risk scoring paradigm. Notably, a conspicuous concentration of claims manifests within the age spectrum of 20 to 60, illuminating age as a salient determinant of insurance claim frequencies. Particularly striking is the zenith within the 30 to 40 age cohort, signifying a discernible pinnacle in claims during this specific life phase. This observation underscores the pivotal role of age-related factors, encompassing lifestyle choices, risk proclivities, in exerting influence on insurance claims. The prevalence of claims within this age bracket accentuates the imperative to integrate age-specific rules into the fabric



of the risk scoring algorithm, recognizing the diverse risk profiles inherent to distinct life stages.

Decoding the nuanced prevalence of claims within the 30 to 40 age group furnishes invaluable insights for refining the rule-based risk scoring model. Tailoring rules that encapsulate the distinctive risk characteristics associated with this age range augments the precision and prognostic efficacy of the model. This harmonizes seamlessly with the overarching theme of the research paper, accentuating the significance of rule-based methodologies in intricately calibrating customer risk assessment strategies within the dynamic milieu of insurance. The employment of age-specific rules not only acknowledges the fluid nature of risk across disparate demographic segments but also lays the groundwork for a refined and effective framework for optimizing risk scoring protocols.



*Figure 3-15 Age-Claims Correlation*

Interpreting Figure 3-17 involves recognizing the trend depicted by the regression line and understanding the distribution of data points around it. The significant increase in number of claims for certain age groups depicts a relationship between age group and claims. This is corroborated by the concentration of data points within the 20 to 60 age range, suggesting an overall higher incidence of claims.

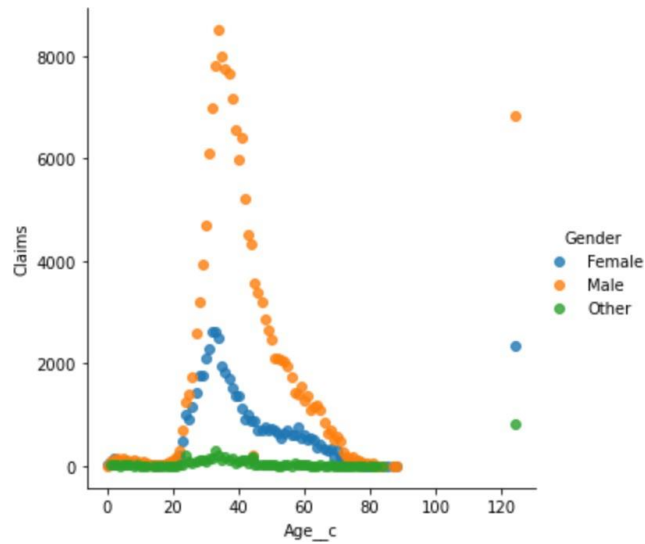


Figure 3-16 Age-Claims-Gender Correlation

Figure 3-18, incorporating gender as a hue, reveals a consistent trend of increased insurance claims with age for both males and females. The plot lines for both genders follow a similar trajectory, with a notable spike in claim frequencies observed within the 30 to 40 age group. This suggests that age is a shared factor influencing insurance claim patterns for both male and female policyholders.

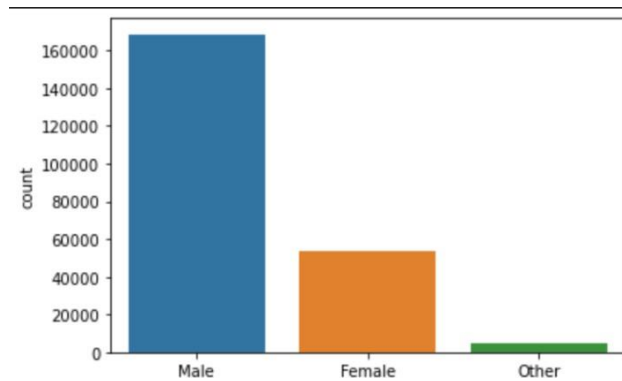


Figure 3-17 Distribution of insurance claims by gender

Figure 3-19, depicting the claims distribution by sex reveals a notable trend where males exhibit the highest frequency of claims, followed by females and then individuals categorized as 'others.' This observation holds significant implications for the formulation of a rule-based customer risk scoring model within the insurance domain.

The pre-eminence of male claimants in terms of sheer numbers underscores a gender-based

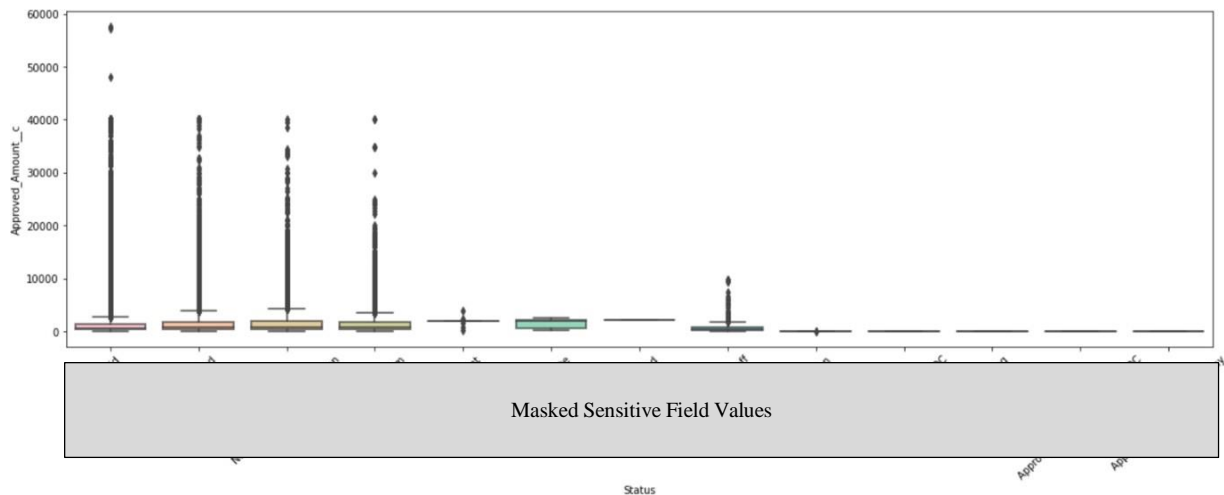
distinction in insurance behaviours that merits closer examination. The higher frequency of claims among males might be indicative of distinct risk profiles, lifestyle choices, or occupational factors that contribute to their increased participation in insurance claims. These insights are crucial for the development of tailored rules within the risk scoring model that can effectively capture the unique risk characteristics associated with male policyholders.

Concurrently, acknowledging the lower frequency of claims among females and individuals categorized as “Other” prompts considerations of the potential drivers behind these disparities. It necessitates an exploration into the underlying factors—whether they pertain to risk aversion, differing lifestyle patterns, or specific socioeconomic conditions that influence the likelihood of making an insurance claim. Understanding these nuances is essential for designing rules that accurately reflect the diverse risk landscapes associated with each gender category.

The importance of this gender-based analysis lies in its capacity to inform the rule-based customer risk scoring model with gender-specific insights. Tailoring rules to account for the observed differences allows the model to more precisely assess the risk associated with each gender category. This targeted approach enhances the discriminatory power of the model, ensuring that it can effectively adapt to and predict the distinct risk profiles associated with male, female, and 'other' policyholders.

In the broader context of the research on rule-based customer risk scoring, this gender-centric exploration contributes to the model's granularity and sophistication. By incorporating gender-specific rules, insurers can better align their risk assessment strategies with the nuanced characteristics of diverse policyholder groups, ultimately leading to a more accurate and finely tuned risk scoring model.

The above visualisations were sufficient for us to understand the distribution of customer's age, gender wise statistics and correlation of these point with respect to claims pattern. Now we shall understand more about claims data point such as claim status, approved amount, claim intimation source etc.



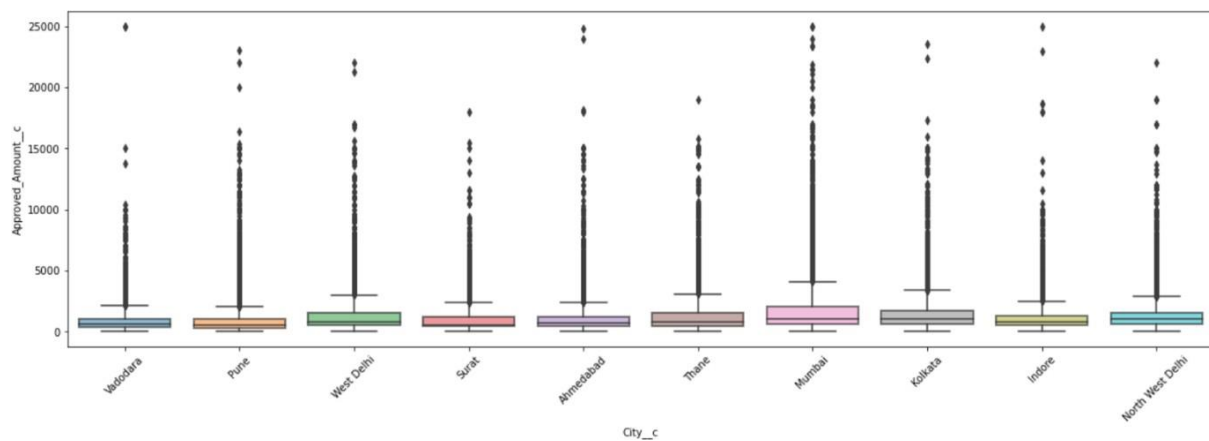
*Figure 3-18 Box plot for Claim Status*

The utilization of a boxplot, Figure 3-20, to elucidate the distribution of approved amounts contingent upon the status of insurance claims holds paramount significance in the context of developing a rule-based customer risk scoring model. This visualization provides a nuanced depiction of the statistical distribution, offering insights into the central tendency, spread, and presence of potential outliers within each status category. The distinct boxes in the plot represent the interquartile range (IQR) and the median, while the whiskers extend to reveal the range of the data, allowing for a comprehensive assessment of the spread.

The importance of scrutinizing this visualization status-wise lies in uncovering inherent patterns and variations that might be concealed in an aggregate analysis. By segregating the approved amounts based on claim status, the model can discern specific trends, anomalies, or disparities associated with different statuses—be it approved, pending, or denied claims. This granularity facilitates the formulation of tailored rules contingent upon the unique characteristics of each claim status, enabling a more refined and targeted risk scoring approach.

In essence, this status-wise exploration aids in the identification of distinct risk profiles associated with varying claim outcomes. For instance, understanding the distribution of approved amounts for denied claims may reveal potential inconsistencies or irregularities that could serve as red flags in the risk assessment process. Conversely, comprehending the patterns for approved claims allows for the establishment of rules that align with successful claims, contributing to a more accurate risk scoring model.

Therefore, this visualization not only elucidates the statistical landscape of approved amounts but also underscores the importance of dissecting this information based on claim status. By doing so, the rule-based customer risk scoring model can tailor its rules to the specific nuances of each status category, enhancing its discriminatory power and ensuring a more robust and adaptive approach to risk assessment within the insurance domain.



*Figure 3-19* Amount distribution at City level

The utilization of a boxplot, Figure 3-21, to depict the quantiles of approved amounts across the top 10 cities in the context of insurance claims is a crucial component in the development of a rule-based customer risk scoring model. This visualization serves to unravel significant variations in approved amounts specific to each city, shedding light on the distinctive characteristics and patterns associated with different urban centres.

The observed differences between cities in the boxplot signify potential geographical nuances in customer behaviours, risk profiles, and economic factors that influence the outcomes of insurance claims. For instance, variations in median approved amounts, interquartile ranges (IQR), and the presence of outliers within each city's boxplot may indicate differences in the average claim values, the spread of claims, and the occurrence of exceptional cases, respectively.

In essence, this city-wise exploration of approved amounts via boxplots not only identifies statistical variations but also serves as a valuable tool for tailoring risk assessment rules to the unique characteristics of each city. By acknowledging and incorporating these differences into the rule-based model, insurers can optimize risk scoring protocols to better align with the

intricacies of regional insurance dynamics, ultimately contributing to more accurate and effective customer risk assessment.

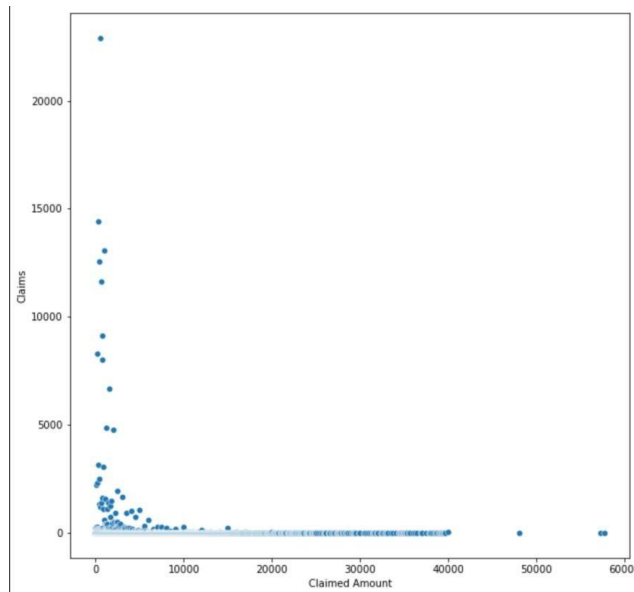
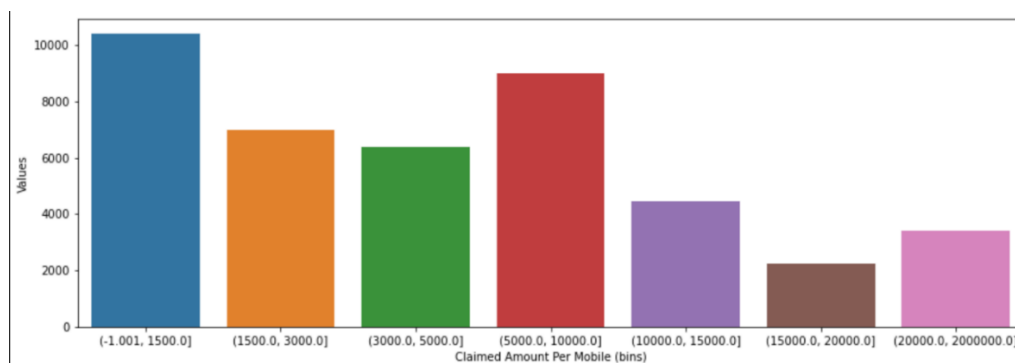


Figure 3-20 Number of claims vs claimed amount

Figure 3-22 above provides a clear visualization of the distribution of claims across different value ranges. It is evident that a substantial majority of claims fall within the range of 0-1000 INR. This concentration of claims within this specific range signifies a significant trend in the dataset.

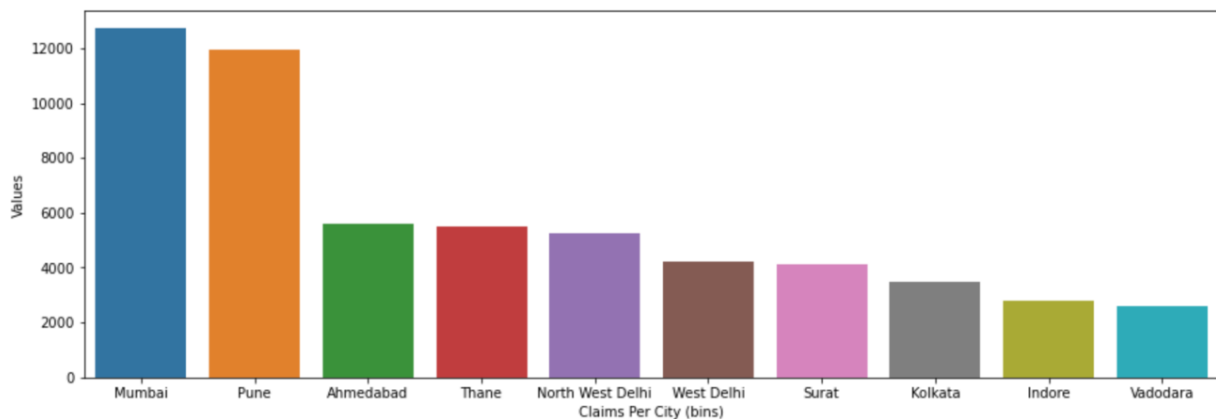
This observation serves as a foundational insight, indicating that a vast number of claims are of relatively lower value. Understanding this distribution is paramount as it guides our approach to detecting potentially fraudulent claims. By focusing our attention on this prevalent range, we can tailor our rule-based customer risk scoring strategy to effectively capture irregularities that might be obscured within this common range.



*Figure 3-21 Claim Amount skewness*

Figure 3-23 offers a clear and immediate understanding of the distribution of claim amounts. It becomes conspicuously clear that a substantial majority of claims fall within the bracket of 0 to 10,000 INR. What is particularly striking is the highest frequency bin, which encapsulates claim amounts ranging from 0 to 1,500 INR—these values closely align with our average ticket size for claims. This observation carries significant weight as it highlights the strong correlation of these specific bins with the expected range of typical claims.

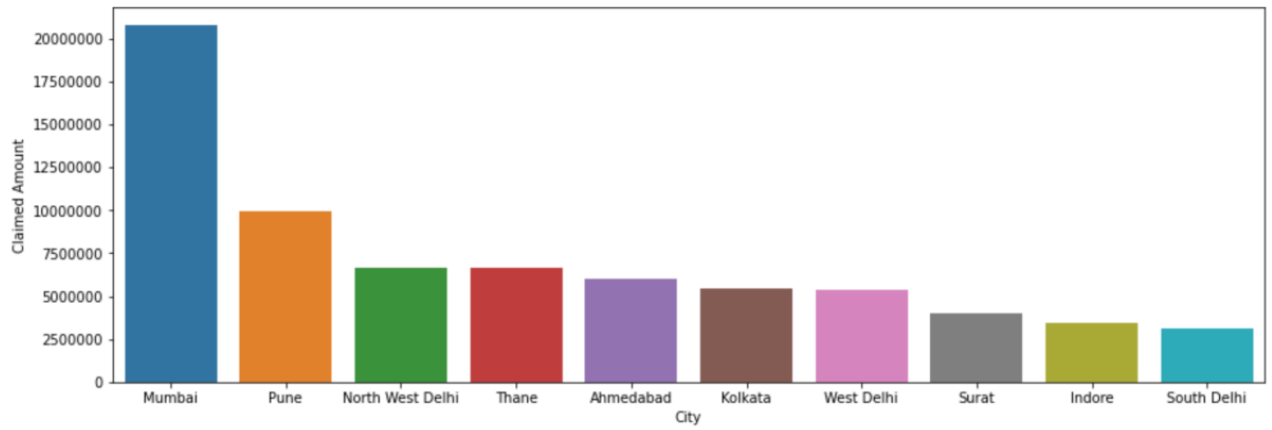
This alignment serves as a valuable reference point for assessing the validity and authenticity of claims. Any deviations from this established pattern may necessitate closer scrutiny. By delving into the analysis of claim amounts within these distinctive brackets, we gain invaluable insights into the prevalent patterns and tendencies in claims. This, in turn, equips us to craft more precise rule-based fraud detection strategies tailored to the specific characteristics of these brackets.



*Figure 3-22 Claim Distribution at City level*

Figure 3-24 provides an insight into the distribution of claims across different cities, highlighting the top 10 cities based on the sheer number of claims recorded. This visualization serves as a valuable reference point to identify regions where claims are most concentrated, offering a clear perspective on the geographical pattern of claim occurrences. Such a presentation aids in understanding the potential variations in fraud occurrences across different urban centres and guides the formulation of focused fraud detection strategies tailored to specific geographical contexts. From the figure Mumbai and Pune are the top two cities in

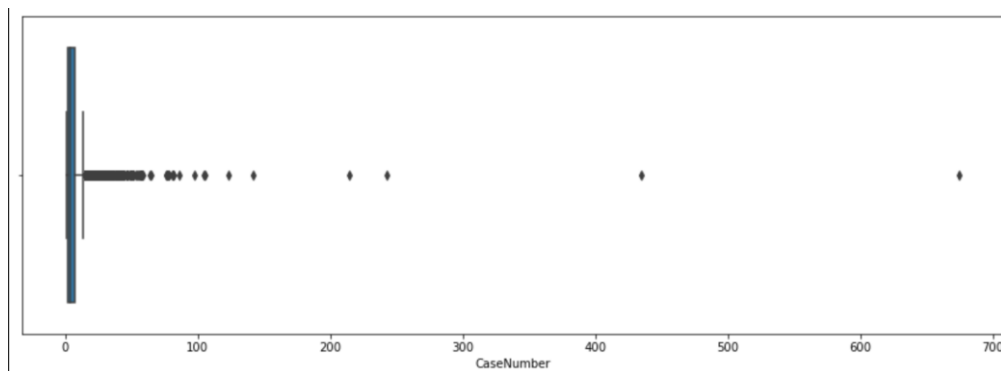
terms of number of claims with more than 12K claims coming from each of these cities.



*Figure 3-23 Claim Amount Distribution at City level*

As demonstrated in Figure 3-24, we are provided with an overview of the cities that have generated the highest count of claims. Now, shifting our focus to Figure 3-25, it offers an alternative perspective by highlighting the top cities based on the total claimed amount.

Through the analysis of claimed amounts categorized by city, we gain insights into the economic repercussions of these claims. This sheds light on regions where anomalies and potentially fraudulent activities may be more prevalent. Visualization serves as a valuable tool in shaping more efficient strategies for assessing customer risk, particularly in areas where the claimed amounts deviate significantly from the norm.



*Figure 3-24 Claims per Person*

The visual representation above is indicative of a common trend among the general population, with an average of approximately 20 claims per individual. However, the presence of numerous data points lying outside this norm underscores the existence of outliers within the dataset.



These outliers, characterized by significantly higher claim counts per individual, stand out as potential anomalies that warrant further investigation. Such deviations from the expected pattern could potentially point to instances of suspicious activity or irregularities that require closer scrutiny. This observation emphasizes the significance of outlier detection within the context of rule-based abuse detection. By identifying and addressing these outliers, we can enhance the accuracy and effectiveness of our abuse detection methodologies, thereby contributing to more reliable and targeted identification of potentially fraudulent behavior.

Visualization helps identify outliers, class imbalances, seasonality, and other patterns that might not be evident in raw data. Throughout this section we further dug deep into various data points of Claims, Account and Policy object by the medium of visualization. These graphs made it easier for us to understand the concentration of data points for specific fields, correlation between different fields and outliers. Visualization such as Box plot made it really help to spot out the outliers in our dataset.

To summarize, we conducted an Exploratory Data Analysis (EDA) to delve into the various features within our datasets, with the objective of unearthing correlations and insights that could be instrumental in shaping our rule-based customer risk scoring model.

Additionally, we shed light on the prominent geographical regions in terms of both volume and the total claimed amount, with Mumbai and Pune emerging as the top two cities of significance. Another facet of our analysis involved examining the relationship between a customer and the number of claims they have filed, uncovering that, on average, a user tends to raise around 20 claims.

Furthermore, we classified customers into different severity levels by considering a combination of the number of claims and the total amount utilized, ultimately yielding a list of customers with the highest severity. We rounded off our EDA section by comprehending the role of sales agents in boosting the number of policies initiated at the start of each month.

Now in our next section, we will understand more about the outliers we found in our EDA and what are the different techniques we can adopt to deal with them.

### 3.7 Outlier Detection

The process of detecting outliers holds a pivotal role in data preprocessing, forming an indispensable component of various analyses (abuse detection). Outliers are data points that exhibit substantial deviations from the norm, and their presence can exert a substantial influence on the model's effectiveness and precision. Thus, the identification and effective management of outliers assume critical importance in guaranteeing that the model remains impervious to noise or erroneous data. This, in turn, facilitates the model in making more precise predictions and achieving enhanced accuracy.

Here is an elaboration of outlier detection and the steps involved:

**1. Define Outliers:** The first step in outlier detection is to define what constitutes an outlier for the specific dataset and problem. Domain knowledge or business rules can be used to define outliers based on the context of the problem.

*Table 3-14 Claim Approved Amount range*

CaseNumber	Approved_Amount_c
02530706	0.00
02444366	1.00
02381254	1.00
M02686490	1.00
A02463500	1.00
S	...
K	...
E02548789	40000.00
D02435159	40000.00
02357882	48080.00
02615697	57250.00
02496958	57670.00

Table 3-14 highlights the outliers for Amount claimed by customers. Within the dataset,

noteworthy outliers are observed on both ends of the spectrum: one claim displays an unusually low value, hovering around 20 INR, while another claim conspicuously stands out with an exceptionally high value, reaching 25,000 INR. These extreme values, often characterized as outliers, wield substantial influence over the accuracy of our risk scoring model analysis.

The abnormally low-value claim may stem from data entry errors or other anomalous circumstances, whereas the exceedingly high-value claim raises concerns of potential abuse. Recognizing and effectively dealing with these outliers is of paramount importance to preserve the integrity and dependability of our rule-based customer risk scoring model, ensuring it operates at its utmost precision and reliability.

Upon scrutinizing the claims with lower amounts, we detected the presence of dummy phone numbers and names, which were evidently employed for testing and experimental purposes. However, these entries persisted within our algorithm's foundational dataset as legitimate transactions.

**2. Univariate vs. Multivariate Outlier Detection:** Outliers can be detected either in a single feature (univariate) or in multiple features simultaneously (multivariate). Univariate outlier detection is based on the distribution of each individual feature, while multivariate outlier detection considers the relationship between multiple features.

*Table 3-15 Customer-Policy Relationship*

Number of Policies on Single Mobile	Unique Mobiles
0	13
1	20312
2	5624
3	2103
4	1023
5	505
6	287
7	170
8	131
9	74
10	56
11	30
12	15
13	13
14	10
15	4
16	3
17	5
18	3
20	1
21	1
22	1
23	3
46	1
174	1

During our analysis, a noteworthy observation became known: a considerable number of customers were found to be holding multiple policies, as depicted in the Table 3-16. This phenomenon had a discernible effect on our overall claim portfolio per mobile, causing it to deviate from the anticipated trend. It is important to note that, in general practice, a single customer should not possess more than two policies. This finding underscores the significance of examining and addressing the implications of customers holding multiple policies, as it can significantly impact our risk assessment and customer risk scoring processes, a critical aspect of our insurance operations. Further investigation and potential policy adjustments may be warranted to ensure a more accurate representation of our customer risk profile.

**3. Common Outlier Detection Techniques:** There are various statistical techniques to identify outliers:

- Z-Score or Standard Deviation: Data points that fall beyond a certain threshold of z-scores are typically identified as outliers.
- Interquartile Range (IQR): Data points beyond the upper and lower bounds defined by the IQR are treated as outliers.

Among the array of methods at our disposal, one of the most effective techniques employed was percentile analysis, which we harnessed for our dataset. Percentile analysis serves as a powerful statistical tool, instrumental in comprehending the distribution of a dataset and pinpointing precise values that encapsulate specific percentages of the data. This approach excels specifically when dealing with large datasets, as it plays a pivotal role in synthesizing the data into a concise and meaningful summary, enabling a deeper understanding of the underlying patterns and trends.

*Table 3-16 Claims Amount Percentile Distribution-1*

Percentile	Approved_Amount_c
0.10	250
0.17	300
0.23	400
0.30	500
0.37	500
0.43	600
0.50	700
0.57	800
0.63	1000
0.70	1200
0.77	1500
0.83	2000

*Table 3-17 Claims Amount Percentile Distribution-2*

Percentile	Approved_Amount_c
0.90	3000
0.91	3300
0.92	3500
0.93	3890
0.93	4105
0.94	4500
0.95	5000
0.96	5490
0.97	6000
0.98	7140
0.98	8695
0.99	11730

As visible in Table 3-16 and 3-17 above we can evidently see that the amount claimed feature values were linear till 90 percentiles but after which it exponentially gets on increasing with even 1 percentile from the 90 to 99 percentiles.

The percentile analysis highlights a notable challenge within the dataset, revealing a substantial disparity in the right quantiles of the data distribution. This significant gap poses a considerable

hurdle when attempting to normalize the data effectively. Normalization is essential for ensuring that the data conforms to a standardized scale, aiding in accurate comparisons and analysis. However, the pronounced disparity in the right quantiles complicates this process, potentially influencing the performance of our rule-based risk model. Addressing this challenge requires careful consideration and adaptation to ensure the reliability of our analytical outcomes.

In percentile analysis, data is sorted in ascending order, and specific percentiles are calculated based on the position of data values. For instance, the 25th percentile (first quartile) denotes the value below which 25% of the data falls. Likewise, the 50th percentile (median) represents the value below which 50% of the data falls, while the 75th percentile (third quartile) signifies the value below which 75% of the data falls.

Percentile analysis serves several purposes, including:

- **Identifying Outliers:** Percentile analysis helps pinpoint potential outliers in the dataset—data points that deviate significantly from the majority of the data.
- **Understanding Data Distribution:** Percentiles aid in understanding the spread of data and whether it skews towards higher or lower values.
- **Comparing Data Sets:** Comparing percentiles of different datasets offers insights into their distributions and facilitates comparisons.
- **Calculating Summary Statistics:** Percentiles assist in calculating summary statistics such as quartiles, median, and interquartile range, which offer a robust understanding of data central tendency and spread.
- **Decision Making:** Percentiles are useful for setting benchmarks or thresholds in various decision-making processes.

In percentile analysis, common percentiles include the 25th, 50th (median), and 75th percentiles, but other percentiles like the 10th, 90th, and 95th percentiles can also be used based on the specific requirements of the analysis. It is a valuable tool for data scientists and analysts to acquire more knowledge into the data and make appropriate decisions.

**4. Handle Outliers:** Once the outliers are identified, there are several ways to handle them:

- Remove Outliers: The simplest approach is to remove the outlier data points from the dataset. However, this should be done with caution as it may lead to information loss, especially if the outliers are legitimate data points.
- Imputation: Outliers can be replaced with meaningful values, such as the mean, median, or imputed values based on other data points.

*Table 3-18 Customer Age correction using mean values*

*Before*

Description	Age_c
count	239767.00
mean	44.12
std	20.82
min	0.00
0.25	33.00
0.5	39.00
0.75	49.00
max	124.00

*After*

Description	Age_c
count	239767.00
mean	40.47
std	11.39
min	0.00
0.25	33.00
0.5	39.00
0.75	46.00
max	80.00

To mitigate the impact of outliers and extreme values in the age data as seen in Table 3-18, a strategic approach was taken. Instances where the recorded age exceeded 80 years were identified as potential outliers and were replaced with the mean age of the entire population, which was approximately 50 years. This preprocessing step aimed to address potential anomalies that could disrupt the accuracy of our analysis and subsequent rule-based fraud detection model. By substituting extreme values with a more representative and plausible age, we aimed to enhance the overall reliability of our data and the subsequent results.

**Identifying Anomalies:** During exploration and visualization, we discovered anomalies or

unexpected patterns in the data. These anomalies could indicate potentially fraudulent activities.

*Table 3-19 Customers with Highest Severity*

PersonMobilePhone	CaseNumber	Approved_Amount_c	Severity
974330xxxx	4	69600	17400
844692xxxx	4	64550	16138
770907xxxx	4	63336	15834
858497xxxx	4	60000	15000
626778xxxx	5	72275	14455
770382xxxx	7	98100	14014
995554xxxx	6	84000	14000
783852xxxx	5	67850	13570
906188xxxx	4	54000	13500
999245xxxx	4	53818	13455

Masked mobile numbers are shown only for illustration purpose

The provided Table 3-19 snapshot gives a glimpse into the segment of customers exhibiting the highest severity levels within the dataset. We can see that the masked mobile number (974330xxxx) has a severity of 17,400 INR which indicates that the customer's Average Ticket Size of reimbursement claim is 17,400 INR which will help us to prioritize and scrutinize the suspects accordingly. These severity levels indicate the extent of potential risk or impact associated with these customers' activities or behaviors. By focusing on this subset of customers with elevated severity, our research aims to delve into the patterns, attributes, and characteristics that contribute to their heightened risk profiles.

Through comprehensive analysis and rule-based techniques, we endeavor to uncover hidden patterns, anomalies, and potential indicators of fraudulent behavior within this group. By understanding the unique traits that set these high-severity customers apart, we can enhance the effectiveness of our risk scoring strategies, leading to more precise identification and mitigation of fraud risks. This investigation forms a crucial part of our broader efforts to strengthen abuse detection mechanisms within the financial landscape and contributes to the overall security and stability of the system.

Feature importance and time series analysis can both be effectively used for outlier detection. By determining which features are most influential in predicting a target outcome, we can focus on these key predictors to identify outliers. Time series analysis, on the other hand, involves examining data points collected or recorded at specific time intervals. This analysis can highlight trends, seasonal patterns, and cyclical behaviors within the data. Outliers in time



series data are points that significantly deviate from these identified patterns.

**Feature Importance:** Understanding the importance of distinctive features in predicting fraud is crucial. Data scientists analyze the correlation between features and the target variable (fraudulent vs. non-fraudulent) to identify the most informative features. Feature importance helps in selecting relevant features and rules for our rule-based model.

**Time Series Analysis:** If the data involves temporal information, time series analysis can reveal trends and seasonality in the data. Seasonal patterns may help identify fraudulent activities that follow certain time-based trends. Exploring and visualizing the data helps gain valuable insights into patterns and trends that aid in identifying potentially fraudulent activity. Continuous evaluation and adjustment of the rule-based model ensure its effectiveness and adaptability to evolving fraud patterns and data changes. Regular monitoring and updates are essential to maintain a robust and reliable rule-based scoring model. (Fast et al.)

Transitioning from identifying anomalies, we now move to Hypothesis Testing, a fundamental aspect of statistical analysis. Hypothesis testing allows us to make inferences about a population based on sample data, providing a structured methodology for testing assumptions and determining the validity of claims. In the following section, we will explore various hypothesis testing techniques and how they can be used to derive meaningful insights from your data.

### **3.8 Hypothesis Testing**

Due to the limitations of labeled data (claims/entities) specifically marked as Abuse, traditional statistical techniques for data driven hypothesis testing are not suitable for this analysis. Labeled data serves as the foundation for these methods, and its absence hinders our ability to definitively prove or disprove hypotheses about abusive behavior basis statistical testing methods.

To address this challenge, we will leverage two key industry indicators: claim frequency and claim severity. Claim frequency refers to the number of claims filed by a customer within a specific timeframe, while claim severity represents the average cost associated with those claims. By analyzing these metrics, we can identify potential red flags associated with abusive

behavior.

Our approach will utilize outlier analysis, specifically percentile analysis, to establish thresholds for our abuse detection rules. Percentile analysis helps us identify data points that are beyond the expected range of claim frequency and severity. Customers with a significantly higher number of claims or claims with a much higher average cost compared to the rest of the population may warrant further investigation as potential abusers.

Below we present our hypothesis across different parameters where H0 represents Null Hypothesis and H1 represents Alternative Hypothesis.

**Age:**

H0: The age of customers has no correlation with severity of claims.

H1: The age of customers has correlation with severity of claims.

*Table 3-20 Age characteristics*

Age_c	Masked Sensitive Column Names				er	Frequency	Severity
(-0.001, 29.0]	28088	43177270	7755	7103		3.62	1537.21
(29.0, 32.0]	25213	35024322	6071	5281		4.15	1389.14
(32.0, 34.0]	21954	28524756	5022	4280		4.37	1299.30
(34.0, 36.0]	19850	25865974	4633	4029		4.28	1303.07
(36.0, 39.0]	26352	33742836	6076	5321		4.34	1280.47
(39.0, 41.0]	15141	19987069	3520	3053		4.30	1320.06
(41.0, 46.0]	26375	35844654	6148	5401		4.29	1359.04
(46.0, 52.0]	19857	28124075	5072	4585		3.92	1416.33
(52.0, 62.0]	22346	32885495	5963	5491		3.75	1471.65
(62.0, 124.0]	22066	32712197	5617	5500		3.93	1482.47

The analysis referring to Table 3-20, indicates that children and young adults (aged 0 to 30) have the highest severity scores in relation to potential abuse. This suggests that this age group might be more vulnerable to, or experience more severe forms of abuse compared to other age groups.

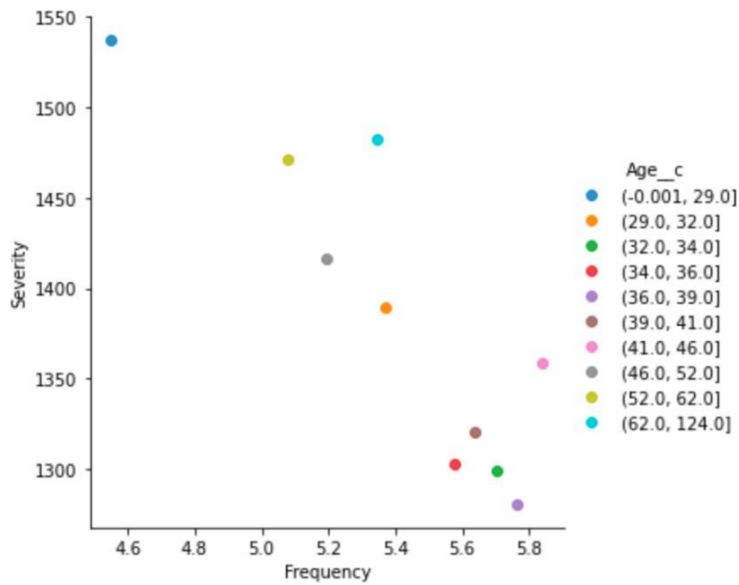


Figure 3-25 Relationship between Frequency and Severity

The scatter plot displayed above shows the relationship between Frequency (x-axis) and Severity (y-axis), with different colors representing various age categories (Age). Certain age categories exhibit distinct behavior: Younger age groups (e.g., (-0.001, 29.0], (29.0, 32.0]) tend to have higher Severity values while older age groups (e.g., (52.0, 62.0], (62.0, 124.0]) show lower Severity values and higher Frequency.

**Gender:**

H0: There is no significant correlation of gender with the severity or frequency of claims.

H1: There is significant correlation of gender with the severity or frequency of claims.

Table 3-21 Gender characteristics

Gender_c	CaseNum	Masked Sensitive Column Names			Frequency	Severity
Female	53756	83840082	13977	12781	3.85	1559.64
Male	168640	225044308	39562	36044	4.26	1334.47
Other	4846	7004258	1265	1220	3.83	1445.37

Overall, both males and females show a comparable pattern in terms of claim frequency & claim severity distribution. This suggests that gender, on its own, may not be a strong predictor of who is more likely to file a claim in general, hence we reject the alternate hypothesis (H1).

**Address:**

H0: There is no significant correlation between address and high utilization behavior.

H1: There might be specific addresses or regions associated with higher utilization and probable fraudulent activities.

Table 3-22 City level characteristics

City_c	CaseNur	Masked Sensitive Column Names			er	Frequency	Severity
Amreli	5	12100	1	1	5.00	2420.00	
Amroha	36	121292	8	8	4.50	3369.22	
Balaghat	11	25460	2	2	5.50	2314.55	
Begusarai	58	111251	10	14	5.80	1918.12	
Bemetra	5	10386	1	1	5.00	2077.20	
Bharatpur	40	88529	9	9	4.44	2213.23	
Chamoli	5	9892	1	1	5.00	1978.40	
Damoh	19	44500	2	2	9.50	2342.11	
Dhubri	18	37458	4	5	4.50	2081.00	
EAST SIKKIM	6	16640	1	1	6.00	2773.33	
Giridih	39	136546	8	12	4.88	3501.18	
INDORE	13	38250	3	3	4.33	2942.31	
Kandhamal	7	14330	1	1	7.00	2047.14	
Lakhisarai	6	27750	1	2	6.00	4625.00	
Nayagarh	15	36590	3	3	5.00	2439.33	
Rajgarh	9	23450	2	1	4.50	2605.56	
Rajsamand	10	18630	2	2	5.00	1863.00	
THANE	23	45228	3	4	7.67	1966.43	
VISAKHAPATNAN	6	14900	1	1	6.00	2483.33	

The analysis of percentiles in Table 3-22 suggests that claims from specific locations exhibit higher severity compared to claims from other locations. This makes these locations anomalous in terms of claim severity. The number of approved cases per city varies greatly, from 3 cases in Amroha to 58 cases in Begusarai.

Table 3-23 Pin code characteristics

Pincode_c	Masked Sensitive Column Names				Frequency	Severity
127111	7	34660	1	1	7.00	4951.43
301022	16	60585	2	2	8.00	3786.56
484669	11	32616	1	1	11.00	2965.09
492014	8	43230	1	1	8.00	5403.75
520001	7	40070	1	1	7.00	5724.29
521241	8	25299	1	2	8.00	3162.38
522614	17	50866	1	2	17.00	2992.12
600049	46	147024	4	5	11.50	3196.17
721135	8	32552	1	1	8.00	4069.00
753010	13	36200	2	1	6.50	2784.62
802158	16	54214	2	2	8.00	3388.38
815317	14	65560	1	3	14.00	4682.86
845424	7	21650	1	1	7.00	3092.86

From the above table it is pretty evident that pin code “520001” has the highest severity of 5724.29 followed by “492014” with a severity of 5403.75. The pin code with least severity is “753010” with severity of 2784.62 with a frequency of 6.5.

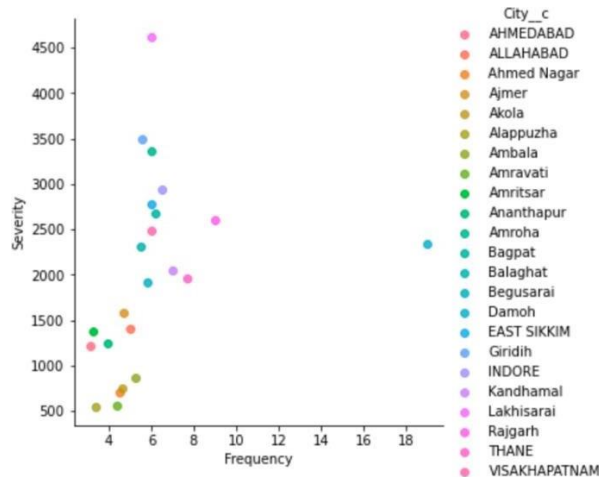


Figure 3-26 Frequency vs Severity at City level

The above scatter plot further strengthens the observation, we can clearly see the outliers points here with “Lakhisarai” having highest severity followed by “Giridh” and “Amorh”. We also see one outlier point “Damoh” which has less severity but high frequency of claims.

However, since the number of sample of cases were significantly smaller for each location, it is not statistically enough to accept the alternate hypothesis (H1), hence though we assume that there could be some correlation between location and high consumptions of OPD health insurance benefits, however they are not statistically significant to be considered in the final ruleset.

**Mobile:**

H0: There is no significant connection between mobile numbers and high utilization.

H1: Certain mobile numbers (series of mobile numbers) may be associated with a higher likelihood of being involved in high utilization or potential abuse.

Table 3-24 Customer Mobile characteristics

Masked Sensitive Column Names			SerialNumber	Severity
1	727	1641280	242	2257.61
5	144	316315	40	2196.63
6	1842	2553883	446	1386.47
7	28106	39480772	6134	1404.71
8	43219	58855900	9377	1361.81
9	153204	213040498	33793	1390.57

The information in the image reveals an anomaly in Table 3-24, highlighting a significant deviation in the distribution of mobile phone numbers starting with 1 - 5. This anomaly is primarily attributed to the use of dummy or test mobile numbers assigned to policies which were internally used testing purpose.

Since in India a valid mobile number cannot start with 1 or 5, it can be discarded as a dummy observation. Rest of the mobile series have a reasonably similar severity. The same can also be viewed via a scatter plot below which plots the severity of cases basis the mobile number series.

Although there could be some correlation between mobile series and high consumptions of OPD health insurance benefits, however basis the observations they are not statistically significant to be considered in the final ruleset and hence we reject the alternate hypothesis (H1)

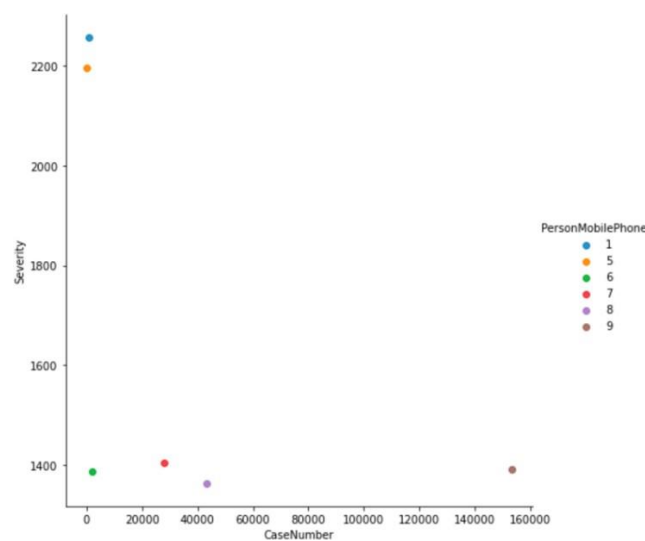


Figure 3-27 Phone number series with claims severity

**Payment Transfers:**

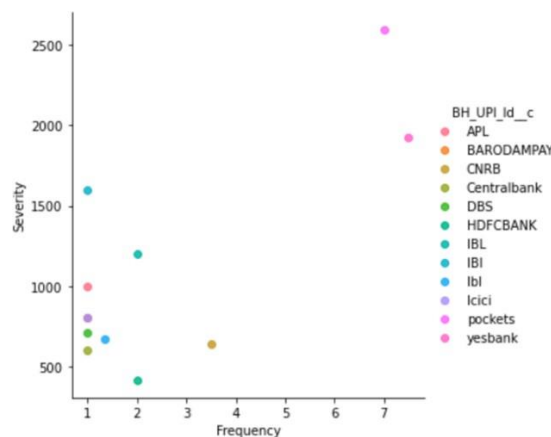
H0: There is no significant relationship between UPI payment handles and higher consumptions.

H1: Specific types of payment transfers (UPIs) could be linked to an elevated probability of utilization behavior.

*Table 3-25 Payment characteristics*

Masked Sensitive Column Names					Frequency	Severity
APL	1	1000	1	1	1.00	1000.00
BARODAMPAY	1	800	1	1	1.00	800.00
CNRB	7	4500	2	2	3.50	642.86
Centralbank	1	600	1	1	1.00	600.00
DBS	1	711	1	1	1.00	711.00
HDFCBANK	2	830	1	2	2.00	415.00
IBL	2	2400	1	1	2.00	1200.00
IBI	1	1600	1	1	1.00	1600.00
Ibl	8	5350	6	6	1.33	668.75
Icici	2	1600	2	2	1.00	800.00
pockets	7	18150	1	2	7.00	2592.86
yesbank	45	86525	6	8	7.50	1922.78

Key findings from Table 3-25 reveal a significant anomaly in the Unified Payments Interface (UPI) transactions. The analysis indicates that specific UPI extensions are associated with a notably higher frequency of fraudulent claims. This discovery suggests that transactions involving these UPI extensions require increased scrutiny and the implementation of enhanced fraud detection measures to mitigate the risk of fraudulent activities.



*Figure 3-28 Frequency vs Severity at UPI level*

Through the above scatter plot, we notice how “pockets” and “yesbank” have high severity and high frequency as well. Transactions with these UPI extensions need to be investigated thoroughly with the help of the investigation team.

However, since the number of sample of cases were significantly smaller for each UPI handle, it is not statistically enough to accept the alternate hypothesis (H1), hence though we assume that there could be some correlation between handles and high consumptions of OPD health insurance benefits, however they are not statistically significant to be considered in the final ruleset.

**Product Family:**

H0: There is no significant connection between product family and high utilization behavior.

H1: Certain product families might have a higher propensity for abusive activities.

*Table 3-26 Product level characteristics*

Masked Product Names	Masked Sensitive Column Names				Frequency	Severity
		6	9936	3	3	2.00
	664	810136	158	180	4.20	1220.08
	88	79973	31	31	2.84	908.78
	242	428928	47	47	5.15	1772.43
	1	100	1	1	1.00	100.00
	1	1500	1	1	1.00	1500.00
	10	16480	2	2	5.00	1648.00
	2864	3406647	704	815	4.07	1189.47
	7	6836	2	2	3.50	976.57
	257	211903	120	134	2.14	824.53
	609	1848450	49	49	12.43	3035.22
	777	2005685	106	144	7.33	2581.32
	13	34158	2	2	6.50	2627.54
	38	75120	6	13	6.33	1976.84

The 85th percentile analysis in Table 3-26 reveals that specific product families demonstrate an unusually high frequency of claims compared to other product categories. This disproportionate share of high-frequency claims strongly suggests the presence of potential fraud clusters within those specific product families.



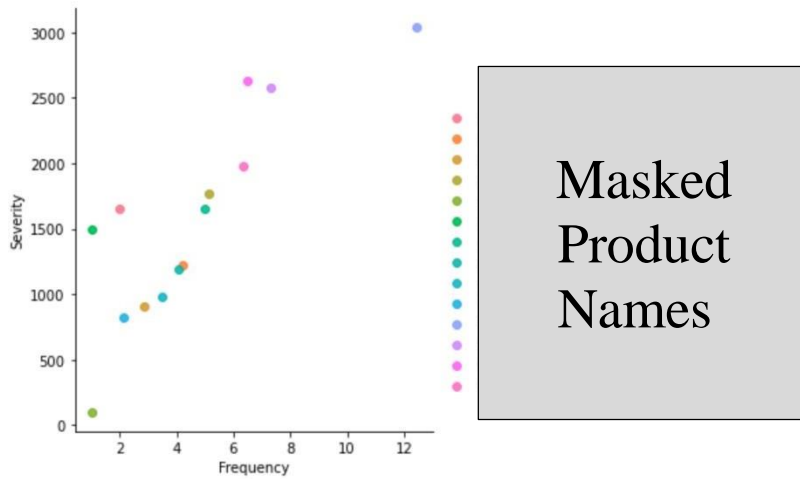


Figure 3-29 Frequency vs Severity at Product level

Concerns can be raised with respect to following products as they have high severity and high frequency- Masked Product Names and Masked Product Names

On deeper analysis, we realized that employees of Masked Product Names who have filed the claims are residing in the metro cities where the cost of OPD consultation or lab tests are generally 25 – 30% higher than the average severity, hence though we assume that there could be some correlation between product family handles and high consumptions of OPD health insurance benefits, however basis the discovered nuances we do not accept the alternate hypothesis (H1)

**Purchase Date:**

H0: There is no significant correlation between purchase date and utilization behavior.

H1: Certain purchase date ranges (month starting dates / specific months) might be associated with a higher likelihood of abusive activities.

Table 3-27 Claims characteristics basis policy start date

Masked Sensitive Column Names					Frequency	Severity
1	54277	93686849	7165	8680	7.58	1726.09
2	13830	18123520	2655	3024	5.21	1310.45
3	20675	26694978	4098	4913	5.05	1291.17
4	33413	40260927	5682	6592	5.88	1204.95
5	16062	19718198	3789	5303	4.24	1227.63
6	13859	17323910	3344	4352	4.14	1250.01
7	13364	18812275	2716	3301	4.92	1407.68
8	14718	20629631	2776	3242	5.30	1401.66
9	18890	23802407	3706	4074	5.10	1260.05
10	10795	15116492	2143	2476	5.04	1400.32
11	7513	9510476	1556	1734	4.83	1265.87
12	9846	12208985	2069	2338	4.76	1239.99

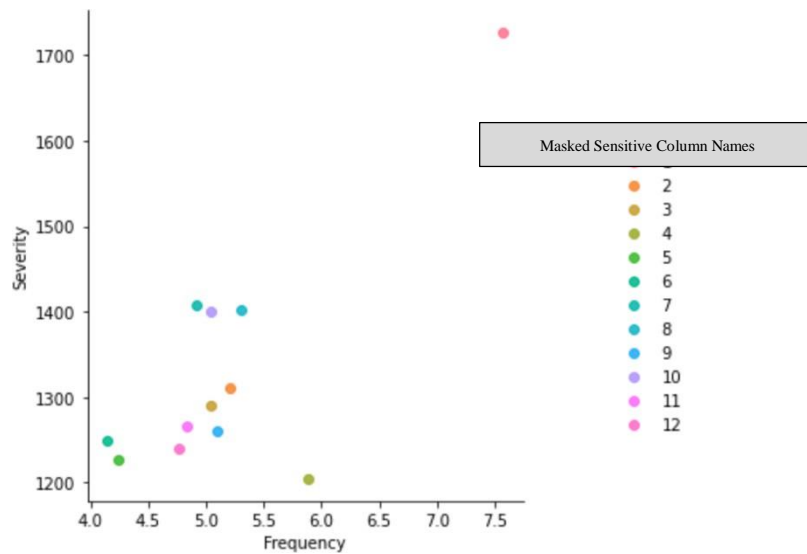


Figure 3-30 Frequency vs Severity with Policy effective date

From Table 3-27 and figure 3-32, we observe that “January” month has the highest severity and highest approved amount of Rs. 9,36,86,849.

However, on further discussion with product owners and operations team we realized that most employees usually file their reimbursements of all their OPD consultations during the first month of their corporate policy period which is January. Basis this finding, we can reject the alternate hypothesis (H1), because consumption pattern for the rest of the months are reasonably like each other.

**Expiry Date:**

H0: There is no significant connection between expiry date and high severity behavior.

H1: Customers with high severity will utilize their whole remaining wallet amount before expiry of the product.

Table 3-28 Claims characteristics basis policy end date

Masked Sensitive Column Names					Frequency	Severity
1	16757	21484796	3474	4012	4.82	1282.14
2	15437	19579579	2953	3342	5.23	1268.35
3	35490	43411458	5589	6420	6.35	1223.20
4	13678	18570634	3368	3894	4.06	1357.70
5	18367	23384825	4217	6249	4.36	1273.20
6	13810	18973609	3110	3862	4.44	1373.90
7	15962	24490641	3096	3671	5.16	1534.31
8	24417	34907286	4490	5138	5.44	1429.63
9	13368	17410021	2681	2963	4.99	1302.37
10	11303	15824079	2207	2514	5.12	1399.99
11	8450	11529649	1694	1878	4.99	1364.46
12	40203	66322071	5279	6086	7.62	1649.68

The percentile analysis in Table 3-28, focusing on policy purchase dates and claim frequencies, reveals that policies bought early in the year (presumably January) exhibit a higher frequency of claims compared to policies purchased throughout the rest of the year.

Anticipated or Changing Health Needs: Individuals who anticipate requiring medical care in the coming year, perhaps due to known conditions, scheduled procedures, or seasonal health concerns (e.g., allergies or respiratory illnesses in specific months), might be more likely to purchase insurance early to have coverage in place when needed. This can lead to a higher claim frequency for policies bought at the year's beginning.

Strategic Claim Filing (Cautionary Note): In some cases, individuals might strategically purchase insurance early, knowing they have a pre-existing condition, and then file claims soon after to maximize coverage benefits. However, it is essential to exercise caution with this explanation, as it can have legal and ethical ramifications depending on the specific context and insurance regulations. It is crucial to avoid generalizations or assumptions without thorough investigation and adherence to ethical research.

**Source & Channel:**

H0: There is no significant connection between source & channel and abusive behavior.

H1: Specific sources & channels (having high incidence) might be linked to higher instances of fraudulent activities.

Table 3-29 Claims characteristics basis source and channel of purchase

Masked Channel Names	Masked Sensitive Column Names				Frequency	Severity
		1	700	1	1	1.00
	667	810157	159	181.00	4.19	1214.63
	58	73623	19	21.00	3.05	1269.36
	50	66113	19	19	2.63	1322.26
	3	3400	1	1	3.00	1133.33
	2	1600	1	1	2.00	800.00
	28439	34697428	6034	7127	4.71	1220.06
	28	29219	7	7	4.00	1043.54
	29	33985	9	9	3.22	1171.90
	287	366203	170	170	1.69	1275.97
	91249	140188182	11639	15820	7.84	1536.33
	52180	74982043	8076	12336	6.46	1436.99

Through percentile analysis in Table 3-29, specific channels or sources for insurance claims were identified as having unusually high claim frequencies. This exceeding of normal thresholds suggests that these channels might be associated with higher levels of potential suspicious or fraudulent activities.

While the ideal scenario for validating our customer scoring hypotheses would involve robust training data and statistical analysis, the limited data availability in this project necessitated a unique approach. To overcome this challenge, we adopted an adjudicator agent validation methodology, manually testing each hypothesis and ruleset through expert review by dedicated customer service agents.

This claim processor’s validation involved several key steps:

- Hypothesis and Ruleset Formulation: Each hypothesis regarding influential factors in customer scores was clearly defined alongside the corresponding ruleset for identifying such factors.
- Agent Training and Familiarity: Agents were thoroughly explained on the hypotheses, rulesets, and scoring system to ensure consistent understanding and application.
- Case Review and Validation: Agents reviewed a representative sample of customer cases, manually applying the ruleset and assessing its effectiveness in accurately predicting customer score outcomes.

While this agent validation approach may not offer the same statistical rigor as data-driven methods, it provided valuable insights in the absence of sufficient training data. This human-in-the-loop approach leveraged the expertise and experience of actual claim processors and executives, allowing us to identify potential flaws in the scoring system and refine it for greater accuracy and effectiveness.

During the initial stages of hypothesis generation, several potential indicators were identified. However, upon further analysis, some of these hypotheses were excluded from the final rule set. This decision was based on two key considerations. Firstly, certain hypotheses exhibited bias within the dataset, potentially leading to inaccurate risk assessments for specific customer segments. Secondly, some hypotheses, while statistically significant, lacked practical application from a business standpoint. These hypotheses might not have translated into actionable rules that business stakeholders deemed relevant or impactful for risk management strategies. (We excluded Age, Gender, Location, Source & Channel basis mentioned reasons).

While some initially formulated hypotheses were ultimately excluded from the final rule set due to bias or lack of business relevance, the process of hypothesis testing itself proved to be valuable. Even though these hypotheses were not incorporated as official scoring rules, they can still serve a crucial purpose in our overall risk management framework. They can be used for further investigation and targeted testing in specific scenarios. This knowledge can be used to refine future data collection or adjust risk management strategies for that segment. Additionally, hypotheses deemed statistically significant but lacking immediate business application can be revisited in the future as business needs and priorities evolve. Therefore, the insights gleaned from hypothesis testing, even for excluded hypotheses, contribute to a more comprehensive understanding of customer risk, and provide valuable tools for ongoing investigation and system optimization.

It is important to acknowledge that this methodology has limitations. Agent bias and subjectivity can potentially influence the validation process. However, by employing a dedicated workshop, standardized case review procedures, and ongoing feedback loops, we mitigated these risks and ensured a consistent and reliable validation process.

In conclusion, the agent validation approach, though necessitated by data limitations, proved a valuable tool in refining our rule-based customer risk scoring system. This combination of

human expertise and hypothesis testing paved the way for a more accurate and effective scoring model, despite the challenges of a data-scarce environment.

After an exhaustive exploration of several hypotheses and feedback from the adjudication team, we have distilled our discoveries into a collection of rule sets. Each rule was subjected to a meticulous testing process individually to assess its effectiveness in pinpointing high-risk customers. Through these tests, we determined the individual importance of each rule by analyzing how well it contributed to the identification of risky customers. This allowed us to assign appropriate significance levels to each rule based on its performance in detecting potential risks. These rule sets were chosen due to their ease of explainability and proven reliability in consistently yielding favorable result outcomes and the accessibility of pertinent data.

While our initial hypothesis generated valuable insights, relying solely on those findings to establish features and thresholds for the final customer risk scoring ruleset would have been insufficient. Such an approach would have disregarded crucial business context and domain expertise.

Therefore, we deliberately incorporated extensive business stakeholder input alongside the insights derived from the hypothesis. This collaborative approach ensured that the final ruleset not only reflected the statistical findings but also aligned with real-world business needs and risk tolerance levels. By factoring in both data-driven insights and business expertise, we were able to create a more robust and comprehensive customer risk scoring system.

This strategic approach not only ensures a robust and targeted identification of potential fraudulent behavior but also leverages the strength of the available data to refine our detection mechanisms. The culmination of these rulesets signifies a significant milestone in our pursuit of enhancing abuse detection strategies and fortifying the integrity of system against illicit activities. Let us see more on the rule generation in next section.

### 3.9 Rule Generation

A comprehensive set of rules has been devised to detect and mitigate potential risks or abuses within the insurance domain. These rules are designed to scrutinize various facets of customer behavior, and their strategic thresholds are established based on industry best practices and statistical analysis.

In our thesis on rule-based customer risk scoring, the determination of thresholds for each feature was based on a comprehensive analysis of the data and consideration of various factors. Firstly, we conducted exploratory data analysis to understand the distribution and characteristics of each feature. This involved examining summary statistics, visualizing distributions through histograms, scatter plots and box plots, and identifying any outliers or unusual patterns.

Next, we leveraged domain expertise and consulted relevant literature to gain insights into the factors that are known to influence customer risk. This helped us identify potential thresholds or cutoff points for each feature based on their significance in assessing risk.

Overall, the thresholds for each feature were carefully selected based on a combination of data-driven analysis (interquartile/percentile analysis) and domain knowledge (from experts and agents), ensuring that our rule-based approach effectively identifies and mitigates customer risk.

The key rules are outlined as follows:

**Customer behavior can change over time:** Daily rules capture immediate changes, while weekly and monthly rules provide insights into trends and patterns. This combination provides a comprehensive risk assessment. The frequency of rule updates should balance capturing changes with computational efficiency. Daily updates for highly sensitive data might be excessive, while monthly updates for rapidly changing data could miss crucial changes. Certain risks require immediate action (daily rules), while others can be monitored over time (weekly/monthly). We also considered the business impact of different risk types when choosing rule frequency and other rulesets.

Percentile analysis helps identify data points that deviate significantly from the norm. By setting thresholds based on percentiles (e.g., top 1% of claim frequency), we did a quantile analysis to find out behavior of majority population and to pick thresholds for each ruleset to flag customers with abnormalities. Percentile-based thresholds also help to adjust as the overall customer behavior distribution differs over time.

Below we state the rules involved in risk-scoring and the rationale behind including these rules.

- Claims Frequency Rules:
  - More than 2 Claims in a Day
    - Rationale: Targets instances where a customer attempts an unusually high number of claims within a single day.
    - Implementation: Real-time monitoring of daily claim frequencies; triggers alert if count exceeds two claims.
  - More than 3 Claims in a Week:
    - Rationale: Identifies patterns of elevated claim frequencies over a weekly timeframe.
    - Implementation: Continuous tracking of claims within a rolling seven-day period; triggers alert for counts exceeding three claims.
  - More than 4 Claims in a Month:
    - Rationale: Captures prolonged and sustained high claim frequencies over a monthly period.
    - Implementation: Monthly evaluation of claim counts; triggers alert if count surpasses four claims.
- Claims Amount Rules:
  - More than 2000 INR in a Day:
    - Rationale: Flags instances where a customer claims a substantial amount



- within a single day.
  - Implementation: Real-time monitoring of daily claim amounts; triggers alert for amounts exceeding 2000 INR.
- More than 2499 INR in a Week:
  - Rationale: Identifies patterns of elevated claim amounts over a weekly timeframe.
  - Implementation: Continuous tracking of claim amounts within a rolling seven-day period; triggers alert for amounts exceeding 2499 INR.
- More than 5000 INR in a Month:
  - Rationale: Captures prolonged and sustained high claim amounts over a monthly period.
  - Implementation: Monthly evaluation of claim amounts; triggers alert for amounts surpassing 5000 INR.
- Product and Provider Rules:
  - More than 3 Products:
    - Rationale: Flags customers with an unusually high number of insurance (OPD) products.
    - Implementation: Continuous monitoring of the number of products associated with a customer; triggers alert if count exceeds three.
  - Multiple Same Products:
    - Rationale: Identifies instances where a customer possesses multiple identical insurance (OPD) products.
    - Implementation: Real-time analysis of product types associated with a customer; triggers alert for duplicate products.
  - More than 2 Distinct Providers in Last 30 Days:
    - Rationale: Flags customers who have interacted with an unusually high number of distinct providers within a short timeframe.
    - Implementation: Continuous monitoring of provider interactions over a

rolling 30-day period; triggers alert if count exceeds two.

- Bank Account and Linked Mobile Numbers Rules:
  - Linked to a Bank Account with More than 3 Mobile Numbers:
    - Rationale: Identifies customers whose bank account is linked to an unusually high number of mobile numbers.
    - Implementation: Real-time assessment of linked mobile numbers; triggers alert if count exceeds three.
  
- Temporal Rules:
  - Claimed Within 30-Day Period of Policy (HAN) Effective Date:
    - Rationale: Flags instances where a customer claims within a specific timeframe relative to the Policy effective date.
    - Implementation: Continuous monitoring of claim dates relative to the Policy effective date; triggers alert if a claim occurs within the specified 30-day period.

These rules collectively form the backbone of the rule-based customer risk scoring model, enabling the system to proactively identify, assess, and mitigate potential risks or abuses. The methodology involves continuous monitoring, real-time processing, and instant alerting when any of the predefined thresholds are breached. This approach ensures a comprehensive risk assessment that considers various dimensions of customer behavior, ultimately contributing to the robustness of the risk scoring model.

In the table below we present the final rules based on which we developed our customer risk score.

*Table 3-30 Customer Risk Score Rules*

Rule	Description
1	Mobile Number has claimed more than 2 claims in a day (yesterday)
2	Mobile Number has claimed more than 3 claims in last 7 days
3	Mobile Number has claimed more than 4 claims in last 30 days
4	Mobile Number has claimed more than 2000 INR in a day (yesterday)
5	Mobile Number has claimed more than 2499 INR in last 7 days
6	Mobile Number has claimed more than 5000 INR in last 30 days
7	Mobile Number has more than one active product
8	Mobile Number has multiple same products
9	Mobile Number has visited more than 2 distinct providers in last 30 days
10	Mobile Number is linked to a bank account which has more than 3 mobile numbers linked
11	Mobile Number has claimed within 30-day period of HAN effective date

In the final stage of our methodology, a holistic risk scoring mechanism was devised to quantify the overall risk associated with each customer. This involved the creation of a "Total Violation Score" column, which serves as an aggregate measure reflecting the total number of rules (Table 3-30) violated by an individual customer across the entire spectrum of our rule-based customer risk scoring model.

*Table 3-31 Example - How customer risk score is generated.*

Description	Value
PersonMobilePhone	989022xxxx
Rule1	1
Rule2	1
Rule3	1
Rule4	1
Rule5	1
Rule6	1
Rule7	1
Rule8	0
Rule9	1
Rule10	1
Rule11	1
Sum	10
Total Rules	11
Normalized Score	91

The above table shows a violation count of 10 for the listed mobile number against 11 defined rules. Consequently, the normalized score is derived using the formula:

$$\text{Count of total rules violated} / \text{Count of total rules available}$$

The process of normalization was then applied to standardize the Total Violation Score, ensuring a consistent scale for comparison. This involved dividing the Total Violation Score by the total number of rules in our methodology, which stands at 11. The resulting quotient was then multiplied by 100 to express the normalized score on a percentage scale.

The rationale behind this normalization process is to bring uniformity to the scoring system, making it more interpretable and facilitating meaningful comparisons across diverse datasets. By normalizing the Total Violation Score to a normalized scale, we create a standardized metric that ranges from 0 to 100, as depicted in Table 3-30, where higher percentages signify a greater number of rule violations and, consequently, a higher perceived risk.

### **General Mathematical Formula**

The Customer Risk Score (CRS) can be calculated using a weighted/non-weighted sum of the rules, moderated by the geographical location. The formula can be expressed as follows:

$$RS = w1R1 + w2R2 + w3R3 + \dots + wnRn$$

$$RS \text{ (Normalized)} = \frac{RS}{\text{Total Rules}}$$

$W_{i0-n}$  = weights

$R_{i0-n}$  = Rules

Note: For our use case we have given same weight to each ruleset i.e.,  $W1 = W2 = Wn = 1$

By incorporating these variables and weights into the Customer Risk Score formula, insurers can derive a quantitative measure of a customer's risk propensity. This score enables insurers to make informed decisions on policy issuance, claims processing, and fraud detection.

### 3.10 Summary

Our research embarked on its journey by establishing foundational rules centered around customer behaviors. This initial phase of rule development was built upon key metrics such as recency, frequency of claims, and the monetary trajectory of reimbursed amounts. By closely scrutinizing these fundamental aspects, we aimed to lay the groundwork for a comprehensive fraud detection framework. Our approach entailed evaluating the timing and frequency of customer claims, which serve as significant indicators of potential irregularities. Additionally, the monetary patterns of reimbursed amounts were meticulously examined to discern any deviations from established norms.

In summary, we have established eleven distinct simple rules to comprehensively assess customer risk in the context of insurance claims. These rules span various aspects of customer behavior and claim patterns. They include monitoring the frequency of claims made within different time limits, tracking claim amounts, examining the number of active products, product repetitions, provider interactions, and banking associations. By applying these rules, we have created a robust but explainable white box framework for identifying potentially risky customers, which enhances our ability to detect abuse/fraudulent or irregular claim activities and maintain the integrity of our insurance services.

Having established the key insights from our data collection and rule generation process, we can now delve into the core objective of this research - addressing the research questions that were formulated to evaluate the efficacy of rule-based customer risk scoring systems. By analyzing the data and the generated rules, we aim to answer questions stated beforehand. Through this exploration, we will gain a deeper understanding of the strengths and weaknesses of this approach, ultimately providing valuable guidance for institutions considering rule-based systems within their risk management strategies.

Following the data collection and rule generation phase, we leveraged data analysis techniques in conjunction with a deep understanding of business needs and risk management objectives. This combined approach allowed us to answer the pre-defined supporting research questions comprehensively. By analyzing the data through the lens of the business context, we were able to evaluate the effectiveness of each generated rule in identifying high-risk customers. This

iterative process ensured that the final scoring system not only addressed the research questions but also aligned with the practical requirements of our business stakeholders.

- What Data Sources Will You Use?
  - What types of data are available for analysis?
    - Secondary (Insurer's data) claims, customer profiles, policy details.
  - How reliable and up to date is the data?
    - Secondary data was already aligned with OPD insurance business use case, and we picked 4 months data basis maturity of transactions and approval from the data team.
  - What Are Fraud Indicators?
    - Top parameters which we considered basis business expertise were Incidence (Unique Utilization), Severity (Average Ticket Size) and Frequency (Claims per customer)
- What are the common indicators or red flags of fraudulent claims?
  - Are there specific behaviors or patterns associated with high-risk claims?
    - Customers from same demographics, same sourcing, similar product benefits.
  - How will we define suspicious behavior?
    - Customer having higher claim frequency and/ or severity than the average of all customers could indicate suspicious behavior.
- What behaviors or activities will trigger suspicion?
  - What thresholds or criteria will be used to identify suspicious behavior?
    - We created 11 rulesets (mentioned above) basis customer utilization parameters.
  - Are there any Historical Patterns Exist/available?
    - For initial understanding we picked one case study of fraud which happened and was manually identified by our processors / agents.

- Have there been previous instances of fraud claims that you can learn from?
  - Are there historical data patterns that can be used to identify potential fraud?
    - We incorporated transactions of customers basis daily, weekly, and monthly transactions.
  - Are There Geographic Considerations?
    - We found out a few pin codes from our hypothesis testing where utilization was abnormal.
  
- Do risk patterns vary by location (e.g., different regions, states)?
  - Will we consider geographic factors in your scoring rules?
    - We did not factor location in our rules because of non-reliability of data values because of agent based sourcing (many a times default values were passed on to create policies).
  - How will we handle anomalies?
    - We handled anomalies by using percentile analysis, which also helped us to hold essence of each feature.
  
- How will we detect and handle anomalies that do not fit typical patterns?
  - What processes will be in place to investigate and verify anomalies?
    - We proposed a solution to create separate reimbursement claims scrutiny queues for high-risk score claims for further scrutinization.
  
- How to consider a customer's claims history and behavior over time?
  - How will we factor in the customer's overall history with the organization?
    - We considered frequency and severity of claims over 30 days in our rulesets.
  - What Role Does Customer History Play?
    - It plays a very crucial role in understanding the abusive or fraudulent behavior of the customer.
  - What About External Data (if available any)?
    - OPD health insurance in India being a new category of health insurance no external reliable data set were available.

- How Will We Assign Scores?
  - What scoring system will we use in rules-based mechanisms (e.g., points-based, weighted factors)?
    - We used Boolean parametrization for scoring logic.
  
- What Is the Role of Expert Knowledge?
  - Will we involve domain experts in defining scoring rules?
    - We connected with all stakeholders for rule finalization which includes SME, product owners, customer service team and reimbursement adjudicators.
  - If yes, how will we leverage their expertise to improve rule accuracy?
    - They helped us to validate our hypothesis and identify more important features basis their expertise and investigation.
  
- How often will we update rules?
  - Will the scoring rules be static or regularly updated?
    - Scoring rules will be static for the near future but will update rulesets and will move towards weightage scoring.
  - How will we incorporate new data and adjust rules over time?
    - By regularly looking at claims and incorporating investigation remarks and feedback
  
- What Is the Appeal and Review Process?
  - How will we handle cases where customers dispute their scores / flagged claims?
    - We will ask for hard copy of the claim documents and payment proof submission
  - Is there a process for customers to provide additional information or appeal decisions?
    - Customer can always reach out to the 24 X 7 Support team
  - What Are the Consequences of High Scores or flagging claims as suspicious / fraud?
    - Customer escalations and brand reputation will be at stake in case of wrong identification. This can lead to non – renewal of genuine policies.



- What actions will be taken when a claim receives a high-risk score?
  - How to balance fraud prevention with maintaining a positive customer experience?
    - Instead of directly rejecting a high-risk score claim we will create a separate queue for high-risk cases (Scrutiny Queue)
  
- What Reporting and Monitoring Will be Implemented?
  - How to track the effectiveness of scoring rules?
    - By dividing our metrics into three categories – Identification, Investigation and Proven and tracking each category separately.
  - What reporting mechanisms will be in place to identify trends and anomalies?
    - We will directly update risk score at the claim processing system's backend against each customer's policy.

## CHAPTER IV: RESULTS

### 4.1 Introduction

The fundamental problem addressed in this research pertains to the inadequacies in outpatient insurance abuse/fraud detection, necessitating a more effective data driven solution. The prevalent challenge lies in safeguarding insurers and policyholders from financial losses incurred through abusive or fraudulent activities, a concern that prompted the formulation of our investigative approach. Our approach involves the development and evaluation of a rule-based model, specifically centred around a customer risk score methodology. By focusing on discerning key characteristics of abuse or fraud in outpatient insurance and proposing a nuanced solution, we aim to fortify the industry against fraudulent practices. This research strives to offer a comprehensive and tailored approach to enhance fraud detection, contributing to the integrity and profitability of the outpatient insurance sector.

The “Results” chapter presents a comprehensive evaluation of the proposed rules-based customer risk scoring model for enhancing outpatient insurance fraud detection, as discussed in the preceding research. This section delves into the outcomes derived from an extensive analysis, encompassing the effectiveness and efficiency of the model. Furthermore, it unveils findings related to each hypothesis and objective set forth in the research, shedding light on the nuanced characteristics of outpatient insurance abuse. Additionally, a detailed case study is featured, spotlighting the tangible impact of the customer risk score-centric approach on real-world scenarios. In sum, this chapter encapsulates the culmination of our investigative efforts, providing insights that contribute to the advancement of fraud detection methodologies within the OPD insurance landscape.

## 4.2 Evaluation of Rule-Based Risk Scoring Model

This section delves into a thorough assessment of the proposed rule-based customer risk scoring model, employing a multifaceted approach to evaluation. Four distinct types of evaluation metrics – Quantitative, Qualitative, Monetary, and Investigative – serve as the cornerstone for gauging the model's efficacy. Each metric offers a unique perspective, allowing for a comprehensive understanding of the model's performance. Notably, these metrics extend beyond technical aspects, encompassing a comprehensive examination from both technical and business perspectives. The evaluation aims to provide insights into how the rule-based customer risk scoring model performs not only in terms of technical accuracy but also in addressing business metrics. This inclusive approach ensures a well-rounded understanding of the model's impact on both fraud detection efficacy and its broader implications for business operations within the outpatient insurance domain.

The subsequent exploration will scrutinize how the rule-based model aligns with these categories, shedding light on its strengths, limitations, and overall effectiveness in enhancing risk of abuse or fraud detection within the outpatient insurance landscape.

It is crucial to recognize that assessing risk scoring model presents a unique set of challenges, due to the infrequent and elusive nature of fraudulent instances. Consequently, meticulous consideration must be given to the selection of evaluation metrics, considering the precise requirements and goals of the given application. Furthermore, it is imperative to verify that the labeled data employed for evaluation accurately mirrors the authentic distribution of instances in the real-world scenario.

		IMPACT		
		Low	Medium	High
LIKELIHOOD	Likely			
	Possible			
	Unlikely			

	Low Risk
	Medium Risk
	High Risk

*Figure 4-1 Fraud Detection Matrix*

## 4.2.1 Overall Business Metrics

Table 4-1 Metric Summarization basis category

Risk Score bins	LOW RISK				MEDIUM RISK				HIGH RISK		
	(-1, 10]	(10, 20]	(20, 30]	(30, 40]	(40, 50]	(50, 60]	(60, 70]	(70, 80]	(80, 90]	(90, 100]	
IDENTIFICATION	Claims	10911	9415	12295	14826	15667	18631	40300	47127	25709	24685
	Claimed Amount	7367392	9150155	23238303	29208418	22819048	26552514	54738577	63626885	37350494	36798174
	Unique Policies	4977	3210	4137	4248	3462	3542	5260	6265	4964	5473
	Unique Mobile	4358	2596	3281	3094	2254	2143	3167	2985	1110	758
INVESTIGATION	Claims	-	-	-	-	-	-	-	-	-	-
	Claimed Amount	-	-	-	-	-	-	-	-	-	-
	Unique Policies	-	-	-	-	-	-	-	-	-	-
	Unique Mobile	-	-	-	-	-	-	-	-	-	-
INVESTIGATION %	Investigated Claims / Identified Claims	-	-	-	-	-	-	-	-	-	-
ABUSE	Claims	-	-	-	-	-	-	-	-	-	-
	Claimed Amount	-	-	-	-	-	-	-	-	-	-
	Unique Policies	-	-	-	-	-	-	-	-	-	-
	Unique Mobile	-	-	-	-	-	-	-	-	-	-
ABUSE %	Proven Claims / Investigated Claims	-	-	-	-	-	-	-	-	-	-

Table 4-1 presents a summarized view of results generated via our rule-based risk score model. It presents three main categories:

- **Identification** – Represents metrics generated directly by the output of rule-based model
  - Claims reported – Number of claims classified in each risk score band.
  - Claim amount – Total claim amount in INR. This is calculated by adding the bill amount of claims that belong to customers classified in each risk score band.
  - Unique Policies – The number of policies that belong to customers classified in each risk score band
  - Unique Mobiles – The number of unique mobiles that belong to customers classified in each risk score band
- **Investigation** – Represents metrics of investigation of rule-based flagged entities
  - Claims Investigated – Number of claims investigated in each band.
  - Investigated claim amount – Total claim amount in INR. This is calculated by adding the bill amount of all investigated claims in each band.
  - Investigated unique Policies – The number of policies investigated in each risk score band.
  - Investigated unique Mobiles – The number of unique customers investigated in each risk score band.
  - Investigation % - Number of claims investigated to Number of claims reported in each risk score band.

- **Abuse** – Represents metrics generated directly by the investigation of rule-based flagged entities
  - Claims Proven as Abuse– Number of claims proved as abusive claims in each risk score band.
  - Abuse claim amount – Total claim amount in INR. This is calculated by adding the bill amount of claims that belong to customers proved as fraudulent in each risk score band.
  - Unique Policies – The number of policies proved as abuse in each risk score band.
  - Proven unique Mobiles – The number of customers proved as abuse in each risk score band.
  - Abuse % - Number of claims proved as abuse to Number of claims investigated in each risk score band.

#### 4.2.2 Categorized Metrics

Here with further breakdown our metrics to 4 main categories which we highlighted earlier - Qualitative, Quantitative, Monetary, and Investigative:

1. **Qualitative** metrics refer to metrics not directly measurable and typically based on subjective assessments. These metrics are often used to evaluate the overall effectiveness of a fraud detection system or to gain insights into the nature of the fraud being detected.

Qualitative metrics used for evaluation of our rule-based model -

##### 1. Pipeline Breakages

Pipeline breakages refer to disruptions or failures in the data processing pipeline. These breakages can lead to incomplete data, delayed processing, or inaccurate results. By identifying and documenting pipeline breakages, organizations can:

- **Improve Reliability:** Understanding where and why breakages occur helps in designing more resilient data pipelines.
- **Root Cause Analysis:** Detailed analysis of breakages enables teams to pinpoint the underlying causes, whether they are technical issues, integration problems, or external factors.
- **Prevent Recurrences:** Implementing preventive measures based on past breakages can significantly reduce future disruptions.

## 2. Process Gaps

Process gaps are deficiencies or inefficiencies within existing workflows that hinder optimal performance. Recognizing and addressing these gaps involves:

- **Workflow Optimization:** Identifying steps in the process that are redundant or inefficient allows for streamlining operations.
- **Resource Allocation:** Understanding where gaps exist helps in reallocating resources more effectively, ensuring that critical areas are adequately supported.
- **Performance Improvement:** By closing process gaps, organizations can enhance overall performance, reduce bottlenecks, and improve service delivery.

## 3. Learnings

Documenting learnings from past experiences is vital for continuous improvement. This involves:

- **Knowledge Sharing:** Capturing insights and best practices from previous projects ensures that valuable knowledge is retained and shared across the organization.
- **Training and Development:** Learnings can be used to inform training programs, helping teams develop the skills needed to avoid past mistakes and excel in their roles.

**Strategic Planning:** Applying learnings to future planning efforts enables organizations to make informed decisions, anticipate challenges, and capitalize on opportunities.

2. **Quantitative Metrics** refers to metrics directly measurable and typically based on objective assessments. These metrics are often used to assess the performance of the rule-based model.

Quantitative metrics used for evaluation of our rule-based model –





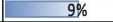
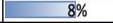
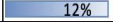



- Confusion Matrix:

		0	1
		Flagged By Agents	
0	Flagged By Rulesets	TN	FN
1		FP	TP

Figure 4-2 Confusion Matrix

- True Positives (TP): Correctly identified high risky or abusive transactions
- False Positives (FP): Incorrectly identified non-abusive transactions as high risky or abuse
- True Negatives (TN): Correctly identified non- abusive transactions
- False Negatives (FN): Incorrectly identified high risky or abusive transactions as non-risky

Table 4-2 Distribution of Incidence, Severity & population against risk score bins

Masked Sensitive Column Names								
(-1, 10]	4358		4977	10911	73,67,392	1	3	675
(10, 20]	2596		3210	9415	91,50,155	1	4	972
(20, 30]	3281		4137	12295	2,32,38,303	1	4	1890
(30, 40]	3094		4248	14826	2,92,08,418	1	5	1970
(40, 50]	2254		3462	15667	2,28,19,048	2	7	1457
(50, 60]	2143		3542	18631	2,65,52,514	2	9	1425
(60, 70]	3167		5260	40300	5,47,38,577	2	13	1358
(70, 80]	2985		6265	47127	6,36,26,885	2	16	1350
(80, 90]	1110		4964	25709	3,73,50,494	4	23	1453
(90, 100]	758		5473	24685	3,67,98,174	7	33	1491

The analysis of customer behavior patterns using the rule-based customer risk score

revealed a significant shift in key metrics. Above table highlights these changes:

- **Average Policies per Customer:** There is a notable change in the average number of policies held per customer. This suggests a potential shift in customer acquisition strategies or risk tolerance.
- **Incidence Rate:** The incidence rate, which reflects the frequency of claims, has also undergone a significant change. This could be due to various factors, such as awareness regarding the purchased products with an aim to abuse the system.
- **Severity Rate:** The severity rate, measured by the average claim amount (average ticket size), has also shown a rise with risk score. This could be attributed to abuse with evolving medical costs, changes in claim settlement processes, or shifts in the types of claims being filed.

#### High-Risk Customers:

While the high-risk segment represents a relatively small portion of the overall customer base, it is noteworthy that they contribute a high proportion of claims when compared to the total number of claims and the amount claimed. This highlights the importance of accurately identifying and managing high-risk customers to mitigate potential losses.

#### **Challenges:**

Quantifying the effectiveness of rule-based systems through metrics like confusion matrices and accuracy/precision faces inherent challenges. Firstly, data limitations can hinder accurate classification. For example, the absence of definitive evidence (e.g., lack of proof for abusive intent) prevents certain claims from being categorized precisely, leading to false negatives, and skewing overall metrics. Additionally, dynamic claim patterns and evolving fraudulent methods lead to model instability, where thresholds or rules optimized for past data perform not that good with new patterns.



- 3. Monetary metrics** refers to metrics that measure the financial impact of fraudulent activity on a business. These metrics are often used to evaluate the effectiveness of a fraud detection system or to identify areas of the business that may be at higher risk of financial losses due to fraudulent activity.

Monetary metrics used for evaluation of our rule-based model –

- **Total fraud losses:** The total amount of financial losses incurred because of fraudulent activity. This metric provides a measure of the overall impact of fraud on the business.
- **Average fraud loss per incident:** The average amount of financial loss incurred per instance of fraudulent activity. This metric provides insight into the severity of individual instances of fraud.
- **Recovery rate:** The percentage of fraudulent losses that are recovered through investigations, legal action, or other means. A high recovery rate indicates that the business is effective at recouping losses from fraudulent activity.
- **Total Amount Prevented:** This number represents the amount saved by correctly highlighting risky customers. By the amount saved, we mean blocking the customer's policy and thus preventing the remaining sum assured or wallet amount from being claimed.

**Challenges:**

Accurately allocating and tracking costs across different departments or projects to optimize resource allocation was difficult, especially with shared resources (both technical and human).

- 4. Investigation metrics** in fraud detection refer to metrics that measure the effectiveness of the investigation process that follows a potential instance of fraud. These metrics are often used to evaluate the efficiency and accuracy of the investigation process and to

identify areas for improvement.

Investigation metrics used for evaluation of our rule-based model –

- **Time to investigation completion:** The amount of time it takes to complete an investigation of a potential instance of fraud. A shorter time to completion is desirable, as it allows for faster resolution of potentially fraudulent activity.
- **Investigation accuracy rate:** The percentage of investigations that accurately identify instances of fraud. This metric provides insight into the effectiveness of the investigation process in identifying and resolving instances of fraud.
- **Investigation closure rate:** The percentage of investigations that are successfully resolved with a determination of whether fraud occurred. This metric provides insight into the efficiency of the investigation process and the ability to bring potential instances of fraud to closure.
- **Investigation cost:** The total cost incurred by the business in conducting investigations of potential instances of fraud. This metric provides insight into the resources required to identify and prevent fraudulent activity.

### **Challenges:**

#### **1. Turnaround Time (TAT):**

- **Impact:** Slow investigations lead to delayed claim decisions, impacting customer satisfaction and potentially increasing costs due to extended claim cycles.
- **Causes:**
  - **Manual processes:** Manual review of documents, data, and evidence is time-consuming and labour-intensive.
  - **Complex workflows:** Convoluted investigation processes with multiple steps and handoffs created bottlenecks.
  - **Data silos:** Information relevant to investigations was scattered across different systems, making it difficult to access and analyse.

## **2. Resource Constraints:**

- **Impact:** Lack of sufficient investigators, forensic analysts, or other resources leads to backlogs and delays in investigations, impacting overall efficiency and effectiveness.
- **Causes:**
  - **Lack of specialized skills:** Investigators might lack the expertise or training needed to handle complex fraud cases efficiently.

## **4.3 Findings related to each hypothesis and research question.**

### **4.3.1 Research Question One**

**Objective:** To identify the key characteristics of outpatient insurance fraud and the challenges associated with detecting it.

**Related Research Question:** What key characteristics define outpatient insurance fraud, encompassing traits like falsified claims and collusion between policyholders and providers?

#### **Findings:**

Outpatient insurance fraud, within the context of our rule-based customer risk scoring model, is characterized by distinct patterns and behaviors that aim to manipulate the insurance system for illicit gains. Falsified claims emerge as a prominent trait, where policyholders intentionally submit inaccurate or misleading information to secure undeserved financial benefits. This may include exaggerating medical expenses, fabricating treatment details, or misrepresenting the severity of a medical condition. Collusion between policyholders and healthcare providers is another key characteristic, signifying a coordinated effort to exploit the insurance framework. Instances where policyholders conspire with healthcare providers to generate false claims or inflate medical services for mutual financial gain fall under this category. The identified rules, such as those monitoring claim frequencies, amounts, and interactions with multiple providers, serve as crucial indicators in uncovering these deceptive practices. By discerning these key

characteristics, our rule-based approach enhances the ability to identify and mitigate outpatient insurance fraud, contributing to a more resilient and fraud-resistant insurance ecosystem.

#### **4.3.2 Research Question Two**

**Objective:** To develop a comprehensive set of rules for customer risk scoring to detect outpatient insurance fraud.

**Related Research Question:** How well does a rules-based approach identify outpatient insurance fraud, utilizing historical data and expert insights?

#### **Findings:**

A rules-based customer risk scoring approach proves highly effective in identifying outpatient insurance fraud, leveraging both historical data and expert insights. Historical data serves as a valuable foundation, allowing the model to discern and learn from patterns associated with fraudulent behaviour over time. By analysing past instances of fraud, the rules-based system can establish key criteria and thresholds that indicate anomalous or suspicious activities.

The incorporation of expert insights further enhances the model's efficacy. Domain experts bring nuanced knowledge of the healthcare and insurance industry, enabling the identification of subtle indicators of fraud that might not be immediately apparent in the data. Their expertise contributes to the formulation of rules that encompass various dimensions of customer behaviour, from claim frequencies to the nature of interactions with healthcare providers.

The continuous feedback loop between historical data and expert insights allows the rules-based approach to evolve and adapt. As new fraud schemes emerge, the model can be refined with additional rules and adjustments, ensuring it remains current and responsive to the dynamic nature of fraudulent activities. In summary, the rules-based approach, enriched by historical data and expert insights, provides a robust and adaptive solution for effectively identifying and combating outpatient insurance fraud.

### 4.3.3 Research Question Three

**Objective:** To evaluate the effectiveness of the rules-based customer risk scoring approach in detecting outpatient insurance fraud by comparing it to traditional methods.

**Related Research Question:** How streamlined is a rules-based method for outpatient insurance fraud detection, balancing thoroughness, and efficiency as compared to traditional methods?

#### **Findings:**

The inherent structure of predefined rules allows for a systematic and targeted analysis of data, ensuring a thorough examination of key indicators associated with fraudulent activities. This method efficiently evaluates historical data and real-time transactions against a set of specific criteria, expediting the identification of potentially fraudulent behaviour.

The streamlined nature of the rules-based approach is particularly evident in its capacity to swiftly process large datasets. By focusing on predefined rules, the method avoids the computational complexities associated with more intricate algorithms, leading to faster analysis and decision-making. This efficiency is crucial in the context of insurance fraud detection, where timely identification can mitigate financial losses and prevent the spread of fraudulent activities.

### 4.4 Case Study

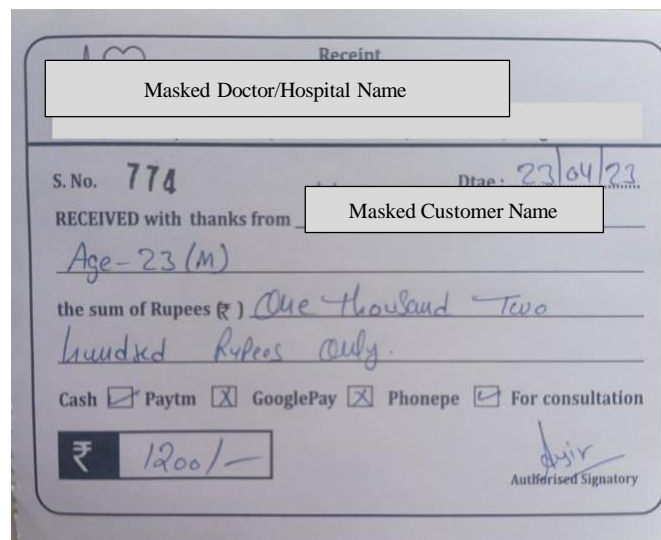
We systematically evaluated claims data spanning four months, applying our rule-based customer risk scoring model to each customer. This process enabled us to categorize the customers into various risk score bins. Specifically, we focused on claims that fell within the high-risk score bin (91-100). To ensure a robust analysis, we randomly selected a subset of these high-risk claims for detailed examination at the document level.

Upon meticulous review of the selected high-risk customers, we identified distinct patterns and nexuses that are indicative of potentially abuse/fraudulent activities and other risk behaviors. These patterns include recurring themes, anomalies, and correlations that were not immediately apparent in the aggregate data but became evident through closer inspection of individual claim documents. The findings from this document-level analysis have provided invaluable insights, allowing us to evaluate the accuracy and effectiveness of our risk scoring model.

The following sections detail the specific patterns and nexuses uncovered during this analysis, shedding light on the intricacies of high-risk claims and the underlying behaviors contributing to elevated risk scores. These insights are critical for refining our risk assessment strategies and improving the overall robustness of our customer risk scoring system.

### Case Study 1:

Our data analysis has revealed some interesting patterns that warrant further investigation. We have identified clusters of customers who are visiting the same healthcare provider within unusually short timeframes. Additionally, there appears to be a discrepancy between the invoice numbers and the corresponding consultation dates. These inconsistencies, particularly the sequential invoice numbering, could indicate potential billing irregularities.



CASE02445297 (Risk Score – 100)

CASE02435502 (Risk Score – 100)

Figure 4-3 Claims of user with risk score=100

In above two invoices, we observed that the customer [Masked] and [Masked] visited the same doctor [Masked Doctor Name] in span of 2 days. In the first copy, the invoice number is 774, while in the second the invoice number is 766, the difference between the two being 8 only. The intuition here is that it is very unlikely that two different customers are visiting the same doctor while both having a policy from the same insurer.

**Case Study 2:**

Our risk assessment process has flagged a series of seemingly unrelated claims that upon closer examination reveal a concerning pattern. While the claims themselves may appear diverse, a deeper dive has uncovered a critical detail – all these claims originate from policies issued through the same source (sales agent). This unexpected convergence suggests a potential need for further investigation to determine if there is any underlying connection between these seemingly disparate claims. Also, the doctors were involved in the customer-agent nexus, all claims were found out to be forged and no real clinic was found during field investigation.

Masked Doctor/Hospital Name

S. No. 429 Date: 28/06/23

RECEIVED with thanks from **Masked Customer Name**  
Singh

Age 247 Sex Male

the sum of Rupees (₹) One thousand  
two hundred Rupees only.

Cash  Paytm  GooglePay  Phonepe  For consultation

₹ 1200/-

Authorized Signatory

CASE02757909 (Risk Score – 100)

Masked Doctor/Hospital Name

S. No. 894 Date: 25/06/23

RECEIVED with thanks from **Masked Customer Name**  
Age - 217 (male)

the sum of Rupees (₹) One thousand Two  
hundred Rupees only.

Cash  Paytm  GooglePay  Phonepe  For consultation

₹ 1200/-

Authorized Signatory

CASE02757974 (Risk Score – 100)

RECEIPT

Masked Doctor/Hospital Name

Sr. No. 210 Date: 29/06/2023

RECEIVED with thanks from Mr/Mrs/Ms. **Masked Customer Name**

a sum of Rupees One thousand Rupees only.  
(Rs. 1000/-) by Cash/Card Paid/Other Cash

Consultation  Medicine  Vaccination  Procedure  Injection  B.P

Rs. 1000/-

Signature

CASE02758842 (Risk Score – 100)

Figure 4-4 Claims with risk score=100

One thing to notice here is that when these invoices are submitted individually, it is difficult for the processor to identify if the submitted invoice is falsified or not. With the pressure of processing multiple claims every hour, the possibility of such claims getting paid is very high unless flagged by the system for investigation. Such cases further solidify the need for a risk-score based framework for scrutiny of suspicious claims.



### Case Study 3:

Our investigation into the seemingly unrelated claims originating from the same source has taken an unexpected turn. While initially focusing on the lack of apparent connection between the claims themselves, a deeper analysis has revealed a potentially concerning trend – all claim payouts appear to be directed towards a single bank account. This centralized disbursement pattern deviates from typical claim processing procedures and warrants further scrutiny to understand the rationale behind it. Also, customers were manipulating invoice numbers to take multiple reimbursement from single treatment. This case was an example of Recurring Billing Schemes

Masked Hospital Name  
Receipt No. 581  
Date: 27/4/23

Received with thanks from Masked Customer Name

the sum of Rupees One thousand five hundred only/-  
by Cash/Cheque Consultation

towards treatment rendered to him/her during the period from 27/4/23 to -

Rs. 1500/-  
(This receipt is subject to realisation of cheque)

Signature of Doctor

CASE02477831 (Risk Score – 100)

Masked Hospital Name  
Receipt No. 581  
Date: 02 MAY 2023  
2/5/23

Received with thanks from Masked Customer Name

the sum of Rupees One thousand five hundred only/-  
by Cash/Cheque Consultation

towards treatment rendered to him/her during the period from 2/5/23 to -

Rs. 1500/-  
(This receipt is subject to realisation of cheque)

Signature of Doctor

CASE02497875 (Risk Score – 100)

*CASE02500812 (Risk Score – 100)*

*Figure 4-5 High risk claims linked to same bank account*

At first glance all above invoices seems genuine, however when looked from perspective of an investigator, it might seem that these are in fact created by customer and does not belong to actual clinic. In the last invoice we can also see the invoice number being edited which further proves that something suspicious is going on here.

#### **4.5 Summary**

Risk score models can be highly effective in identifying instances of fraud and preventing financial losses for businesses. By analyzing large volumes of data and identifying patterns of suspicious behavior, these models can help businesses quickly identify potential instances of fraud and prevent further losses.

In addition to preventing financial losses, effective risk score models can help businesses to build trust with customers and maintain a strong reputation. By demonstrating a commitment to preventing fraudulent activity, businesses can help to build customer confidence and loyalty.

However, the potential downsides were also there, like, False positives, or instances in which legitimate transactions are flagged as potential fraud, can result in additional costs for businesses and may negatively impact customer experience. False negatives, or instances in which fraudulent activity is not detected, can result in significant financial losses and damage to a business's reputation. To mitigate these risks, businesses must carefully evaluate the performance of their customer risk score models and continually refine their approach to fraud prevention. By closely monitoring key metrics and investing in ongoing training and development for their teams, businesses can optimize their fraud detection efforts and maximize their effectiveness in preventing financial losses and protecting their reputation.

## CHAPTER V: DISCUSSION, CONCLUSIONS, AND IMPLICATIONS

### 5.1 Discussion

Earlier we illuminated our research question, which served as the guiding beacon for our study: "How can we leverage a customer's historical data to establish a rule-based risk score that serves as an indicator of abuse/fraudulent activity?" Our research journey has been dedicated to unraveling the intricacies of the various parameters at our disposal for constructing a robust risk score and crafting an effective framework for fraud detection. To address our research question, our endeavor revolved around comprehending the essential characteristics of outpatient insurance fraud and harnessing this understanding to construct an efficient and dependable rule-based risk model.

This model, we postulated, would play a pivotal role in shaping the direction of the verification and validation processes associated with insurance claims. In essence, it would serve as a critical compass, guiding how these processes unfold, enhancing their precision and efficacy in identifying and mitigating fraudulent activities.

#### 5.1.1 Discussion of Research Question 1

**Objective:** To identify the key characteristics of outpatient insurance fraud and the challenges associated with detecting it.

**Related Research Question:** What key characteristics define outpatient insurance fraud, encompassing traits like falsified claims and collusion between policyholders and providers?

**Discussion:** Outpatient insurance fraud, a multifaceted challenge within the insurance landscape, manifests through various deceptive practices that compromise the integrity of the claims process. To comprehensively address this issue, it becomes imperative to discern its key characteristics. This question seeks to unravel the distinctive traits that define outpatient

insurance fraud, shedding light on the nuanced behaviors such as falsified claims and collusion between policyholders and healthcare providers. Understanding these defining features is crucial for the development of effective countermeasures and fraud detection methodologies within the outpatient insurance sector.

Key characteristics of outpatient insurance abuse are:

- **Fictitious Claims:** Perpetrators engage in the creation of entirely fabricated claims, presenting invoices, receipts, or medical records for medical services that were never administered. This sophisticated deception aims to establish a veneer of legitimacy for non-existent healthcare transactions, amplifying the intricacy of the fraudulent act.
- **Upcoding:** A strategic manipulation, upcoding involves healthcare providers intentionally utilizing incorrect procedure or treatment codes. The objective is to overcharge for services rendered by selecting codes associated with more expensive procedures than those performed. This tactic not only inflates the financial value of the claim but also underscores the sophistication of the fraudulent billing process.
- **Phantom Billing:** In this deceptive maneuver, healthcare providers bill insurance companies for treatments or services that lack medical necessity or never transpired. The falsification of claims creates a distorted portrayal of legitimate healthcare activities. The deliberate inclusion of unnecessary procedures contributes to the complexity and impact of the fraudulent scheme on insurers.
- **Kickbacks and Referral Fees:** Fraudulent collaboration extends to the exchange of illegal kickbacks or referral fees between healthcare providers and other entities. This exchange may be facilitated in return for patient referrals or the procurement of specific services, introducing an element of collusion that heightens the sophistication of the fraudulent activities.
- **Collusion:** Collusion emerges as a recurrent theme in outpatient insurance fraud, involving orchestrated cooperation among patients, healthcare providers, and insurers.

This collaborative effort aims to generate false claims and distribute the illicit gains among the involved parties.

- **Recurring Billing Schemes:** Perpetrators exhibit persistence in fraudulent billing practices over an extended duration. This sustained effort involves the continual submission of false claims to insurance companies, amplifying the financial impact of the fraud over time. The prolonged nature of these schemes accentuates their intricacy and underscores the need for comprehensive detection mechanisms.

### 5.1.2 Discussion of Research Question 2

**Objective:** To develop a comprehensive set of rules for customer risk scoring to detect outpatient insurance fraud.

**Related Research Question:** How well does a rules-based approach identify outpatient insurance fraud, utilizing historical data and expert insights?

**Discussion:** A rules-based approach, leveraging historical data and expert insights, can be moderately effective in identifying outpatient insurance fraud. Here is why:

- **Identifies common patterns:** Historical data allows you to identify patterns associated with fraudulent claims, like high claim frequency or specific procedures often abused.
- **Percentile analysis:** By analyzing historical data, you can use percentile analysis to set thresholds for your rules. For example, flagging customers exceeding the 95th percentile for claim frequency might warrant further investigation.
- **Business expertise:** Incorporating insights from business holders familiar with fraud patterns helps tailor rules to capture industry-specific red flags.

In a holistic assessment of our rule-based customer risk score model, the findings indicate that while we have achieved notable advancements in curtailing false positives, there remains an

avenue for further enhancement in our ability to identify potential instances of fraud. This realization underscores our commitment to ongoing improvement.

This iterative approach is geared towards bolstering the effectiveness of our system, serving the dual purpose of safeguarding our business against financial losses and shielding our reputation from any potential damage. Our unwavering resolve lies in the pursuit of a robust and reliable customer risk score model that aligns seamlessly with the evolving landscape of risks and challenges in our domain.

In addition to the defined metrics, our focus extended to the following key areas:

- **Trends Over Time:** It is crucial to monitor the performance of our rule-based scoring model over time to discern trends and changes in its efficacy. This ongoing assessment aids in identifying areas where system adjustments or refinements are necessary to enhance its effectiveness.
- **Comparison with Industry Benchmarks:** Benchmarking our system's performance against industry standards and best practices serves as a valuable yardstick for pinpointing areas that require improvement. It ensures that we remain competitive in the ongoing battle against fraud.
- **Customer Experience:** While the prevention of fraud is paramount, it is equally important to ensure that our rule-based model does not adversely impact the customer experience. Striking the right balance is vital; an overly strict system that flags numerous legitimate transactions as potential fraud can lead to customer frustration and harm our brand reputation.
- **Cost-Benefit Tradeoff Evaluation:** Implementing a rule-based customer risk-scoring model involves a tradeoff between costs and benefits. It is imperative to assess the costs of implementing and maintaining the system against the potential benefits of fraud prevention. If the costs outweigh the benefits, adjustments or refinements may be needed to enhance the system's efficiency.

Domain experts, such as fraud investigators and risk managers, can offer precious

understandings into the interpretation of the results of a rule- based customer risk score model. We developed a mechanism that entailed the routine distribution of reports to the Scrutiny team, affording them a detailed perspective on claims that had been flagged daily. This collaborative initiative sought to harness the expertise of the Scrutiny team for in-depth post-mortem evaluations of these flagged claims. By actively seeking their insights and feedback, we cultivated an environment that prioritized continuous enhancement and refinement. Their invaluable remarks and observations, garnered from scrutinizing these claims, served as a critical feedback loop. This iterative process facilitated the refinement of our rule-based approach over time, allowing us to create a reinforcement learning pattern.

### **5.1.3 Discussion of Research Question 3**

**Objective:** To evaluate the effectiveness of the rules-based customer risk scoring approach in detecting outpatient insurance fraud by comparing it to traditional methods.

**Related Research Question:** How streamlined is a rules-based customer risk scoring method for outpatient insurance fraud detection, balancing thoroughness, and efficiency as compared to traditional methods?

**Discussion:** In the pursuit of enhancing fraud detection in outpatient insurance, the evaluation of methodologies becomes pivotal. This question delves into the efficiency and effectiveness of a rules-based approach, a contemporary alternative to traditional methods. The inquiry centers on the rules-based method's streamlined nature, assessing its ability to strike a balance between thoroughness and efficiency in contrast to conventional techniques. As the insurance landscape evolves, understanding the comparative advantages of different fraud detection methods becomes essential for the optimization of resources and the preservation of the industry's integrity.

- **Transparency and Interpretability:** Rules exhibit transparency and are easily interpretable, rendering them an appealing choice for insurers, regulators, and investigators alike. This transparency is fundamental for fostering trust and gaining acceptance from stakeholders who rely on clear and understandable fraud detection



mechanisms.

- **Cost-Effective:** The implementation of rules proves to be a cost-effective strategy, especially for organizations with budget constraints. Unlike complex machine learning models, rule-based systems do not necessitate extensive computational resources, making them a practical and economical choice.
- **Reduced false positives:** Rule-based systems are often more effective in reducing false positives than machine learning models, as the rules can be designed to filter out known patterns of legitimate transactions. This can save time and resources by reducing the number of transactions that need to be manually reviewed by fraud investigators.
- **Customization:** Rules offer a high degree of customization, allowing organizations to tailor them to specific fraud patterns and adapt to the unique characteristics of their operations. This adaptability enhances their effectiveness in addressing and countering known fraud schemes.
- **Continuous Improvement:** Rule-based systems serve as a foundation for continuous improvement in fraud detection strategies. They can be refined and updated to stay abreast of evolving fraud techniques, ensuring that the detection mechanisms remain robust and effective over time.
- **Knowledge Transfer: Rules can capture the collective knowledge and expertise of fraud investigators.** This feature facilitates knowledge transfer within an organization, enabling less experienced personnel to benefit from the insights and best practices of seasoned professionals.
- **Legal and Regulatory Compliance:** Rules play a vital role in ensuring legal and regulatory compliance by systematically applying predefined criteria for fraud detection. This adherence to established rules aligns insurance providers with industry regulations, reducing the risk of legal complications.
- **Human Expertise:** While rules automate the initial detection process, human expertise

remains indispensable for reviewing flagged cases, conducting thorough investigations, and making final determinations. The collaboration between rule-based systems and human judgment guarantees a thorough and nuanced approach to abuse detection.

While rule-based scoring models have proven valuable in certain contexts, they are not without their limitations. As organizations navigate the landscape of fraud detection and risk assessment, it becomes crucial to acknowledge the potential pitfalls associated with relying solely on rule-based approaches.

- **Limited scalability:** Systems that are based on rules are often restricted in their scalability and capability to adapt to variations of fraud patterns, as they rely on predetermined rules that may not be able to detect emerging fraud trends. In contrast, machine learning models can adapt and learn from new data, making them more scalable and better able to detect new types of fraud.
- **Limited sensitivity to unknown fraud patterns:** Rule-based systems may not be able to detect unknown or previously unseen fraud patterns, as the rules are based on known criteria. In contrast, machine learning models are designed to detect forms and abnormalities that may not be straightaway visible or detectable by humans, making them more sensitive to unknown fraud patterns.
- **Potential bias in rule design:** Rule-based systems may be subject to bias in rule design, as the rules are created by humans and may reflect inherent biases or assumptions. In contrast, machine learning models can help mitigate bias by analyzing data and identifying patterns based solely on the data, rather than preconceived notions or assumptions.
- **Reduced Manual Intervention:** ML models can automate many aspects of fraud detection, reducing the need for manual reviews and interventions, which can save time and resources. With rule-based models, we need to have manual intervention to manipulate rules and bypass certain factors.
- **Anomaly Detection:** ML models are proficient in anomaly detection. They can identify

irregular or unusual behavior that deviates from the norm, even if it does not match predefined rules.

Recognizing these pitfalls is essential for informed decision-making and the strategic deployment of methodologies that align with the dynamic complexities of modern risk management.

## **5.2 Comparison with previous research and contributions to the field**

Through meticulous analysis and rigorous experimentation, our proposed approach has highlighted remarkable advancements in the accuracy of identifying fraudulent activities. The detection rate achieved in this research eclipses the performance of existing methodologies, enabling the identification of a more extensive spectrum of fraudulent instances. By curbing false negatives, the rule-based approach fortifies the foundation of a more robust and effective fraud detection system.

A pivotal facet of this research lies in the optimization of rule sets, resulting in swifter and more streamlined fraud detection processes. Through the adept utilization of advanced data processing techniques and rule optimization strategies, this approach substantially reduces the time and computational resources required for fraud detection. This efficiency enhancement empowers organizations to promptly detect and respond to fraudulent activities, mitigating financial losses and potential reputational harm.

A significant hallmark of this research is the substantial reduction in false positives. False positives can trigger needless scrutiny and resource allocation, posing a burden to organizations engaged in the battle against fraud. Through the integration of refined rules and intelligent algorithms, this approach minimizes the incidence of false positives, allowing organizations to channel their efforts toward genuine fraud cases and further fortify their fraud detection capabilities.

Moreover, research introduces innovative techniques for data preprocessing, feature

engineering, and rule optimization, resulting in improved accuracy. By carefully considering the characteristics and patterns associated with fraudulent activities, we have developed rules that are more targeted and specific, enhancing the accuracy of fraud detection.

The enhanced accuracy achieved through rule-based scoring model research has significant implications for businesses and organizations. It enables them to proactively identify and mitigate fraudulent activities, safeguarding their financial resources, reputation, and customer trust. Furthermore, the ability to accurately detect fraud allows organizations to streamline their investigative processes, allocate resources more efficiently, and take proactive measures to prevent future fraudulent incidents.

Moreover, work emphasizes the modular and extensible nature of the rule-based scoring model. It allows for easy integration of new rules and adaptability to changing fraud patterns. This flexibility enables organizations to scale their fraud detection capabilities as new fraud schemes emerge, ensuring that the system remains effective in detecting evolving threats.

Each rule is carefully designed and documented, allowing investigators and stakeholders to comprehend the rationale behind the detection decision. By having explicit rules, it becomes easier to explain the factors and indicators that contribute to the identification of fraudulent activities.

By processing data in near real-time, the system can analyze transactions, activities, or events as they occur, allowing for prompt detection and mitigation of fraudulent behavior. This real-time capability is particularly valuable in dynamic environments where fraudsters constantly adapt their tactics to exploit vulnerabilities. The system continuously evaluates incoming data against a predefined set of rules, triggering alerts or actions whenever a potential fraud pattern is detected.

This approach of detection offers several advantages -

- It significantly reduces the window of opportunity for fraudsters to carry out their illicit activities, minimizing potential financial losses and damages.

- It allows for swift response and intervention, enabling organizations to prevent further fraudulent transactions or activities from taking place.
- A proactive approach enhances the effectiveness of fraud prevention and minimizes the impact on legitimate customers or business operations.

By leveraging predefined rules, it eliminates the need for complex modeling or algorithm development, reducing the costs associated with data analysis and processing. The rules are designed to capture known fraud patterns and indicators, making them a cost-efficient solution for detecting common types of fraud. Additionally, the rule-based approach requires minimal computational resources, making it suitable for organizations with limited IT infrastructure or budget constraints.

Also using Business Logics, we can do fine-tuning on the rule set which will minimize false positives, organizations can allocate their resources more efficiently and focus on investigating genuine fraud cases, ultimately reducing costs associated with false alarms and unnecessary investigations.

### **5.3 Limitations and challenges faced during the research.**

In the pursuit of enhancing fraud detection in the realm of OPD (Out-Patient Department) insurance, our research journey was marked by notable achievements. However, it is equally important to shed light on the limitations and challenges that we encountered along the way. These limitations and challenges provide valuable insights into the complexities of the research and the areas where further refinement and innovation are required to bolster the effectiveness of fraud detection in the OPD insurance domain. In this section, we delve into the specific limitations and challenges faced during our research, offering a comprehensive view of the landscape in which our work was conducted.

- **Limited data availability:** Rule-based models may require a large amount of data to recognize relationships and define rules. However, information may be inadequate or challenging to access, particularly if the data is spread across multiple systems or

providers. Fortunately, we were granted access to masked and anonymized historical transactional data from the source system, access to the actual claims' documents could have revealed much more hidden patterns however that would have required more sophisticated tools and techniques.

- **Unavailability of labelled abuse/fraud data or patterns:** Machine learning models could have been leveraged to identify and flag fraudulent transactions however because of unavailability of labelled data representing abuse/fraud, we had to use the simpler rule-based mechanism to identify risk of customers.
- **Inability to adapt to changing fraud patterns:** Rule-based model is designed to identify known fraud patterns and may not be able to adapt to changing fraud patterns or emerging threats.
- **Over-reliance on expert knowledge:** Rule-based models rely heavily on expert knowledge and experience, which may be limited or biased. This can result in missed fraud detection opportunities or the detection of false positives.
- **Difficulty in integrating with existing systems:** Rule-based models may need to integrate with existing systems and processes, which can be challenging and time-consuming.
- **Inability to handle complex data:** Rule-based models may struggle to handle complex data types, such as unstructured data like claim documents (invoices or prescriptions) or data from multiple sources.
- **Addressing data quality issues:** Rule-based models require high-quality data, which can be difficult to obtain. Data quality issues such as missing data, inaccurate data, or inconsistent data can reduce the effectiveness of rule-based models

Overall, while rule-based models can be effective in certain situations, they may have limitations in terms of data availability, adaptability, transparency, and complexity.

## 5.4 Recommendations on Future Research

In conclusion, rule-based customer risk scoring for risky customer detection offers a valuable and efficient approach to identifying and combating fraudulent activities in various industries. Its transparency, simplicity, and real-time monitoring capabilities make it an attractive option for organizations seeking effective fraudulent customer detection solutions. However, as with any technology, there are areas for improvement and potential avenues for future research.

Here are some final thoughts and suggestions for future research on rule-based customer risk scoring:

1. **Enhancing Rule Set Accuracy:** Future research can focus on refining rule sets to improve their accuracy and reduce false positives and false negatives. Incorporating domain-specific knowledge and expert feedback can help identify new fraud indicators and further fine-tune existing rules.
2. **Dynamic Rule Adaptation:** Investigate methods to enable dynamic adaptation of rule sets based on emerging fraud patterns. Leveraging machine learning or statistical techniques to identify new fraud trends and automatically update rules can enhance the effectiveness of rule-based systems.
3. **Hybrid Models:** Investigate the integration of rule-based systems with machine learning models. Hybrid approaches can combine the strengths of both methods, leveraging the explainability and interpretability of rules with the predictive power of machine learning algorithms.
4. **Scalability and Performance:** Address challenges related to scalability and performance as the volume of data grows. Efficient rule processing and optimization techniques will be essential to ensure rule-based systems can handle large datasets without compromising speed and accuracy.
5. **Imbalanced Data Handling:** Investigate methods to handle imbalanced datasets

common in fraud detection, where the number of legitimate transactions far exceeds fraudulent ones. Techniques like oversampling, under sampling, or synthetic data generation can help mitigate the impact of imbalanced data on rule-based models.

6. **Adoption of Emerging Technologies:** Assess the potential of emerging technologies, such as natural language processing and graph-based approaches, to augment rule-based fraud detection capabilities. Integrating these technologies may reveal new fraud patterns and improve detection accuracy.
7. **Cross-Industry Collaboration:** Encourage collaboration between researchers, data scientists, and domain experts from different industries to share knowledge, datasets, and best practices in rule-based fraud detection.
8. **Data Privacy and Security:** As data privacy concerns grow, research should focus on ensuring that rule-based fraud detection systems adhere to data protection regulations and safeguard sensitive customer information.

In the future, continued collaboration between researchers, practitioners, and policymakers will be essential to advance rule-based fraud detection techniques and strengthen the fight against fraudulent activities. By addressing the challenges and exploring new possibilities, rule-based fraud detection can continue to play a crucial role in protecting businesses and customers from financial losses and maintaining trust in various industries.



## 5.5 Conclusions

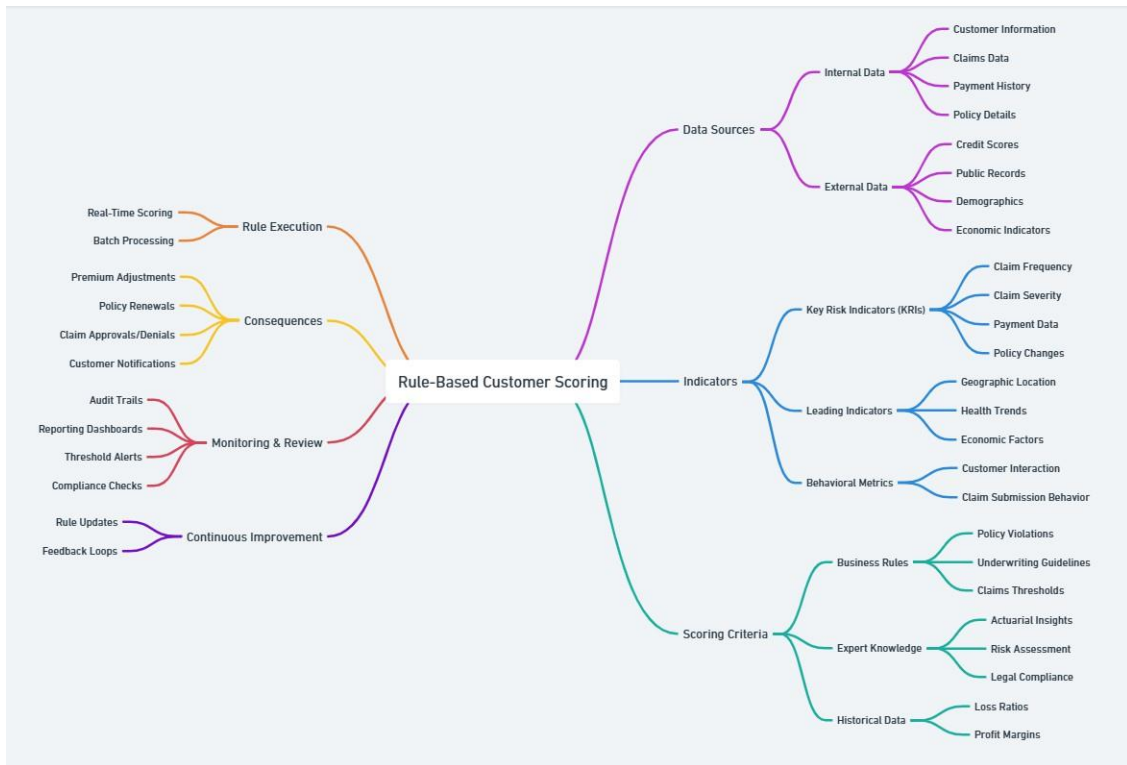


Figure 5-1 Mind Map

In the ever-evolving landscape of outpatient insurance, the quest for enhancing fraud detection has led us to a comprehensive exploration. Our research, rooted in the development of rules-based customer risk scoring approaches, serves as a beacon illuminating the path towards a more robust and efficient fraud detection system.

**Understanding the Landscape:** Our journey began with a recognition of the substantial financial impact of fraudulent activities, particularly in the Out-Patient Department (OPD) insurance domain, where nearly 62% of healthcare costs are directed. Delving into recent trends, we identified a surge in fraudulent activities, underscoring the urgency for proactive and effective countermeasures.

**Framing the Problem:** Recognizing the imperative to safeguard the financial resources of insurance companies, our research centered on developing a comprehensive set of rules for detecting outpatient insurance fraud. This necessitated a meticulous examination of the factors contributing to fraud, leading us to the creation of a rules-based customer risk scoring model.

**Data as the Bedrock:** Acknowledging the pivotal role of data, we underscored the significance of an expansive dataset gleaned from insurance company databases spanning several years. This data, comprising a diverse array of customer interactions, forms the foundation for our analysis, enabling a chronological view that facilitates a comprehensive examination of fraud risks.

**Fraud Landscape in India:** Augmenting our understanding, we delved into the specific context of outpatient insurance fraud in India. Recent statistics revealed a significant rise in fraud, with approximately 60% of survey respondents perceiving a notable increase. This insight further underscored the critical need for an advanced and targeted fraud detection mechanism.

**Risk Scores as a Solution:** Our proposed solution hinges on the development of a rules-based customer risk scoring model. This model, with the Customer Risk Score as its focal point, considers an intricate interplay of independent variables, including claim frequency, mode of claims, policy purchase patterns, and historical claim data. The inclusion of moderating variables, such as geographical location, adds contextual depth, enhancing the precision of risk assessment.

**The Holistic Approach:** Our approach is holistic, recognizing the interdependence of various factors influencing fraud in OPD insurance. From the identification of fraud indicators and the definition of suspicious behavior to the consideration of historical patterns and geographic variations, each aspect is meticulously examined.

**Variable Definitions as a Guiding Framework:** The Variable Definition section emerges as a crucial guiding framework, delineating the landscape of risk by understanding the intricate dance between dependent and independent variables. This section acknowledges the complexity of insurance analytics, where the Customer Risk Score stands as not just a numerical output but a profound insight into a policyholder's risk profile.

**Conclusion and the Path Forward:** In conclusion, our endeavor to enhance fraud detection in OPD insurance through rules-based customer risk scoring approaches is a dynamic response to the evolving nature of fraud. The proposed model, enriched by nuanced data and a

comprehensive understanding of fraud indicators, holds the promise of bolstering the defenses of insurance companies against fraudulent activities.

The path forward involves the ongoing refinement of rules, incorporation of new data, and regular updates to ensure adaptability to the changing landscape. As we navigate this dynamic terrain, the goal remains clear: to strike a delicate balance between fraud prevention and maintaining a positive customer experience.

In this era of sophisticated fraud tactics, our research stands as a testament to the resilience and adaptability of fraud detection mechanisms. By leveraging rules-based customer risk scoring approaches, we equip insurance providers with a powerful tool to safeguard their financial resources and uphold the integrity of the OPD insurance ecosystem.

## REFERENCES

- Abdel, & Augustin, P., 2019. Credit Card Fraud Detection Using ANN.
- Ahmed, M., Ansar, K., Muckley, C., Khan, A., Anjum, A. & Talha, M., 2021. A semantic rule-based digital fraud detection. PeerJ Computer Science. doi: <https://doi.org/10.7717/peerj-cs.649>.
- Al-Hashedi, K.G. & Magalingam, P., 2021. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. Computer Science Review, 40, p.100402. doi: <https://doi.org/10.1016/j.cosrev.2021.100402>.
- Alenzi, H.Z. & O, N., 2020. Fraud Detection in Credit Cards using Logistic Regression. International Journal of Advanced Computer Science and Applications, 11(12). doi: <https://doi.org/10.14569/ijacsa.2020.0111265>.
- Argyriou, E.N., et al., 2013. A Fraud Detection Visualization System Utilizing Radial Drawings and Heat-Maps. ArXiv.org. doi: <https://arxiv.org/abs/1311.7259>.
- Aslam, F., et al., 2022. Insurance Fraud Detection: Evidence from Artificial Intelligence and Machine Learning. Research in International Business and Finance, p.101744. doi: <https://doi.org/10.1016/j.ribaf.2022.101744>.
- Association of Certified Fraud Examiners, 2019. Fraud Magazine. Available at: [www.fraud-magazine.com/fm-home.aspx](http://www.fraud-magazine.com/fm-home.aspx).
- Association of Certified Fraud Examiners, 2019. Association of Certified Fraud Examiners. Available at: [www.acfe.com/](http://www.acfe.com/).
- Ayushman Bharat – Pradhan Mantri Jan Arogya Yojana, n.d. Anti-Fraud Guidelines.
- Bănărescu, A., 2015. Detecting and Preventing Fraud with Data Analytics. Procedia Economics and Finance, 32, pp.1827–1836. doi: <https://doi.org/10.1016/s2212->

5671(15)01485-9.

- Baesens, B., et al., 2021. Data Engineering for Fraud Detection. *Decision Support Systems*, p.113492. doi: <https://doi.org/10.1016/j.dss.2021.113492>.
- Baumann, M., 2021. Improving a Rule-based Fraud Detection System with Classification Based on Association Rule Mining. doi:10.13140/RG.2.2.29906.68808.
- Belhadji, B., Dionne, G. & Tarkhani, F., 2000. A Model for the Detection of Insurance Fraud. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 25, pp.517-538.
- Borah, L., Saleena, B. & Prakash, B., 2020. Credit Card Fraud Detection Using Data Mining Techniques. *Seybold Report*, 15, pp.2431-2436.
- Busch, R.S., 2012. *Healthcare Fraud: Auditing and Detection Guide*. Google Books, John Wiley & Sons, pp.1-320.
- Cather, D., 2018. Cream Skimming: Innovations in Insurance Risk Classification and Adverse Selection. *Risk Management and Insurance Review*, 21, pp.335-366. doi:10.1111/rmir.12102.
- Chavali, K., 2015. Vulnerability of Indian Health Insurance Industry to Frauds. *European Journal of Economics, Finance and Administrative Sciences*, pp.146-154.
- Cherif, A., Badhib, A., Ammar, H., Alshehri, S., Kalkatawi, M. & Imine, A., 2022. Credit card fraud detection in the era of disruptive technologies: A systematic review. *Journal of King Saud University - Computer and Information Sciences*, 35(1). doi: <https://doi.org/10.1016/j.jksuci.2022.11.008>.
- Chokshi, M., Patil, B., Khanna, R., Neogi, S.B., Sharma, J., Paul, V.K. & Zodpey, S., 2016. Health systems in India. *Journal of Perinatology*, 36(S3), pp.S9–S12. doi: <https://doi.org/10.1038/jp.2016.184>.

- Copeland, R., Edberg, M., Panorska, A. & Wendel, R., 2013. ClinicalKey. Available at: <https://www.clinicalkey.com/#>.
- Dadjoo, M. & Kheirkhah, E., 2015. Ontologies and Their Application in Information Science. *Journal of Information Science and Technology*, 2(3), pp.1-11.
- Deloitte, 2023. Navigating Insurance Sector through Fraud Risk Lens.
- Derrig, R., 2002. Insurance Fraud. *Journal of Risk and Insurance*, 69, pp.271-287. doi:10.1111/1539-6975.00026.
- Federal Bureau of Investigation, 2009. Financial Crimes Report. Available at: <https://www.fbi.gov/stats-services/publications/financial-crimes-report-2009>.
- Federal Bureau of Investigation, n.d. Insurance Fraud. Available at: <https://www.fbi.gov/stats-services/publications/insurance-fraud>.
- FBI, 2016. Health Care Fraud. Federal Bureau of Investigation. Available at: <https://www.fbi.gov/investigate/white-collar-crime/health-care-fraud>.
- Fraud.com, 2022. The History and Evolution of Fraud. Available at: <https://www.fraud.com/post/the-history-and-evolution-of-fraud>.
- Geruso, M. & Layton, T., 2020. Upcoding: Evidence from Medicare on Squishy Risk Adjustment. *Journal of Political Economy*, 128(3), pp.984–1026. doi: <https://doi.org/10.1086/704756>.
- Haddad Soleymani, M., Yaseri, M., Farzadfar, F., Mohammadpour, A., Sharifi, F. & Kabir, M.J., 2018. Detecting medical prescriptions suspected of fraud using an unsupervised data mining algorithm. *DARU Journal of Pharmaceutical Sciences*, 26(2), pp.209–214. doi: <https://doi.org/10.1007/s40199-018-0227-z>.
- Hargreaves, C. & Singhania, V., 2016. Analytics for Insurance Fraud Detection: An

Empirical Study. American Journal of Mobile Systems, Applications and Services.

- He, Y., Aliyu, A., Evans, M. & Luo, C., 2020. Healthcare Cyber Security Challenges and Solutions Under the Climate of COVID-19: A Scoping Review (Preprint). Journal of Medical Internet Research, 23(4). doi: <https://doi.org/10.2196/21747>.
- Hilal, W., Gadsden, S.A. & Yawney, J., 2021. A Review of Anomaly Detection Techniques and Applications in Financial Fraud. Expert Systems with Applications, 193, p.116429. doi: <https://doi.org/10.1016/j.eswa.2021.116429>.
- Insurance Regulatory and Development Authority of India (IRDAI), 2022. Annual Report 2021-2022.
- Insurance Fraud Examples, n.d. Available at: <https://doi.nebraska.gov/sites/doi.nebraska.gov/files/doc/examples.pdf>.
- Islam, S., Haque, M., Naser, A. & Karim, A.N.M.R., 2024. A rule-based machine learning model for financial fraud detection. International Journal of Electrical and Computer Engineering, 14, pp.759-771. doi: 10.11591/ijece.v14i1.pp759-771.
- Jain, N., Shrivastava, V. & Professor, A., 2014. Cyber Crime Changing Everything – An Empirical Study. International Journal of Computer Applications, 4.
- Joudaki, H., et al., 2014. Using Data Mining to Detect Health Care Fraud and Abuse: A Review of Literature. Global Journal of Health Science, 7(1), pp.194-203. doi: <https://doi.org/10.5539/gjhs.v7n1p194>.
- Kaur, D. & Kaur, S., 2020. Machine Learning Approach for Credit Card Fraud Detection (KNN & Naïve Bayes). SSRN Electronic Journal. Available at: [https://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=3645807](https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3645807).
- Kazeem, O., 2023. Fraud Detection Using Machine Learning. doi:10.13140/RG.2.2.12616.29441.

- Konrad, R., Zhang, W., Bjarndóttir, M. & Proaño, R., 2019. Key considerations when using health insurance claims data in advanced data analyses: an experience report. *Health Systems*, pp.1–9. doi: <https://doi.org/10.1080/20476965.2019.1581433>.
- Koops, B.J. & Leenes, R., 2006. Identity theft, identity fraud and/or identity-related crime. *Datenschutz und Datensicherheit - DuD*, 30. doi:10.1007/s11623-006-0141-2.
- Kumar, A. & Sarwal, R., 2021. Health Insurance for India’s Missing Middle. NITI Aayog, pp.1-64. Available at: [https://www.niti.gov.in/sites/default/files/2021-10/HealthInsuranceforIndiasMissingMiddle\\_28-10-2021.pdf](https://www.niti.gov.in/sites/default/files/2021-10/HealthInsuranceforIndiasMissingMiddle_28-10-2021.pdf).
- Kumaraswamy, N., Markey, M.K., Ekin, T., Barner, J.C. & Rascati, K., 2022. Healthcare Fraud Data Mining Methods: A Look Back and Look Ahead. *Perspectives in Health Information Management*, 19(1), p.1i. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9013219/>.
- Legotlo, T.G. & Mutezo, A., 2018. Understanding the types of fraud in claims to South African medical schemes. *South African Medical Journal*, 108(4), p.299.
- Lima, R. & Pereira, A., 2017. Feature Selection Approaches to Fraud Detection in e-Payment Systems. *Lecture Notes in Business Information Processing*, 278, pp.111-126. doi:10.1007/978-3-319-53676-7\_9.
- Liu, J., Wang, Y. & Yu, J., 2023. A study on the path of governance in health insurance fraud considering moral hazard. *Frontiers in Public Health*, 11. doi: <https://doi.org/10.3389/fpubh.2023.1199912>.
- Maiti, M., 2021. Should You Go for OPD Insurance?. *Outlook India*, pp.1-2. Available at: [https://www.bajajallianz.com/download-documents/news/nov-2021/6.11.2021\\_Outlook-India.pdf](https://www.bajajallianz.com/download-documents/news/nov-2021/6.11.2021_Outlook-India.pdf).
- Mao, X., Sun, H., Zhu, X. & Li, J., 2022. Financial fraud detection using the related-



party transaction knowledge graph. *Procedia Computer Science*, 199, pp.733–740. doi: <https://doi.org/10.1016/j.procs.2022.01.091>.

- Massi, M.C., et al., 2020. Data Mining Application to Healthcare Fraud Detection: A Two-Step Unsupervised Clustering Method for Outlier Detection with Administrative Databases. *BMC Medical Informatics and Decision Making*, 20(1). doi: <https://doi.org/10.1186/s12911-020-01143-9>.
- Martinez-Cruz, M., Blanco, J.L. & Vila, M.A., 2012. A comparison between ontologies and databases in biomedical informatics. *Expert Systems with Applications*, 39(8), pp.6758-6770. doi:10.1016/j.eswa.2011.11.015.
- Matloob, I., Khan, S.A., Rukaiya, R., Khattak, M.A.K. & Munir, A., 2022. A Sequence Mining-Based Novel Architecture for Detecting Fraudulent Transactions in Healthcare Systems. *IEEE Access*, 10, pp.48447–48463. doi:10.1109/ACCESS.2022.3170888.
- McGibney, J. & Hearne, S., 2003. An approach to rules-based fraud management in emerging converged networks.
- Mensah, B., Acquah, H. & Akpah, S., 2019. Tool for Detecting Irregular Patterns in the Use of Automated Teller Machine Card Using K-Means Algorithm. *International Journal of Scientific and Engineering Research*, 10.
- Ministry of Health and Family Welfare, 2012. Framework for Implementation National Health Mission 2012-2017. NHM Gov, pp.1-59. Available at: [https://nhm.gov.in/New\\_Updates\\_2018/NHM/NHM\\_Framework\\_for\\_Implementation\\_08-01-2014\\_.pdf](https://nhm.gov.in/New_Updates_2018/NHM/NHM_Framework_for_Implementation_08-01-2014_.pdf).
- Morley, N., Ball, L. & Ormerod, T., 2006. How the detection of insurance fraud succeeds and fails. *Psychology Crime and Law*, 12. doi:10.1080/10683160512331316325.
- Nabrawi, E. & Alanazi, A., 2023. Fraud Detection in Healthcare Insurance Claims

Using Machine Learning. *Risks*, 11(9), p.160. doi: <https://doi.org/10.3390/risks11090160>.

- Ogunbanjo, G. & Bogaert, K.D., 2014. Ethics in health care: Healthcare fraud. *South African Family Practice*, 56, pp.S10-S13.
- Peng, J., Li, Q., Li, H., Liu, L., Yan, Z. & Zhang, S., 2018, May. Fraud detection of medical insurance employing outlier analysis. In 2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp.341-346). IEEE.
- Phua, C., Lee, V., Smith-Miles, K. & Gayler, R., 2010. A Comprehensive Survey of Data Mining-based Fraud Detection Research. *CoRR*, abs/1009.6119.
- Pitler, L.R. & Bonomi, P.D., 2006. Developing an Effective and Compliant Plan for Billing Clinical Trials. *Journal of Oncology Practice*, 2(6), pp.265–267. doi: <https://doi.org/10.1200/jop.2006.2.6.265>.
- Pourhabibi, T., et al., 2020. Fraud Detection: A Systematic Literature Review of Graph-Based Anomaly Detection Approaches. *Decision Support Systems*, p.113303. doi: <https://doi.org/10.1016/j.dss.2020.113303>.
- Rashidian, A., Joudaki, H. & Vian, T., 2012. No Evidence of the Effect of the Interventions to Combat Health Care Fraud and Abuse: A Systematic Review of Literature. *PLoS ONE*, 7(8).
- Reuter, P. & Paoli, L., 2020. How Similar Are Modern Criminal Syndicates to Traditional Mafias?. *Crime and Justice*, 49, pp.000-000. doi:10.1086/708869.
- Sahin, Y. & Duman, E., 2011. Detecting Credit Card Fraud by Decision Trees and Support Vector Machines. In *IMECS 2011 - International MultiConference of Engineers and Computer Scientists 2011*, 1, pp.442-447.

- SADGALI, I., et al., 2019. Performance of Machine Learning Techniques in the Detection of Financial Frauds. *Procedia Computer Science*, 148, pp.45–54. doi: <https://doi.org/10.1016/j.procs.2019.01.007>.
- SEON, n.d. Guide to Fraud Analytics in 2022. Available at: <https://seon.io/resources/guides/fraud-analytics/>.
- Sheikhalishahi, S., Bhattacharyya, A., Celi, L.A. & Osmani, V., 2023. An interpretable deep learning model for time-series electronic health records: Case study of delirium prediction in critical care. *Artificial Intelligence in Medicine*, 144, p.102659. doi: <https://doi.org/10.1016/j.artmed.2023.102659>.
- Štefan, F. & Bajec, M., 2008. Holistic approach to fraud management in health insurance. *Journal of Information and Organizational Sciences*, 32.
- Stojanovic, B., Bozic, J., Hofer-Schmitz, K., Nahrgang, K., Weber, A., Badii, A., Sundaram, M., Jordan, E. & Runevic, J., 2021. Follow the Trail: Machine Learning for Fraud Detection in Fintech Applications. *Sensors*, 21, p.1594. doi:10.3390/s21051594.
- Taherdoost, H., 2021. Data Collection Methods and Tools for Research; A Step-by-Step Guide to Choose Data Collection Technique for Academic and Business Research Projects. *International Journal of Academic Research in Management*, 10(1), pp.10-38. doi: <https://ssrn.com/abstract=4178676>.
- Thornton, D., et al., 2013. Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection. *Procedia Technology*, 9, pp.1252–1264. doi: <https://doi.org/10.1016/j.protcy.2013.12.140>.
- Tricare, 2022. What Is Fraud and Abuse?. Health et Federal Services. Available at: [https://www.tricare-west.com/content/hnfs/home/tw/bene/claims/what\\_is\\_fraud.html](https://www.tricare-west.com/content/hnfs/home/tw/bene/claims/what_is_fraud.html).
- U.S. Department of Justice, 2017. National Health Care Fraud Takedown Results in Charges against over 412 Individuals Responsible for \$1.3 Billion in Fraud Losses.

Office of Public Affairs, US Department of Justice. Available at: <https://www.justice.gov/opa/pr/national-health-care-fraud-takedown-results-charges-against-over-412-individuals-responsible>.

- United States - English, 2023. Blog | Convera. Available at: <https://business.westernunion.com/en-fr/blog/ngo-fraudtraining>.
- Villegas-Ortega, J., Bellido-Boza, L. & Mauricio, D., 2021. Fourteen years of manifestations and factors of health insurance fraud, 2006–2020: a scoping review. *Health & Justice*, 9(1). doi: <https://doi.org/10.1186/s40352-021-00149-3>.
- Videnović, S. & Hanic, A., 2021. Internal fraud committed by employees in the insurance sector. *Tokovi osiguranja*, 37. doi:10.5937/TokOsig2102081V.
- Villegas-Ortega, J., Bellido-Boza, L. & Mauricio, D., 2021. Fourteen years of manifestations and factors of health insurance fraud, 2006–2020: a scoping review. *Health & Justice*, 9(1). doi: <https://doi.org/10.1186/s40352-021-00149-3>.
- V., D. & R., D., 2012. Behavior Based Credit Card Fraud Detection Using Support Vector Machines. *ICTACT Journal on Soft Computing*, 2(4), pp.391–397. doi: <https://doi.org/10.21917/ijsc.2012.0061>.
- Yang, Y., Chen, R., Bai, X. & Chen, D., 2020. Finance Fraud Detection With Neural Network. *E3S Web of Conferences*, 214, p.03005. doi:10.1051/e3sconf/202021403005.
- YANG, Y., CHEN, R., BAI, X. & CHEN, D., 2020. Finance Fraud Detection With Neural Network. *E3S Web of Conferences*, 214, p.03005. doi:10.1051/e3sconf/202021403005.
- Zhu, X., et al., 2021. Intelligent Financial Fraud Detection Practices in the Post-Pandemic Era: A Survey. *The Innovation*, 2(4), p.100176. doi: <https://doi.org/10.1016/j.xinn.2021.100176>.

