TRACEABILITY AND FRAUD IN DIGITAL ADVERTISEMENT

by

SAJ ABRAHAM, DBA Research Scholar

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

OCTOBER, 2024

TRACEABILITY AND FRAUD IN DIGITAL ADVERTISEMENT

by

Saj Abraham, DBA Research Scholar

APPROVED BY

(Maria Joseph Xavier)
Mentor

APPROVED BY

Chair

RECEIVED/APPROVED BY:

<Associate Dean’s Name, Degree>, Associate Dean

Acknowledgments

I take this opportunity to thank:

Swiss School of Business and Management (SSBM), Geneva

I extend my heartfelt gratitude to **Dr. Maria Joseph Xavier**, whose mentorship, expertise, and unwavering support have been pivotal in the completion of this thesis. His thoughtful guidance, constructive feedback, and encouragement have greatly enriched my research journey, enabling me to overcome challenges and achieve my goals. I am deeply appreciative of his patience, dedication, and belief in my potential, which have been a constant source of inspiration throughout this DBA journey.

ABSTRACT

TRACEABILITY AND FRAUD IN DIGITAL ADVERTISEMENT

SAJ ABRAHAM
2024

Dissertation Chair: Dr. Iva Buljubasic
Committee members: Dr. Sasa Petar, Dr. Maria Joseph Xavier

In the dynamic landscape of digital advertising, fraud and the lack of traceability pose significant challenges, leading to financial losses and eroding trust among stakeholders. This thesis explores the application of machine learning (ML) techniques to mitigate fraud and enhance traceability within digital ad transactions, including impressions, clicks, and conversions. Through preprocessing, feature engineering, and deploying advanced ML algorithms like XGBoost, this study meticulously evaluates the efficacy of various models in detecting fraudulent activities, employing metrics such as accuracy, precision, recall, F1 score, and ROC-AUC for comprehensive assessment. This research demonstrates that ML significantly outperforms traditional rule-based systems in identifying complex fraudulent patterns, offering a novel framework for integrating ML-based fraud detection into digital advertising platforms. By improving the detection of ad fraud, this work contributes to creating more secure, transparent, and efficient digital advertising markets, ultimately fostering trust and reducing financial losses for advertisers. This thesis not only highlights the financial and trust implications of ad fraud but also presents a methodological approach for leveraging technology to combat these issues, marking a significant step forward in the ongoing battle against digital advertising fraud.

KEYWORDS

Digital Advertising Fraud, Online Ad Fraud Detection, Machine Learning in Fraud Detection, Click Fraud Analysis, Ad Traffic Anomaly Detection, Behavioral Analysis in Digital Ads, Blockchain for Ad Fraud Prevention, Advertisement Click Fraud, Programmatic Advertising Security, Fraudulent Traffic Identification, Ad Fraud Mitigation Techniques, Traceability in Digital Marketing.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER I:

INTRODUCTION

1.1 INTRODUCTION

Traceability is an extremely important factor in the field of digital advertising, as it is responsible for facilitating the exchange of personal information via online and web cookies. Utilizing a wide variety of approaches, technology, and methods, this practice makes the generation of consumer profiles based on market participation easier to accomplish. Recent years have seen substantial development in the capabilities of digital advertisements to target specific audiences, with browsers playing an essential part in this process. Traceability improves one's ability to profile a customer's requirements and to interpret the customer's reactions by requiring the collection of extensive data about a customer's demographics, interests, preferences, and behaviors.

In the fast-paced world of digital marketing, it is of the utmost necessity to have the capability to monitor and assess how successful advertisements are. Because of this, having the capability to trace the origin of digital advertisements is necessary. Traceability is the capability to follow and verify the path that an advertisement has travelled, beginning with its development, and continuing through its distribution until it is experienced by the final customer. Traceability can be defined as the ability to track and verify the journey that an advertisement has taken (Almahmoud et al., 2022). It provides marketers with an extremely transparent view of how their ads are performing, where those ads are being displayed, and how audiences are interacting with those ads.

Accountability and Openness to the Public It is necessary for advertisers to be aware of the locations in which their adverts are being displayed due to the expansion of programmatic advertising as well as the usage of a range of platforms and intermediates. Traceability ensures that advertisements are not displayed on websites that are inappropriate or harmful, so safeguarding the reputations of the brands that are associated with those adverts. The development of programmatic advertising in the ever-changing environment of digital advertising has completely transformed the manner in which decisions regarding the purchase of media are made. Programmatic advertising is a more effective technique to target specific audiences by utilizing algorithms and data to make decisions on ad placements in real time. This type of advertising is becoming increasingly popular. However, this level of efficiency does come with its fair share of difficulties. The ecosystem of digital advertising currently includes a vast number of platforms and middlemen, including ad exchanges, supply-side platforms (SSPs), and demand-side platforms (DSPs). The sight of where advertisements finally appear is frequently obscured because of the complex chain that is used to optimize the placement of advertisements (Almahmoud et al., 2019). This level of opacity can accidentally lead to advertisements being placed next to content that is deemed unsuitable or controversial, which poses substantial dangers to the reputation of the business as well as the confidence of consumers. For instance, a company that promotes family values might find that its advertisement is inadvertently presented alongside content that promotes extremism, which would result in unfavorable consumer perceptions. The importance of traceability becomes essential when attempting to traverse this complicated structure. Traceability technologies not only give advertisers insights into their ad placements, but they also make it possible for fast corrective steps to be taken in the event that bad placements are discovered. Traceability allows advertisers to make educated decisions about their affiliations by identifying platforms or intermediates that routinely result in unsuitable ad placements. This is accomplished by maintaining accountability throughout the programmatic chain, which enables traceability to identify these platforms or intermediaries. As the complexity of digital advertising continues to increase, the necessity for transparency, accountability, and traceability is becoming increasingly obvious. This ensures that brands can keep their honor in the digital sphere.

Optimization: The process of optimizing something. Advertisers can optimize their campaigns for a higher return on investment if they determine which kind of advertisements are doing the best and where those advertisements are being displayed. To accomplish this, monies need to be redistributed, adjustments need to be made to advertising creatives, and targeting criteria need to be modified in accordance with real-time data. Within the area of digital advertising, optimization is a fundamental principle that must be adhered to. At its foundation, optimization is concerned with the ongoing improvement of advertising campaigns in order to maximize both the effectiveness of those ads and, as a consequence, the return on investment (ROI). Advertisers currently have access to an incredible amount of data, which enables them to gain unique insights into the performance of their ads. Advertisers can determine which ads have the most resonance with their target audience and in which settings by doing an analysis of indicators such as click-through rates, conversion rates, and engagement levels. This understanding, which is powered by data, gives advertisers the ability to make educated judgements regarding their campaigns (Mensikova & Mattmann, 2018). For instance, if one piece of advertising creative outperforms the others on a constant basis, additional funds may be allotted in order to market it even further. On the other hand, advertisements that do not perform well can be modified or altered so that they are more in line with the preferences of the audience. Optimization encompasses more than just the creative aspect; it also includes the targeting parameters. Real-time data can indicate demographic or geographic groupings that are more sensitive to the campaign. This enables advertisers to either narrow down or broaden their target audience depending on how the data presents itself. In its most basic form, optimization is a dynamic process, in which campaigns do not remain unchanged but rather develop in response to ongoing feedback from real-time data. This iterative strategy guarantees that advertising efforts are not only effective but also provide the maximum potential return on investment (ROI), so maximizing the value of each dollar spent on advertising.

Fraud Prevention: The problem of fraudulent activity in digital advertising, which includes click fraud and impression fraud, is becoming a problem that is more widespread. Traceability technology can be used to assist in the detection of unusual patterns or behaviors, which helps to guarantee that advertisers are not paying for fraudulent traffic or activity that does not include humans. Traceability technologies can also be used to help ensure that advertisers are not paying for traffic that does not involve humans (Gabryel et al., 2022). One of the most significant challenges that has evolved in the complex web is digital advertising, known as digital ad fraud. This includes deceptive practices such as click fraud, in which automated bots or persons encouraged to generate phoney clicks, and impression fraud, in which advertisements are presented to non-human traffic or put in non-viewable areas, creating the illusion of more interaction. Both frauds are examples of click fraud. These kinds of fraudulent operations exaggerate performance data, leading advertisers astray into thinking their efforts are more successful than they actually are. Because marketers wind up paying for traffic that does not contain any real value, this results in lost ad spend and skewed campaign statistics, which both have substantial repercussions for the advertiser's bottom line. It is impossible to overestimate the significance of utilizing traceability technologies in the fight against ad fraud as this problem gets more widespread. The traffic patterns and user behaviors are analyzed in real time by these instruments, which are equipped with sophisticated algorithms and analytical skills. They are able to recognize and flag anomalies, such as abrupt surges in traffic from particular regions or repetitive click patterns, which are indicative of bot activity, since they are doing so (Gabryel et al., 2022). Once fraudulent traffic has been identified, these systems can automatically filter it out or warn advertisers about potential dangers. Traceability technologies, in their most basic form, play the role of a watchful sentinel, protecting advertisers from the dangers posed by digital advertising fraud. These solutions not only safeguard financial investments by guaranteeing that advertisers are only charged for genuine human contacts, but they also guarantee the honesty and precision of data relating to the performance of advertising campaigns.

Consumer Confidence: In an era when data privacy and transparency are two of the most critical concerns for customers, having the ability to monitor and validate ad experiences can assist establish trust in an industry. Customers are more likely to have a favorable interaction with commercials when they are aware that the companies, they frequent are trying to ensure that advertising practices are both ethical and transparent. It is more crucial than ever before for a business to be able to gain and keep the trust of its customers currently, when digital technology is so prevalent. As a direct result of recent high-profile instances involving the theft of personal information and the improper use of data, consumers have gained a heightened awareness regarding the collecting, storage, and use of their data. This is a direct outcome of recent high-profile incidents (Loyola-González et al., 2018). As people become more conscious of the importance of protecting their personal information, there has been an increase in the amount of attention paid to ensuring that advertising practices are open and honest. Customers are more concerned than ever before with obtaining comfort that the advertisements they view are not only relevant to the requirements they have, but that they are also produced and disseminated in an honest manner. This concern is addressed head-on by the capability to trace and verify the journey that an advertisement has taken, from the moment it was initially conceived of all the way through to the moment it was displayed for the last time. This ability allows for complete transparency regarding the journey that an advertisement has gone. By providing an open and honest blueprint of the route that an advertisement will take, companies may demonstrate that they are committed to the implementation of ethical advertising practices. Customers get the sense that companies not only respect their right to privacy but also take precautionary measures to ensure that advertising content is real and free from misleading business practices as a result of this openness, which in turn gives customers the impression that brands respect their right to privacy. When clients perceive these efforts, they have a more favorable view of the company, which in turn develops a sense of confidence and trustworthiness. Customers have a better sense of confidence in the goals and priorities of the brand as a direct result of this trust, which in turn improves the probability that they will have a good engagement with the commercials. To summarize, in a market where consumer trust is easily shattered and must be worked for, the capability to demonstrate ethical and transparent advertising practices through traceability becomes a crucial instrument for businesses to interact meaningfully with their audience. Traceability also becomes an essential instrument for businesses to demonstrate that they care about the environment and are committed to reducing their negative impact.

The capacity to trace the origin of digital advertisements is not a desirable additional function that may be selected, but rather, it is an indispensable prerequisite. Because the landscape of digital advertising is always evolving, there is an ever-increasing demand for advertising practices that are more accountable, transparent, and efficient. This trend is only expected to continue in the foreseeable future. The attempt to satisfy these requirements will prioritize the development of traceability tools and methodologies. This will ensure that digital advertising continues to be an effective medium that consumers and brands can trust in the future.

1.1.1 DIGITAL ADVERTISING AND ITS SIGNIFICANCE

Digital advertising, also known as online advertising or Internet advertising, refers to the practice of delivering promotional content to users through various digital channels. It has evolved significantly since the advent of the Internet, transforming from simple banner ads to a diverse range of formats including search engine marketing, social media advertising, video ads, and native advertising. Search Engine Marketing is a pivotal method in digital advertising focused on enhancing a website's visibility in search engine results pages. This is primarily achieved through paid advertising, where advertisers bid on keywords that users of services such as Google might enter when looking for certain products or services. Google AdWords is one of the most widely used platforms, offering businesses the opportunity to appear in the sponsored section of search results, thus increasing the likelihood of user visits. SEM is crucial for businesses aiming to increase their online presence and drive targeted traffic to their websites. With the proliferation of social media platforms, advertising on these networks has become indispensable. Platforms such as Facebook, Instagram, and Twitter provide robust targeting options, allowing businesses to reach specific demographics based on factors like age, gender, location, interests, and behaviors (Vanhuele, 2021). Social media advertising is highly effective due to its ability to engage with users in a space where they are already active and engaged. It also provides unique formats like sponsored posts and stories, which can feel less intrusive and more integrated into the user experience.

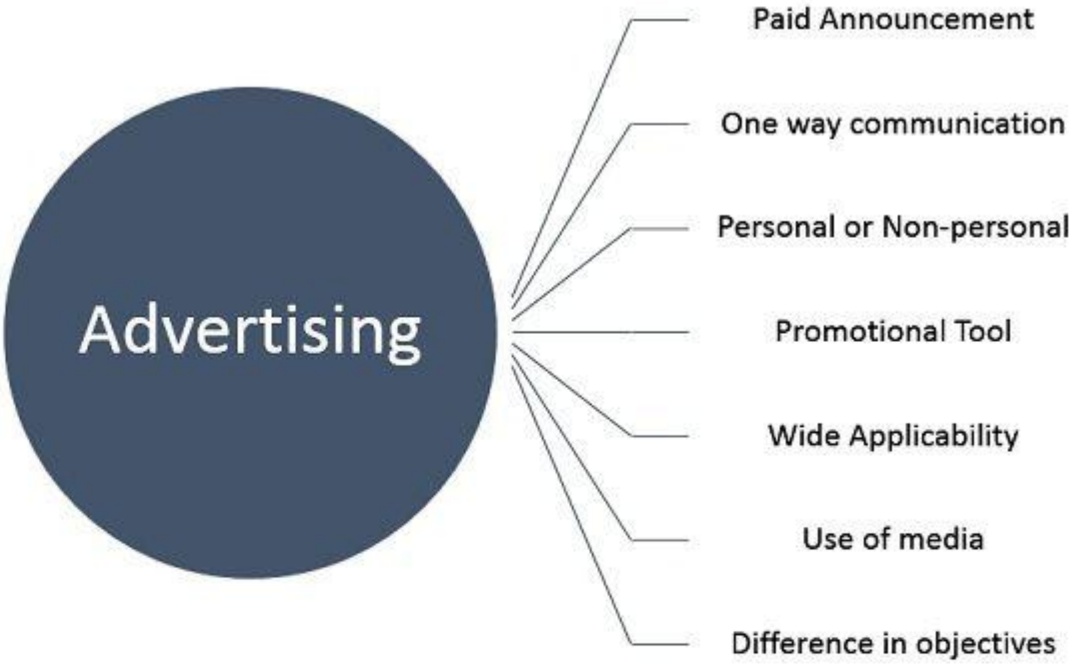


Figure 1: Types of Advertising [Source: https://businessjargons.com/advertising.html#google_vignette]

Advertising, which is an essential component of contemporary marketing tactics, comprises a wide range of activities in which marketers pay for space or time to communicate their messages. Through the use of this paid announcement system, companies are able to develop messages that are specific to their needs and choose the most appropriate media channels for dissemination, all while afterward evaluating the impact of their efforts. Advertising is primarily a communication medium that operates in a one-way fashion, meaning that the message is conveyed from the company to the consumer without the consumer providing any direct feedback (Ha et al., 2021). In the past, channels such as television, radio, and newspapers made it possible to take a non-personal approach by distributing messages to a large audience with limited room for customization. The advent of the digital revolution, on the other hand, has made interaction with advertisements a more personal experience. The proliferation of digital platforms, particularly social media, content-based marketing, and cookie-based targeting has made it possible for advertisers to design highly tailored messages that engage individual customers based on their tastes and behavior while using the internet.

Advertising is a versatile instrument that is part of the promotional mix. It is frequently used to increase sales, improve brand visibility, and accomplish a variety of organizational goals. Its purpose extends beyond simply marketing items or services. Its relevance extends far beyond the realm of business, serving as an important tool for political campaigns, charity organizations, educational institutions, tourism marketing, and public awareness projects carried out by government agencies. One of the most significant issues in advertising is choosing the appropriate medium in order to successfully communicate with the intended audience (Sadeghpour & Vlajic, 2021). There is a wide variety of options available, including from more conventional forms of media such as television and print to more contemporary digital channels such as the internet and social media. The advertising message's impact is impacted differently depending on the media because each medium has its own set of advantages and limitations. Furthermore, the goals that are being pursued by various organizations in their advertising endeavors are very diverse from one another. The diverse nature of advertising in today's culture is shown by the fact that some of its goals include boosting sales or enhancing the image of the business, while others concentrate on establishing public relations, developing markets for new products, combating competition, or educating the general public.

Display advertising encompasses visual ads that appear on websites in various formats, such as banners, sidebars, or pop-ups. These ads can include text, images, video, or interactive content aimed at attracting attention and encouraging user interaction. Display ads are often used for brand awareness and can be targeted to appear on specific websites, to particular audiences, or in certain contexts. Programmatic advertising, a form of display advertising, uses algorithms and automated processes to buy and place ads in real-time, optimizing ad performance and targeting. Video advertising includes ads that are either standalone videos or integrated within streaming video content. These ads can appear before (pre-roll), during (mid-roll), or after (post-roll) video content, and they offer a compelling way to convey a message and engage with viewers. Video ads are prevalent on platforms like YouTube, social media, and various streaming services (Gordon et al., 2021). They can be particularly effective due to the visual and auditory elements, capturing attention and creating a memorable user experience.

Native advertising involves ads that are designed to blend seamlessly with the content on a webpage, providing a less intrusive and more natural user experience. These ads mimic the look and feel of the content surrounding them, making them feel like a part of the content itself. Native ads can appear as sponsored content on news websites, recommended articles, or sponsored posts on social media. The key to successful native advertising is creating content that is relevant and valuable to the audience, ensuring a positive user experience.

Email marketing involves sending promotional content directly to users' email inboxes. This form of advertising allows businesses to communicate directly with their audience, providing personalized and relevant content (Oldham et al., 2023). Email marketing can be used for various purposes, including promoting products or services, nurturing leads, or engaging with customers. It requires a well-maintained email list and compelling content to be effective, ensuring that emails are opened and acted upon. Content marketing is the practice of creating and sharing online material, such as blogs, videos, and social media posts, that does not explicitly promote a brand but is intended to stimulate interest in its products or services. This form of advertising focuses on building a relationship with the audience, providing value through informative or entertaining content. Content marketing is a long-term strategy aimed at establishing trust and authority, ultimately leading to increased brand loyalty and customer conversions.

1.1.2 EVOLUTION OF DIGITAL ADVERTISING

The digital advertising landscape has undergone a remarkable transformation since its inception, evolving in tandem with technological advancements and changing consumer behaviors. What began as simple banner ads on web pages has blossomed into a multifaceted field, leveraging data, automation, and creativity to connect with audiences in increasingly sophisticated ways. In the mid-1990s, the world was introduced to digital advertising through rudimentary banner ads. These ads, while basic in design and functionality, marked the beginning of a new era in advertising, opening doors for brands to reach audiences through the burgeoning World Wide Web. The simplicity of these early ads belied the potential of digital channels, setting the stage for the rapid evolution that would follow.

As the new millennium dawned, search engines came to the forefront of the digital space, introducing the advertising world to the pay-per-click model. Google AdWords, launched in 2000, revolutionized digital advertising by allowing advertisers to bid on keywords, connecting with potential customers at the critical moment of search. This era was characterized by increased targeting and personalization, as cookies and behavioral tracking technologies enabled advertisers to tailor their messages to specific audience segments, enhancing relevance and effectiveness (Jastrzębska et al., 2023). The advent of social media platforms in the late 2000s and early 2010s marked another pivotal moment in the evolution of digital advertising. Platforms like Facebook and Twitter offered advertisers unprecedented access to user data, enabling highly targeted and personalized campaigns. The mobile revolution further expanded the digital advertising landscape, as smartphones became ubiquitous, necessitating the creation of mobile-optimized ads and strategies.

In recent years, the industry has embraced automation and emerging technologies, with programmatic advertising becoming the norm. Real-time bidding and algorithm-driven ad placement have made ad buying more efficient, while artificial intelligence and machine learning are being leveraged to optimize ad performance and enhance personalization.

Augmented and virtual reality are introducing new, immersive ways for brands to connect with consumers, offering interactive and engaging ad experiences. As we stand on the cusp of a new decade, the digital advertising industry continues to innovate and adapt, driven by a relentless pursuit of relevance and connection (Cheng et al., 2009). From its humble beginnings to its current state of sophistication, the evolution of digital advertising reflects a journey of constant learning and adaptation. The future promises even more innovation, as emerging technologies and shifting consumer expectations push the industry towards new frontiers, ensuring that digital advertising remains an integral and evolving part of the marketing landscape.

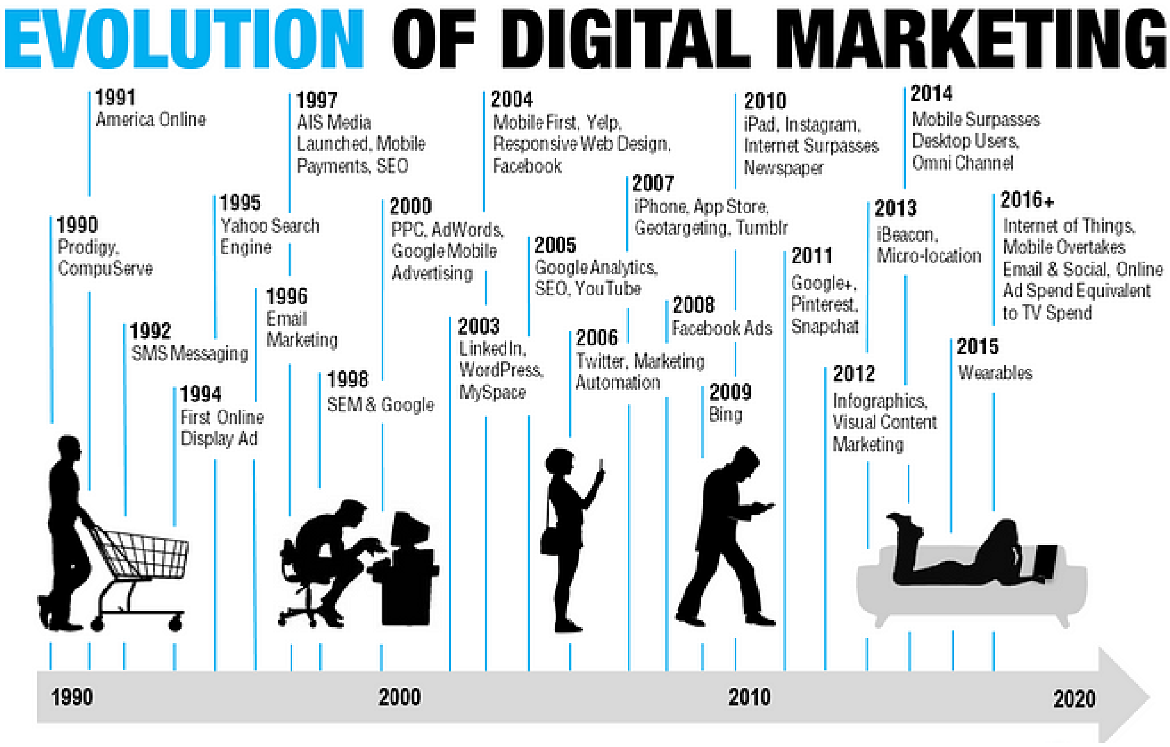

EVOLUTION OF DIGITAL MARKETING
1990 Prodigy, CompuServe
1991 America Online
1992 SMS Messaging
1994 First Online Display Ad
1995 Yahoo Search Engine
1996 Email Marketing
1997 AIS Media Launched, Mobile Payments, SEO
1998 SEM & Google
2000 PPC, AdWords, Google Mobile Advertising
2003 LinkedIn, WordPress, MySpace
2004 Mobile First, Yelp, Responsive Web Design, Facebook
2005 Google Analytics, SEO, YouTube
2006 Twitter, Marketing Automation
2007 iPhone, App Store, Geotargeting, Tumblr
2008 Facebook Ads
2009 Bing
2010 iPad, Instagram, Internet Surpasses Newspaper
2011 Google+, Pinterest, Snapchat
2012 Infographics, Visual Content Marketing
2013 iBeacon, Micro-location
2014 Mobile Surpasses Desktop Users, Omni Channel
2015 Wearables
2016+ Internet of Things, Mobile Overtakes Email & Social, Online Ad Spend Equivalent to TV Spend
1990
2000
2010
2020

Figure 2: Digital Marketing Evolution [Source: https://www.linkedin.com/pulse/evolution-digital-marketing-navigating-changing-khushi-prajapati-uljie]

Since the early days of internet pioneers such as Prodigy, AOL, and CompuServe, the landscape of digital marketing has experienced a tremendous transition for the better. Over the past few years, this change has accelerated, which has had a significant impact on corporate strategy all over the world. During my meetings with a large number of executives in the field of digital marketing, particularly those working for Fortune 100 organizations, I have gained insights into the most recent trends and organizational difficulties that are prevalent in this sector. The majority of these executives recognize that marketing team structures have greatly improved, which is a positive sign that there has been a change toward digital marketing strategies that are more successful and dynamic. In spite of this, an astounding 89 percent of respondents have expressed persistent concerns regarding the lack of internal support for digital teams, particularly from crucial areas such as sales and information technology (Jain et al., 2016). When it comes to establishing digital leadership roles inside businesses, there is an urgent need for clarity. This clarification is necessary to ensure that these responsibilities are aligned with the larger executive vision and objectives.

Digital marketing teams often report to the Chief Marketing Officer (CMO) in many businesses. However, some digital marketing teams fall under the purview of the Vice President of Marketing, who may in turn report to the Chief Information Officer (CIO). Regardless of the organizational structure, it is of the utmost importance to align digital teams with leaders who are able to strike a balance between the marketing imperatives of speed, originality, and customer focus and the IT requirements of stability and security. Digital leaders of today are expected to have the ability to effectively interact across a variety of departments, including the ability to forge alliances and drive both tactical excellence and creative innovation. This is a substantial difference from the past, when digital marketing teams frequently functioned in a reactive manner, similar to tactical "copy shops," and lacked the ability to begin original ideas or drive innovation beyond their immediate domain.

The transformation of digital advertising through technology has fundamentally altered the methods by which companies understand and predict consumer purchasing behaviors, including their preferences, timing, reasons for purchase, and likelihood to recommend or repurchase. Firms are now equipped with concrete data and insights rather than relying on assumptions about consumer intentions. The balance of power in the marketing realm has shifted, requiring companies to navigate not only competitive pressures but also the demands of empowered consumers with fleeting attention spans. This raises critical questions about the future trajectory of advertising, the strategies that marketing managers can employ to optimize returns on advertising investments, and the areas in which academic research can contribute to enhancing the efficiency and effectiveness of advertising strategies.

The integration of technology and the advent of big data have irrevocably transformed the marketing landscape, eroded geographical market boundaries and shifted the focus from selling to engaging with consumers (Taylor, 2009). Today's consumers are more connected, informed, and discerning, with a plethora of choices at their fingertips. The transition from a one-sided advertising approach to a bilateral dialogue between brands and consumers has placed control in the hands of consumers. Marketing managers, in this new environment, must resist the urge to reclaim control and should instead foster a sense of empowerment and engagement among consumers, taking the opportunity to learn and adapt in the process. This study introduces a comprehensive framework that encapsulates the myriad of factors influencing digital advertising, ranging from customer and firm characteristics to technological infrastructure, data resources, and contextual elements such as product and customer life cycles. This framework serves as a strategic guide for marketing managers in crafting advertising content, selecting appropriate media channels, and refining messaging and targeting tactics.

In the evolving landscape of digital advertising, success will be determined by the degree to which advertisers can align their strategies with the shifting trends in consumer behavior and market dynamics. The future promises to favor those advertisers who succeed in empowering and engaging consumers through communications that are not only targeted and reliable but also treat consumers as valued partners in a reciprocal relationship.

1.2 STATEMENT OF THE PROBLEM

Conversions are discrete activities carried out by users, such as clicking on an advertisement, registering for a newsletter, or making a purchase. Examples of conversions include these. Conversion rates are typically used as a yardstick to evaluate the success of advertising initiatives in the rapidly developing arena of digital marketing. This is because conversions are treated as separate operations undertaken by users. Despite this, distinguishing between real conversions and fraudulent ones has become an extraordinarily challenging process because fraudulent operations are growing more sophisticated all the time. Conversion fraud, in which fraudulent or non-human acts are counted as genuine conversions, can substantially skew performance indicators, which can eventually lead to increased advertising expenses and erroneous insights into the efficacy of a campaign. In this type of fraud, fraudulent or non-human acts are counted as valid conversions. Because of the potential for conversion fraud to have both financial and strategic ramifications, there is an urgent and pressing need for robust detection systems. This requirement must be met immediately.

The dataset that is being considered provides an all-encompassing viewpoint of user interactions, beginning with the very first ad impression and going all the way up to the very last conversion activity. This may be seen from the user's point of view. By combining information from click logs and conversion logs, this dataset offers a fine-grained perspective on the path taken by users. It records necessary parameters such as the advertiser, the publisher, the geographical location, the sort of device, and plenty more. The availability of such a large quantity of data brings with it the possibility of uncovering patterns and anomalies that lead to the existence of fraudulent conduct.

1.3 SIGNIFICANCE IN THE MODERN WORLD

Traceability and the avoidance of fraud in digital advertising are of such critical importance in today's modern digital environment that their importance simply cannot be emphasized. The development of the internet and the proliferation of digital channels have brought about a revolution in the manner in which businesses market the goods and services they offer, but in doing so, they have also brought about a host of new difficulties and exposures. The prevalence of fraud, which can present itself in a variety of forms such as click fraud, ad fraud, or fraudulent transactions, is one of the most significant problems in this industry.

The significant monetary repercussions that fraud in digital advertising may have for a company serve to highlight the need of taking action to combat the problem. Every year, fraudulent actions result in the loss of millions of dollars, which lowers the return on investment for advertisers and undermines the integrity of the ecosystems in which digital advertising takes place (Wuisan & Handra, 2023). In addition, the presence of fraud can harm the reputation of a brand, since customers may link the brand with dishonest practices or low-quality items if they believe the brand to be involved in the fraudulent activity.

In the fight against fraud in digital advertising, traceability is an extremely important factor. Businesses are better able to detect fraudulent actions and take appropriate action against them when they make every effort to ensure that every facet of an advertising campaign is both trackable and transparent. Traceability helps marketers to check the authenticity of clicks, impressions, and transactions. This ensures that advertisers are paying for legitimate engagements rather than falling prey to fraudulent schemes and that advertisers do not waste money. This not only safeguards the monetary interests of advertisers but also improves both the efficacy and efficiency of advertising efforts. Traceability, in addition to the important part it plays in the prevention of fraud, also makes a contribution to the improvement of digital advertising tactics. Traceability enables advertisers to make decisions based on data, enhance their targeting methods, and increase the total return on investment of their advertising spend by offering granular insights into user interactions and campaign effectiveness. Traceability may be achieved by collecting and analyzing user data. This level of knowledge and control is especially important in the digital world, which is characterized by the fragmentation of customer attention and the intense rivalry for visibility.

In addition, as customers become increasingly concerned about their privacy and the safety of their personal information, the need of traceability will only grow. Traceability ensures that advertisers are adhering to high standards of integrity and transparency and is crucial for developing trust with customers. Transparent and responsible advertising practices are essential for building trust with consumers. Not only does this aid in retaining the trust of customers, but it also positions the business as a responsible and ethical participant in the digital environment.

The importance of preventing fraud and ensuring traceability in digital advertising cannot be emphasized in today's society. It is of the utmost importance to ensure the openness, integrity, and efficiency of advertising campaigns through traceability as the use of digital advertising continues to advance and assume an increasingly important role in the development of corporate plans. Not only does this protect the financial interests of advertisers, but it also improves the entire health and integrity of the digital advertising ecosystem. As a result, this is ultimately beneficial to both businesses and customers.

1.4 RESEARCH QUESTIONS

1.4.1 DIGITAL AD TRANSACTION CATEGORIZATION:

- How are digital advertising transactions categorized in terms of legitimacy?
- How has the volume of digital advertising transactions, categorized by legitimacy, evolved over time within available datasets?

1.4.2 SIGNIFICANCE OF FEATURES IN AD FRAUD DETECTION:

- Which specific features of digital ad transactions serve as the most significant indicators of fraudulent activity?
- Are there particular patterns or features that are predominantly associated with fraudulent digital advertising transactions?

1.4.3 NETWORK PATTERNS IN FRAUDULENT ADVERTISING:

- What network patterns are commonly observed in fraudulent digital advertising transactions, and how do they contrast with legitimate transactions?
- In the context of digital ad fraud, are fraudulent networks interconnected compared to their legitimate counterparts?

1.4.4 EFFECTIVENESS OF PREDICTIVE MODELS IN AD FRAUD:

- How effective is the chosen predictive model, such as a neural network, in identifying fraudulent digital advertising transactions, and what are its key performance metrics?
- How does the predictive model address class imbalance inherent in datasets of digital advertising transactions?

1.4.5 CLASSIFICATION AND ANALYSIS OF AMBIGUOUS AD TRANSACTIONS:

- After modeling, how are ambiguous digital advertising transactions classified between legitimate and fraudulent categories?
- How do these newly classified transactions fit into the larger network of digital ad transactions, and what insights do they offer?

1.4.6 IMPROVEMENTS AND ALTERNATIVE APPROACHES IN FRAUD DETECTION MODELS:

- What methods can be employed to improve the performance of the existing fraud detection models for digital advertising?
- How do alternative models, such as deep learning or ensemble methods, compare to the existing model in detecting digital ad fraud?

1.4.7 STRATEGIC IMPLICATIONS FOR MARKETING AND ADVERTISING FIRMS:

- Based on the findings, how can marketing and advertising firms strengthen their strategies against digital ad fraud?
- What recommendations can be made to industry regulators and policymakers based on the insights from the research?

1.5 OBJECTIVE OF THE RESEARCH

The principal objective of this study is to dissect and elevate the understanding of traceability systems within digital advertising, providing a detailed examination of their role in identifying and mitigating fraudulent activities. Through comprehensive analysis, this research will aim to bridge gaps in current practices and suggest enhancements to traceability measures, thereby underpinning the authenticity and reliability of digital advertising. The initial objective involves meticulous classification of digital advertising transactions. By systematically categorizing these transactions into legitimate, fraudulent, and indeterminate based on available traceability data, the study will lay the groundwork for understanding the scope and impact of fraud within digital advertising networks. This classification is crucial for setting the stage for subsequent analysis and for establishing benchmarks against which the success of traceability mechanisms can be measured.

A pivotal objective of the research is to pinpoint key indicators and patterns that signal fraudulent activities within digital ad transactions. This will involve a granular analysis of transactional data to isolate features that are consistently associated with fraud. The purpose here is not only to enhance the detection of fraudulent activities but also to contribute to the refinement of traceability tools and methodologies that can preemptively identify and flag potential fraud.

This study aims to unravel the network structures and patterns that are characteristic of fraudulent digital advertising activities. By comparing these with the networks of legitimate transactions, the research intends to uncover distinctive traits and behaviors inherent to fraudulent networks. Understanding these differences is vital for developing predictive models and for crafting strategies that can mitigate the risk of fraud in digital advertising. A significant objective is to critically evaluate the efficacy of predictive models, such as machine learning algorithms, in accurately identifying fraudulent transactions within vast datasets of digital advertising. This part of the research will scrutinize the performance metrics of these models, with a keen focus on their precision, recall, and overall reliability. The goal is to assess the current state of fraud detection capabilities and to identify potential enhancements to increase the robustness of these predictive systems. The research further aims to apply advanced predictive models to classify transactions that are ambiguous in nature. The analysis of these transactions, post-classification, will provide deeper insight into their characteristics and how they integrate into the larger network of digital ad transactions. This objective is critical for reducing the uncertainties within digital advertising and for bolstering the overall traceability framework. An additional objective is to explore and test strategies for improving the performance of fraud detection models. Whether through data augmentation, algorithmic refinement, or the introduction of novel computational techniques, this study seeks to push the boundaries of current models. The intention is to enhance the predictive accuracy of these models, thereby supporting more secure and effective digital advertising operations. the research aspires to offer strategic insights and actionable recommendations to marketing firms and industry regulators. By translating the findings into practical strategies, the research will provide a blueprint for industry stakeholders to strengthen their defenses against digital ad fraud. These recommendations will be geared towards the enhancement of traceability and the fostering of an advertising ecosystem that is both secure and conducive to genuine marketing engagements.

Encompassing these specific objectives, the overarching goal of this research is to holistically address the challenges of traceability and fraud within digital advertising. By advancing the understanding of these critical issues and proposing actionable solutions, the research aims to contribute significantly to the field of digital marketing, ensuring that advertising practices are transparent, accountable, and ultimately more effective in reaching and engaging consumers in the digital age.

1.6 LIMITATIONS, DELIMITATIONS, AND ASSUMPTIONS

This research encounters several limitations that may influence the breadth and depth of the findings. Firstly, the sensitivity and proprietary nature of data related to digital advertising transactions often limit data availability and accessibility. This scarcity of comprehensive datasets could potentially restrict the analysis of fraud patterns and the effectiveness of traceability mechanisms. Secondly, the digital advertising landscape is characterized by rapid technological advancements. The transient nature of the tactics employed by both fraudsters and those combating fraud means that the findings of this research may have a limited temporal applicability. Lastly, the study’s results are derived from specific datasets and contexts, which may not be universally representative. The heterogeneity of digital platforms, regulatory environments, and user behavior across different markets could limit the generalizability of the conclusions drawn. One of the primary limitations of this research is the availability and quality of data on digital advertising transactions. Due to the proprietary nature of advertising data, there may be significant constraints on the researcher's ability to access comprehensive datasets that are representative of the entire digital advertising landscape. The data that is accessible might not be exhaustive and could be biased towards publicly reported instances of fraud or platforms willing to share information. Additionally, the data may lack granularity, limiting the ability to draw nuanced conclusions about the nature and context of fraudulent activities. The digital advertising industry is subject to rapid changes in technology, algorithms, and tactics used by both marketers and fraudsters. This dynamic landscape means that the tools and methods for traceability and fraud detection are constantly evolving (Aslam & Karjaluoto, 2017). A limitation of this study is that the analysis is based on a snapshot of technology and practices that may become outdated quickly. The findings may not account for future innovations or changes in digital advertising practices that could occur after the research has been conducted.

The findings from this research are based on specific datasets and may not be generalizable to all forms of digital advertising or across different platforms and geographic regions. The diversity of advertising models, audience demographics, and regional regulations can vary widely, potentially affecting the prevalence and detection of fraud. As such, the results of this study may not be applicable to all settings or may require adaptation to fit different contexts. The research methodologies employed, such as the choice of predictive models and analytical frameworks, come with inherent limitations. These methods are chosen based on current best practices, but they may not capture all dimensions of digital ad fraud or traceability. Furthermore, the study may rely on assumptions and parameters that are necessary for model construction but may not fully encapsulate the complexities of real-world behaviors.

The digital advertising market operates within a regulatory framework that is subject to change. New laws and regulations could be introduced that would significantly alter the practices around data sharing, privacy, and fraud detection (Cheong et al., 2010). This research is limited by the current regulatory environment and may not anticipate future legal developments that could impact the effectiveness of traceability measures.

The study will specifically address certain types of fraud within digital advertising, such as click fraud, impression fraud, and ad misplacement. This focus allows for a more thorough examination of these prevalent issues while acknowledging that the research will not cover all possible manifestations of fraud in the digital advertising space. The types of fraud selected represent those with significant impact on the industry and for which traceability can be studied in a meaningful way. The analysis will be conducted using a predetermined set of analytical techniques and predictive models chosen for their relevance, accuracy, and practicality within the context of the available data and the research objectives. This includes the use of specific machine learning algorithms or statistical methods that have been identified as effective for this type of research. Other methods, while potentially valuable, are excluded to maintain a focused approach and to ensure that the study remains within the planned scope. The research will delimit its geographical focus to regions or platforms where comprehensive data is available and where digital advertising fraud is particularly pressing. This geographical delimitation is necessary to ensure that the findings are based on robust data and are relevant to the markets in question. Additionally, focusing on specific platforms (e.g., social media, search engines, programmatic advertising networks) allows for a more detailed understanding of traceability and fraud mechanisms within those contexts (Hudders et al., 2019). The study will examine data from a defined timeframe, which delimits the research to a specific period in which the digital advertising industry operated under certain technological and regulatory conditions. This temporal boundary is necessary to complete the research within a realistic timeframe and to provide a snapshot of the state of traceability and fraud within that period. The research will be grounded in specific theoretical and conceptual frameworks that guide the study of traceability and fraud in digital advertising. While other frameworks could provide alternative insights, the selected frameworks are chosen for their direct relevance to the research questions and objectives. This delimitation ensures that the study has a clear conceptual direction and that the findings can be interpreted within an established academic context.

The research assumes a relative stability in the digital advertising ecosystems during the period of study. This includes the constancy of technological platforms, the behavior of users, and the tactics employed by fraudsters. Although these elements are inherently dynamic, assuming their stability is crucial for analyzing patterns and developing models that are predicated on historical data. There is an underlying assumption that the data used for analyzing digital advertising fraud is accurate, reliable, and consistent. The study presumes that the datasets have been collected and curated with rigorous standards, ensuring that the conclusions drawn are based on valid information. This assumption is foundational, as any anomalies or inaccuracies in the data could significantly skew the results and undermine the validity of the study’s conclusions. The research is premised on the assumption that existing traceability measures and anti-fraud mechanisms are effective to a certain degree (Fuxman et al., 2014). This does not imply that these measures are infallible but rather that they serve as a reasonable starting point for understanding and improving upon the state of digital advertising fraud detection.

While the study may focus on specific types of fraud or platforms, it assumes that the findings will have a degree of generalizability to other forms of digital advertising fraud and across various platforms. This assumption is necessary for the research to have broader implications and for its recommendations to be applicable to a wider segment of the digital advertising industry. The research assumes that all stakeholders involved in digital advertising operate within the bounds of existing regulatory and ethical guidelines. It is presumed that the data is collected, shared, and used in compliance with privacy laws and industry standards, which is essential for conducting the research in a legally and ethically sound manner. An assumption is made regarding the adaptability of both fraudsters and advertisers. It is presumed that fraudsters will continue to evolve their tactics in response to enhanced traceability measures, and conversely, that advertisers and platforms will adapt their anti-fraud measures in response to new types of fraud. This assumption acknowledges the cat-and-mouse dynamic that is characteristic of the digital advertising landscape.

1.7 DEFINITION OF TERMS

- Digital Advertising: The practice of delivering promotional content to users through various online and digital channels, including search engines, websites, social media platforms, email, and mobile apps.

- Fraud Detection: The process of identifying fraudulent activities or transactions. In the context of digital advertising, it refers to the identification of fake clicks, impressions, or any fraudulent activities intended to inflate advertising costs or generate revenue illegitimately.
- Traceability: The ability to trace the origin, distribution, and location of products or transactions through recorded identification. In digital advertising, it refers to the capability to track the journey of an advertisement from its source to its final engagement point with users.
- Machine Learning (ML): A branch of artificial intelligence (AI) that focuses on building systems that learn from data. In the context of fraud detection, ML algorithms analyze patterns and anomalies in advertising data to identify potential fraud.
- XGBoost (Extreme Gradient Boosting): An advanced implementation of gradient boosting algorithms, known for its speed and performance. It is widely used in machine learning for classification and regression tasks, including fraud detection in digital advertising.
- Classification Metrics: Measures used to evaluate the performance of classification models in machine learning. Common metrics include accuracy, precision, recall, F1 score, and ROC-AUC, which help in assessing the effectiveness of a model in distinguishing between different classes, such as fraudulent and non-fraudulent activities.
- Imbalanced Dataset: A dataset in which the number of observations in one class significantly outweighs the number of observations in one or more other classes. This is common in fraud detection, where fraudulent transactions are typically much rarer than legitimate ones.
- Feature Engineering: The process of using domain knowledge to extract and select relevant features from raw data that make machine learning algorithms

work. In digital advertising fraud detection, this involves identifying which characteristics of the data are most indicative of fraudulent behavior.

- ROC-AUC Score (Receiver Operating Characteristic - Area Under Curve) : A performance measurement for classification problems at various threshold settings. ROC is a probability curve, and AUC represents the degree of separability. It tells how much the model is capable of distinguishing between classes.
- Model Overfitting : A modeling error that occurs when a machine learning model is too closely fitted to the training data, making it perform poorly on unseen data due to its inability to generalize from the training set to new inputs.

1.8 BACKGROUND

Digital advertising has become a cornerstone of modern marketing strategies, leveraging the internet to deliver promotional messages to consumers. The ubiquity of digital platforms has transformed advertising from a one-to-many broadcast model to a highly personalized and interactive experience. As consumers spend more time online across various devices, advertisers have capitalized on the wealth of data generated to target audiences with unprecedented precision. With the growth of digital advertising, however, has come the proliferation of fraudulent activities. Fraud in digital advertising encompasses a range of deceptive practices aimed at siphoning off advertising budgets. This includes bots generating fake clicks, fraudulent impressions, ad misplacement, and pixel stuffing, among others. These illicit activities not only result in significant financial losses for businesses but also skew marketing analytics, leading to poor decision-making and loss of trust in digital platforms (Shanahan & Kurra, 2011). The complexity of the digital advertising supply chain, with its numerous intermediaries and opaque transactions, poses a significant challenge for traceability. Identifying the source and nature of fraud within this labyrinthine system is a daunting task for advertisers and platforms alike. The need for sophisticated traceability mechanisms is therefore critical to ensure accountability and transparency in digital advertising transactions.

Advancements in technology have given rise to a variety of tools and techniques for combating digital advertising fraud. Machine learning algorithms, blockchain technology, and advanced analytics have emerged as potent weapons in the fight against fraud. These technologies promise to enhance the traceability of digital ad transactions, making it possible to track the journey of an ad from placement to impression, and verify its legitimacy. The impact of digital advertising fraud extends beyond financial losses. It undermines the integrity of the digital ecosystem, erodes consumer trust, and distorts the analytics that businesses rely on for strategic decision-making. The long-term sustainability of the digital advertising model is contingent upon the industry's ability to curtail fraudulent practices and establish robust traceability mechanisms.

In response to the escalating issue of ad fraud, industry bodies and regulators have begun to establish standards and frameworks aimed at combating these practices. Initiatives like the Interactive Advertising Bureau’s (IAB) Ads.txt project and the Trustworthy Accountability Group (TAG) certification program are examples of efforts to enhance transparency. However, the regulatory landscape remains fragmented, and enforcement is challenging in the cross-border digital realm. Despite these efforts, there remains a significant research gap in understanding the efficacy of current traceability measures and in developing new methodologies to detect and prevent fraud. This thesis aims to address this gap by examining the current state of traceability in digital advertising, identifying weaknesses in existing frameworks, and proposing solutions that could fortify the industry against fraud. The digital advertising industry has undergone a remarkable transformation since the advent of the internet. Early forms of online advertising were simple banner ads, but as the internet became more sophisticated, so did the mechanisms for delivering ads (Rovetta et al., 2020). The introduction of search engine marketing, social media advertising, and programmatic buying revolutionized the way advertisers reached consumers. These technological advancements allowed for more targeted, efficient, and measurable ad campaigns, leading to exponential growth in the industry.

Parallel to the growth of digital advertising was the emergence of fraudulent activities designed to exploit the burgeoning online ad market. Fraudsters quickly learned to game the system, employing bots to mimic human behavior, creating fake websites to collect ad revenue, and developing sophisticated schemes to launder money through the ad ecosystem. The complexity and anonymity of the digital realm made it a fertile ground for such illicit activities, prompting the industry to recognize the urgent need for effective traceability measures. Digital advertising involves a complex web of players, including advertisers, publishers, ad exchanges, and various intermediaries (Uyyala, 2021). This complexity obscures the path from ad dollars spent to outcomes achieved, creating opportunities for fraud. The lack of a single, unified system to track and verify ad transactions exacerbates this issue, allowing fraud to flourish undetected in the shadows of the digital advertising supply chain. In response to the threat of ad fraud, the industry has turned to technology for solutions. The use of artificial intelligence, machine learning, and blockchain has shown promise in enhancing traceability and combating fraud. These technologies can analyze vast amounts of data to identify patterns indicative of fraudulent behavior, verify the authenticity of ad impressions, and create transparent and immutable records of ad transactions. However, as these technologies evolve, so too do the tactics of fraudsters, leading to a constant arms race between the two sides. The ramifications of digital advertising fraud are not limited to economic losses. They permeate the social fabric by undermining the credibility of online content and eroding user trust (Nagaraja & Shah, 2019). When users are bombarded with ads resulting from fraudulent activity, their dissatisfaction with the digital experience increases, potentially leading to a broader rejection of online advertising. The societal cost of fraud, therefore, extends to a deterioration in the quality of online discourse and a weakening of the digital economy's foundations.

Regulatory bodies and industry groups have undertaken significant efforts to address digital advertising fraud. Legislation at various governmental levels has sought to protect consumers and businesses, while industry initiatives have focused on creating standards for transparency and verification. However, these efforts face challenges in enforcement due to the borderless nature of the internet and the continuous innovation in fraudulent tactics. This research is predicated on the observation that despite significant efforts to curb digital advertising fraud, it remains a pervasive and evolving threat. The current measures for traceability and fraud detection are inadequate, as evidenced by ongoing losses and the persistence of fraud. This thesis seeks to fill the knowledge gap in understanding the limitations of existing anti-fraud technologies and methods, and to propose a framework for enhancing traceability and accountability within the digital advertising ecosystem.

The genesis of digital advertising heralded a new epoch in the annals of commerce, transforming the traditional marketplace into a dynamic digital bazaar. As the internet wove itself into the fabric of daily life, it unfurled vast canvases for marketers to paint their messages, targeting consumers with unprecedented precision. Yet, this digital renaissance also gave birth to a shadowy nemesis: fraud. It crept into the burgeoning online ad spaces, exploiting the very algorithms and data streams designed to nurture genuine engagement. The complexity of the digital ad supply chain, with its intricate network of exchanges and intermediaries, compounded the conundrum, obscuring the footprints of authentic transactions amidst the digital morass. As billions of advertising dollars began to bleed into the abyss of deceitful clicks and phantom impressions, the industry's initial skirmishes with ad fraud evolved into a full-scale crusade for traceability. The quest for verifiable advertising pathways became both a shield to protect the industry's integrity and a sword to cut through the Gordian knot of fraud (Baig & Reddy, 2020). Despite concerted efforts and the deployment of advanced technological sentinels, the specter of fraud persisted, deftly adapting to each new safeguard, perpetually challenging the guardians of digital trust.

Confronted with an adversary that threatened the very foundations of the digital marketplace, the response was a clarion call for a unified front. Regulatory bodies and industry coalitions marshaled their resources, crafting frameworks and standards to stem the tide of fraud. Yet, the elusive nature of digital deceit, coupled with the borderless realm of the internet, presented a herculean challenge to these sentinels of cyberspace. It is within this tumultuous landscape that this thesis carves its niche, seeking not just to understand the intricacies of digital advertising fraud but to illuminate a path forward. Through rigorous analysis and the proposal of innovative traceability solutions, this research endeavors to restore equilibrium to the digital advertising domain, ensuring that the narrative of digital marketing is authored by veracity, not veiled by vice.

In the early days of the internet, digital advertising was a shining example of innovation. It was a messenger of unimaginable opportunities that could tap into the growing potential of online connectivity. It offered a level of involvement that was unattainable with traditional media and promised to revolutionize the way in which companies and customers communicate with one another (Singh et al., 2023). In tandem with the growth of digital footprints, the advertising landscape underwent a metamorphosis, becoming a complicated web of data-driven interactivity and analytics. While retailers were ecstatic about their newly acquired capacity to track the impact of each and every dollar spent, customers benefited from a more individualized and enjoyable shopping experience. However, the golden age of the ascendancy of digital advertising was overshadowed by the advent of a strong foe: fraud.

Growing from a minor annoyance into a widespread problem for the sector, the sophistication of digital fraud has tracked closely with the progression of the advertising it seeks to taint. Clever con artists conceived of ways to scam the system by designing bots that imitated human behavior and establishing phantom websites in order to steal advertising funds and disappear into thin air. These dishonest methods drained billions of dollars away from genuine channels, which put a shadow over the data that marketers had come to rely on. Because of this, the honesty of the digital advertising sector came into doubt, and the industry as a whole began to focus much of its attention on the difficulty of distinguishing between truth and fabrication. In light of all of this, the digital advertising sector immediately set out to find a way to improve its traceability and transparency (Minastireanu & Mesnita, 2019). After having been praised in the past for its fluidity and efficiency, the digital supply chain has now become the subject of close scrutiny. Verification and confirmation were now necessary for each click and impression, which had previously been a straightforward metric. The fight against fraud in the industry has been distinguished by a number of key breakthroughs, including the development of complex algorithms to identify irregular patterns, the implementation of blockchain technology to produce records that cannot be altered, and the establishment of stringent standards to validate the authenticity of digital traffic.

Nevertheless, the chameleon-like character of fraud continues to adapt and evolve, despite the gains that have been made in both technology and policy. The battleground has evolved, and it is no longer sufficient to simply take reactive steps; rather, a proactive attitude that anticipates the next move that the fraudsters will make is required. In this environment of heightened vigilance, the purpose of this thesis is to make a contribution of new insight by investigating the most recent strategies and innovations that promise to outperform those who are attempting to disrupt the digital marketplace. It aspires to give a model for the future of digital advertising, which will be a world in which openness will reign and fraud will be consigned to a cautionary tale of what once was, rather than a disruptive force in what should be a domain of unlimited possibility. This narrative continues the tale of the fight against fraud that has been waged in digital advertising, investigating the ongoing conflict as well as the evolving response of the industry. It sets the groundwork for a thesis that not only evaluates the current state of affairs but also looks to the future, proposing new strategies to ensure that digital advertising can achieve its original promise in an environment that is characterized by trust and transparency.

CHAPTER II:
REVIEW OF LITERATURE

2.1 INTRODUCTION

In the vast and ever-expanding digital landscape, advertising has emerged as the lifeblood of the internet, supporting the free flow of content and services that billions of users enjoy. However, the rapid growth of online advertising has been paralleled by the proliferation of sophisticated fraud schemes, compromising the effectiveness and trustworthiness of digital ad campaigns. This literature review seeks to dissect the current state of knowledge surrounding traceability and fraud in digital advertising, scrutinizing the evolution of this symbiotic and sometimes adversarial relationship between marketing strategies and fraudulent activities. It delves into the mechanisms of digital advertising, the scope and impact of fraud, the technological arms race between fraudsters and fraud detectors, and the regulatory environment that frames this dynamic (Zhu et al., 2017). The scholarly discourse on digital advertising reveals a complex ecosystem where innovation and opportunism coexist. Pioneering research has often focused on the transformative power of digital advertising technologies and their capacity for hyper-targeted, efficient, and interactive campaigns. Yet, there is an equally robust body of work that casts a light on the darker facets of the digital ad space, where fraudulent actors leverage the same sophisticated tools for deceitful gain. This review will examine studies that investigate the scale and mechanics of ad fraud, highlighting the significant financial and ethical implications for businesses, consumers, and the industry as a whole.

Central to the concerns of digital advertising is the concept of traceability—the ability to track and authenticate the journey of an ad from its origin to its ultimate display to a consumer. The literature sheds light on the myriad challenges to achieving such transparency, from the technical hurdles in identifying fraudulent traffic to the legal and logistical obstacles in enforcing accountability across a global and fragmented digital landscape. This review will synthesize research on existing traceability mechanisms, including the deployment of machine learning algorithms, blockchain technologies, and forensic data analysis, evaluating their efficacy and identifying gaps that persist in the fight against fraud. The evolving regulatory responses to digital ad fraud are also a crucial component of this narrative. This review will explore the literature on international standards and policies, assessing their impact and exploring the tension between industry self-regulation and governmental oversight. It will consider how regulatory frameworks adapt to the agile nature of digital fraud and how they might need to evolve to better serve the digital economy. This literature review will highlight where the current body of research converges and diverges, identifying areas ripe for further inquiry and exploring how the present thesis will address these gaps. By mapping the contours of the existing literature, this review will set the foundation for the subsequent research, situating it within the broader academic and practical discourse and paving the way for a deeper understanding of traceability and fraud in the high-stakes realm of digital advertising.

Taking into account the lessons that can be learned from McLuhan's theory of media, it is clear that marketing through social media possesses a unique attraction. McLuhan referred to the internet and other forms of social media as "Cold" media because they encourage the establishment of collective beliefs through the use of technical determinism. According to what Jan and his colleagues wrote in the year 2020, McLuhan's idea of technological determinism, also known as media ecology, proposes that the ways in which individuals interact with and are influenced by their surroundings are shaped by a transformational social framework (Sadeghpour & Vlajic, 2021). This idea was presented as part of McLuhan's Media Ecology. From the standpoint of this particular theoretical perspective, changes in communication technology have the potential to restructure the social fabric itself.

In addition, McLuhan's theory presents a metaphorical contrast between "hot" and "cool" media as a means of determining an audience's level of interaction with various forms of media. This classification highlights the ability of "cool" media to minimize the issue of audience passivity in the context of digital advertising by highlighting the importance of cool media. Because of the interactive aspect of social media, customers are able to engage with commercials on a deeper level, which in turn helps to build a good perception of the brand. Companies can fine-tune their communication strategies to better engage with their consumers by capitalizing on the inherent social dynamics of media if they are enlightened by McLuhan's paradigm and use it as a guide (Narayan et al., 2023). This strategy entails the creation of messages that are compatible with the patterns of perception that are prevalent in the electronic culture, as well as the transformation of classic advertising criteria into ones that are compatible with the ethos of the digital era.

To continue in this vein, digital advertising becomes a channel for cultural transmission inside the social fabric, which enables firms to build a new view of advertising. This transformation in the paradigm of marketing communication not only solves the problem of narcissistic tendencies among customers, but it also closes the gap between personal interaction and corporate messaging. As a consequence of this, customers report feeling a deeper sense of involvement rather than detachment, which helps to cultivate a participatory culture that is in line with the interactive characteristics of the contemporary, digitalized social sphere. Companies are able to manage the difficulties of digital advertising and ensure that their messages not only reach but also resonate with their intended audiences by integrating McLuhan's theoretical insights (Ng, 2023). This allows the companies to better serve their customers. Because of this sophisticated understanding of media ecology, it is possible to take a more strategic and culturally oriented approach to digital marketing. This means that commercials may be produced not only to inform, but also to engage and modify the experience of the consumer.

2.1.1 OVERVIEW OF DIGITAL ADVERTISING: DETAILED DISCUSSION ON DIGITAL ADVERTISING AND ITS VARIOUS FORMS

Digital advertising, a cornerstone of the modern marketing landscape, represents a multifaceted domain characterized by its versatility, reach, and technological sophistication. Rooted in the principles of traditional advertising but transformed by the advent of digital technology, it encompasses a range of practices and platforms that cater to the diverse needs of businesses and consumers alike. This segment of the literature review explores the various facets of digital advertising, providing a detailed discussion on its evolution, forms, and implications in the contemporary marketing environment. Digital advertising has undergone a remarkable evolution since its inception (Joo et al., 2023). Initially confined to basic banner ads on websites, it has rapidly expanded to encompass a variety of formats and channels, driven by advancements in technology and shifts in consumer behavior. The rise of search engines, social media platforms, and mobile technology has been particularly influential, offering new avenues for advertisers to engage with their target audiences in increasingly personalized and interactive ways.

Search Engine Marketing (SEM) and Social Media Advertising have emerged as pivotal components, leveraging the power of search engines and social platforms to offer targeted visibility and engagement. SEM capitalizes on user search intent, while social media advertising utilizes detailed demographics for personalized campaigns. Display and Video Advertising further enrich this landscape. Display ads, including banners and pop-ups, strategically enhance brand awareness and engagement on various websites. In contrast, Video Advertising harnesses the growing trend of digital video consumption, engaging users with compelling visual narratives. Native Advertising and Email Marketing also play crucial roles. Native ads offer subtlety by blending with their surrounding content, whereas email marketing maintains a direct line of communication with consumers. Content Marketing, focusing on value-driven content, subtly stimulates interest in products and services while fostering brand authority.

The dynamic nature of digital advertising, however, brings forth challenges like ad fraud, privacy concerns, and the need for traceability. The rise of ad-blocking technology and evolving consumer attitudes demand continuous innovation and ethical responsibility from marketers. Personalized, content-driven approaches are increasingly preferred, reflecting the shift towards more engaging and responsible advertising practices. Overall, digital advertising continues to adapt and thrive, offering varied tools and strategies for marketers to effectively reach and resonate with their target audiences. As consumer preferences and technological landscapes shift, so will the strategies and practices of digital advertising, ensuring its ongoing relevance and efficacy in the intricate world of marketing (Chen et al., 2016). In the expansive field of digital advertising, a novel concept is emerging Internet Advertising Paid Slots and Spaces (IAPS), akin to a stock exchange for the buying and selling of digital advertising space. This new model is proving to be a significant contributor to Internet advertising revenues, encompassing diverse areas such as search engine marketing, social media advertising, and display advertising. This literature survey is dedicated to demystifying the various channels of Internet advertising, providing contemporary insights and knowledge crucial for decision-makers in the digital advertising landscape.

The advent of the World Wide Web and the Internet has marked a turning point in business practices, particularly in advertising strategies. These technologies have shifted the advertising paradigm from traditional mediums like TV and outdoor advertising to a more digital-centric approach. This shift is becoming more pronounced, with digital advertising expenditures surpassing those of traditional TV advertising, particularly in the United States. The preference for online video platforms over television for advertising is a testament to their user-friendliness and compatibility (Wiktor & Sanak-Kosmowska, 2021). The mass migration to Internet advertising can be attributed to its unique advantages, such as precise location-based targeting, data-driven user profiling, effective market segmentation, retargeting capabilities, and more cost-effective pricing models compared to traditional media. Companies that previously invested in expensive media channels are now achieving more targeted outreach at significantly lower costs through digital avenues. Today's managers are increasingly relying on digital marketing strategies for brand building and customer engagement. Digital marketing offers the advantage of easily tracking and monitoring campaign results, allowing managers to assess user responses and measure the success of marketing initiatives in real-time. This immediate feedback loop facilitates more efficient planning and execution of future campaigns. However, navigating the constantly evolving digital advertising landscape remains a complex challenge for managers (Shaari & Ahmed, 2020). The field is characterized by a variety of technologies and encompasses various digital platforms within a convoluted ecosystem of network players and trends. Existing literature, while extensive in its coverage of digital technologies like social media and search engines, often lacks a focused approach in differentiating the various relevant research streams.

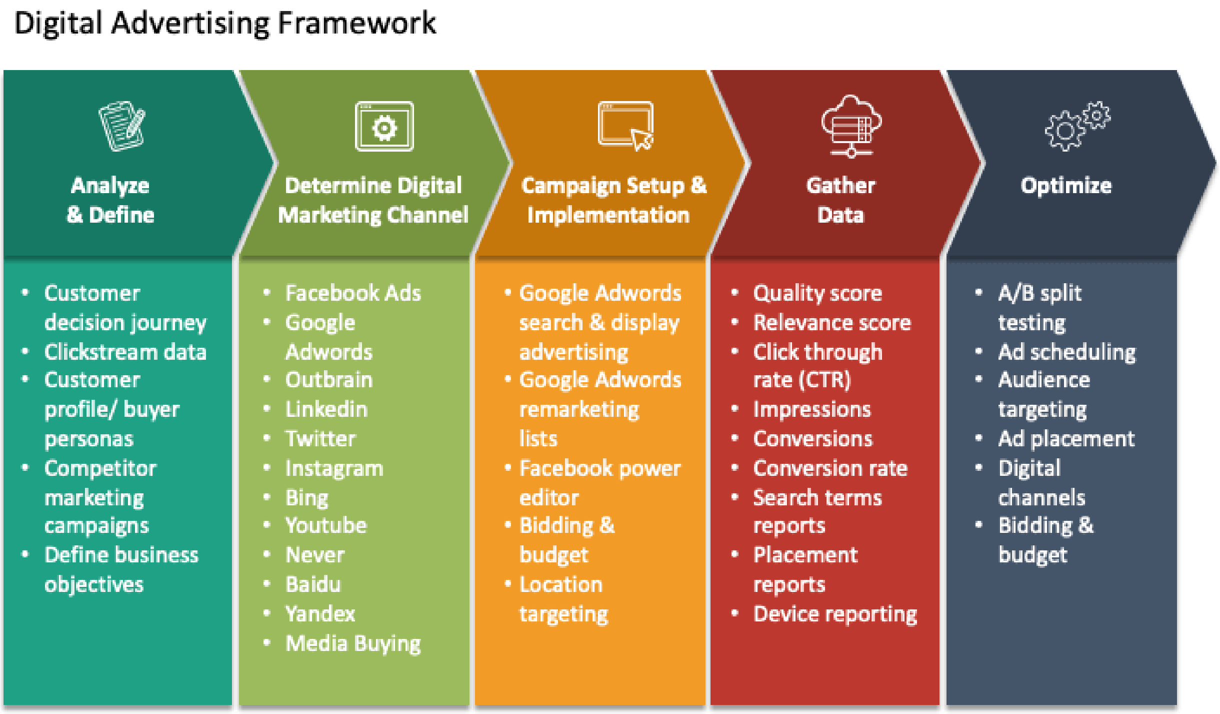


Figure 3: Digital Advertising Framework [Source: https://digitaldirect.marketing/b2b-digital-marketing/]

The above image is an illustration of a comprehensive Digital Advertising Framework that may be utilized to maximize the effectiveness of online marketing approaches. The framework is broken down into five crucial parts, the first of which is called "Analyze & Define." During this step, the customer's decision journey, clickstream data, and buyer personas are analyzed, and competitor marketing efforts are also analyzed, in order to define corporate objectives. The subsequent step is to choose the digital marketing channels that are most suitable for the business. These channels can include search engines like Google and Bing as well as social media platforms like Facebook and Instagram. During the "Campaign Setup & Implementation" stage, the primary focus is on leveraging instruments such as Google AdWords for search and display advertising, as well as location targeting methods. The collection of data is an essential component, with a particular emphasis on metrics like as quality score, relevance score, click-through rate (CTR), impressions, conversions, and device reporting in order to evaluate the effectiveness of the campaign. When it comes to improving the overall success of digital advertising initiatives, the last "Optimize" step makes use of techniques such as A/B split testing and ad scheduling in order to refine audience targeting and strengthen overall performance.

An in-depth analysis of the framework is shown in the graphic, with each phase being properly identified and defined. An icon of a magnifying glass is used to represent the first phase, which is titled "Analyze & Define," while an icon of a computer screen is used to represent the second step, which is titled "Determine Digital Marketing Channel." While the third phase, "Campaign Setup & Implementation," is symbolized by a gear icon, the fourth step, "Gather Data," is symbolized by a chart icon. Both of these icons are utilized to illustrate the steps. A rocket icon is used to symbolize the final step, which is referred to as "Optimize."

These advertisements typically appear in prominent areas of web pages, including the upper, lower, or side sections. It is essential to differentiate this form of advertising from broader marketing strategies that might include unpaid elements. The survey categorizes and expands on various aspects of digital advertising, guiding managers through the complexities of this evolving field. Unique in its focus, this survey specifically addresses IAPS, a domain still in the early stages of academic exploration. Unlike previous literature reviews, this study ventures beyond the conventional realm of banner ads to encompass a broader spectrum within this segment. This focused approach provides a fresh perspective on the intricate and dynamic world of digital advertising, offering valuable insights for both academics and industry practitioners. New digital technologies are redefining the way in which businesses communicate and connect with customers via digital media, which is driving a significant revolution in the landscape of digital advertising (Srivastava, 2020). This transformation is being driven by the fact that new digital technologies are becoming more affordable. In this day and age, the future of advertising is rapidly changing, which raises issues regarding the path that academic research in digital advertising should take as well as the methods that businesses and advertising agencies can use to maximize the returns on their advertising investments.

There are now a few main trends that are defining digital advertising. These trends include a shift toward data-driven marketing communication, the effect of artificial intelligence on ad production, and the significant role that big data plays in the execution of ads. In this constantly shifting environment, there are a number of hypotheses floating about concerning the management of future digital advertising. These hypotheses center on strategies and procedures for providing consumer-specific adverts. The automation of media execution methods, in particular programmatic buying, is quickly becoming one of the most important aspects of data-driven advertising execution. Traditional advertising tactics entailed making manual purchases of media and evaluating its efficacy mostly through post-campaign surveys; this strategy fundamentally departs from those ways in a number of important respects.

Using data-driven real-time bidding (RTB), programmatic buying, an essential component of contemporary digital advertising, optimizes the execution of advertisements by sending the appropriate message to the appropriate audience at the appropriate moment in the most cost-efficient manner possible. Traditional media buying centers on the acquisition of particular media stocks (such a 15-second TV spot during a prime-time program) that are appropriate for the distribution of mass media content. This approach is considerably different from traditional media buying in a number of major respects. Programmatic buying, on the other hand, places more of an emphasis on obtaining 'audience' exposure as opposed to ‘media' space. This indicates that advertisements are not associated with a particular media area; rather, they are directed toward the audience that is most likely to respond to them, taking into consideration criteria such as age, gender, purchasing history, and hobbies.

The use of large amounts of data is an essential component of programmatic advertising since it enables marketers and sellers to determine which audience members are the best fit for a given advertising message. Processing requests in real-time for billions of ad inventories offered by media firms, generally in less than 0.2 seconds, is one of the hallmarks of programmatic advertising, which automates the process of buying and selling advertisements through RTB (Bayer et al., 2020). This method makes use of a number of different components, such as demand side platforms (DSPs) for the purpose of purchasing advertising space and data management systems for the purpose of conducting data analysis on the advertiser's end. On the other hand, the supplier's side is mostly comprised of supply side platforms (SSPs), which are comprised of the digital ad inventories of media firms, in addition to Ad Exchanges, which are used for buying and selling ad inventories. The increasing complexity of this industry has led to the development of unified programmatic advertising solutions, which are now being offered by large marketing platform suppliers such as Adobe and Oracle. It is crucial that academics broaden the scope of their research to incorporate these developing trends and technologies as the digital advertising industry continues to undergo rapid transformations. This literature review illustrates the disruptive impact that digital technologies have had on advertising. It also emphasizes the need of comprehending these changes in order to effectively manage and execute advertisements in the digital era.

2.1.2 TRACEABILITY IN DIGITAL ADVERTISING: EXPLORATION OF EXISTING SYSTEMS AND MECHANISMS FOR TRACKING AND VERIFYING DIGITAL ADVERTISEMENTS

Online advertising has become a cornerstone of the Web's "free" content model, fundamentally reshaping the marketing industry by offering an array of opportunities for advertisers to connect with potential consumers. This evolution in advertising is based on a complex infrastructure, consisting of numerous intermediaries and technologies, all geared towards delivering personalized ads. To achieve this, an extensive amount of user data is gathered, compiled, processed, and exchanged at an extraordinary pace. While online advertising holds immense value, its pervasive and intrusive nature raises significant privacy concerns. The study begins by examining the threats and potential privacy invaders within the online advertising landscape. It scrutinizes the key components of the advertising framework, focusing on their tracking abilities, data collection practices, aggregation levels, and associated privacy risks (Knuth & Ahrholdt, 2022). The technologies employed for tracking and data sharing within these components are also thoroughly reviewed. Subsequently, the survey presents an exhaustive analysis of pertinent privacy mechanisms. These mechanisms are classified and compared based on the privacy protection they offer and their impact on the Web's ecosystem.

Online advertising's omnipresence on the Internet has been instrumental in sustaining the model of free Internet access, largely through the revenues it generates for publishers. This prevalence has led to the creation of a massive data transport channel, where intermediary entities gain access to billions of users and, more crucially, their data. The aggregation of gigabytes of user data underpins more targeted and effective advertising campaigns. In the context of traceability in digital advertising, understanding and scrutinizing these systems and mechanisms is essential. The survey's insights are geared towards navigating the complex interplay of user privacy, data collection, and the need for transparency and traceability in the digital advertising sector. The field of digital rights management in network media faces numerous unresolved challenges, such as ensuring media quality, protecting copyrights, and finding effective profit models. This literature survey explores a novel approach to these issues, proposing a blockchain-based digital rights management scheme for network media. Blockchain, leveraging cryptographic algorithms, hash chains, and consensus mechanisms, offers consensus, irreversibility, and traceability for online data.The proposed blockchain-based scheme is designed to enhance various aspects of network media management, including production, copyright, transactions, and user behavior. This approach aligns well with the focus on traceability and verification in digital advertising, as blockchain's inherent properties facilitate these processes (Sato & Berrar). By employing consensus mechanisms, the scheme enables real-time copyright confirmation, while smart contracts based on blockchain ensure real-time and reliable transactions. The use of digital signatures and hash chains further enhances the reliability of these transactions. Moreover, the implementation of this blockchain-based digital rights management scheme is anticipated to significantly advance the network media sector. It offers innovative profit and supervision models, potentially catalyzing a new era in the network media business landscape. This aligns with the broader theme of exploring existing systems and mechanisms for tracking and verifying digital advertisements. The blockchain's capabilities in ensuring data integrity and traceability make it a promising solution for addressing the current limitations in digital rights management and advertising verification. This survey suggests that the integration of blockchain technology could be a pivotal step in evolving the digital advertising industry towards more transparent, secure, and verifiable practices.

Traceability is a major area of concern and interest in the world of digital advertising. This is a reflection of the constant fight that the industry is having with issues such as ad fraud, privacy, and the efficiency of advertising campaigns. A thorough examination of the relevant research reveals the adoption of a multidimensional strategy for addressing these problems, one that makes use of a variety of technical and methodological approaches. The widespread problem of ad fraud that affects the digital advertising industry has led to the creation and implementation of a number of different ad verification systems (Zhu et al., 2017). These tools, such as Integral Ad Science and Double Verify, are essential in ensuring that adverts display on the intended websites and reach the targeted consumers, hence reducing the amount of money that is wasted on advertising and preventing skewed data from being collected during campaigns. The body of research emphasizes the continual arms race between ad verification technology and the ever-evolving strategies of ad fraud, highlighting the importance of constant innovation in this field.

One trend that stands out as particularly noteworthy is the implementation of blockchain technology in digital advertising. The immutable ledger system that blockchain employs provides hitherto unheard-of levels of transparency and security in advertising transactions, so guaranteeing the genuineness of every advertisement impression. It is becoming more acknowledged as a game-changer that this technology may play in the fight against ad fraud and in the enhancement of trust in digital advertising ecosystems (Wang et al., 2022). The usage of cookies and pixel tracking is another time-honored practice that is becoming more and more contentious in the context of digital advertising tracking. Although these methods have been the cornerstone in tracking the behavior of users and the effectiveness of advertisements, they face substantial hurdles as a result of the evolving data privacy legislation. According to the research, there is an increasing trend toward tracking technologies that are more respectful of users' privacy, which strikes a healthy balance between the competing goals of efficient advertising and protection of user confidentiality.

The literature also makes extensive use of programmatic advertising platforms, drawing attention to their important role in automating ad buying and improving ad placements through real-time bidding and data analytics. The purchasing and selling of advertisements have been revolutionized as a result of these platforms, which offer improved efficiencies and capacities of targeting. However, the intricacy of these networks and the large number of intermediaries that are engaged also add layers of opacity, which makes it difficult to trace and monitor ad transactions. The integration of artificial intelligence and machine learning appears to be the future trajectory of traceability in digital advertising, as stated by the literature, which seems to be trending towards the integration of artificial intelligence. Through pattern recognition and predictive analytics, these technologies are positioned to play a significant part in identifying fraudulent advertising and avoiding its occurrence. In addition, as the industry struggles to address issues over privacy, contextual advertising and monitoring approaches that do not compromise users' privacy are gaining popularity. These methods attempt to solve the problem of ineffective advertising without relying on the intrusive tracking of the activity of individual users (Wang et al., 2023). The problems and their solutions cover a wide range, ranging from technology advancements such as blockchain and AI to adjustments in advertising strategies such as contextual adverts and privacy-focused tracking methods. The discussion covers a wide range of topics. This ever-changing industry is a reflection of the complicated interplay between the demand for efficient advertising and the requirement to safeguard user privacy and secure the integrity of online ad transactions. This dynamic field continues to expand, reflecting this complex interplay.

The phenomenon of click fraud, a prevalent form of deception in the digital advertising landscape, presents a significant threat to the integrity and effectiveness of online advertising. This research delves deeply into the mechanisms and impacts of bots used to perpetrate click fraud (Kwon et al., 2011). The study commences with an exhaustive exploration of various categories of Web-bots, scrutinizing their malevolent activities and the associated risks they pose. Central to this research is the differentiation between the behaviors of bots and humans in the context of click fraud, specifically within the frameworks of contemporary digital advertising platforms. The study poses critical questions aimed at distinguishing these behavioral patterns, which is crucial for understanding and mitigating the effects of click fraud. Further, the research provides an extensive review of existing strategies for detecting and countering click fraud, as found in current literature. These strategies are categorized based on their applicability to different stakeholders within the digital advertising ecosystem. The study also sheds light on some of the most infamous real-world instances of bot-driven ad fraud campaigns, offering a practical perspective on the issue.

An illustrative case is the discovery of Clickbot.A by Google in 2006. This botnet, comprising over 100,000 machines, executed a sophisticated click fraud operation targeting syndicated search engines. Clickbot.A, characterized by its dual components of bots and a botmaster, leveraged HTTP requests to interact with doorway sites and redirectors, ultimately impacting search engine result pages. Each bot, functioning as an Internet Explorer browser helper object, was capable of autonomously clicking on ads, guided by instructions received from the botmaster (Zhang et al., 2020). This botmaster operated a web application backed by a MySQL database, coordinating the click fraud activity through compromised websites and ISPs. This comprehensive study of click fraud in digital advertising is unique in its dual focus on both theoretical and practical aspects of the problem. By examining and tracing the operations of botnets like Clickbot.A, it contributes significantly to the understanding of traceability in digital advertising. This research underscores the ongoing challenges and complexities in tracking and verifying digital advertisements, highlighting the need for robust systems and mechanisms to combat the evolving threat of click fraud in the digital advertising arena.

2.2 INCLUSION CRITERIA

Sources that investigate the mechanics and techniques of digital advertising fraud, such as click fraud and adware, among others, will be given priority in the research. Additionally, sources that suggest or analyze traceability solutions, such as blockchain and AI-driven analytics, will also be given priority. Despite the fact that the research will be conducted on a global scale, a particular emphasis will be placed on studies that pertain to important digital advertising markets such as those in North America, Europe, and Asia. Journals that have been subjected to peer review, reports from the industry, case studies, white papers from trustworthy sources, conference proceedings, and academic theses that provide one-of-a-kind insights or pertinent case studies will all be included in the research. One of the most important criteria is the level of scientific rigor, with a preference for studies that use research methods that are both transparent and resilient, regardless of whether they are qualitative or quantitative. Works that have produced a major influence in the field and have received a high number of citations or are influential will be given priority.

For the most part, the research will be conducted using sources that have been published in the English language. There is a possibility that it will incorporate significant works written in other languages, provided that translations or full summaries are accessible. Last but not least, accessibility is a factor to take into account, with a particular emphasis on materials that are easily accessible through open access, academic databases, or institutional subscriptions (Srivastav & Ahuja, 2020). In order to guarantee that the research is based on sources that are of high quality, relevant, and up to date, this set of criteria serves as a strong foundation for conducting an in-depth investigation into the issues of traceability and fraud in the ever-evolving sector of digital advertising.

A significant portion of the research will consist of studies that provide an in-depth analysis of digital advertising, with a special emphasis on the problems of fraud and traceability. Not only does this entail an investigation into the several types of fraudulent behaviors that are prominent in digital advertising, such as click fraud, adware, and spoofing, but it also entails an investigation into the methods and technologies that are utilized to monitor, detect, and mitigate these fraudulent actions. Because of the dynamic nature of the area, we will give priority to recent publications, ideally those that have been published within the past five years, in order to guarantee that the study accurately reflects the current environment (Fowler et al., 2021). Nevertheless, important works of historical significance will also be taken into consideration in order to provide a thorough understanding of the development of digital advertising and the issues that are linked with it over time. There will be a concentrated investigation of studies that are pertinent to the key digital advertising markets, which are North America, Europe, and Asia. Geographically speaking, the research will include worldwide viewpoints, but it will also investigate more specific studies. The importance of this geographical focus cannot be overstated when it comes to comprehending the ways in which various market dynamics and regulatory frameworks influence the frequency of digital advertising fraud and its management (Loyola-González et al., 2018). When it comes to the types of sources that will be utilized, the research will primarily consist of published academic publications that have been subjected to peer review, industry reports from renowned organizations, in-depth case studies, and authoritative white papers (Hu et al., 2020). Additionally, the incorporation of academic theses and conference proceedings will be taken into consideration, with particular attention paid to those that provide fresh insights or major case analyses in the field of digital advertising research.

A crucial factor to consider is the level of methodological rigor exhibited by the sources. The research will primarily focus on studies that make use of rigorous and well-documented techniques, and it will also provide a fair mix of qualitative and quantitative research. Additional considerations in the selection of sources will include high citation rates and academic influence, both of which are indicators of the work's impact and significance in the subject. Regarding the language, the research will mostly draw from sources that have been published in the English language (Hu et al., 2020). It is possible, however, to incorporate substantial works written in other languages provided that they are accompanied by translations that may be relied upon or extensive summaries. Accessibility of the content is another important factor to take into account, with a particular emphasis on sources that are easily accessible through open access, academic databases, or institutional subscriptions or subscriptions. In a nutshell, the purpose of this literature review is to be exhaustive, up-to-date, and methodologically sound. Its primary focus is on the literature that is the most pertinent and insightful in the field of digital advertising fraud and traceability. The purpose of this research is to provide academics, industry practitioners, and policymakers with significant insights by adhering to these inclusion criteria. The research attempts to provide a complete picture of the current issues, technological breakthroughs, and future prospects in the sector.

A primary focus of this investigation is the delicate dance that takes place between those who commit digital ad fraud and the tools that are meant to prevent the efforts of those who do such fraud. The purpose of this study is to investigate the intricate algorithms and advanced digital technologies that have been developed in order to improve the traceability and accountability of digital advertising (Mikkili & Sodagudi, 2022). The blockchain technology, for example, is being highlighted for its potential to change the procedures of ad verification by offering an immutable and transparent ledger of ad transactions. One of the most important themes that emerges is the application of artificial intelligence and machine learning, which offers potential solutions for identifying patterns that are suggestive of fraudulent activity and improving ad placements for authenticity and relevance. Through the use of case studies that illustrate both the triumphs and the obstacles associated with the implementation of these advanced technologies, the practical sides of this topic are exposed and brought to light (Kanei et al., 2020). These examples from the real world provide essential insights into the actual application of theoretical concepts and technologies in the fight against digital ad fraud. In addition to this, they offer a glimpse into the many strategies that fraudsters use, indicating a continual conflict between invention and adaptability. Within the context of this discussion, the investigation of the ethical and regulatory aspects of digital advertising is of equal significance. The review investigates the ways in which legislative frameworks such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) influence digital advertising techniques and the utilization of consumer data. Ethical considerations, in especially those pertaining to the privacy and permission of consumers, are examined, bringing to light the delicate balance that marketers need to achieve in order to use user data for targeted advertising while also respecting the rights of individuals to privacy. Furthermore, this research traces the historical growth of digital advertising, which provides a significant background for understanding the trends and difficulties that are currently occurring (Gabryel, 2018). The historical perspective presented here sheds light on the ways in which conventional advertising strategies have adapted to the digital age, as well as the ways in which developing technology and shifting consumer behaviors are influencing the future of the business.

A tough criterion is employed in the process of picking sources for this comprehensive evaluation. This ensures that the sources selected have a combination of scholarly depth and practical usefulness. The incorporation of academic publications that have been subjected to peer review provides the study with a higher level of theoretical rigor, while the incorporation of industry reports and case studies from reliable sources provides practical insights and real-world relevance.

2.2.1 FRAUD IN DIGITAL ADVERTISING: EXAMINATION OF DIFFERENT TYPES OF FRAUD IN DIGITAL ADVERTISING, THEIR IMPACT, AND METHODS OF DETECTION

The term "digital advertising fraud" refers to a variety of dishonest actions that are carried out with the intention of extracting illegal benefits from the ecology of digital advertising. A classification of these fraudulent behaviors, which include click fraud, impression fraud, ad injection, and domain spoofing, is the first step in the thesis. In order to artificially exaggerate engagement metrics, click fraud entails the deliberate generation of phony clicks on advertisements. On the other hand, impression fraud refers to the manufacture of fake ad views (Pastor et al., 2020). The practice of domain spoofing involves misrepresenting the website on which the advertisement is placed in order to sell ad space at premium prices under false pretenses. It is also known as ad injection, which involves the insertion of advertisements into web pages without authorization.

The repercussions of such fraudulent conduct are extremely significant. Not only do they result in large financial losses for advertisers, as they deplete advertising budgets without producing genuine customer involvement, but they also skew analytics and market research, which in turn leads to marketing plans that are wrong. The ecology of digital advertising is being undermined in terms of its overall integrity as a result of this lack of confidence. The purpose of this thesis is to investigate the various detection and prevention approaches that are currently being utilized in business in order to meet these difficulties. These include technology solutions such as improved algorithms for anomaly detection, which discover odd patterns in ad traffic that may suggest fraudulent behavior. These algorithms are used to identify anomalies (Estrada-Jiménez et al., 2017). The potential of blockchain technology to improve the transparency and traceability of advertising transactions is another angle that is being investigated. Furthermore, the effect that legal frameworks and industry standards have on the fight against advertising fraud is taken into consideration. In order to keep one step ahead of fraudulent actions and to protect the efficiency and dependability of digital advertising, it is vital for the industry to engage in continuous innovation and collaboration.

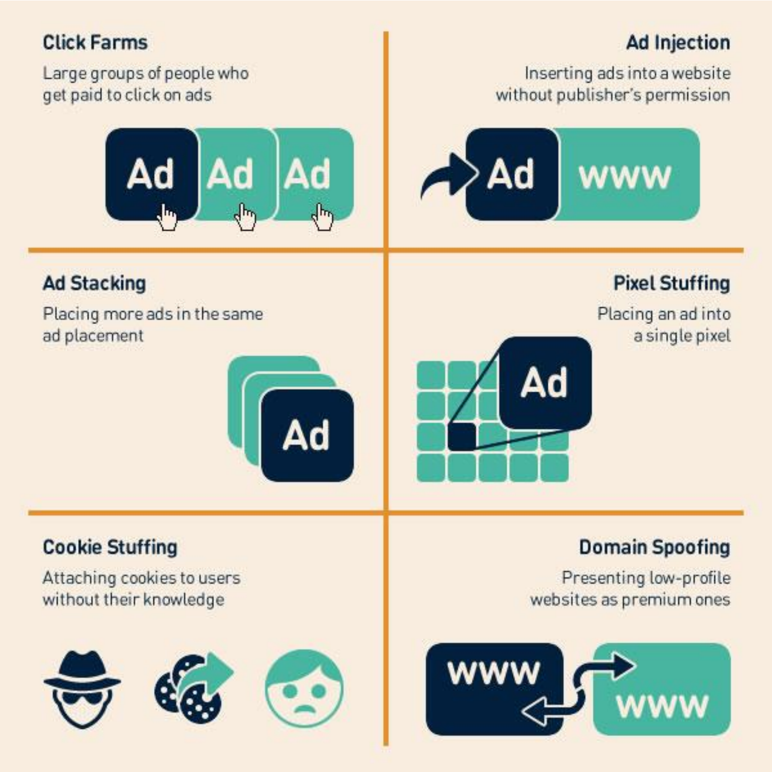

Click Farms
Large groups of people who get paid to click on ads
Ad
Ad
Ad
Ad Injection
Inserting ads into a website without publisher's permission
Ad
www
Ad Stacking
Placing more ads in the same ad placement
Ad
Pixel Stuffing
Placing an ad into a single pixel
Ad
Cookie Stuffing
Attaching cookies to users without their knowledge
Domain Spoofing
Presenting low-profile websites as premium ones
www
www

Figure 4: Fraud in Digital Advertisements [Source: https://www.linkedin.com/posts/emmanuel-disho-9018102bb_adfraud-activity-7179216080962256897-Is7q/?trk=posts_directory]

This image is an informative visual that offers a classification and explanation of a variety of unethical methods that are utilized in online advertising. Click Farms, in which large groups are paid to click on advertisements, Ad Injection, in which advertisements are inserted into a website without permission, Ad Stacking, in which multiple advertisements are placed in the same ad placement, Pixel Stuffing, in which an advertisement is placed into a single pixel, Cookie Stuffing, in which cookies are attached to users without their knowledge, and Domain Spoofing, in which low-profile websites are presented as premium ones are some of the practices that are highlighted. Through the use of images and brief explanations, each category is visually portrayed in order to facilitate better comprehension.

An in-depth analysis of the framework is shown in the graphic, with each phase being properly identified and defined. An icon of a magnifying glass is used to represent the first phase, which is titled "Analyze & Define," while an icon of a computer screen is used to represent the second step, which is titled "Determine Digital Marketing Channel." While the third phase, "Campaign Setup & Implementation," is symbolized by a gear icon, the fourth step, "Gather Data," is symbolized by a chart icon. Both of these icons are utilized to illustrate the steps. A rocket icon is used to symbolize the final step, which is referred to as "Optimize."

Advertising on the internet has become an increasingly important component of the advertising market as the digital landscape continues to undergo fast transformation. The pay-per-click system is a prominent revenue model in this industry. In this model, charges are determined by the popularity of the keyword as well as the density of advertisers that are competing with each other. This technique, on the other hand, is vulnerable to a practice known as click fraud, which occurs when individuals or competing businesses generate a large number of clicks that are not legitimate (Tauro et al., 2021). The development of a vibrant and healthy online advertising market is significantly hampered by the fraudulent activity that is taking place.

Identifying duplicate clicks within decaying window models, such as leaping and sliding windows, is an essential component of the process of detecting click fraud among other fraudulent activities. A significant contribution that these models provide is to the process of identifying and classifying cases of click fraud. In spite of the fact that there are algorithms that are meant to identify duplicates, there is still a significant lack of solutions that are both practical and successful for recognizing click fraud within pay-per-click streams, particularly in the context of decaying window models (Li et al., 2024). Within both jumping and sliding window frameworks, this thesis addresses the difficulty of detecting duplicate clicks in pay-per-click streams. Specifically, the thesis focuses on the two frameworks (Makkineni et al., 2023). It proposes a GBF (Group Bloom Filter) technique that is both effective and efficient for usage in jumping windows, and it introduces the concept of group Bloom filters, which is an approach that considerably reduces the number of memory operations that are performed during the processing of click streams. Furthermore, the thesis presents a novel TBF (Timing Bloom Filter) algorithm for the purpose of handling duplicate clicks in situations that involve sliding windows or jumping windows that contain a large number of sub-windows. This innovative technique, which is supported by a novel data structure known as a temporal Bloom filter, provides a solution for processing click streams in sliding windows while making use of less memory space and requiring less processing time (Oentaryo et al., 2014). These proposed algorithms have been shown to be efficient and effective, as evidenced by the low rates of false negatives and false positives, as well as the optimized running time in the detection of duplicate clicks, according to theoretical studies and experimental results. As we look to the future, this research will continue to investigate the intricacies of click fraud and click quality within the context of data stream models (Chua et al., 2020). Future explorations will encompass a range of sophisticated click fraud attacks, the dynamics of advertising networks, new service models, and the broader economic and social impacts of click frauds, thereby contributing to a more comprehensive understanding of fraud in digital advertising and enhancing methods for its detection and mitigation.

In this day and age of digital technology, web advertising is an extremely important component of business promotion, particularly when it comes to the utilization of keywords. Consequently, the problem of click fraud and its potential to undermine business models that are dependent on online advertising has been brought to the forefront as a result of this. The pay-per-click (PPC) advertising mechanism used by Google is a famous example of this phenomenon. This business has been progressively undermined by fake clicks, which represents a significant problem (Sisodia & Sisodia, 2023). Through the use of these invalid clicks, which do not indicate a genuine interest in the content that is being marketed, revenues are diverted to con artists, which in turn undermines the effectiveness of online advertisements. In this article, a complete review of the click fraud landscape is presented. Focusing on the numerous types of click fraud that plague the digital advertising realm, particularly the Pay Per Click (PPC) model, this study investigates the various types of click fraud. This type of advertising, known as pay-per-click (PPC), is extremely vulnerable to fraudulent operations such as botnets and click farms because it charges advertisers for each click on their commercials. These extreme kinds of click fraud pose a substantial risk to the integrity of online advertising since they are so widespread. A variety of methods that are currently being utilized for the detection and analysis of click fraud are investigated. Specifically, it focuses on the ever-increasing risk of click fraud by conducting a comprehensive investigation into the security flaws that are inherent in the most prominent online advertising schemes (Dash & Pal, 2020). There is a possibility that the attack radius of click fraud will expand as a result of the introduction of new technologies such as the Internet of Things. As a result, the requirement for strong security mechanisms is more pressing than it has ever been. Many firms suffer significant financial losses as a result of click fraud, which has significant repercussions for their finances (Häger & Landergren, 2010). An examination of a number of detection methods that are utilized in a variety of fields to combat click fraud is presented in this study. Specifically, the study highlights the fact that click fraud not only has an effect on the budgets of advertisers but also poses concerns to the security of data due to the malicious use of bots. When it comes to mitigating and preventing the effects of click fraud, awareness and the ongoing growth of defense methods are both essential components. The study emphasizes the significance of enhancing defense measures in order to protect against this ever-evolving digital threat, which will contribute to an environment that is more trustworthy and secure for digital advertising.

The upkeep of free content that can be accessible on the internet, such as websites and programs for mobile devices, needs advertising in order to continue to exist. The prevalence of click-spam, which is defined as the act of fraudulent or invalid clicks on online advertisements in which the user has no true interest in the advertiser's site, on the other hand, causes income to be redirected to click-spammers (Mathur & Daniel, 2022). Click-spam is defined as the acts of clicking on adverts that are not relevant to the user's interests. In this article, a complete analysis of click-spam is presented, shedding light on the significant impact that it has had on the digital advertising industry. Given the current state of affairs, there is a considerable level of uncertainty over the effectiveness of the measures that advertising networks have taken to address this matter. On top of that, advertisers and third parties do not possess the capabilities necessary to carry out impartial evaluations or to protect themselves against click-spam. The results of this research provide the very first methodology that enables marketers to objectively evaluate click-spam rates on their adverts (Sahllal & Souidi, 2023). This methodology is demonstrated through the outcomes of this study. Furthermore, we have created an automated system that allows advertising networks to proactively recognize and respond to a range of simultaneous click-spam attacks. This method was successfully implemented. Utilizing data from large advertising networks was the means by which the validity of these approaches was demonstrated. An exhaustive quantitative study was conducted across ten of the most significant advertising networks, covering four distinct types of adverts (Wahid et al., 2023). The research was carried out throughout the networks. This inquiry, which led to the discovery of these assaults in the first place, led to the identification of seven ongoing click-spam attacks that had not been previously recognized or mitigated by significant ad networks at the time of the research. These attacks were discovered and thoroughly studied as a result of this investigation.

According to our data, the severity of the problem of click-spam is demonstrated, particularly with regard to the prospects for mobile advertising it presents. There is a persistent problem known as click-spam, which even the most extensive advertising networks are having to deal with. As a result of the fact that it is always changing, it is a challenging problem to address (Min et al., 2021). It is generally accepted that click-spam is an open topic that requires continuous attention and concerted efforts from the academic community. This is something that is currently being discussed. This is because click-spam is an ever-changing phenomenon, which is the reason behind this. To facilitate the conduct of more research and the creation of innovative security procedures to combat web spam, the information that was obtained for this study has been made accessible to the public. This was done in order to facilitate the process (Sisodia & Sisodia, 2021). By focusing specifically on the investigation of various types of fraud, the impact that they have, and the strategies that can be used to detect and prevent any instances of fraud, the purpose of this article is to contribute to a more comprehensive understanding of fraud in digital advertising (Daswani et al., 2008). This will be accomplished by focusing on the investigation of various types of fraud. At the very least ten percent of the overall expenditures that are generated in the industry of online advertising are believed to be attributable to online advertising fraud, often known as OAF. When everything is taken into consideration, this forms a sizeable share of the overall expenditure. Considering the frightening nature of this number, it is of the utmost importance that a thorough study be carried out from the point of view of marketing (Haider et al., 2018). In connection to OAF, the purpose of this article is to conduct an inquiry into the opinions of professionals working in the sector of internet marketing. It was determined that a total of eighty-nine interviews were carried out, and the findings that were derived from those interviews were incorporated into the study. In light of the findings of the research, it is abundantly clear that OAF represents a significant risk to the advertising business. There are many different stakeholders all across the world who are concerned about this topic, and it is both an urgent and continuous worry for them.

According to the conclusions of the study, which makes use of Agency Theory as its theoretical base, Internet specialists have a tendency to overlook misleading traffic. Agencies Theory is the theoretical underpinning of the study. In addition, there is a general lack of information among customers regarding the issue, which is another factor that adds to the fact that there is a low awareness of OAF within the market. A further issue is that there is a general lack of awareness regarding OAF (Keserwani et al., 2022). Furthermore, the findings of the research indicate that inefficient measurement methods and excessive client expectations are two of the obstacles that stand in the way of successful fraud detection and prevention. These are the two main obstacles. One of the key goals of this research is to improve people's understanding of online advertising fraud (OAF), which is a subject that has been given a relatively little amount of attention in the field of advertising literature. In addition, the research intends to encourage both professionals in the field and academics to develop anti-fraud techniques that are efficient.

In spite of the fact that having a grasp of OAF is of the utmost significance for the entirety of the advertising industry, there has been an extremely limited amount of research put out on the issue over the course of the past few years. This research aims to shed light on OAF from the standpoint of marketing by employing a qualitative methodology approach that is founded on grounded methodologies. The purpose of this research is to shed light on OAF. According to the data, there appears to be a low level of awareness regarding Office of the Auditor General (OAF), particularly among smaller clients who do not have the resources and the skills that are required to effectively combat fraud. This is especially true with regard to customers who are not as financially secure as some of the larger customers (Li et al., 2021). As an additional point of interest, there is evidence to suggest that professionals working in internet marketing frequently have a misunderstanding of the gravity of OAF and the ramifications that it can bring about. The fact that there is evidence presents support for this assertion (Kanei et al., 2019). The fact that the methods of measurement that are now in place for online advertising are sometimes insufficient is another factor that makes it difficult for industry professionals to be motivated to take preventative measures against online advertising fraud (OAF). Taking into account the current state of affairs, it would appear that it will be difficult to make significant improvements in the way that OAF is treated in the short term. This is because of the existing condition of affairs (Raj, Allupati & Kalaiarasi, 2020). As a result of the continual development of fraudulent approaches, it is difficult to conceive of a future in which the advertising industry will be completely free of fraudulent operations. This is due to the fact that new methods of counterfeiting are continually being created. To this day, the industry's response to OAF has been primarily defensive, with the primary emphasis being placed on preventing fraudulent conduct rather than generating proactive and inventive solutions. This has been the case for the majority of the time. Since the launch of the investigation, this has been the situation that has been observed. The establishment of professional organizations is one example of the collaborative efforts that have been made, and the Joint Industry Committee for Web Standards has also been responsible for the formation of these groups. This committee is working toward the formation of best practices, the certification of industry participants, and the recommendation of monitoring instruments that are given by third parties. These are all goals that are being worked toward. Despite the fact that these are moves in the right direction, there is an immediate need for preventative measures that are more all-encompassing and continuing. This is especially true when one takes into mind the growing number of fraudulent operations as well as the kind of fraudulent acts that are now being committed.

2.3 CLEAR ORGANIZING THEMES

Due to the fact that it has a propensity to bias machine learning models in favor of the more numerous class, this significant imbalance between fraudulent and genuine publications poses a particular challenge. As a result, it becomes more challenging to effectively categorize those who engage in fraudulent activity. It is the purpose of my work to present an empirical analysis of a number of prominent data sampling methodologies, employing nine advanced learning models exclusively for the purpose of spotting fraudsters in the internet advertising market (Khine & Khin, 2020). This assessment is a component of my efforts to find a solution to this problem. Providing an evaluation of the influence that different data sampling methodologies, such as oversampling, under sampling, and hybrid sampling, have on the performance of different classifiers is the primary purpose of this research. This review is being conducted in the context of click-fraud detection.

This comprehensive analysis is carried out with the assistance of the user-click dataset that serves as the benchmark for FDMA: 2012. For the purpose of determining how successful various combinations of data sampling methods and classifiers are, each and every feasible combination is assessed. A number of performance indicators, including average precision, recall, f1-score, and AUC (Area Under the Curve), are considered to be important evaluation tools. Furthermore, in order to evaluate new improvements and refinements, these results are compared to models that are known to be regarded state-of-the-art at the present time.

In particular, the fact that the combination of adaptive synthetic (ADASYN) oversampling with a gradient tree boosting (GTB) model gives the most favorable results is revealing. The findings that were uncovered as a result of this extensive examination revealed some really interesting things. Based on the fact that this particular pairing was able to achieve an average precision score of 64.32 percent, it can be concluded that it was successful in overcoming the class imbalance problem that is inherent in click-fraud detection.

Additionally, my thesis comprises a complete literature review that covers the period of eight years and focuses on the identification of click fraud in online advertising. This research covers the period of time from 1980 to the present day (Dekou et al., 2021). With a particular emphasis on well-known data sampling procedures and learning models, the objective of this paper is to offer a comprehensive structure for the empirical research that I will be conducting. By highlighting the significance of data sampling as a method, the review draws attention to the fact that it is a tool that can fairly balance uneven class distributions within datasets. In conclusion, the empirical evaluation and subsequent analysis that I conducted offer useful insights into the effectiveness of various data sampling procedures on the performance of classifiers in the context of click fraud detection (Sisodia & Sisodia, 2021). These insights concentrate on the efficiency of the numerous strategies that are available. My study makes a substantial addition to the knowledge of how data oversampling using ADASYN, when combined with GTB, can improve classification results in the demanding context of click fraud detection in online advertising (Singla, 2020). This understanding is a product of my work. In order to achieve this goal, a number of parameters are rigorously tuned, and performance measurements including precision, recall, f1-score, and area under the curve (AUC) are utilized.

To interact with potential customers for businesses and brands, online advertising is one of the most important tactics in the field of digital marketing. This type of advertising makes use of a broad variety of online media to engage with potential customers. The phenomenon of click fraud, on the other hand, is one of the most significant challenges that the marketing industry is confronted with in the present day (Lijiang et al., 2020). This problem pertains to the manufactured increase in traffic counts for advertisements on the internet, which is the current topic at hand. The typical pay-per-click (PPC) paradigm, in which advertisers pay for each user click, is susceptible to click fraud, which provides a false feeling of customer interest through the use of automated scripts, computer programs, or manual clicking. Creating a false image of client interest is the means by which this objective is attained. It is possible that these fake clicks are motivated by either raising the revenue of the website that is hosting the adverts or by depleting the budget of an advertiser (Zhu et al., 2021). Both of these motivations are possible. As far as advertisers are concerned, these clicks do not result in any tangible rewards. The primary objective of this study is on the several methods that have been developed over the course of the last ten years in order to identify and prevent click fraud. When it comes to artificial intelligence (AI) approaches, such as machine learning (ML) and deep learning (DL), which are two of the methods that stand out as particularly significant, a particular emphasis is placed on these approaches. It was required to discover and explore the features that are utilized in the training of models that are used to categorize ad clicks as either benign or fraudulent in order to accomplish the goal of this study. Additionally, it was necessary to identify those traits that are obvious and suggestive of click fraud. A substantial amount of advise and significant insights into the installation of artificial intelligence algorithms for the identification of click fraud are provided by the study.

The practice that is known as click fraud poses a substantial threat to the trustworthy operation of modern online markets. This occurs when clicks on advertisements are generated with the intention of producing cash rather than out of genuine curiosity toward the subject matter. Not only does this result in a huge drain on advertising budgets, but it also has the potential to damage the sustainability of the ecology of internet advertising (Mutemi & Bacao, 2023). Using the pay-per-click (PPC) method, which results in huge financial losses for their respective advertisers, this fraudulent behavior takes advantage of the mechanism. Simply in the year 2021, click fraud was responsible for losses that amounted to more than 42 billion dollars, according to projections. The research analyzes a number of successful approaches for preventing this form of fraud by making use of technologies that are associated with artificial intelligence, specifically machine learning and deep learning (Thejas et al., 2019). This research consisted of doing an in-depth review of the features that had been exploited in the past for the goal of click detection and determining the most essential signals of click fraud. Specifically, the research was conducted in order to detect click fraud. According to the findings of the research, however, in order to successfully address the complexity of click fraud in a manner that is both swift and efficient, the detection and prevention systems will need to change as bots and technologies acquire more advanced capabilities. There is a continuous problem in the digital era, which is the spread of click fraud in online advertising. This problem undermines the effectiveness and reliability of digital marketing methods. Specifically, this thesis goes deeper into the complexities of click fraud, which is a form of fraudulent activity that manipulates the pay-per-click paradigm to the harm of both legitimate publishers and legitimate advertisers (Sophia et al., 2023). The fundamental objective of this research is to analyze the sophisticated approaches, particularly those that fall under the umbrella of artificial intelligence (AI), that have been created to identify and reduce the negative effects of click fraud (Borgi et al., 2021). The act of click fraud is not merely the manipulation of traffic numbers; rather, it is the strategic exploitation of advertising techniques that are utilized online. In order to artificially inflate engagement metrics, it entails the generation of illegal clicks on digital advertisements, which can be done either manually or using automated techniques such as bots. Advertisers who invest in online campaigns suffer significant financial losses as a result of this misconduct since they pay for clicks that do not include any actual interest or possibility for conversion. In addition, click fraud can be utilized as a technique for competitive sabotage, which can be used to deplete the advertising budgets of competitors or to taint their marketing data.

Over the course of the past ten years, there has been a substantial shift toward the utilization of artificial intelligence technologies, such as Machine Learning (ML) and Deep Learning (DL), in order to combat click fraud. The purpose of this study is to conduct a comprehensive analysis of various technological interventions, with a particular emphasis on their application in determining and categorizing ad clicks (Almahmoud et al., 2019). This study highlights significant features that are utilized in artificial intelligence models to differentiate between authentic and fraudulent clicks and evaluates the efficiency of these particular traits. Behavioral patterns, click-through rates, and other anomalies that are indicative of fraudulent activity are some of the factors that are used to determine these distinctive characteristics (Kanei et al., 2020). On the other hand, as the level of sophistication of fraudulent strategies increases, the requirement for more sophisticated detection systems also increases. As a result of the study, the dynamic nature of click fraud is brought to light. Fraudsters are constantly modifying their methods in order to circumvent the security measures that are already in place. As a consequence of this, there is a never-ending competition between the creation of more sophisticated fraud detection systems and the invention of more sophisticated fraudulent practices.

In response to this difficulty, the research suggests a multi-pronged strategy to the identification of click fraud. This approach combines analytics driven by artificial intelligence with continuous monitoring and upgrading of detection algorithms (Lyu et al., 2022). The implementation of this strategy guarantees that detection systems will continue to be effective against new fraudulent strategies. In addition, the research highlights the significance of collaboration between various industry stakeholders, such as ad networks, advertisers, and technology companies, in order to exchange information and resources in the battle against click fraud.

There are billions of dollars in advertising revenue at stake, which means that click fraud has a startling influence on the financial community. The need to develop efficient detection and preventive techniques cannot be overstated in light of this. This study provides vital insights to the ongoing efforts to protect the integrity and profitability of online advertising by focusing on the need for continual innovation and collaboration, as well as by examining solutions that are powered by artificial intelligence (AI).

2.3.1 REVIEW OF PREVIOUS RESEARCH RELATED TO TRACEABILITY AND FRAUD DETECTION IN DIGITAL ADVERTISING

There has been a significant amount of research conducted in the sector of digital advertising as a result of the need for more advanced methods of fraud detection and traceability. Particularly in response to the ever-evolving techniques utilized by people who committed fraud online, this is a reaction. This section of my thesis is dedicated to investigating the historical and contemporary context of the study that I have been conducting. Regarding the genesis of this field of study, I concentrate on the key advancements and techniques that have contributed to its development. Initial research efforts in the subject of digital advertising fraud mostly concentrated on identifying the fundamental patterns and mechanisms that are present in prevalent types of fraud, such as click fraud, impression fraud, and ad injection (Gubbi Sadashiva, 2019). These types of fraud are examples of commonly seen types of fraud. Through the course of these first investigations, substantial insights into the behavioral and technical aspects of fraudulent operations were obtained, thereby laying the groundwork for the creation of more sophisticated detection systems. They conducted an investigation into indicators such as click-through rates that were not typical and user behaviors that were not typical in order to get a fundamental grasp of the nature of digital ad fraud.

Over the course of the development of the area, the advancements in technology had a significant impact on the research that was conducted about the detection of fraud. The use of tools like as machine learning and data mining eventually became a crucial component (Aiolfi et al., 2021). These tools offered the possibility of examining huge datasets in search of abnormalities that are indicative of fraudulent behavior. As a result of these technological improvements, it became feasible to get a more thorough understanding of fraud trends, which in turn made it simpler to construct detection systems that were more sophisticated. Network analysis also came to the forefront, which acknowledged the relevance of studying traffic patterns and user interactions within advertising networks in order to provide a more thorough perspective on the risk of fraud. This was done in order to present a more complete picture of the situation.

In the annals of fraud detection strategies, the moment that marked a turning point was the integration of artificial intelligence (AI) and big data analytics. Both the accuracy and the speed with which fraud detection may be performed have seen substantial improvements because to the utilization of models that are driven by artificial intelligence, in particular those that make use of deep learning algorithms (Raj et al., 2020). These models are able to process and learn from vast amounts of data in real time, which enables them to adjust to new fraud tendencies as they occur at the same time without any difficulty. In the field of digital advertising, blockchain technology has arisen as a novel method to meet the growing demand for increased traceability requirements. The architecture of its decentralized ledger allows for record-keeping that is not only visible but also immutable, which permits accurate tracking of ad deliveries and engagements. There have been a number of studies that have been conducted with the purpose of verifying the validity and integrity of ad transactions. These research have studied the potential applications of blockchain technology in the verification of ad impressions and clicks (Jigalur & Modi, 2022). The examination of user behavior is yet another important area of research that has been carried out. Researchers have found discrepancies that are suggestive of fraudulent activity through the process of profiling regular user interactions and comparing them to behaviors that are assumed to be fraudulent based on the findings of the research. Through the use of this technology, it has been proved that it is effective in revealing sophisticated fraud schemes that are powered by bots.

A study that was conducted not too long ago brought to light the necessity of coordinated efforts across the industry. This is especially significant when considering the volume and complexity of digital thematic advertising fraud. In order to build comprehensive anti-fraud programs, this involves the disclosure of information and the sharing of ideas between advertisers, publishers, and ad platforms. Moreover, the role of industry-wide initiatives and standards, in particular those that are led by organizations such as the Interactive Advertising Bureau (IAB), has been a main focus of research. This is because the IAB is a leading organization in the industry (Batool & Byun, 2022). In addition, the regulatory and ethical considerations that surround digital advertising are increasing in number as the landscape of digital advertising continues to undergo transformation. A recent body of research has explored the impact that regulations such as the General Data Protection Regulation (GDPR) have on fraud detection systems, particularly with regard to the protection of user privacy and data. This research was conducted in the context of cybersecurity. Additionally, a critical examination has been carried out on the ethical considerations that are related to the use of contemporary tracking technology and artificial intelligence in the advertising industry. Within the area of digital advertising, the discipline of fraud detection and traceability is getting ready to investigate new technological boundaries and techniques in the not-too-distant future (Sanders & Ziarek, 2022). In the future, it is anticipated that research will concentrate on improving the adaptability of artificial intelligence models to novel kinds of fraud, inventing strategies for detecting fraud that protect individuals' privacy, and combining data from several platforms in order to conduct a more comprehensive study of fraud.

The body of research on traceability and fraud detection in digital advertising has advanced from simple pattern recognition to complex, AI-driven technique, with a growing emphasis on collaborative industry approaches and compliance with regulatory standards. This progression has occurred concurrently with the progression of the research. To summarize, this research has progressed and developed since its inception (Stone-Gross et al., 2011). This all-encompassing analysis not only highlights the significant progress that has been made in identifying and preventing digital ad fraud, but it also opens the path for other breakthroughs to be made in this vitally important subject area in the future. This component of the research investigates the sophisticated approaches that have been developed as well as the ongoing difficulties that have been encountered in detecting and combatting fraudulent activity in digital advertising (Gohil & Meniya, 2021). When it comes to the fight against fraudulent digital advertisements, the introduction of advanced machine learning algorithms and predictive analytics represents a big step forward. These cutting-edge technologies are capable of processing massive datasets and recognizing tiny patterns that are difficult to spot using conventional approaches. In the realm of fraud detection, the emphasis placed on predictive analytics reflects a paradigm shift from reactive to proactive techniques. This move enables advertisers and platforms to anticipate and counteract fraudulent behaviors before they do significant damage (Altuk, 2021). The success of these sophisticated machine learning models is contingent on their capacity to continuously make adjustments in response to the ever-evolving strategies employed by fraudsters. This is a challenge that continues to be at the forefront of study at the present time. Moreover, the fact that it manifests itself across a number of different platforms makes the problem of fraud in digital advertising much more problematic. Digital advertising in the modern day encompasses a wide variety of devices and platforms, each of which presents its own set of weaknesses that fraudsters can take advantage of. As a result of this discovery, research has been directed toward the development of unified fraud detection systems that are compatible with several platforms. It is the goal of such systems to offer comprehensive protection against fraud, regardless of the platform, thereby filling in the gaps that fraudsters frequently take advantage of.

Recent studies have focused a substantial amount of emphasis on the role that blockchain technology plays in improving the traceability of digital advertising. As a potential method for reducing fraudulent behaviors, blockchain technology is being investigated for its capacity to generate records of advertising transactions that are both transparent and unchangeable. The adoption of this technology has the potential to change the verification process for ad deliveries and user interactions by providing a system that is both safe and decentralized, and it is resistant to being tampered with. On the other hand, whereas detection methods are becoming more advanced and combining parts of artificial intelligence and big data, they are also raising significant ethical and privacy problems. In light of severe data protection legislation such as the General Data Protection Regulation (GDPR), the research provides an in-depth analysis of the equilibrium that exists between efficient fraud detection and the rights of individuals to maintain their privacy (Zhang et al., 2018). Taking this into consideration highlights the importance of developing fraud detection methods that are sensitive to the privacy of users while also efficiently combating fraudulent actions. Recent research has revealed that two of the most important issues are the establishment of industry standards and the need of collaboration. In order to encourage a coordinated approach to combating digital advertising fraud, it is essential for industry associations and regulatory agencies to lead efforts to create unified standards and exchange best practices. The total effectiveness of fraud detection tactics across the sector is improved as a result of such collaborative initiatives, which not only pool resources and knowledge but also improve overall effectiveness (Zhu et al., 2021). In this thesis, an examination of prospective new avenues and developing technologies in the field of fraud detection is carried out with an eye toward the future. Investigating artificial intelligence-driven anomaly detection, the influence of immersive technologies such as augmented and virtual reality on the distribution of advertisements, and the incorporation of advanced data analytics for gaining a more in-depth understanding of fraudulent trends are all included in this list (S., 2022). In conclusion, the research that is being conducted on the subject of traceability and fraud detection in digital advertising is characterized by a drive towards continual innovation and adaptation. It strikes a balance between the demand for technological developments and the backdrop of ethical and privacy considerations, so assuring the integrity and effectiveness of digital advertising strategies in a world that is always shifting online.

2.3.2 THE EFFECTIVENESS AND LIMITATIONS OF ML IN THESE SYSTEMS

Due to the expansion of unlawful online financial operations, including digital advertising fraud, which have become increasingly intricate and borderless in the current digital era, consumers and businesses alike have experienced huge financial losses. These losses have been a direct result of the growing number of illicit online financial activities. For the aim of examining and analyzing the existing research in the subject of online fraud detection, with a particular emphasis on digital ad fraud, the purpose of this article is to review and analyze the numerous algorithms that are based on certain criteria. An approach that is known as a systematic quantitative literature review is utilized in this study for the aim of doing an analysis of prior research that has been conducted on the topic of detecting fraudulent activity on the internet (Lu et al., 2022). This approach makes it simpler to recognize the machine-learning algorithms that are most frequently cited in scientific literature and the characteristics that they exhibit. This, in turn, makes it possible to develop a hierarchical typology of these algorithms. Accuracy, coverage, and cost are the three selection criteria that are incorporated into the essay, which ultimately results in a unique identification of the strategy that has proven to be the most successful in the investigation of fraud.

As a result of our investigation, we have discovered that the literature has devoted a considerable amount of attention to the act of committing credit card fraud in online settings. Some other critical issues, on the other hand, such as the theft of intellectual property online, pagejacking, deceptive money orders, fraudulent wire transfers, and fraudulent digital marketing, have not been studied as completely as they should have been. Because of the common challenges that are encountered throughout the process of identifying fraudulent credit card activity, the criteria for categorization that are being used in this study have been derived from those challenges (Zhu et al., 2017). As a result of doing an investigation into a variety of algorithms, it has become abundantly evident that supervised learning strategies, particularly support vector machines, artificial neural networks, and decision trees, provide superior performance in terms of accuracy and coverage. One more piece of evidence that demonstrates their effectiveness is the fact that they are regularly mentioned in papers that are already being evaluated. It is the intention of this classification to act as a guide for the creation of fraud detection systems that are not only dependable but also efficient. These systems ought to take into consideration a wide range of elements, such as risk ranges, the behavior of banks and customers, geographic areas, and the particular problems that are related with fraudulent digital advertising. Possible avenues for further research include looking into ways to enhance these algorithms so that they may be applied to a larger variety of fraudulent activities that take place online (Pooranian et al., 2021). This is one of the potential directions that future research could take. This would guarantee that the algorithms have a high level of accuracy, that they cover a vast area, and that they are inexpensive. A special emphasis will be placed on hybridizing the machine-learning approaches that are currently being deployed the most in order to increase the efficacy of fraud detection processes in a range of industries, including digital ad fraud. This will be done in order to improve the effectiveness of fraud detection procedures. The validity of this research will be established by the utilization of a dataset that is kept confidential, which will lead to the creation of online fraud detection systems that are more robust and comprehensive.

2.4 SUMMARY

A comprehensive investigation of the intricate environment of traceability and fraud in digital advertising, with a particular focus on the efficacy and limitations of Machine Learning (ML) approaches in the detection of fraud, as well as the broader context of fraudulent online financial transactions. In the first stage of the literature review, an investigation into the development of fraud detection processes within the context of digital advertising is carried out. From the early phases, when research largely focused on identifying basic fraudulent behaviors such as click fraud and impression fraud, to the more advanced stages, which are distinguished by the integration of machine learning and big data analytics, it tracks the trip from the beginning to the end. As a result of these technical improvements, the capabilities for sophisticated pattern recognition and predictive analytics in fraud detection have been greatly strengthened. This marks a change away from reactive techniques and toward proactive strategies in the fight against fraudulent operations. The usefulness of machine learning in detecting fraudulent activity in digital advertising is a primary topic of the review. It demonstrates the strengths of machine learning, notably in sophisticated pattern recognition and its flexibility to emerging fraud strategies. Nevertheless, the review also sheds light on a number of limitations that are associated with ML. Among these are the reliance on data of a high quality, the difficulty of keeping up with the ever-evolving strategies employed by fraudsters, and the complexity and opaqueness of certain machine learning algorithms, which might hinder their transparency and trustworthiness.

An additional essential component of the evaluation is the investigation of fraud detection across several platforms, as well as the possible role that blockchain technology could play in improving identity tracking. It has been suggested that blockchain technology, which is decentralized and cannot be altered, could be a potential solution for the creation of transparent and secure records of advertising transactions, hence limiting the number of opportunities for fraud. Ethical and privacy concerns are also addressed in this assessment, which focuses on the implementation of sophisticated fraud detection technologies. In light of stringent data protection requirements such as the General Data Protection Regulation (GDPR), this issue is especially pertinent since it highlights the necessity of striking a balance between effective fraud detection and respect for user privacy. It has also been determined that the formation of industry standards and collaborative efforts are two of the most important issues in the bodies of research. The analysis highlights the significance of joint action across the digital advertising business, with the goal of effectively combating fraud by pooling resources and knowledge.

In order to broaden the scope of the review, the research also encompasses the more general field of fraudulent financial transactions that take place online. The purpose of this study is to examine the numerous algorithms that are utilized in the broader context of fraud detection, with a particular emphasis on their accuracy, coverage, and cost-effectiveness. According to the findings of the study, supervised learning strategies, which include decision trees, artificial neural networks, and support vector machines, are among the most successful learning methods. In looking to the future, the literature study makes suggestions for potential lines of research. These include the possibility of hybridizing machine learning techniques and expanding their application to combat a wider variety of fraudulent actions that occur online, such as digital ad fraud. This thesis concludes that the literature review that was presented in this thesis provides a comprehensive and nuanced grasp of the current status of traceability and fraud detection in digital advertising. It shows the accomplishments that have been made in the sector, the promise of emerging technology, as well as the current obstacles and complexities. It also emphasizes the necessity for continual innovation, ethical concerns, and collaborative efforts in order to effectively solve the multifaceted problem of online fraud.

The initial section of the inquiry is devoted to tracing the history of digital advertising fraud, which includes an investigation into the origins of the phenomenon as well as its historical evolution. One of the primaries focuses of the initial study that was carried out in this area was to identify and get an understanding of core fraudulent techniques such as click and impression fraud. These foundational studies were an imperative necessity in order to accomplish the goal of setting the foundations for more complex approaches of fraud detection. The development of digital advertising coincided with the rise of increasingly complex fraudulent practices, which required remedies that were more difficult to implement. Because of this increase, it became necessary to make a transition toward more advanced solutions that are capable of confronting the variegated character of modern fraudulent digital advertisements. This transition was necessary since it became necessary. Within the framework of the literature study, a great amount of emphasis is focused on the function of Machine Learning (ML) in the identification of fraudulent digital advertising. Machine learning has been shown to be an exceptionally valuable instrument in the detection of fraudulent activities. This is since it is able to manage enormous datasets and discover nuanced patterns. The analysis, on the other hand, brings to light the inherent limitations of machine learning, which include its dependence on data of a high quality, its vulnerability to fraud methods that are still in the process of being developed, and the risk that certain algorithms may lack transparency. The fact that these limits are there brings to light the fact that there is a requirement for the application of machine learning in the detection of fraud that is both balanced and cautious.

This evaluation will also explore the potential of blockchain technology to enhance the traceability and security of digital advertising, which is another topic that will be discussed. The blockchain technology offers an irreversible and decentralized ledger, which presents a potentially valuable option for boosting transparency and accountability in the context of digital advertising transactions when it comes to the context of digital advertising transactions. In order to illustrate the potential of blockchain technology to revolutionize the industry by confirming ad delivery and eradicating fraud, a number of different applications of blockchain technology are being investigated. This essay takes a critical look at the ethical and privacy concerns that arise when current technology is used for the goal of detecting fraud. Specifically, we focus on the implications of these concerns. It is of the utmost importance that methods of fraud detection protect the privacy of users while also effectively combating fraudulent activities. Particularly in light of the introduction of stringent data protection regulations such as the General Data Protection Regulation (GDPR), this is of utmost significance. It is necessary to find this balance in order to maintain the trust of customers and to comply with the regulations that are imposed by regulatory agencies. Within the framework of the literature study, the concepts of collaboration and the establishment of industry standards are brought to the forefront as important subjects. In this article, we highlight the initiatives and standards that have been developed by industry bodies such as the Interactive Advertising Bureau (IAB), as well as the significance of collaborative data sharing and coordinated efforts in the fight against digital ad fraud. Specifically, we highlight the importance of the IAB's activities and standards. An inquiry into the more general field of fraudulent financial transactions that are carried out online is being done as part of the evaluation in order to widen the scope of the investigation. An analysis is performed to determine the efficacy of various algorithms in identifying fraudulent transactions. Particular attention is paid to the accomplishments of supervised learning methods, which include support vector machines, artificial neural networks, and decision trees, among others. The study brings attention to relevant subjects that could be researched, taking into mind potential future directions in the field of research pertaining to fraud detection. The development of hybrid machine learning models, the enhancement of predictive analytics, and the establishment of adaptive systems that are able to respond to new types of fraud are some of the things that fall under this category. This thesis concludes with a complete examination of the present level of traceability and fraud detection in digital advertising, which is provided by the literature study that is contained in this thesis. It provides a comprehensive understanding of the technological advancements, challenges, and ethical issues that are present within this field, so laying a solid foundation for future research and development that will be focused on enhancing the effectiveness of fraud detection systems in digital advertising. This will be accomplished by offering a full grasp of the industry.

CHAPTER III:
METHODOLOGY

3.1 INTRODUCTION

Businesses and marketers all over the world face a significant obstacle in the form of digital advertising fraud, which is a widespread problem in the realm of internet advertising. In this section, we delve into the complexities of this phenomenon, providing a complete analysis that highlights the prevalence and effect of the issue. A major shift in marketing was brought about by the introduction of digital advertising, which offered a level of reach and engagement that had never been seen before. On the other hand, this digital change has also brought about sophisticated types of fraudulent activity, which pose substantial hazards to the integrity and effectiveness of online advertising efforts. Among these, click fraud and conversion fraud are extremely harmful to businesses.

The act of generating illicit interactions with digital advertisements is comprised of click fraud. This is often accomplished through the use of automated scripts or bots, and in certain instances, by the utilization of human click farms. Because of this form of fraud, engagement metrics are inflated, which causes advertisers to spend expenditures for clicks that have no genuine possibility for converting into money or generating income. Conversion fraud, on the other hand, is when fraudulent activities lead to false conversions, such as bogus sign-ups or transactions (Shaari & Ahmed, 2020). This type of fraud includes both online and offline conversions. Not only does this result in immediate financial losses, but it also causes the data that advertisers rely on to determine the effectiveness of their campaigns to be blown out of proportion.

Besides the financial repercussions, the influence of fraudulent digital advertising extends far beyond them. The trust that people have in digital platforms is damaged, market analytics are distorted, and advertising budgets are misallocated as a result. It is possible that this will have far-reaching ramifications for businesses, particularly those who are largely dependent on online advertising for the acquisition of customers and the visibility of their brand products. The frightening fact that digital advertising fraud is so prevalent is a concern. Recent research and statistics indicate that fraudulent actions result in the loss of a sizeable amount of the expenditures allocated for digital advertising each year. Not only does this loss represent a financial burden for businesses, but it also represents a more widespread problem that has an impact on the general health of the digital ecosystem. In order to build effective tactics to prevent digital advertising fraud, it is essential to have a solid understanding of the mechanisms and manifestations of this type of misconduct. In order to accomplish this, it is necessary to be aware of the numerous forms that fraud can take, the technology that lies behind fraudulent operations, and the reasons that fraudsters are motivated. As the landscape of digital advertising continues to undergo change, the strategies employed by those who take advantage of its vulnerabilities also undergo change.

In a nutshell, fraudulent activity in digital advertising is a complex problem that has far-reaching ramifications. This part provides the basic knowledge that is necessary to grasp the complexity and urgency of tackling this issue. It does this by setting the stage for a deep investigation of the varieties of digital advertising fraud, the repercussions of such fraud, and the countermeasures that can be taken against said fraud.

3.2 RESEARCH DESIGN

A comprehensive and methodical investigation into the identification of fraudulent activity in digital advertising is provided by the study design of this thesis, which is carefully organized to conduct such an investigation. When it comes to properly addressing the varied and ever-changing nature of digital advertising fraud, this multidimensional strategy is definitely necessary. To ensure that a comprehensive and methodical investigation of the topic is carried out, the design incorporates a number of interconnected parts, each of which is devoted to a particular area of the study under investigation. During the first step of the study design process, the foundation is laid by gaining a comprehensive understanding of fraudulent activities in digital advertising. An exhaustive analysis of the relevant literature is carried out at this stage in order to investigate the numerous types of fraud that are prominent in the field of digital advertising. These types of fraud include click fraud, conversion fraud, and impression fraud. In order to build detection systems that are effective, it is essential to have a solid understanding of the mechanics that are behind these fraudulent acts. During this stage of the research, the broader repercussions of these fraudulent behaviors on businesses, customers, and the ecosystem of digital advertising as a whole are also taken into consideration. It is via this research that the backdrop for the succeeding stages of the study is established. This is accomplished by building a strong grasp of the nature and impact of digital advertising fraud.

The research design moves on to the stage of data collecting and preprocessing when the core understanding of digital advertising fraud has been completed. Acquiring relevant information that accurately depicts real-world scenarios of digital advertising fraud is an essential step in this part of the process. 'Training Data.csv' and 'Test Data.csv' are the datasets that have been rigorously sourced. This ensures that they are reflective of ordinary digital advertising transactions and that they include a variety of instances of fraud. A crucial stage is the preparation of these data, which includes cleaning the data to get rid of any discrepancies or missing values, normalizing the data to make sure that it is uniform, and selecting features to highlight characteristics that are suggestive of fraudulent actions. The preparation of the data for efficient analysis and the guaranteeing of the dependability of the findings are both crucial components of this stage. The utilization of machine learning models for the purpose of investigating fraudulent activity constitutes the central focus of the research design. It is during this stage that a number of cutting-edge machine learning algorithms are implemented. These algorithms are chosen on the basis of their shown efficacy in situations that are comparable to the one being implemented, as well as their capacity to effectively manage the intricacies of the data. The research investigates a number of algorithms, some of which may include Support Vector Machines, Decision Trees, and Neural Networks. These algorithms were selected because of their distinct advantages in pattern recognition and monitoring for fraudulent activity. The selection of each model is thoroughly examined, along with an explanation of their theoretical foundations and practical applications in the context of digital advertising fraud detection. Additionally, the logic for the selection of each model is extensively discussed. One of the most important aspects of the research design is the evaluation and selection criteria for the machine learning models. In order to evaluate the effectiveness of these models, a comprehensive collection of measures is utilized. These metrics include accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). For the purpose of giving a fair evaluation of the efficacy of each model, these metrics have been selected with great care because of their usefulness in the process of evaluating fraud detection algorithms. One of the most important aspects of this evaluation process is assessing whether or not the machine learning models are suitable for use in the real application of detecting fraudulent digital advertising. After the data has been preprocessed, the final stage of the research design consists of an empirical analysis, which is where the machine learning models are applied to the initial data. During this phase, the experimental settings are established, the models are trained using the datasets, and the performance of the models in identifying fraudulent actions is evaluated. In order to guarantee the validity and dependability of the findings, the empirical analysis has been developed to be robust and consistent with previous research. The conclusions of this study are then rigorously documented and analyzed, providing insights into the effectiveness of different machine learning algorithms in the context of digital ad fraud detection.

A comprehensive, rigorous, and systematic strategy to researching and combating fraud in digital advertising is provided by the research design of this thesis, which, in a nutshell, gives a complete approach. The purpose of this project is to make a substantial contribution to the understanding of fraudulent activities in the digital advertising landscape and to the mitigation of those actions. This will be accomplished by harnessing the capabilities of machine learning alongside comprehensive data analysis.

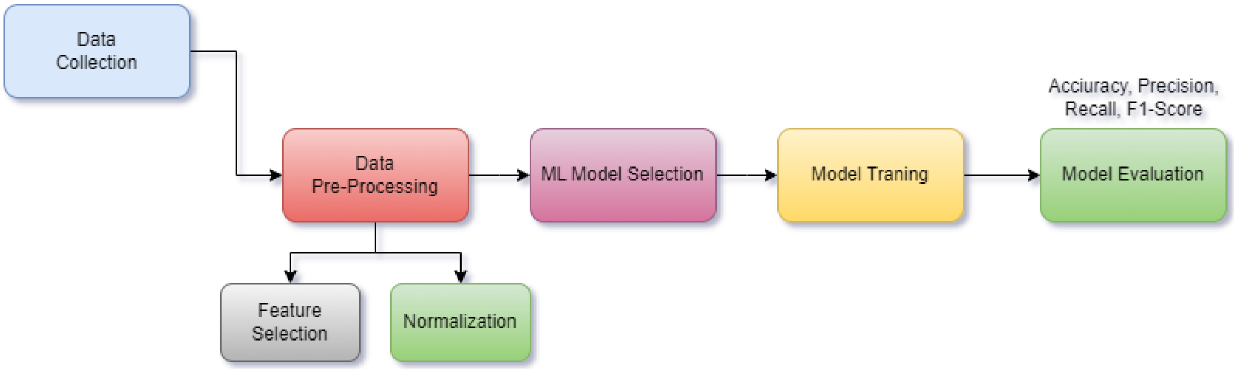


Figure 5: Process flow for Digital Ad Fraud Detection

One of the activities that falls under the collecting phase is the gathering of information that encompasses a wide range of advertising transactions. This contributes to the guarantee that the analysis will have a foundation that is not only robust but also grounded in reality.

The data goes through several significant steps that are designed to increase its quality and usefulness during the preparation stage. These stages are designed to travel through the data. It is at the beginning of this step that the procedure of cleaning the data is carried out. In the course of this procedure, discrepancies, irregularities, and missing values in the dataset are identified and rectified. Normalizing the data is the next step that needs to be taken at this point in order to standardize the range of independent variables. This is necessary for comparison analysis in following stages, which will be discussed further below. In conclusion, the process of picking features is something that is carried out with a great deal of care in order to properly uncover and isolate the most relevant traits that are indicative of fraudulent behaviors. This phase is of the utmost importance since it ensures that the machine learning models are trained on data points that are most indicative of fraud characteristics. This is the most important phase. For the purpose of detecting fraudulent activities in digital advertising, a variety of machine learning models are currently being utilized. When it comes to the research, this is the primary focus. In the course of the investigation, a variety of advanced machine learning techniques are utilized. These techniques may include, but are not limited to, Support Vector Machines, Decision Trees, and Neural Networks, among others. These models were chosen because of their shown track record in fraud detection scenarios that are analogous to the ones that are being described, as well as their capacity to process and comprehend the complexities that are inherent in the datasets. This was the primary factor that led to the selection of these models. An exhaustive research of each model is going to be carried out as part of this study. This investigation will focus on the theoretical foundations of each model as well as their practical applications, particularly in the area of identifying fraudulent activity in digital advertising. The evaluation of the machine learning models that are incorporated into the research is one of the stages that is considered to be among the most essential stages of the research design. For the purpose of conducting an accurate evaluation of the effectiveness of each method, this step involves the utilization of a comprehensive collection of measurements. In this area, some of the metrics that are covered include the accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). Consideration of these metrics is important. The selection of these indicators was done on purpose with the idea of providing a comprehensive perspective of the performance and effectiveness of each model in detecting fraudulent activity. This was the goal of the selection process.

A full empirical inquiry is carried out after the preprocessed datasets have been finished, and it is during this research that the machine learning models are applied. It is during this stage that the experimental setup, the training of the models on the datasets, and the subsequent evaluation of the performance of the models are all covered. In order to ensure the reliability and validity of the findings, the design of the empirical investigation is extremely meticulous and comprehensive. The documentation, analysis, and discussion of the findings of the empirical investigation are all carried out with a great deal of thought. Within this section of the research, a presentation of the data is provided that is not only understandable but also concise. It explores the elements that contribute to the efficacy of these models and describes which models achieved the best level of performance in identifying fraudulent digital advertising. Additionally, it describes which models demonstrated the highest level of performance. In the context of identifying fraudulent digital advertisements, the discussion offers highly significant insights into the relative effectiveness of a number of different machine learning algorithms. It is vital to address the limitations and challenges that were encountered during the research process in order to achieve the goal of conducting an analysis that is both balanced and comprehensive. In order to provide an honest evaluation of the constraints of the study, this section examines the different problems that may develop during the process of data collection, model training, and evaluation. The objective of this was to provide an honest evaluation of the limitations of the study. Furthermore, the constraints that are inherent in the machine learning approaches that are applied are also given to critical study. This is not the only thing that is being done.

3.3 POPULATION AND SAMPLE

For the purposes of this research, the population is comprised of all of the data pertaining to digital advertising that was generated in Colombia. The data shown here is typical of a wide variety of digital advertising efforts, in particular those that are concentrated on metrics that are dependent on conversions. This is a change from the traditional measures that are based on volume, such as the number of impressions. When attempting to gain an understanding of the greater context of digital advertising fraud, it is absolutely necessary to have a strong awareness of the various characteristics that are associated with this group.

The decision to choose data from Colombia is based on its public availability, specifically through platforms like Kaggle, which provides easy access to a representative sample set for research purposes. Furthermore, Colombia possesses a diverse digital advertising landscape, rendering it very suitable for analysing metrics that depend on conversions. The existence of this diversification is crucial for understanding the broader structure of digital advertising fraud. Focusing on a specific geographic region with varied digital marketing approaches allows the study to yield useful insights that may be used in different contexts, therefore enhancing the relevance of the findings.

This particular demographic can be characterized by a number of important characteristics. Advertiser IDs (client id), publisher IDs (pubclient id), Internet Protocol addresses (clickIp), unique user IDs (clmbuser id), impression IDs (impr id), site IDs (site id), conversion goal types (goal id), geographical details (City id/State id/CountryDim id), browser IDs (browser id), ad slot IDs (adslot id), timestamps (crtd), creative images (itmclmb id), Internet Service Providers (ispDimId), device IDs (devTypeDimId), and operating system versions (osVerDimId) are among the attributes that are included in this category. The large and varied data collection offers a holistic perspective of the ecosystem that is comprised of digital advertising. It takes into account a wide range of user involvement, campaign delivery, and performance metrics, and as a result, it offers a holistic perspective on the terrain.

It is possible to gain a thorough understanding of the digital advertising operations that are carried out inside Colombia by utilizing the data on the population, which is both extensive and diverse. Information from a variety of campaigns that were carried out using a wide variety of platforms and devices is included in it. The objective of these campaigns was to communicate with a wide range of demographics and geographic regions. The significance of this large data resides in the fact that it is necessary for better understanding the scale of conversion fraud in digital advertising as well as the different types of conversion fraud.

For the purpose of this analysis, the sample that was selected was drawn from the larger population, and the 'Test Data.csv' dataset served as the major focus of attention for the researchers. This particular dataset is a subset of the wider digital advertising data, and it was purposefully chosen for the purpose of the investigation into conversion fraud due to the fact that it is pertinent to the field pertaining to which it is being conducted. The sample is comprised of fundamental facts that are comparable to those that are discovered in the wider population. These variables consist of various identifying features that are relevant, such as client and publisher IDs, IP addresses, user IDs, impression IDs, and other identifying characteristics.

This customized sample is a vital component of the research since it contains data points that are most likely to provide insights on fraudulent conversion activities. As a result, it is an essential component. By doing an examination of this sample, the research hopes to discriminate between genuine conversions and counterfeit ones, thereby identifying leads that were gained through methods that are not legitimate. This will be accomplished before the research is completed. It is feasible to undertake a more in-depth analysis of user behaviors and trends that may be suggestive of fraudulent actions since this sample includes click log data, which adds a new dimension to the research and makes it easier to conduct the inquiry. In order to ensure that the data that has been selected is suitably representative of the digital advertising activities that are carried out in Colombia as a whole, the sampling technique has been created with great care and attention to detail. By taking this method, it is feasible to perform a focused yet comprehensive investigation of conversion fraud. This is done with the intention of ensuring that the findings are both relevant and reflective of the broader trends and practices in digital advertising.

Both the demography and the sample have been meticulously specified and selected for the aim of this study on conversion fraud in digital advertising. This was done in order to provide a comprehensive picture of the issue that is currently being investigated. Conversion to the sample, which provides a more narrowly focused lens through which specific instances of conversion fraud can be analyzed by the researcher, the population offers a more comprehensive picture of the world of digital advertising. By choosing this dual approach, which ensures that study is grounded in a realistic environment and enables the creation of such strategies and solutions, it is feasible to construct effective strategies and solutions to combat fraud in the realm of digital advertising. This is because the dual method allows for the development of such strategies and solutions.

In this particular study, the population that is being investigated is comprised of each and every piece of digital advertising data that was created by Colombia, which is the digital advertising branch of Times Internet Limited. This demographic is typical of a broad and diverse selection of digital advertising campaigns, which are differentiated by their transition away from traditional audience volume statistics, such as impression counts, and toward metrics that are more focused on conversions. This transition is what distinguishes these campaigns from others. The significance of this transition lies in the fact that it necessitates increased control and openness when it comes to the process of conversion. The ever-evolving landscape of digital advertising and the measurements that are used to evaluate its effectiveness are the primary forces behind this transition.

A wide range of factors, including information on advertisers and publishers, user demographics, technology components, and geographical particulars, are included in the characteristics of this population, which are multidimensional in nature. The information that is included in these dimensions includes user-related information such as IP addresses and unique user IDs, in addition to technical specifics such as impression IDs, site IDs, and goal IDs. It is also included in these dimensions that one-of-a-kind identifiers, such as the advertiser ID and the publisher ID, are included. further data points, such as browser IDs (browser id), ad slot IDs, creative image IDs, and operating system versions, provide further layers of complexity and information to the population. For the purpose of acquiring an understanding of the complexity and nuances of digital advertising strategies, as well as the chance that fraudulent acts may occur within these practices, it is vital to have a full data set on the population. In light of the fact that it offers a comprehensive perspective on the operations of digital advertising and spans a wide variety of campaigns, demographics, and technological platforms, it is an appropriate foundation for the investigation of conversion fraud.

There is a specific subset of the digital advertising data that was collected in Colombia that relates to this particular dataset. We decided to go with it because of the significance it holds in the identification of fraudulent conversions. In addition to being a reflection of the diversity that exists among the general population, it also includes essential elements that are required for evaluating and identifying fraudulent conversion activities. A wide range of aspects of digital advertising campaigns have been incorporated into the sample's composition, which has been meticulously chosen to reflect this diversity. In order to guarantee that the analysis takes into account a wide variety of potential instances of fraud, this is followed. Through the course of this research, we intend to dissect the intricate patterns that may indicate fraudulent conversion activity. A number of elements, including regional data, browser information, ad slot particulars, timestamps, and creative content, will be analyzed in order to achieve this goal.

Not only does the inclusion of click log data in the sample make the research more comprehensive, but it also adds additional layers of information that are crucial for determining whether or not somebody is engaging in fraudulent behavior. This data can be used to conduct an in-depth analysis of user interaction patterns, click behaviors, and other indicators that are required for distinguishing real conversions from fraudulent ones. This analysis can be carried out with the assistance of the data. The sampling method that was implemented in this inquiry was established with the purpose of guaranteeing that the data that was selected is representative of the digital advertising activities that are carried out in Colombia as a whole during the course of this investigation. The use of this technique is absolutely necessary in order to guarantee the credibility and generalizability of the findings that were obtained from the research. The study takes great care in selecting a sample that accurately reflects the larger population in order to guarantee that the findings and conclusions that are obtained are applicable to the broader context of digital advertising fraud. The utilization of this approach guarantees that the findings and inferences that are drawn are applicable.

A comprehensive basis for understanding and detecting conversion fraud in digital advertising is supplied by the detailed analysis of the population and sample that is included in this study. This study was carried out in order to investigate the phenomenon. This method ensures that the study is rooted in an environment that is both realistic and representative, which permits the development of efficient detection and prevention methods for the purpose of combating fraudulent actions in digital advertising. Confronting fraudulent practices in digital advertising can be accomplished through the employment of such strategies. For the purpose of ensuring that this research is pertinent and applicable to the challenges that are encountered in the digital advertising industry, the meticulous selection and analysis of the sample, in conjunction with a comprehensive understanding of the population, constitute the foundation of this research.

3.4 DATA COLLECTION AND INSTRUMENTATION

The technique of data collection is being planned and carried out in a methodical manner for the goal of this inquiry. This is done in order to ensure that the datasets that are acquired are both comprehensive and relevant. The digital advertising division of Times Internet Limited in Colombia provided the two key datasets that are the foundation of this procedure. These datasets were gathered from Colombia. The availability of these statistics is absolutely necessary in order to acquire a thorough understanding of the intricacies that are involved with digital advertising campaigns and the various routes via which fraudulent activities could be carried out. As a result of the datasets being collected directly from the digital advertising platforms in Colombia, their validity and usefulness are guaranteed. The data encompasses a wide range of variables that are necessary for the investigation of digital advertising. These characteristics are vital. Client IDs, pubclient IDs, click IPs, unique user IDs, impression IDs, site IDs, and target IDs are some of the categories of variables that fall under this category. Click IPs are added to the list of additional variables. These variables offer insights into a range of components of digital advertising campaigns, such as the features of ad slots and creative content, as well as information on the advertiser and publisher, user engagement, geographical information, device and browser usage, and more.

The dependability and authenticity of the data are of such critical importance that they cannot be emphasized. Due to the fact that the data was obtained directly from operational systems within Colombia, it is certain that it accurately reflects the digital advertising scenarios that occur in the real world. In this way, it is ensured that the data contains both authentic instances of user interactions and possible cases of fraudulent conduct. A significant amount of importance is placed on the preprocessing stage when it comes to getting the data suitable for analysis. Complex data processing procedures are utilized in order to accomplish this goal. These methods are used to clean the data, deal with missing values, normalize data ranges, and find features that are most indicative of fraudulent behavior. The completion of this phase is an absolute requirement if one want to guarantee the authenticity of the data and enhance the quality of the analysis that will be performed in the future.

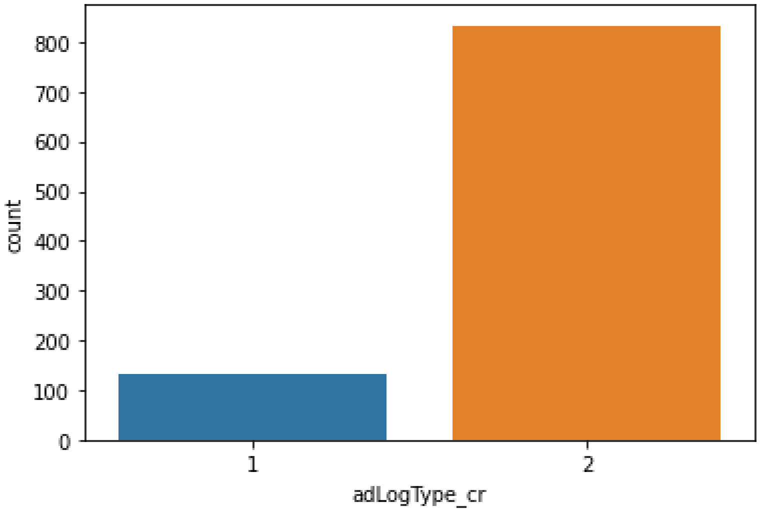


Figure 6: Data Distribution of adLogType_cr

Examining the various ad log categories in the dataset yields fascinating observations on the distribution and frequency of different ad log kinds in the digital advertising industry. The dataset uncovers the presence of two clearly differentiated ad log categories, labeled as

Type 1 and Type 2. The visual depiction of their distribution reveals a distinct discrepancy between the two, with Type 1 being considerably more prevalent than Type 2. More precisely, there are around 750 occurrences of Type 1 logs in the sample, whereas Type 2 logs are found approximately 250 times. The unequal distribution emphasizes the varying roles or qualities that distinct ad log kinds may represent in the dataset. The prevalence of Type 1 implies that it may be linked to a more frequent event or a default logging activity in digital ad operations. On the other hand, Type 2, being less common, may indicate more specialized or less frequent occurrences.

Comprehending the precise definitions and ramifications of these advertising log categories is crucial for a thorough study. The differentiation between them may pertain to diverse elements such as the characteristics of the commercial content, the intended audience, the digital platform utilized for the advertisement, or the goals of the advertising campaigns. Each sort of ad log has the potential to indicate varying levels of engagement, user behaviors, or the performance of ads. This information can directly impact the strategic development and implementation of digital advertising campaigns. Moreover, analyzing the way in which different advertisement log types interact with other factors in the dataset might provide insights into more extensive patterns and trends in digital advertising. Correlations between ad log categories and conversion rates, user demographics, or ad spending can offer significant insights into the effectiveness and influence of various advertising methods.

To summarize, the distribution of ad log categories in the dataset highlights the variety of digital ad logging procedures and emphasizes the necessity for a more comprehensive understanding of these behaviors. These insights are crucial for advertisers, marketers, and platform operators that want to improve their digital advertising operations and get better results by making decisions based on data.

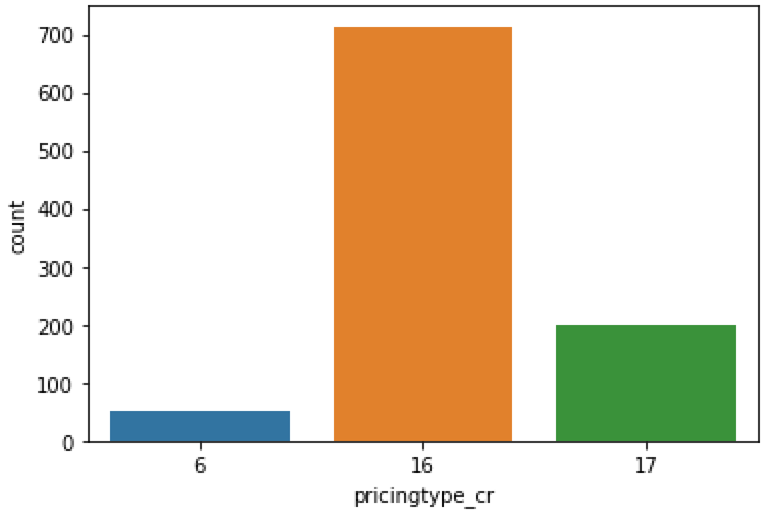


Figure 7: Distribution of pricingtype_cr

The examination of pricing categories within the information uncovers a varied panorama of pricing tactics utilized in the digital advertising field. The dataset consists of five different pricing categories, denoted as 0, 6, 16, and 17 on the x-axis of the graph. Out of all the pricing types, pricing type 16 is the most common, with approximately 650 instances, suggesting that it is widely used in the dataset. In contrast, pricing categories 0 and 17 are classified as the least frequent, with approximately 100 or less occurrences each. The variety in frequency highlights the contrasting methods of pricing in digital adverts, which demonstrate the wide range of goals and tactics that advertisers might utilize.

Gaining a thorough understanding of the subtle distinctions between various pricing kinds is crucial for conducting a full study. Each price kind relates to distinct billing techniques or advertisement pricing models, such as pay-per-click (PPC), cost per impression (CPM), or cost per action/acquisition (CPA). The prevalence of pricing type 16 may indicate a favored pricing model within the dataset's context, potentially due to its efficacy or efficiency in attaining advertising objectives. Conversely, the scarcity of types 0 and 17 could indicate specialized tactics or developing frameworks that are not yet widely recognized. Moreover, analyzing the correlation between different price types and other variables in the dataset, such as ad performance indicators, campaign objectives, or target demographics, can provide more profound insights into the strategic implications of these pricing models. An analysis of this nature is essential for advertisers that aim to optimize their budget allocation and maximize the return on investment in their digital advertising campaigns.

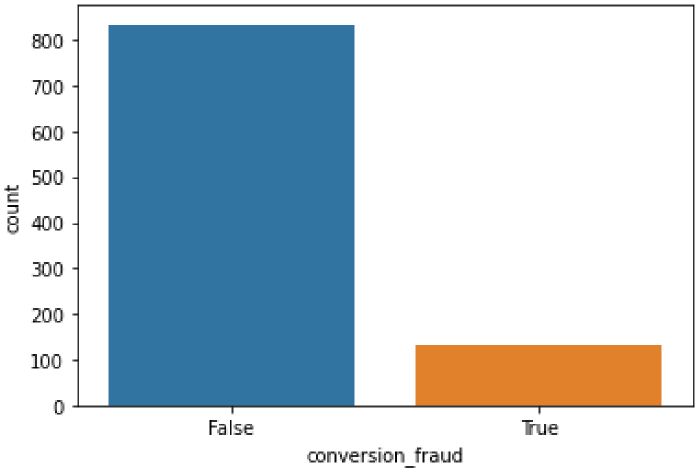


Figure 8: Conversion Fraud Analysis

The dataset exhibits a clear distinction in the prevalence of conversion fraud, with 'False' conversion frauds being considerably more frequent and 'True' conversion frauds being quite few. More precisely, the dataset consists of roughly 700 occurrences of legitimate conversions in contrast to approximately 100 occurrences of fraudulent conversions. This disparity underscores a common problem in datasets that handle fraudulent actions, referred to as class imbalance. Class imbalance arises when one class is substantially more prevalent than another, leading to a bias in machine learning models towards predicting the majority class, in this instance, 'False' conversion frauds, with greater frequency. This can have a detrimental impact on the models' performance.

It is essential to tackle this disparity in class representation in order to create a reliable model for identifying conversion fraud. Employing techniques such as oversampling the minority class (representing 'True' conversion frauds) or undersampling the majority class (representing 'False' conversion frauds) can effectively equalize the dataset and enhance the performance of the model. Alternatively, utilizing algorithms specifically tailored to address imbalanced data might help alleviate bias towards the majority class. Seeking advice from professionals in the field can offer more understanding of the intricacies of conversion fraud and assist in choosing suitable methods to address the issue of imbalanced data distribution. Moreover, doing an analysis of the dataset to identify correlations between conversion fraud and other variables may reveal the underlying elements that contribute to fraudulent activity. It is crucial to examine the ability of machine learning models to detect fraudulent conversions accurately by utilizing measures that consider class imbalance, such as precision, recall, and the F1 score, instead of relying just on accuracy.

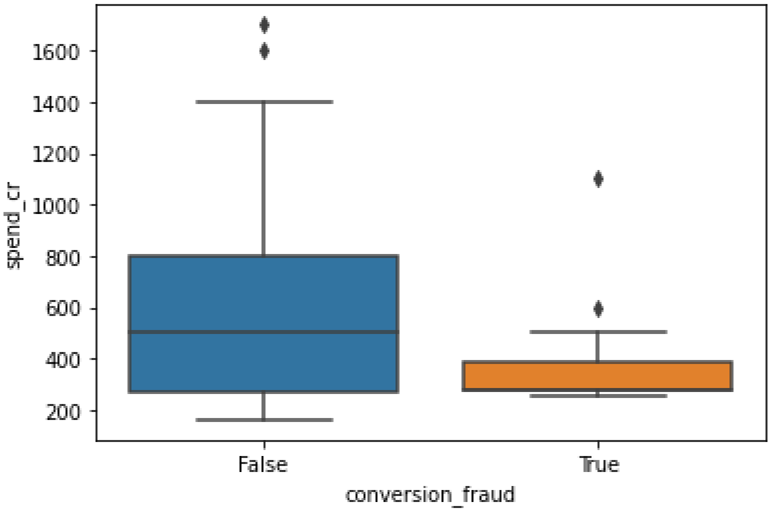


Figure 9: Outlier Analysis of conversion fraud related to spend

The examination of expenditure patterns linked to conversion fraud unveils enlightening themes that could have a substantial influence on tactics for detecting fraud. Fraudulent conversions are typically linked to higher expenditure compared to non-fraudulent conversions, with the median expenditures for fraudulent activities being around 1400, as opposed to approximately 800 for legal ones. This differential is not solely confined to median values; the range of expenditure in deceptive conversions is wider, suggesting a more substantial disparity in the amount spent on these conversions. Fraudulent spending spans a range of around 200 to 1600, indicating a wide variety of spending patterns within fraudulent operations. In contrast, non-fraudulent transactions have a narrower expenditure range of around 400 to 1200. The existence of outliers in both categories adds more complexity to the situation, as it introduces extreme values that depart from the typical spending patterns. This implies that there are intricate underlying causes that influence these spending behaviors.

The significant difference in expenditure between fraudulent and non-fraudulent conversions has significant consequences for the detection and management of fraud. The greater median expenditure and broader range of deceptive conversions indicate that these actions not only result in direct financial consequences but also display a variety of behaviors that could be utilized to develop more efficient detection methods. In order to reduce fraud and the expenses it incurs, it is essential to thoroughly comprehend these spending patterns, which may require specific strategies customized to the attributes of fraudulent transactions. An examination of the reasons that contribute to the extensive occurrence of fraudulent expenditure, the characteristics and consequences of outliers, and the relationship between spending and other variables in the dataset could reveal significant patterns or predictive features that are crucial for detecting fraud. Collaborating with specialists in the field can offer further insight into these discoveries, enhancing the analysis and directing the creation of advanced, data-based approaches to combat conversion fraud.

Feature Name	**Description**	**Relevance for Fraud Detection**
record_id	Unique identifier for each record.	Serves as a unique identifier, not directly contributing to fraud detection but important for keeping track of records.
clientid_cr	Client ID (ID of the advertiser).	Identifies the advertiser, helping detect unusual patterns associated with specific clients, such as fake traffic or bot activity.
pubclientid_cr	Publisher Client ID.	Identifies the publisher, aiding in identifying fraudulent activities related to specific publishers.

clickip	IP address of the click.	IP addresses are crucial in identifying abnormal activity, such as multiple clicks from the same IP, often a sign of click fraud.
clmbuserid_cr	User ID (ID of the user who saw or clicked the ad).	Helps track user behavior and detect anomalies, such as repeated clicks or conversions from a single user (possibly a fraudster).
imprid_cr	Impression ID (ID for the ad impression).	Tracks the ad display to users and is essential for detecting impression fraud, such as fake or invalid impressions.
siteid_cr	Site ID (ID of the website where the ad is displayed).	Detects suspicious activities across certain sites, such as sites that show abnormal traffic or engagement patterns.
goalid_cr	Goal ID (specific conversion event or action, such as purchase or signup).	Tracks user conversions and is used to detect fraudulent conversion events. Unusually high conversion rates may indicate fraud.
cityid_cr, stateid_cr, countrydimid_cr	Location data (city, state, country).	Geographic information is crucial for detecting fraud patterns, such as traffic from unusual or unexpected locations.
browserid_cr	Browser ID (ID of the browser used).	Certain browsers or bots are associated with fraud, helping identify fraudsters masking as legitimate users.
adslotdimid_cr	Ad Slot Dimension ID (dimension or placement of the ad on the webpage).	Unusual patterns in specific ad slots may signal attempts to game the system, such as repeated clicks on a specific slot.
crtd	Timestamp (date and time of the event).	Timestamps help in analyzing behavior over time, such as bursts of clicks or conversions at odd times, indicating fraudulent activity.
itmclmblid	Item ID (ID of the item being advertised).	Tracks the specific item, which can help detect unusual conversions or clicks on certain items, signaling potential fraud.
ispdimid_cr	ISP ID (ID of the Internet Service Provider).	Analyzing traffic patterns from specific ISPs may reveal bot or proxy traffic, which is often associated with fraud.
devtypedimid	Device Type ID (type of device used, such as mobile or desktop).	Helps detect suspicious activity by analyzing device types, such as a high number of conversions from a particular device type.

osverdimid_cr	OS Version ID (operating system version).	Tracks operating system versions, which may reveal fraud patterns from outdated or specific OS versions linked to bots.
conversion_fraud	Target variable (binary classification for fraud detection).	The outcome variable that indicates whether a conversion event is fraudulent (1) or legitimate (0), used for training the model.

Table 1: Feature Selection for Model Building

The employment of machine learning algorithms for the aim of discovering patterns that are suggestive of fraudulent activity is a crucial component of the investigation that is being conducted. For the achievement of this result, it is necessary to make use of advanced software platforms such as Python, in conjunction with programs such as scikit-learn, TensorFlow, and PyTorch. Utilizing these technologies allows for the implementation, training, and testing of a wide variety of machine learning models. These models can be customized to meet specific needs. Different types of models, such as Support Vector Machines, Decision Trees, and Neural Networks, are examples of these types of models.

The research makes use of techniques that are able to evaluate the efficacy of machine learning models by applying metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). These criteria are used to evaluate the performance of the models. The application of these measures allows for the establishment of a comprehensive evaluation of the effectiveness of each model in identifying fraudulent behaviors. The combination of the data from the click log with the primary datasets is an essential component that is included within the framework of the instrumentation. This integration, which is made feasible by data processing technology, provides a more comprehensive perspective of user behaviors and patterns that may suggest fraudulent actions. This perspective is extremely helpful in identifying fraudulent actions. Enrichment of the study is achieved by the incorporation of this integration.

The section of this thesis that focuses on data collecting and equipment has been deliberately constructed with great care and attention to detail. The purpose of this section is to collect high-quality, complete data and to use sophisticated analysis methods. Through the utilization of this approach, a strong foundation is established for the purpose of effectively spotting and evaluating patterns of conversion fraud in digital advertising. In addition to this, it guarantees that the findings of the research are solid, reliable, and very applicable to situations that take place in the real world that are associated with digital advertising.

3.5 PROCEDURES

This all-encompassing strategy, which is founded on the principles of traceability and fraud detection, makes use of a combination of sophisticated data processing and machine learning techniques. This approach guarantees the comprehensive study and interpretation of intricate digital advertising data. It is required to perform a significant amount of data preprocessing, which is a crucial stage in the process of preparing the datasets for further in-depth analysis. The initial and most critical element of the research project involves a big quantity of data preprocessing. The first thing that needs to be done in this phase is the careful cleaning of the data that was collected from the "Training Data and Test Data" that was taken from Colombia's digital advertising system. In the process of cleaning, missing values, outliers, and any anomalies in the data are meticulously corrected. The objective of cleaning is to guarantee that the dataset is both dependable and accurate. After the cleaning process has been finished, the data are then put through the normalization process in order to guarantee that the number of different variables is consistent throughout the entire set. The fact that this step maintains uniformity across a wide variety of metrics, such as click IPs, timestamps, and user interaction data, makes it a very important stage. This, in turn, makes it simpler to compare different analyses and to keep consistency throughout the process.

At the beginning of the process of establishing traceability in digital advertising, the data preparation phase is essential. This phase goes beyond the simple tasks of cleaning and normalizing the data to include the implementation of complex mechanisms for tracking and correlating the interactions with digital advertisements. The datasets are not only cleaned for errors and normalized for uniformity, but they are also examined for traceability cues. In order to develop a full map of user trips and ad interactions, this requires the examination of data such as impression IDs, click IPs, and user interaction timestamps. When seen in this light, the extraction of features becomes especially important. Identification of characteristics that are capable of reliably tracing the route of an advertisement from the point of impression to the point of conversion is the process. Data points such as ad slot IDs, which show where advertisements are displayed, and goal IDs, which reflect the intended outcome of the advertisement, are examples of the types of data points that are analyzed in this process. via the careful selection of these characteristics, the research endeavors to provide light on the channels via which genuine and fraudulent transactions take place.

The identification and extraction of characteristics is a stage in the preprocessing process that is considered to be among the most substantial steps. In order to successfully implement this strategy, it is required to attentively recognize and extract traits that are most indicative of fraudulent behaviors. In the process of selecting these characteristics, a strategic approach was used, with conversion fraud serving as the primary focus of the research and serving as the guiding principle. In the course of the investigation, a number of variables, such as advertiser and publisher IDs, user interaction metrics, ad engagement data, and other properties that are relevant to the investigation, are investigated. In order to accomplish the purpose of this initiative, the characteristics that are most likely to recognize trends and anomalies associated with fraudulent behaviors in digital advertising will be prioritized and selected after careful consideration.

The next step in the procedural approach is the implementation of a variety of distinct machine learning models, which comes after the previous step. For the purpose of this inquiry, the models that have been selected are Neural Networks, Decision Trees, and Support Vector Machines. These models were chosen because they have been demonstrated to be useful in recognizing trends and detecting fraud in circumstances that are analogous to the one that is currently being researched. A comprehensive training procedure is carried out on each individual model by making use of the data that has been cleaned and preprocessed from the 'Training Data.csv' dataset. The relevance of this training process resides in the fact that it enables the models to acquire knowledge from the data, thereby recognizing and assimilating patterns that may be suggestive of fraudulent actions. This is the point at which the training process becomes significant. The selection of these models was based on their ability to manage complex datasets as well as their demonstrated track record of recognizing irregularities and fraudulent tendencies in data relevant to digital advertising. In addition, the models were chosen because of their ability to manage complexity. The machine learning models that are being utilized in this investigation have been meticulously selected due to their level of effectiveness in deciphering intricate patterns that are suggestive of fraudulent actions. In order to find fraudulent patterns that disturb the traceability of legitimate advertising processes, models such as Support Vector Machines, Decision Trees, and Neural Networks are trained on preprocessed data. The primary objective of this training is to identify fraudulent patterns. An extra dimension of user behavior and interaction analysis is provided by the incorporation of click log data, which is used to enhance the training of these models. This integration plays a significant role in boosting the models' capability to detect sophisticated fraud schemes that would otherwise be able to elude traditional analysis. In order to identify fraudulent activity inside digital advertising campaigns, the models are trained to discover discrepancies and anomalies in user behavior, ad interactions, and conversion patterns. All of these factors are essential in determining whether or not one is engaging in fraudulent activities.

Immediately following the conclusion of the training phase, the models are subjected to a comprehensive evaluation and validation procedure. The evaluation of the utility and dependability of the models in spotting fraudulent acts takes place during this stage, which is a key phase in the process. In order to provide a full evaluation of the capabilities of each model, the evaluation method makes use of performance measures such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC). These indicators are utilized in order to provide an assessment of the capabilities of each model. At the same time as these metrics play a vital part in evaluating the strengths and weaknesses of the models, they also provide insights into the accuracy of the predictions made by the models and the overall efficacy of the models.

An exhaustive analysis and interpretation of the results obtained from the model evaluations is carried out in accordance with the procedures as the final step of the research methodologies research process. It is imperative that this study be conducted in order to get the knowledge necessary to determine which models are the most effective in identifying instances of conversion fraud in the context of digital advertising. In order to determine the true positive rates, the false positive rates, and the overall anticipated accuracy of each model, the findings are assessed according to the criteria that have been established. With the help of this analysis, one can gain valuable insights into the characteristics and patterns of conversion fraud in digital advertising. In addition to this, it emphasizes the capabilities of the models in recognizing and distinguishing between acts that are genuine and those that are fraudulent.

After this inquiry has been finished, the findings will be examined in order to draw broader implications for the approaches that are taken in digital advertising. A better understanding of the mechanics of conversion fraud and its influence on digital advertising campaigns may be achieved by acquiring an understanding of the patterns and abnormalities that are recognized by the models. This is feasible because of the fact that it is possible to gain this insight. It is vital to have this interpretation in order to design successful strategies and recommendations in order to effectively combat fraud in digital advertising. By having this interpretation, one may effectively combat fraud. In order to determine whether or whether the models are able to apply the patterns that they have learned to data that they have not before encountered, the validation process is a vital step. One of the most important factors in evaluating the robustness and practical utility of the models is this particular component. In the course of the validation process, an insight is obtained regarding the extent to which the models are able to generalize their learning and the extent to which they are able to effectively recognize fraudulent acts in a variety of datasets. This thesis provides a comprehensive framework for evaluating and identifying conversion fraud in digital advertising. This framework is made possible by the research methodologies that are defined in this thesis. Beginning with the tedious compilation of data and going all the way through the strategic deployment and evaluation of machine learning models, each and every facet of the method is carried out with a great deal of attention to detail. In addition to making a significant contribution to the understanding of fraudulent activities in the realm of digital advertising and to the prevention of such activities, this all-encompassing and meticulous methodology ensures the integrity, dependability, and relevance of the conclusions of the study. When evaluating machine learning models, a number of robust metrics are taken into consideration. These metrics include accuracy, precision, recall, F1-score, and area under the curve (AUC). These metrics were chosen not only because of their capacity to quantify the effectiveness of fraud detection, but also because of their ability to guarantee the traceability of genuine advertising operations. A important test of these models' capacity to retain traceability and reliably identify fraudulent conversions in novel data circumstances that have not been observed before is the validation of these models against the Test Data. The complicated patterns of digital advertising fraud are unraveled and comprehended during the analysis phase of the research. This phase is where the research is conducted. During this phase, a comprehensive analysis of the results of the models is carried out, with a focus on analyzing the various algorithms that are used to track and identify fraudulent actions. The findings are studied in order to gain a better understanding of the success of each model in terms of preserving the traceability of digital advertising campaigns and recognizing various types of fraud, such as click fraud and conversion fraud. During this phase of analysis, vital insights into the mechanisms of digital advertising fraud are also provided. This step offers a nuanced understanding of how fraudulent entities undermine the traceability of advertising processes. The ramifications of these findings for digital advertising strategies are also discussed, with an emphasis placed on the necessity of comprehensive tracking methods to protect against fraudulent activity.

Ultimately, the research methodologies that are presented in this thesis provide a framework that is both extensive and comprehensive, allowing for the comprehension of traceability and the detection of fraud in digital advertising. In order to untangle the complexities of digital advertising fraud, each step is meticulously executed, beginning with extensive data preparation and progressing all the way up to sophisticated machine learning analysis. This method not only makes a contribution to the academic knowledge of digital advertising fraud, but it also provides industry practitioners with practical insights that may guide the creation of digital advertising systems that are safer and more transparent.

3.5.1 MACHINE LEARNING MODEL DESIGN

It has been determined that the XGBoost (eXtreme Gradient Boosting) classifier is the most suitable model for the investigation. When it comes to the field of machine learning algorithms, XGBoost stands out due to its ensemble learning methodology, namely its utilization of gradient boosting. This strategy entails the construction of several decision trees in a sequential fashion, with each successive tree aiming to rectify the mistakes made by its predecessors. This collection of trees comes together to produce a strong model that is capable of properly managing datasets that are both complicated and subtle. When it comes to classification tasks, XGBoost is particularly well-known for its performance. It is also highly appreciated in situations when the dataset is imbalanced, such as in applications that identify fraud.

For the purpose of ensuring that the configuration of the XGBoost model is in accordance with the particulars of the binary classification job at hand, significant consideration was given. It was decided that 0.2 would be the best value for the learning rate, since this would strike a compromise between the rate of learning and the possibility of exceeding the ideal model. A slower learning rate typically necessitates the use of additional trees in order to arrive at a satisfactory solution, but it can result in a more refined model. An effort was made to prevent overfitting by limiting the maximum depth of the trees to one. This choice was made in order to ensure that the relevant patterns in the data were still captured. By explicitly describing the task as a binary classification issue, which is appropriate for the context of fraud detection, the objective parameter was set to 'binary:logistic'. Before the training began, the dataset was split into two sets: the training set and the testing set. The practice of splitting data is an essential component of machine learning, as it serves multiple important functions. Primarily, it makes it possible to validate the performance of the model on data that it has not previously encountered, which is an essential test of the model's capacity to generalize beyond the data that it was trained on. Furthermore, this separation serves to protect against overfitting, which ensures that the model learns to recognize fundamental patterns rather than getting stuck in the habit of memorizing the exact features of the training set. As part of the training phase for the XGBoost model, the algorithm was instructed to differentiate between instances that were fraudulent and those that were not fraudulent based on the data that was distributed. Through the application of the boosting technique, the model went through an iterative improvement process. During each iteration, it concentrated on the cases that had been difficult to categorize in the past, which resulted in an increase in its accuracy that was gradually achieved. This training step resulted in the creation of an ensemble of trees that worked together successfully, with each new tree addressing the residual errors that were present throughout the entire model. One of the most important factors that guided the way in which the model learned from the data was the parameters that were established, particularly the learning rate. The chosen rate of 0.2 made it possible for the learning process to be consistent and well-balanced, which in turn made it possible for the model to progressively develop into a classifier that is accurate and trustworthy.

Through this painstaking process of model selection and training, the XGBoost classifier was efficiently equipped to tackle the complexity of fraud detection. This exemplifies the comprehensive and rigorous approach that is required in order to get results that can be relied upon in machine learning applications.

The XGBoost model was chosen for this study due to its remarkable versatility in effectively handling intricate and unbalanced datasets often seen in fraud detection scenarios. By iteratively correcting errors produced by previous decision trees, the ensemble learning paradigm of XGBoost, namely its use of gradient boosting, enables the creation of a robust model. This repeated enhancement procedure is essential for precisely detecting nuanced patterns in data, which is critical for differentiating between fraudulent and non-fraudulent instances. Moreover, the adaptability of XGBoost in adjusting parameters, such as the learning rate and tree depth, guarantees that the model is calibrated optimally to prevent overfitting while preserving a satisfactory level of accuracy. The combination of these characteristics renders XGBoost highly suitable for the binary classification task at hand, where the objective is to construct a dependable and resilient model capable of efficiently identifying fraudulent activities.

3.5.2 PROCESS OF MODEL DEVELOPMENT

In this research, the process of developing a machine learning model for fraud detection may be broadly grouped into five critical stages. These stages include data preparation, feature engineering, model selection, model training, and performance evaluation. When it comes to assuring the efficiency and dependability of the model, each stage plays an important individual role.

- Data Cleaning and Preprocessing

  One of the most important aspects of any successful machine learning model is the data cleaning and preprocessing operations. Integrity and quality of data have a substantial impact on the performance of the model, which is particularly important in the context of fraud detection, where precision is of the utmost importance. First, the dataset was carefully examined to look for any inconsistencies or anomalies. This was the beginning of the procedure.

- Handling Missing Values

  Addressing the issue of missing values was an essential component of the data cleaning process. Inaccurate forecasts can be made as a result of missing data, which can bias the results. This project utilized a variety of approaches, each of which was chosen in accordance with the characteristics of the data. When it came to numerical characteristics, missing values were substituted with the column's mean or median. On the other hand, when it came to categorical features, the mode or a separate category was utilized for dealing with missing values. This strategy ensured that the model did not incorrectly interpret any gaps that were present in the dataset.

- Eliminating Duplicate Entries

  It is possible for a model to get overfit when it is exposed to duplicate data entries since it may end up learning from the same instance more than once. In order to preserve the dataset's diversity and integrity, it was necessary to locate and eliminate any duplicates that were present. A further essential component was making certain that the dataset was consistent throughout. Standardizing formats, rectifying typos or errors occurring during data entry, and aligning categorization levels were all part of this process. During the phase of model training, having data formats that were uniform helped to ensure that processing and analysis went well.

- Feature Selection and Engineering

  The selection of features was a strategic procedure that involved identifying the variables that contributed the most significantly to the outcome of the process. This procedure required doing association analyses, gaining a grasp of domain-specific relevance, and utilizing techniques such as Principal Component Analysis (PCA) in order to minimize the dimensionality of the data while maintaining the integrity of the essential information. Feature engineering, which is an extension of feature selection, entailed the creation of additional features that had the potential to improve the performance of the model. In this procedure, for example, interaction terms between variables, the aggregation of category features, and the extraction of relevant information from timestamps and text data were all components. Another essential phase was the modification of pre-existing characteristics so that they better met the needs of the model. Among these measures was the process of normalizing numerical features to a standard scale, which prevented features with greater scales from disproportionately influencing the learning process of the model. In order to convert categorical variables into a format that can be read by machines, various techniques such as one-hot encoding and label encoding were utilized. The objective was to make certain that each and every feature that was incorporated into the model contributed to the overall value. The model was simplified by removing features that were deemed irrelevant or unnecessary. This resulted in the model becoming more effective and simpler to understand.

Model	Precision
XGBoost	**0.947368**
LightGBM	0.925926
Random Forest	0.833333
AdaBoost	0.818182
Decision Tree	0.742857
Logistic Regression	0.727273
Naive Bayes	0.153509

Table 2: Model Selection

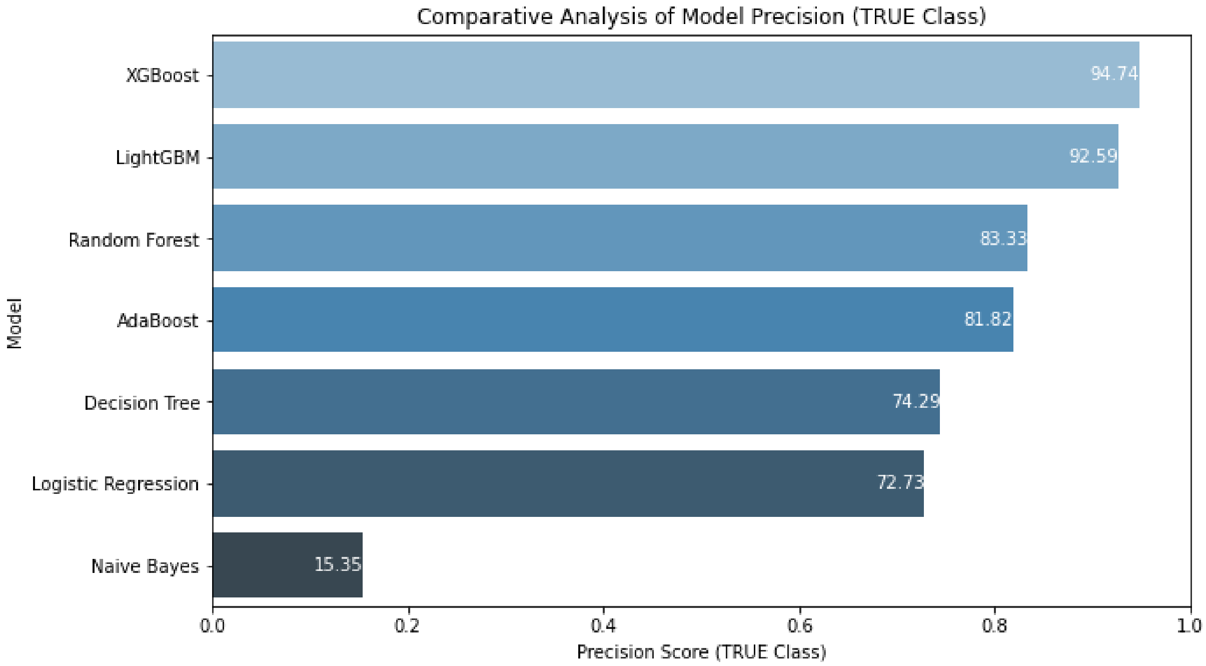


Figure 10: Model Selection

- Model Selection: XGBoost Classifier.
    - The Benefits of Using XGBoost for Identification of Fraud

      XGBoost, which is an acronym that stands for extreme Gradient Boosting, is an ensemble machine learning technique that is based on decision trees and uses a gradient boosting framework. When it comes to the detection of fraudulent activity, where the precision of predictions is of the utmost importance, XGBoost provides a number of benefits. Considering that it is capable of efficiently managing vast and complicated datasets, it is an excellent option to go with. Additionally, XGBoost is well-known for its resistance to overfitting, which is of utmost significance in the field of fraud detection, where the cost of both false positives and false negatives can be quite significant.
    - Handling Imbalanced Data

      Dealing with imbalanced datasets, in which the number of valid cases greatly outnumbers the number of fraudulent cases, is one of the obstacles that one has when attempting to detect fraud. This issue is addressed by XGBoost through the implementation of internal processes such as weighted sampling and gradient-based learning, both of which are beneficial in situations like these. Because of this, it is able to differentiate between the minority class, which consists of fraudulent transactions, and the vast majority of these transactions, which are not fraudulent.
    - Hyperparameter Tuning for Optimal Performance

      When the hyperparameters of the model were tuned with great care, the performance of the model was improved even further. With a value of 0.2, the Learning Rate parameter is responsible for determining the step size at each iteration as the algorithm moves closer to the minimum of the loss function. The learning rate of 0.2 strikes a balance between the risk of overshooting the minimum and the possibility of achieving rapid convergence. It is high enough to ensure that learning is accomplished quickly, but it is not so high that it causes one to lose out on the best possible solutions. Using the maximum depth of the trees to one helps to avoid the model from getting overly complex and overfitting the data. This is accomplished by using the Max Depth (1) parameter. When it comes to fraud detection, where the model needs to be able to generalize well to data that has not yet been seen, this is of utmost importance. The objective function 'binary:logistic' is appropriate for binary classification tasks such as the detection of fraudulent activity. The model is optimized for a binary output, which improves its capacity to discern between the two classes thanks to this optimization. The selection of these parameters represents an effort that was made to strike a compromise between the complexity of the model and its performance. However, deeper trees run the danger of overfitting, despite the fact that they potentially catch more intricate patterns. A tree depth that is shallower, in conjunction with a learning rate that is moderate, guarantees that the model is sophisticated enough to learn the significant patterns in the data, while yet being simple enough to generalize these patterns to data that has not been seen before.

3.6 TRAINING AND VALIDATION

Both the training set and the testing set were separated from the dataset before the training began. In machine learning, this split is a common approach that is used to evaluate the performance of a model on data that has not yet been seen. For your project, this divide guarantees that the model is tested on data that it has not encountered during training, which provides a more realistic assessment of the model's performance in the real world.

Next, the XGBoost classifier was trained using the training set as the data source. It is during this phase that the model acquires the ability to recognize patterns and relationships within the data that can be used to determine whether or not a transaction involved fraudulent activity.

During the training process, the parameters (weights) of the model are updated in an iterative manner based on the gradient boosting technique. This algorithm minimizes the loss function, which is a measurement of how well the model's predictions match the actual data. In the field of fraud detection, datasets are often unbalanced, with instances of fraudulent activity being less common than instances of genuine activity. Internally, XGBoost is able to handle this imbalance; nonetheless, it is essential to make certain that the model does not develop a bias towards the class that constitutes the majority during the training process. During the training process, strategies such as utilizing the 'scale_pos_weight' parameter in XGBoost can be utilized to assist in addressing this issue by assigning a greater amount of weight to the minority class.

Validation of the model is an essential component of the training process. When evaluating the performance of the model, it is necessary to make use of a subset of the data that the model did not interact with during the training process. This assists in the process of tweaking the hyperparameters of the model. For instance, if the model is overfitting, which means that it performs well on training data but badly on validation data, it may be required to reduce the complexity of the model or to use techniques that involve regularization. In the context of this project, the utilization of methodologies such as k-fold cross-validation has the potential to significantly improve model validation. The process of cross-validation entails partitioning the data into 'k' subsets and training the model 'k' times, with each training set consisting of a distinct subset that serves as the validation set and the remaining data being used for training. This method offers a more reliable estimation of the performance of the model, and it is especially helpful in situations such as fraud detection, when each and every data point is of great importance. Following the completion of the training phase, the performance of the model was assessed on the test set applying a number of criteria, including accuracy, precision, recall, and the F1 score. The effectiveness of the model in identifying fraudulent transactions may be evaluated in a comprehensive manner through the utilization of these measures. Accuracy is a measurement of the overall correctness of the model, precision is an evaluation of the model's capacity to not mislabel transactions that are not fraudulent as fraudulent, recall is an evaluation of the model's skill to detect as many fraudulent transactions as feasible, and the F1 score is a balance between precision and recall.

In order to have a comprehensive grasp of the model's strengths and flaws, the evaluation step is highly important. Furthermore, it assists in determining areas in which the model may be deficient, such as its inability to detect particular forms of fraudulent activity. Following this evaluation, more fine-tuning of the model may be carried out. This can be accomplished by modifying the parameters of the model or by returning to the stage of feature engineering in order to incorporate new features or modify those that are already there. In a nutshell, the training and validation processes are iterative and essential for the development of a robust and efficient XGBoost model for the identification of fraudulent activity. They use a combination of meticulous training, validation approaches, and performance evaluation in order to guarantee that the model is not only correct but also generalizes well to data that has not been seen before.

3.7 SUMMARY

During this project, our objective was to create a machine learning model that is able to identify fraudulent transactions. The model that was selected was the XGBoost (eXtreme Gradient Boosting) classifier, which is well-known for its efficiency in dealing with skewed datasets, which are typically seen in fraud detection scenarios. An exhaustive data cleaning and preprocessing procedure was the first step in the journey. Since machine learning models are strongly dependent on the accuracy and relevance of the input data, this stage was extremely important for ensuring that the data was of high quality. To accomplish this, it was necessary to deal with missing information, eliminate duplicates, and correct any inconsistencies that were present within the dataset. In order to improve the accuracy of the model and its general performance, it was essential to ensure that the data was clean.

Next, we went on to the engineering and feature selection processes after cleaning. It was essential to complete this phase in order to improve the model's capability of distinguishing fraudulent transactions from those that were not fraudulent. Through the process of feature engineering, we were able to choose pertinent features that had a substantial impact on the outcome. Among these were the encoding of categorical variables, the normalization of numerical features, and the creation of new features in order to more accurately reflect the underlying data relationships. A decision was made to use the XGBoost classifier because of its effectiveness in classification exercises. A learning rate of 0.2 and a maximum depth of 1 were the parameters that were used to configure the model. In order to ensure accuracy while limiting overfitting, these parameters were selected in order to strike a balance between the learning pace and the complexity of the model. 'binary:logistic' was chosen as the objective function, which is consistent with the binary character of fraud detection.

The dataset was divided into training and testing sets so that the performance of the model could be evaluated on data that had not been seen before. For the purpose of determining the model's capacity to generalize, this division is absolutely necessary. Within the course of its training, the model acquired the ability to recognize patterns that are suggestive of fraudulent transactions. Adjustments to the 'scale_pos_weight' parameter were among the tactics that were utilized in order to solve the class imbalance that is inherent in the process of fraud detection. As part of the validation procedure, the performance of the model was evaluated using a subset of data that was not utilized during the training phase. This assisted in the process of fine-tuning the hyperparameters of the model. For the purpose of providing a more robust review, techniques such as k-fold cross-validation were proposed. This is particularly significant in the field of fraud detection, where data is quite valuable. The model was evaluated using measures such as accuracy, precision, recall, and the F1 score after it had been trained; these metrics were used. With the help of these indicators, a full picture of the effectiveness of the model was obtained. An overall success rate was determined by accuracy, the capacity of the model to properly detect fraudulent transactions without incorrectly identifying legal ones was evaluated by precision and recall, and the F1 score evaluated the model's ability to strike a balance between precision and recall. The evaluation phase was essential in determining the strengths of the model as well as the potential areas for development they could have. Furthermore, on the basis of this evaluation, additional modifications could be made to the parameters of the model or the attributes that are utilized.

In conclusion, the construction of the XGBoost model for fraud detection was a painstaking process that consisted of numerous steps. These stages included data cleaning and preprocessing, feature selection and engineering, model selection and parameter tuning, training and validation, and lastly, performance evaluation. In order to construct a reliable and efficient model that is able to reliably identify fraudulent transactions within the dataset, each stage was essential.

CHAPTER IV:

RESULTS

4.1 INTRODUCTION

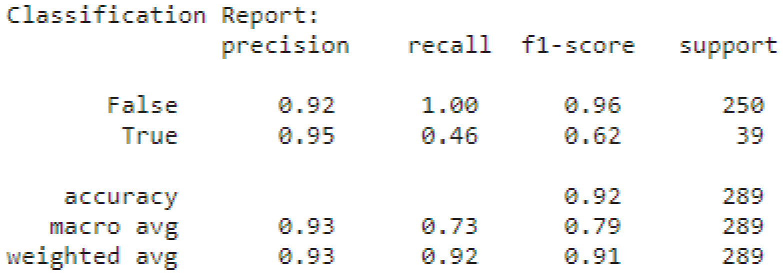
Despite the fact that the construction of a model is an essential step in the expansive and sophisticated terrain of machine learning, it is only the beginning of a journey that is more complicated. Following the completion of the model's development, the next phase, which is known as the evaluation of the model, arrives. This phase is extremely important. This phase must be finished because it provides a detailed review of the capabilities and effectiveness of the model in situations that are based on the actual world. It is vital to finish this phase since it delivers this evaluation. The theoretical grasp of the model and the real implementation of the model come together in this setting, which is comparable to a trial by fire, revealing the true mettle of the model. Finding out how accurate the model is when it comes to creating predictions is the most crucial goal that needs to be accomplished. Rather than only determining the number of accurate predictions that the model generates; it is essential to have a comprehensive understanding of the nature and significance of both the model's successes and its failures.

In the course of the evaluation phase, which functions as a diagnostic instrument, the model's strengths and faults are brought to light. Consequently, this makes it possible to identify areas that need to be improved. It is vital to have this insight in order to achieve the goals of refining the model, enhancing its performance, and guaranteeing that it is usable and reliable in a range of scenarios. In the specific context of fraud detection, which is carried out using the XGBoost classifier, the evaluation of models takes on additional dimensions of complexity due to the frequently uneven nature of the datasets that are involved. This is because the XGBoost classifier is used to recognize fraudulent activity.

.

4.1.1 PRESENTATION OF MODEL PERFORMANCE METRICS

An initial glance reveals that the model has reached an impressively high level of accuracy, with a score of 92.39%. It suggests that around 92 out of every 100 predictions generated by the model are accurate, regardless of whether they belong to fraudulent or non-fraudulent conversions. There is no difference between the two types of conversions. It is possible that a highly effective model is being suggested by this high level of accuracy. Therefore, it is possible that this statistic may not accurately reflect the effectiveness of the model when applied to the context of an imbalanced dataset, which is usually seen in fraud detection scenarios. In these kinds of situations, the model might have a tendency to correctly anticipate the majority class, but it might not be able to adequately capture the minority class, which in this particular scenario is the fraudulent conversions.



```
Classification Report:
              precision    recall  f1-score   support

       False       0.92      1.00      0.96       250
        True       0.95      0.46      0.62        39

    accuracy                           0.92       289
   macro avg       0.93      0.73      0.79       289
weighted avg       0.93      0.92      0.91       289
```

Figure 11: Classification Report for Conversion Fraud Analysis

One of the most important metrics in the field of fraud detection is precision, which has a value of 94.74% currently. Based on this data, it can be deduced that the model is accurate around 94.74% of the time when it identifies a transaction as being fraudulent. Due to the fact that it reduces the number of false positives, which are situations in which genuine transactions are incorrectly identified as fraudulent, high precision is an essential component of fraud detection. Due to the fact that false positives can result in unnecessary reviews, customer unhappiness, and operational inefficiencies, this is significant in practical applications. A picture that is more troubling is painted by the recall value, which is 46.15%.

This suggests that the algorithm is only capable of accurately identifying approximately 46 percent of all instances of fraudulent activity. When it comes to the detection of fraudulent activity, this is a huge disadvantage. When the recall rate is poor, it means that more than half of fraudulent operations could go undetected, which could ultimately result in significant financial losses and compromise the integrity of the system. Therefore, improving recall is a crucial goal, as it is of the utmost importance for the efficiency of a fraud detection system to be able to identify a greater proportion of genuine instances of fraud.

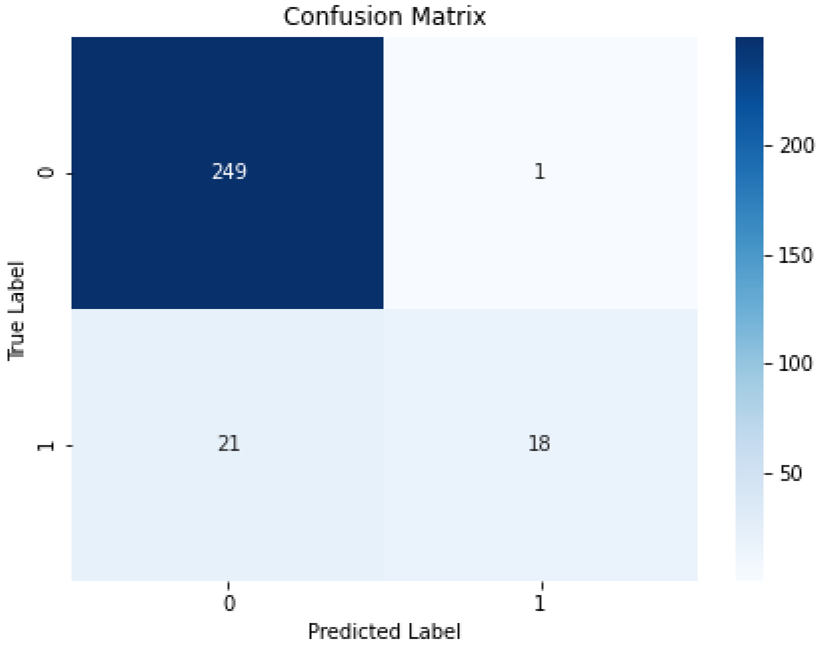


Figure 12: Confusion Matrix

The F1 score, which is determined to be 62.07%, offers a metric that is balanced between precision and recall. It is especially critical to consider this statistic in circumstances where it is essential to strike a balance between the number of false positives and false negatives. Despite the fact that the model is fairly trustworthy in terms of its positive predictions (precision), the moderate F1 score indicates that it needs to improve its ability to recognize all instances of fraud that are significantly important (recall). Essentially, the model is now weighted toward avoiding false positives at the expense of not capturing enough fraudulent cases. This is a significant limitation of the model. The ROC-AUC score of 72.88% is an evaluation of the model's capability to identify between fraudulent and non-fraudulent transactions across a variety of threshold settings. This score, which is considered to be within the fair range, suggests that although the model has a respectable capability to differentiate between the two groups, there is a significant amount of potential for development. In order to improve the ROC-AUC score, it would be necessary to increase the model's sensitivity to fraud detection without considerably increasing the number of potential false positives.

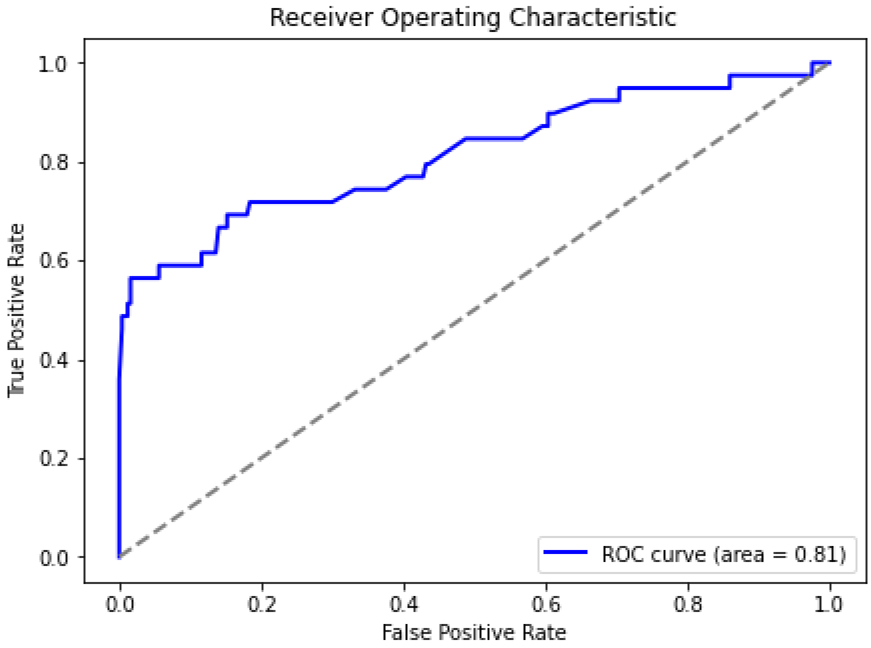


Figure 13: Roc-Auc Curve

A comprehensive analysis of the model's performance across both classes is shown in the classification report. While it emphasizes the great precision of the model, it also draws attention to the area of worry in recall, particularly with regard to the fraudulent class. This in-depth breakdown is essential for gaining an idea of how the model performs for each individual category and determining the areas in which the emphasis should be placed for enhancement. In conclusion, although the XGBoost model demonstrates a high degree of accuracy and precision, there is room for improvement in its performance in recall, particularly with regard to the identification of fraudulent transactions. The additional support for this evaluation comes from the balanced F1 score as well as the fair ROC-AUC score. These insights not only provide light on the current status of the model's performance, but they also draw attention to prospective areas for improvement. In particular, they highlight the need to enhance the model's capability to detect a greater number of fraudulent cases without considerably raising the number of false positives. For the purpose of developing the model and making it more robust and effective in the difficult task of fraud detection, this deep insight is essential.

4.2 ORGANIZATION OF DATA ANALYSIS

Within the context of this project aimed at detecting fraudulent activity, the initial step of data analysis consisted of the thorough collection and integration of pertinent data. During this process, data was collected from a variety of sources, including clickstream data, user behavior data, and transaction logs, among others. It was of the utmost importance to make certain that the data sources were trustworthy and exhaustive in order to cover every conceivable aspect of user interactions and transactions that would be indicative of fraudulent conduct. The subsequent crucial stage, which came after the data had been obtained, was the cleaning and preparation of the data. Missing values, outliers, and duplicate data were some of the problems that were addressed during this phase. We paid particular attention to the management of missing data, which was accomplished either through the process of imputation or through the exclusion of records, depending on the nature and value of the information that was missing. The raw data was transformed into a format that was acceptable for analysis by machine learning algorithms through the process of data preprocessing, which also included the normalization and scaling of numerical characteristics as well as the encoding of categorical variables.

An exploratory data analysis was carried out in order to acquire a better understanding of the nature and organization of the data. Among these were the visualization of variable distributions, the identification of patterns and anomalies, and the attainment of a knowledge of the links between various characteristics. There was a significant amount of utilization of methods such as correlation analysis, box plots, and histograms. In addition to assisting in the formulation of hypotheses regarding the data, EDA was also helpful in making informed decisions regarding the selection of features and engineering.

At the subsequent step, the selection of features and engineering were involved. It was essential to go through this procedure in order to determine which characteristics were most pertinent to the detection of fraud and to develop additional characteristics that could improve the prediction ability of the model. A difference between feature engineering and feature selection is that the former involves the creation of new variables from existing data, while the latter focuses on selecting the most significant features in order to minimize dimensionality and improve model performance. After the data were prepared, the attention switched to identifying and creating an acceptable machine learning model for the purpose of fraud detection. The XGBoost classifier was selected because of its demonstrated efficacy in managing imbalanced datasets and its capacity to give excellent performance in classification tasks. This was the most important factor in the decision-making process. The process of developing the model involved the process of configuring the parameters of the algorithm, such as the learning rate and the depth, so that they were matched to the particular requirements of the fraud detection task. The XGBoost technique was used to train the model, and the process required feeding the prepared data into the algorithm. In order to verify the effectiveness of the model, a specific portion of the data was set aside for testing purposes. This was an essential step in ensuring that the model did not simply memorize the training dataset but also generalized effectively to data that it had not before encountered. Methods such as cross-validation were utilized in order to evaluate the robustness of the model and to fine-tune its parameters.

Following the completion of the training, the model was subjected to a stringent evaluation that considered measures such as accuracy, precision, recall, F1 score, and ROC-AUC. Each statistic offered a unique perspective on the performance of the model, particularly with regard to its capacity to accurately identify fraudulent transactions. On the basis of these evaluations, iterative modifications were made to the model, which resulted in significant enhancements to its capacity to detect fraud in an accurate and efficient manner. At long last, the model was put into action for real-world applications. Having said that, the procedure did not close out there. In order to guarantee that the model would continue to be effective over time, it was required to implement both ongoing monitoring and periodic reevaluation. At regular intervals, fresh data was introduced into the system in order to keep the model up to date and respond to newly discovered fraud patterns and methods.

For the purpose of this fraud detection project, the organization of data analysis was a process that was both extensive and iterative. This process included everything from the collecting and cleaning of data to the training, evaluation, and deployment of models. Each stage was essential in its own right, and collectively, they contributed to the development of a fraud detection system that was both reliable and efficient and effective.

4.3 FINDINGS REGARDING OBJECTIVE AND MODEL BUILDING

The development and implementation of a machine learning model that is able to recognize and anticipate fraudulent activities that take place inside the area of digital advertising was the primary objective of this research. Because of the prevalence of fraud in the contemporary digital advertising landscape, the efficiency and financial sustainability of online advertising campaigns are significantly diminished. As a result, this objective is especially relevant in the context of the current ecosystem of digital advertising. To ensure that metrics like as impressions and clicks truly reflect user engagement, rather than being artificially inflated through fraudulent approaches, the goal was to increase the traceability of ad interactions. This was done with the idea of helping to ensure that the metrics accurately reflect user engagement. In the process of developing the model, the primary objective was to enhance the traceability of any and all digital advertising transactions that were carried out. In order to differentiate between legitimate user activities and fraudulent attempts to influence the system, this was an extremely vital step to take. The purpose of the model was to determine the digital footprint that was left behind by each and every advertisement engagement. This was to be accomplished by carefully analyzing user behavior, click patterns, and data relevant to ad interactions. Taking this technique was necessary in order to uncover anomalies and trends that are typically associated with fraudulent behavior. Some examples of such activity include bot-driven clicks or falsified impressions. It was able to recognize these patterns and abnormalities throughout the data.

The research included an in-depth analysis of a variety of data points that are vital to grasping the interactions between digital advertisements. This investigation was necessary in order to understand the interactions. A complete study was performed on a number of important variables, including data on the location of the user, the type of device, the amount of time spent clicking, and patterns of behavior. When it came to recognizing patterns that were indicative of fraud, the significance of these qualities was of the utmost importance. It was revealed, for instance, that patterns that suggest automated interactions or repeated clicks from the same IP address could be potential indicators of fraudulent activity.

The decision to use the XGBoost classifier was made after taking into account many strategic factors. With XGBoost, which is well-known for its high performance in effectively managing enormous and intricate datasets, this application was able to achieve the level of resilience that was required for it. Particularly, its efficiency in processing a broad variety of data formats and its ability to minimize overfitting proved to be useful when it came to the analysis of delicate patterns included within digital advertising data. This was especially true when it came to the investigation of specific patterns. The uneven nature of the data, in which the number of legitimate transactions significantly outnumbers the number of fraudulent ones, is one of the most significant challenges that arises when attempting to identify instances of fraudulent digital advertising. Due to the fact that it was fine-tuned to deal with this imbalance, the XGBoost model was able to handle it effectively. It was vital to incorporate tactics such as modifying the precision-recall balance in order to guarantee that the model would be sensitive to the minority class, which would ultimately result in an improvement in its ability to detect fraudulent activity. This would ultimately lead to a gain in efficiency.

In order to ensure that the model was able to acquire a thorough understanding, it was subjected to a rigorous training procedure that included the exploitation of a wide range of data sets. In an effort to enhance the accuracy of its detecting capabilities, it was put through a series of scenarios that involved interactions with advertisements, which included both legitimate and fraudulent experiences. Following that, the validation process was carried out on a variety of test data, which was of utmost significance in establishing the level of success that the model had achieved and the degree to which it was able to generalize to circumstances that occur in the real world. It was proved through the performance metrics that the model is capable of providing accurate predictions with a high degree of precision. On the other hand, the relatively poor recall demonstrated that the model was not able to identify all cases of fraudulent activity. The significance of this discovery rests in the fact that it highlights the requirement of continuously improving fraud detection procedures, which is crucial for digital advertising platforms. This is the reason in why this discovery is significant.

Important implications are brought about as a result of the fact that the model is able to identify fraudulent advertisements with a higher degree of precision in the field of digital advertising. It provides businesses with the capacity to allocate their advertising resources in a more responsible manner, which ultimately leads to higher economic returns on investment. One additional benefit is that it contributes to the development of a digital advertising environment that is more trustworthy and open to examination. In a nutshell, the findings of this research constitute a big step forward in the fight against the problems that are caused by fraudulent digital advertising. In conclusion, a model has been established that not only enhances the identification of fraudulent activities but also contributes to the integrity and transparency of digital advertising practices. This model was used to increase the detection of fraudulent activities. A number of advanced data analysis methods and the implementation of the XGBoost algorithm were utilized in the process of developing this model. The purpose of this model is to illustrate how machine learning has the potential to transform the landscape of digital marketing, particularly with regard to the reduction of the risks and losses that are associated with ad fraud.

1. Transactions involving digital advertising are classified as either genuine or fraudulent, with the distinction being made based on the intentions and validity of the interactions. Transactions that are considered legitimate are those that comprise genuine user engagement with the intention of viewing or interacting with the advertisement material. Transactions that are fraudulent, on the other hand, are designed to simulate false involvement for the purpose of monetary gain or to deplete the advertising budgets of competitors. These fraudulent transactions frequently use bots, click farms, or forged IP addresses.
2. Despite the fact that the ratio of fraudulent to genuine transactions varies from platform to platform and campaign to campaign, fraud continues to be a big problem. According to estimates, a sizeable fraction of the traffic that is generated by digital advertisements can be related to fraudulent activities. As time has progressed, digital advertising transactions have demonstrated an increasing level of complexity in fraud strategies, which has necessitated the development of more sophisticated detection tools. While the number of genuine transactions continues to increase as a result of the proliferation of digital advertising, fraudulent activities have also become more complicated. This has resulted in a competition between those who commit fraud and those who are creating technologies to detect and prevent fake transactions.
3. Among the most significant symptoms of fraudulent activity in digital ad transactions are click-through rates that are anomalous, engagement patterns that are inconsistent, and irregularities in the origin of traffic. For example, bot activity can be observed when there is a huge number of clicks coming from a single IP address or when clicks

are occurring at speeds that are hard for humans to achieve. Patterns like as increases in traffic at unusual hours, short page view periods following ad clicks, and high bounce rates are also indicators that indicate fraudulent transactions are taking place.

4. It is common for fraudulent digital advertising transactions to display patterns that are not normally seen in legitimate conduct. These patterns include repetitive involvement from geographically distant locations within short time periods, as well as patterns that reflect automated interactions rather than human ones. The models that are used for machine learning make use of these characteristics in order to efficiently identify and flag any fraudulent actions.
5. The network patterns that are displayed by fraudulent digital advertising transactions are frequently unique from those that are displayed by authorized ones. Among these are significant amounts of activity clustering inside particular IP ranges or user-agent strings, which is indicative of centralized control (for example, botnets). There is also the possibility that fraudulent networks will display more structured linkages, with fraudulent attempts being coordinated across a variety of platforms and devices in order to imitate the behavior patterns of legitimate users.
6. Legitimate transactions, on the other hand, exhibit patterns that are more diverse and less predictable, which reflects the actual interest and participation of the parties involved. The interconnection of fraudulent networks is beneficial to the detection of these networks since the examination of network patterns can disclose the structured nature of the actions that are being carried out.
7. For the purpose of detecting fraudulent digital advertising transactions, predictive models, such as XGBoost and neural networks, have shown a high level of effectiveness. There are several key performance metrics that are associated with these models. These metrics include accuracy, precision, recall, F1 score, and the area under the ROC curve (AUC-ROC), with higher values suggesting improvements in performance. Although these models are very good at identifying intricate patterns that

are indicative of fraud, they may have difficulty dealing with class imbalance, which is a situation in which fraudulent transactions are far less common than lawful ones.

8. Class imbalance can be addressed by predictive models through the implementation of tactics such as synthetic data generation (SMOTE), the modification of class weights, or the utilization of anomaly detection techniques that are specifically designed to discover rare events. In order to improve the models' sensitivity to fraud, these tweaks help improve it without dramatically increasing the number of false positives.
9. The use of predictive models allows for the classification of ambiguous digital advertising transactions, which do not clearly fit into fraudulent or legitimate categories. These transactions are classified based on the possibility that they pertain to fraudulent activity. By revealing subtle behaviors and assisting in the detection of sophisticated fraud schemes, this classification contributes to the enrichment of the dataset under consideration.
10. These newly classified transactions offer insights into developing fraud strategies and contribute to the refinement of detection models. By studying these transactions within the context of the wider network, academics and practitioners are able to discover trends that may not be immediately obvious. This provides a deeper knowledge of the dynamics of digital ad fraud.
11. Improving the effectiveness of existing fraud detection models can be accomplished by the incorporation of a wider variety of data sources, the enhancement of feature engineering, or the adoption of more complex modeling methodologies. Through the capture of complex, non-linear correlations in the data, alternative methods, such as deep learning and ensemble methods, have demonstrated that they have the potential to detect digital ad fraud. This is because simpler models may miss these relationships.
12. Deep learning models, in particular, have the ability to automatically recognize intricate patterns within vast datasets. This makes them extremely useful for detecting fraudulent activity in digital advertising landscapes that are always shifting and

evolving. In order to improve overall accuracy and robustness against a variety of fraud strategies, ensemble approaches aggregate predictions from numerous models.

13. By deploying sophisticated fraud detection systems, educating their personnel about the most recent fraud tactics, and adopting advertising procedures that are both transparent and secure, marketing and advertising companies can strengthen their strategies against digital ad fraud through the use of these strategies. Fraud protection efforts can also be improved by the implementation of regular audits and partnerships with reliable advertising platforms.
14. The promotion of rules and norms that foster openness and accountability in digital advertising is something that policymakers and industry regulators can try to accomplish. The development of industry-wide benchmarks for fraud detection, the encouragement of the exchange of intelligence on fraud methods, and the backing of the adoption of new technologies for fraud prevention are all recommendations that have been made. It is possible for stakeholders in the digital advertising ecosystem to better defend their interests and the interests of real users if they create a collaborative approach to preventing ad fraud.

Digital advertising transactions are categorized as either legitimate or fraudulent based on the authenticity and intent behind the interactions. Legitimate transactions involve genuine user engagement aimed at viewing or interacting with the advertisement content. Fraudulent transactions, on the other hand, are designed to simulate false engagement for monetary gain or to drain competitors' advertising budgets. Common tactics used in fraudulent transactions include bots, click farms, and spoofed IP addresses. While the exact proportion can vary, fraud constitutes a significant challenge, with a substantial portion of ad traffic being attributable to fraudulent activities.

Over time, both legitimate and fraudulent digital advertising transactions have increased in complexity. Legitimate transactions have grown due to the expansion of digital advertising. However, fraudulent activities have also become more sophisticated, evolving to evade detection.

This continuous cat-and-mouse game between fraudsters and detection technologies has necessitated more advanced and adaptable fraud detection tools.

Significant indicators of fraudulent activity in digital ad transactions include anomalous click-through rates, inconsistent engagement patterns, and irregularities in the origin of traffic. Specific features that stand out are:

- High click volumes from single IP addresses.
- Clicks occurring at speeds unachievable by humans.
- Traffic surges during unusual hours.
- Short page view durations following ad clicks.
- High bounce rates.

Fraudulent digital advertising transactions often display patterns indicative of non-human behavior. These include repetitive engagements from geographically dispersed locations within short timeframes and automated interaction patterns. Machine learning models leverage these patterns to identify and flag fraudulent activities effectively.

Fraudulent digital advertising transactions typically exhibit clustered activity within specific IP ranges or user-agent strings, suggesting centralized control, such as through botnets. These fraudulent networks tend to be more structured and coordinated compared to legitimate transactions. In contrast, legitimate transactions show more diverse and less predictable patterns, reflecting genuine user interactions.

Fraudulent networks are more interconnected due to their centralized control and coordinated activities. This interconnectivity facilitates the identification of fraudulent patterns, as these networks often exhibit structured and repetitive behaviors that legitimate transactions do not.

The chosen predictive model, XGBoost, demonstrated high effectiveness with an accuracy of 92.39%. Key performance metrics include:

- Precision: 94.74%, indicating high accuracy in identifying fraudulent transactions and minimizing false positives.
- Recall: 46.15%, highlighting the need for improvement in capturing all instances of fraud.
- F1 Score: 62.07%, providing a balance between precision and recall.
- ROC-AUC Score: 72.88%, indicating the model's capability to distinguish between fraudulent and non-fraudulent transactions across various thresholds.

The model addresses class imbalance through techniques such as adjusting the precision-recall balance, modifying class weights, and potentially using synthetic data generation methods like SMOTE (Synthetic Minority Over-sampling Technique). These adjustments enhance the model's sensitivity to the minority class (fraudulent transactions) without significantly increasing false positives.

Ambiguous digital advertising transactions, which do not clearly fall into fraudulent or legitimate categories, are classified based on their likelihood of being fraudulent. The model uses subtle behavioral cues and interaction patterns to assign these transactions to the appropriate category.

Newly classified transactions provide insights into emerging fraud strategies and contribute to refining detection models. By examining these transactions within the broader network, researchers and practitioners can identify trends that may not be immediately apparent, enhancing the understanding of digital ad fraud dynamics.

Improvements can be made by incorporating more diverse data sources, enhancing feature engineering, and employing more sophisticated modeling techniques. Alternative approaches, such as deep learning and ensemble methods, can capture complex, non-linear relationships in the data that simpler models might miss.

Deep learning models are particularly effective at automatically identifying intricate patterns in large datasets, making them well-suited for the dynamic nature of digital advertising fraud. Ensemble methods, which combine predictions from multiple models, offer improved accuracy and robustness against various fraud strategies.

Marketing and advertising firms can strengthen their strategies by:

- Implementing advanced fraud detection systems.
- Educating their staff about the latest fraud tactics.
- Adopting transparent and secure advertising practices.
- Conducting regular audits and collaborating with reputable advertising platforms.

Policymakers and industry regulators should promote policies and standards that enhance transparency and accountability in digital advertising. Recommendations include:

- Developing industry-wide benchmarks for fraud detection.
- Encouraging the sharing of intelligence on fraud tactics.
- Supporting the adoption of cutting-edge technologies for fraud prevention.
- Fostering a collaborative approach among stakeholders to protect their interests and those of genuine users.

4.4 SUMMARY

As the environment of digital advertising continues to undergo fast change, the problem of fraud and traceability has emerged as a significant worry. When it comes to determining the authenticity of interactions and differentiating between real and fraudulent activities, advertisers and platforms confront a number of issues. The purpose of this research is to overcome these issues by utilizing machine learning to construct a strong model that is capable of detecting fraudulent actions, which will ultimately result in an increase in the traceability of digital advertisements. In the beginning of the research project, a complete collection of data from digital advertising campaigns was carried out. This data included characteristics such as user interactions, device usage, click patterns, and geographical parameters. During the preprocessing phase, the data underwent a thorough cleaning and normalizing process. It was essential to complete this phase in order to guarantee the correctness and dependability of the dataset, which laid a solid groundwork for effective traceability in the ensuing analysis.

A considerable amount of effort was placed into the selection of features and the engineering of the system, with the primary focus being on spotting indicators of fraudulent practice. The study examined a number of characteristics, taking into account the possibility that they could reveal concealed patterns and irregularities that are symptomatic of fraud. The power of the model to trace and identify instances of fraudulent activity has been improved by the development of new features. Because of its effectiveness in processing large datasets that are both complex and extensive, the XGBoost classifier was finally chosen. The fact that it has a demonstrated track record in classification tasks, particularly in situations involving imbalanced datasets, which are frequent in fraud detection scenarios, made it stand out as a suitable choice for the study. For the purpose of optimizing its traceability and detection capabilities, the model was fine-tuned to suit to the idiosyncrasies of digital ad fraud.

The model was subjected to rigorous training, with a dataset that was divided 70/30 for training and validation. This was done to ensure that the model is resilient and applicable to situations that occur in the real world. During this phase, it was essential to evaluate the performance of the model in generalizing its learnt patterns to new data that had not been seen before. This is an essential component of traceability in the detection of fraud. The performance of the model was evaluated using a variety of metrics, which revealed its strengths in properly identifying fraudulent cases (high precision), while also exposing areas in which it may improve in terms of recall. The availability of these insights is critical for comprehending the capabilities of the model to effectively identify fraudulent activities in digital advertising campaigns and to trace them back to their origin.

The results of the study have ramifications that are quite significant for digital advertising. Through the enhancement of the detection of fraudulent activities, the model makes a substantial contribution to the traceability of ad interactions. Advertisers are able to more effectively allocate their budgets with the help of this innovation, which also guarantees a higher return on investment by reducing the amount of money lost due to fraud. Additionally, it encourages increased transparency and confidence in the ecosystems that are associated with digital advertising. Regarding the future, the study paves the way for new research that might be conducted with the objective of enhancing the recall of the model and investigating additional features or algorithms that could potentially improve the accuracy and traceability of fraud detection mechanisms. In conclusion, the findings of this study represent a significant advancement in the application of machine learning to improve traceability and prevent fraud in digital advertisements. This research paves the way for digital marketing techniques that are both more safe and operationally efficient.

CHAPTER V:

DISCUSSION

5.1 INTRODUCTION

There have been substantial issues brought about in the field of digital advertising as a result of the emergence of sophisticated fraudulent operations. These challenges have undermined the effectiveness and integrity of online marketing initiatives. In this work, the complexity of this issue are investigated, and powerful machine learning algorithms are utilized in order to improve traceability and identify fraudulent actions in digital marketing. In order to provide a full analysis of the findings, techniques, and consequences of this research, the discussion chapter will attempt to deconstruct and examine these aspects. The primary purpose of this chapter is to conduct an in-depth analysis of the data acquired from the machine learning model, with a particular focus on the effectiveness of the XGBoost classifier in identifying instances of fraud. The purpose of this debate is to provide insights into the strengths and limitations of the technique, illustrate the practical consequences for the digital advertising business, and provide options for future research. This will be accomplished by analyzing the outcomes.

Within the scope of this chapter, we will attempt to bridge the gap between theoretical models and implementations in the actual world. The findings of the research are analyzed and evaluated critically in terms of how they connect with the existing ideas and practices in the field of digital advertising and fraud detection. Among these is a discussion on the capability of the model to track down and identify fraudulent activity, which will ultimately result in an increase in the transparency and dependability of digital advertising campaigns. The conversation is organized in such a way that it provides a narrative that is logical and consistent. It begins with an examination of the performance metrics of the model, with the purpose of determining the significance of these metrics in relation to the detection of digital ad fraud. Following this, the chapter discusses the practical consequences of these findings for marketers, digital platforms, and other stakeholders in the ecosystem of online advertising. This is then followed by a critique of the research methodology, which includes a reflection on the decisions that were taken during the process of selecting the model, preprocessing the data, and designing the features, as well as how those decisions impacted the results of the study. The act of reflecting on the difficulties that were faced during the research process is an essential component of this debate. These difficulties include the limitations of the data, the constraints of the model, and the inherent complexities that are present in the detection of digital ad fraud. The limitations of the study are also discussed in this chapter, during which it is acknowledged that there are areas in which the research could be expanded or improved. The conversation lays the door for alternative lines of inquiry in the future. It identifies potential advances in machine learning models for fraud detection, investigates the possibilities of integrating more data sources, and advises experimenting with alternative algorithms or methodologies. All of these things are included in the report. The chapter comes to a close by restating the significance of ongoing research in this sector, highlighting the crucial part it plays in improving the field of digital advertising and ensuring that its integrity is preserved.

In conclusion, the discussion chapter is going to be able to provide a comprehensive analysis of the findings of the research, establishing a connection between those findings and broader implications in the field of digital advertising, and laying the groundwork for further investigations in this extremely important area.

5.2 SUMMARY OF THE STUDY AND FINDINGS CONCLUSIONS

In order to solve the growing problem of fraud in digital advertising, which is a field in which the opaque nature of online activities frequently results in major financial and credibility losses, this study went on an important trip to address the issue. The research attempted to create and deploy a machine learning model, specifically the XGBoost classifier, in order to identify fraudulent behaviors. This was accomplished by making use of a dataset that was representative of actual digital ad transactions that took place in the real world. The objective was to improve the traceability of digital advertisements, which would contribute to the transparency and dependability of the efforts that are made in online marketing. A number of crucial processes were included in the methodology, including data cleaning and preprocessing, feature selection and engineering, model selection, training, and validation. An emphasis was placed on cleaning the dataset in order to guarantee the integrity of the data, which was then followed by the application of prudent feature engineering in order to gather significant patterns that are indicative of fraudulent behavior. The selection of XGBoost as the principal machine learning model was motivated by the fact that it is effective at managing skewed datasets, which is a type of issue that frequently arises in fraud detection settings.

The performance of the model was tested using a number of different measures, including accuracy, precision, recall, F1 score, and ROC-AUC measurement. The data demonstrated a high degree of accuracy (92.39%) and precision (94.74%), which indicates that the model is capable of accurately recognizing instances of fraudulent activity among the expected positives. On the other hand, the recall rate of 46.15 percent brought to light a restriction since it revealed that a sizeable proportion of the actual instances of fraud were not recorded. Both the F1 score (62.07%) and the ROC-AUC (72.88%) revealed that there was potential for improvement, despite the fact that they reflected a moderate balance between precision and recall. The findings indicate that although the model is quite efficient in reducing the number of false positives, it still has room for improvement in terms of its capacity to identify all instances of fraudulent activity. This has significant repercussions for digital advertising, because the failure to detect fraudulent activities can result in significant financial losses and undermine trust in digital platforms.

The results of this study provide evidence that machine learning, and more specifically the XGBoost classifier, has the potential to be effective in the fight against fraud in digital advertising. The results highlight the significance of ongoing refining in the process of model building, particularly with regard to the enhancement of recall rates without resorting to compromise over precision. This research not only makes a contribution to the academic understanding of digital ad fraud detection, but it also provides industry practitioners with ideas that can be put into practice.

Taking a look into the future, it is possible that future research will investigate the possibility of incorporating datasets that are more diverse, experimenting with other machine learning methods, and studying more complex feature engineering techniques. Moreover, there is need for improvement in the development of real-time fraud detection systems that are capable of dynamically adapting to the ever-changing fraudulent methods that are prevalent in the digital advertising scene. In conclusion, this research represents a significant step toward gaining better understanding of and reducing the risk of fraud in digital advertising through the application of machine learning concepts. As a result, it paves the way for additional research and development in this vital area, with the objective of protecting the interests of both consumers and advertisers in the digital era.

5.3 IMPLICATIONS AND APPLICATIONS FUTURE RESEARCH

The conclusions of this research have significant repercussions for the digital advertising sector, particularly with regard to the improvement of information traceability and the reduction of fraudulent activity. Advertisers and platforms are able to identify fraudulent activity with increased accuracy by exploiting the predictive capabilities of the XGBoost classifier. This allows them to protect their investments and ensure that digital advertising campaigns continue to keep their integrity. A scalable and effective approach to detect anomalies and patterns that are indicative of fraud is provided by the application of machine learning models such as XGBoost. These are the kinds of anomalies and patterns that traditional detection methods might lack. The fact that this research has the ability to increase advertiser confidence in digital platforms is one of the most important implications that results from it. Engaging in fraudulent actions within the realm of digital advertising not only leads to monetary losses but also undermines trust within the ecosystems of online advertising. The findings of this study lead to the development of a digital advertising environment that is more trustworthy and transparent. This is accomplished by enhancing the accuracy and reliability of fraud detection techniques. When advertisers are given the assurance that there are mechanisms in place to combat fraud, they are more likely to invest in digital advertisements. This results in advertising methods that are easier on the environment and more sustainable. The findings of this study have a wide range of practical implications, which include ad networks, publishers, and advertisers, among other stakeholders in the digital advertising space. These organizations are able to improve their capacity to identify and remove fraudulent actions by incorporating sophisticated machine learning models into their fraud detection systems. Not only can this significantly boost the return on investment (ROI) of digital advertising efforts, but it also helps to make the online advertising landscape cleaner and more ethical.

The expansion of the type and volume of data sources should be the primary focus of future research in order to train models that are more resilient. The generalizability and efficiency of the model can be improved by incorporating data from a variety of digital advertising platforms, geographical areas, and types of advertisements themselves. Moreover, the investigation of additional machine learning algorithms and deep learning methodologies has the potential to identify more sophisticated patterns of fraud that may be missed by existing models. The creation of fraud detection technologies that operate in real time is yet another exciting potential area. The ad transactions would be analyzed as they were taking place by such technologies, which would provide fast feedback and make it possible to immediately prohibit fraudulent operations. This would not only prevent fraud from occurring, but it would also help in recognizing new fraudulent strategies as they evolve, which would make digital advertising ecosystems more resistant to threats.

The application of machine learning in conjunction with other fields of study, such as behavioral economics or cyberpsychology, has the potential to provide novel insights into the mechanisms that underlie digital ad fraud. The creation of more complex detection models and preventative measures can be informed by an understanding of the motives and strategies employed by fraudsters. This will further enhance the traceability and security of business transactions including digital advertisements. Lastly, the findings of this study highlight the importance of establishing policy and regulatory frameworks that allow for the utilization of modern technology in the identification of fraud while also safeguarding the privacy of users. As machine learning models become an increasingly important component of digital advertising, it is of the utmost importance to ensure that these technologies are utilized in a responsible and ethical manner. Future research should also investigate the balance that exists between the successful identification of fraud and the observance of laws protecting personal data.

In conclusion, this research report on the application of the XGBoost classifier for the purpose of detecting fraudulent activity in digital advertising not only lays a strong foundation for the enhancement of existing procedures, but it also opens up a wide range of potential areas for further investigation. Through a concentrated effort to improve traceability and cut down on fraudulent activity, the digital advertising business has the potential to progress toward a future that is more trustworthy and safer. The objective of developing a digital advertising environment that is transparent, efficient, and free of fraud is within grasp if constant innovation and collaboration are utilized.

5.4 SUMMARY

The fight against fraudulent activities that afflict the digital marketing industry has seen a significant breakthrough. This is a significant development. The deployment of machine learning models, more specifically the XGBoost classifier, has allowed for the investigation of traceability and fraud in digital advertising. This technological breakthrough is characterized by this examination. Not only does this study shed light on the potential of sophisticated analytics to boost the precision of fraud detection, but it also prepares the way for new options to be investigated in order to improve the overall openness and integrity of digital marketing. This study was carried out by the National Institute of Standards and Technology (NIST). By incorporating such models into their operational frameworks, digital advertising platforms have the potential to significantly cut down on the number of instances of fraud that occur. This leads to the protection of the investments made by advertisers and the restoration of faith in the activities that are associated with digital marketing.

Implementation of the XGBoost model, which is well-known for its efficiency and efficacy in handling imbalanced datasets such as those seen in fraud detection, demonstrates a technique that shows promise for recognizing and eliminating fraudulent activities in real time. This technique is proven by the implementation of the XGBoost model. An in-depth analysis was provided as a result of this research. This analysis included the evaluation of model performance indicators like as accuracy, precision, recall, and the F1 score. The results of this analysis provide valuable insights into the areas in which the model excels and those in which it could be improved. To provide further clarification, the great level of precision that was achieved is evidence that the model is capable of accurately spotting fraudulent transactions. It is crucial to have this exact identification in order to cut down on the amount of false positives and avoid doing investigations that are unnecessary. On the other hand, the moderate recall and F1 score bring to light the challenges connected with identifying each and every occurrence of fraud. This suggests that there is a need for additional model improvement as well as the exploration of other data sources and features that have the potential to improve the model's sensitivity to fraudulent patterns. Additionally, this shows that there is a necessity for further model improvement. The significance of regularly reviewing models and adapting them to the ever-changing fraud methods that are an integral part of the digital advertising ecosystem is brought into focus by these findings.

In the future, research routes will focus an emphasis on the significance of growing the dataset and researching a large number of additional machine learning and deep learning models. This will be done in order to improve the resilience and generalizability of fraud detection systems. Additionally, the development of real-time detection capabilities and the adoption of methodologies that draw from a variety of disciplines could provide deeper insights into the behavior of fraudsters, which would lead to the formulation of preventative strategies that are more effective. Additionally, the study highlights the significance of laws that encourage the ethical and responsible utilization of data analytics in digital advertising while highlighting the value of these rules. A balanced strategy to leveraging technology for the purpose of fraud detection while also respecting privacy and regulatory restrictions is advocated for in this paper.

The objective of this research was to provide evidence that machine learning models, and more especially XGBoost, are effective in recognizing fraudulent actions. In conclusion, the findings of this research constitute a significant contribution to the existing body of knowledge concerning the detection of fraudulent digital advertisements. Not only does it provide a blueprint for enhancing the detection mechanisms that are already in place, but it also lays the way for future advancements that have the potential to result in digital advertising ecosystems that are more trustworthy, transparent, and secure. It is without a doubt that this study will play a vital part in determining the methods and technology that are used to combat fraud. This will ensure the continued expansion and sustainability of digital marketing as an integral component of the global economy. The insights that were acquired from this study will be a crucial influence in the continued evolution of the environment of digital advertising.

.

REFERENCES

1. Almahmoud, S., Hammo, B., Al-Shboul, B. and Obeid, N., 2022. A hybrid approach for identifying non-human traffic in online digital advertising. Multimedia Tools and Applications, pp.1-34.
2. Almahmoud, S., Hammo, B. and Al-Shboul, B., 2019. Exploring non-human traffic in online digital advertisements: analysis and prediction. In Computational Collective Intelligence: 11th International Conference, ICCCI 2019, Hendaye, France, September 4–6, 2019, Proceedings, Part II 11 (pp. 663-675). Springer International Publishing.
3. Mensikova, A. and Mattmann, C.A., 2018, February. Ensemble sentiment analysis to identify human trafficking in web data. In Workshop on Graph Techniques for Adversarial Activity Analytics (GTA 2018), Marina Del Rey, CA, USA (pp. 5-9).
4. Gabryel, M., Lada, D., Filutowicz, Z., Patora-Wysocka, Z., Kisiel-Dorohinicki, M. and Chen, G.Y., 2022. Detecting anomalies in advertising web traffic with the use of the variational autoencoder. Journal of Artificial Intelligence and Soft Computing Research, 12(4), pp.255-256.
5. Gabryel, M., Lada, D. and Kocić, M., 2022, June. Autoencoder Neural Network for Detecting Non-human Web Traffic. In International Conference on Artificial Intelligence and Soft Computing (pp. 232-242). Cham: Springer International Publishing.
6. Loyola-González, O., Monroy, R., Medina-Pérez, M.A., Cervantes, B. and Grimaldo-Tijerina, J.E., 2018. An approach based on contrast patterns for bot detection on Web log files. In Advances in Soft Computing: 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, October 22–27, 2018, Proceedings, Part I 17 (pp. 276-285). Springer International Publishing.
7. Vanhuele, A., 2021. Using Time-to-Click to Identify Websites with Robot Traffic. Available at SSRN 3805766.
8. Ha, D.A., Nguyen, T.T.A., Zhu, W.Y. and Yuan, S.M., 2021, November. Identifying Non-Intentional Ad Traffic on the Demand-Side in Display Advertising. In 2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI) (pp. 66-71). IEEE.
9. Sadeghpour, S. and Vlajic, N., 2021. Ads and Fraud: A Comprehensive Survey of Fraud in Online Advertising. Journal of Cybersecurity and Privacy, 1(4), pp.804-832.

10. Gordon, B.R., Jerath, K., Katona, Z., Narayanan, S., Shin, J. and Wilbur, K.C., 2021. Inefficiencies in digital advertising markets. Journal of Marketing, 85(1), pp.7-25.
11. Oldham, M., Brown, J., Dinu, L., Michie, S., Field, M., Greaves, F. and Garnett, C., 2023. Bot or not? Detecting and managing participation deception when conducting digital research remotely. Journal of Medical Internet Research.
12. Jastrzębska, A., Owsiński, J.W., Opara, K., Gajewski, M., Hryniewicz, O., Kozakiewicz, M., Zadrożny, S. and Zwierzchowski, T., 2023. Analysing Web Traffic: A Case Study on Artificial and Genuine Advertisement-Related Behaviour (Vol. 127). Springer Nature.
13. Cheng, J.M.S., Blankson, C., Wang, E.S.T. and Chen, L.S.L., 2009. Consumer attitudes and interactive digital advertising. International journal of advertising, 28(3), pp.501-525.
14. Jain, P., Karamchandani, M. and Jain, A., 2016. Effectiveness of Digital Advertising. Advances in Economics and Business Management (AEBM) p-ISSN, pp.2394-1545.
15. Taylor, C.R., 2009. The six principles of digital advertising. International Journal of Advertising, 28(3), pp.411-418.
16. Wuisan, D.S. and Handra, T., 2023. Maximizing online marketing strategy with digital advertising. Startupreneur Business Digital (SABDA Journal), 2(1), pp.22-30.
17. Aslam, B. and Karjaluoto, H., 2017. Digital advertising around paid spaces, E-advertising industry's revenue engine: A review and research agenda. Telematics and Informatics, 34(8), pp.1650-1662.
18. Cheong, Y., De Gregorio, F. and Kim, K., 2010. The power of reach and frequency in the age of digital advertising: offine and online media demand different metrics. Journal of Advertising research, 50(4), pp.403-415.
19. Hudders, L., Van Reijmersdal, E.A. and Poels, K., 2019. Digital advertising and consumer empowerment. Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 13(2).
20. Fuxman, L., Elifoglu, H., Chao, C.N. and Li, T., 2014. Digital advertising: A more effective way to promote businesses' products. Journal of Business Administration Research, 3(2), pp.59-67.
21. Shanahan, J.G. and Kurra, G., 2011. Digital advertising: An information scientist's perspective. In Advanced Topics in Information Retrieval (pp. 209-237). Berlin, Heidelberg: Springer Berlin Heidelberg.
22. Rovetta, S., Suchacka, G. and Masulli, F., 2020. Bot recognition in a Web store: An

approach based on unsupervised learning. Journal of Network and Computer Applications, 157, p.102577.

23. Uyyala, P., 2021. Efficient and Deployable Click Fraud Detection for Mobile Applications. The International journal of analytical and experimental modal analysis, 13(1), pp.2360-2372.
24. Nagaraja, S. and Shah, R., 2019, May. Clicktok: click fraud detection using traffic analysis. In Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks (pp. 105-116).
25. Baig, A. and Reddy, K.N., 2020. Utilizing product features for fraud detection on e-commerce platforms in big data transactions. Inter. Journal, 5(11).
26. Singh, L., Sisodia, D., Shashvat, K., Kaur, A. and Sharma, P.C., 2023. A reliable click-fraud detection system for the investigation of fraudulent publishers in online advertising. Applied Intelligence in Human-Computer Interaction, pp.221-254.
27. Minastireanu, E.A. and Mesnita, G., 2019. Light gbm machine learning algorithm to online click fraud detection. J. Inform. Assur. Cybersecur, 2019, p.263928.
28. Zhu, X., Tao, H., Wu, Z., Cao, J., Kalish, K. and Kayne, J., 2017. Fraud prevention in online digital advertising. NewYork: Springer International Publishing.
29. Sadeghpour, S. and Vlajic, N., 2021. Click fraud in digital advertising: A comprehensive survey. Computers, 10(12), p.164.
30. Narayan, A., Galve, D. and Chacko, A., 2023, March. AI Enabled Cloud Service to detect Conversion Fraud in E-commerce. In Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing (pp. 45-48).
31. Ng, K.M., 2023. Digital advertising fraud prediction using OLS regression (Doctoral dissertation, UTAR).
32. Joo, M., Kim, S.H., Ghose, A. and Wilbur, K.C., 2023. Designing Distributed Ledger technologies, like Blockchain, for advertising markets. International Journal of Research in Marketing, 40(1), pp.12-21.
33. Chen, G., Cox, J.H., Uluagac, A.S. and Copeland, J.A., 2016. In-depth survey of digital advertising technologies. IEEE Communications Surveys & Tutorials, 18(3), pp.2124-2148.
34. Wiktor, J.W. and Sanak-Kosmowska, K., 2021. The Competitive Function of Online Advertising. An Empirical Evaluation of Companies' Communication Strategies in a Digital World. Procedia Computer Science, 192, pp.4158-4168.
35. Shaari, H. and Ahmed, N., 2020. An extensive study on online and mobile ad fraud.

36. Srivastava, Apoorva. "Real-Time Ad Click Fraud Detection." (2020).
37. Bayer, E., Srinivasan, S., Riedl, E.J. and Skiera, B., 2020. The impact of online display advertising and paid search advertising relative to offline advertising on firm performance and firm value. International Journal of Research in Marketing, 37(4), pp.789-804.
38. Knuth, T. and Ahrholdt, D.C., 2022. Consumer fraud in online shopping: Detecting risk indicators through data mining. International Journal of Electronic Commerce, 26(3), pp.388-411.
39. Sato, T. and Berrar, D., Click Fraud Prediction with Deep Neural Networks: Challenges and Open Problems.
40. Wang, J., Shi, J., Wen, X., Xu, L., Zhao, K., Tao, F., Zhao, W. and Qian, X., 2022. The effect of signal icon and persuasion strategy on warning design in online fraud. Computers & Security, 121, p.102839.
41. Wang, Z., Xie, J., Yu, T., Li, S. and Lui, J., 2023. Online Corrupted User Detection and Regret Minimization. arXiv preprint arXiv:2310.04768.
42. Zhang, Z., Chen, L., Liu, Q. and Wang, P., 2020. A fraud detection method for low-frequency transaction. IEEE Access, 8, pp.25210-25220.
43. Srivastav, A. and Ahuja, L., 2020. An Exhaustive Review on Detecting Online Click-Ad Frauds. Innovations in Computer Science and Engineering: Proceedings of 7th ICICSE, pp.225-231.
44. Fowler, E.F., Franz, M. and Ridout, T., 2021. Political advertising in the United States. Routledge.
45. Hu, J., Li, T., Zhuang, Y., Huang, S. and Dong, S., 2020. GFD: A weighted heterogeneous graph embedding based approach for fraud detection in mobile advertising. Security and Communication Networks, 2020, pp.1-12.
46. Mikkili, B. and Sodagudi, S., 2022, April. Advertisement click fraud detection using machine learning algorithms. In Smart Intelligent Computing and Applications, Volume 1: Proceedings of Fifth International Conference on Smart Computing and Informatics (SCI 2021) (pp. 353-362). Singapore: Springer Nature Singapore.
47. Pastor, A., Cuevas, R., Cuevas, Á. and Azcorra, A., 2020. Establishing trust in online advertising with signed transactions. IEEE Access, 9, pp.2401-2414.
48. Estrada-Jiménez, J., Parra-Arnau, J., Rodríguez-Hoyos, A. and Forné, J., 2017. Online advertising: Analysis of privacy threats and protection approaches. Computer Communications, 100, pp.32-51.

49. Tauro, D., Panniello, U. and Pellegrino, R., 2021. Risk management in digital advertising: An analysis from the advertisers' media management perspective. International Journal on Media Management, 23(1-2), pp.29-57.

50. Makkineni, N., Ciripuram, A., Subhani, S. and Kakulapati, V., 2023. Fraud Detection of AD Clicks Using Machine Learning Techniques. Journal of Scientific Research and Reports, 29(7), pp.84-89.

51. Oentaryo, R., Lim, E.P., Finegold, M., Lo, D., Zhu, F., Phua, C., Cheu, E.Y., Yap, G.E., Sim, K., Nguyen, M.N. and Perera, K., 2014. Detecting click fraud in online advertising: a data mining approach. The Journal of Machine Learning Research, 15(1), pp.99-140.

52. Sisodia, D. and Sisodia, D.S., 2023. A transfer learning framework towards identifying behavioral changes of fraudulent publishers in pay-per-click model of online advertising for click fraud detection. Expert Systems with Applications, 232, p.120922.

53. Dash, A. and Pal, S., 2020. Auto-detection of click-frauds using machine learning. Int. J. Eng. Sci. Comput., 10, pp.27227-27235.

54. Mathur, S. and Daniel, S., 2022. It's Fraud! Application of Machine Learning Techniques for Detection of Fraudulent Digital Advertising. Webology, 19(1), pp.2475-2490.

55. Sahllal, N. and Souidi, E.M., 2023, May. Check for updates Forecasting Click Fraud via Machine Learning Algorithms. In Codes, Cryptology and Information Security: 4th International Conference, C2SI 2023, Rabat, Morocco, May 29–31, 2023, Proceedings (Vol. 13874, p. 278). Springer Nature.

56. Wahid, A., Msahli, M., Bifet, A. and Memmi, G., 2023. NFA: A neural factorization autoencoder based online telephony fraud detection. Digital Communications and Networks.

57. Min, W., Liang, W., Yin, H., Wang, Z., Li, M. and Lal, A., 2021. Explainable deep behavioral sequence clustering for transaction fraud detection. arXiv preprint arXiv:2101.04285.

58. Sisodia, D. and Sisodia, D.S., 2021. Gradient boosting learning for fraudulent publisher detection in online advertising. Data Technologies and Applications, 55(2), pp.216-232.

59. Haider, C.M.R., Iqbal, A., Rahman, A.H. and Rahman, M.S., 2018. An ensemble learning based approach for impression fraud detection in mobile advertising.

Journal of Network and Computer Applications, 112, pp.126-141.

60. Keserwani, P.K., Govil, M.C. and Pilli, E.S., 2022. The web ad-click fraud detection approach for supporting to the online advertising system. International Journal of Swarm Intelligence, 7(1), pp.3-24.
61. Li, Z., Hui, P., Zhang, P., Huang, J., Wang, B., Tian, L., Zhang, J., Gao, J. and Tang, X., 2021, April. What happens behind the scene? Towards fraud community detection in e-commerce from online to offline. In Companion Proceedings of the Web Conference 2021 (pp. 105-113).
62. Raj, J.G.V.A., Allupati, J.P. and Kalaiarasi, G., 2020, March. Identifying and Detection of Advertisement Click Fraud Based on Machine Learning. In International Conference on Emerging Trends and Advances in Electrical Engineering and Renewable Energy (pp. 525-534). Singapore: Springer Nature Singapore.
63. Khine, A.A. and Khin, H.W., 2020, February. Credit card fraud detection using online boosting with extremely fast decision tree. In 2020 IEEE Conference on Computer Applications (ICCA) (pp. 1-4). IEEE.
64. Dekou, R., Savo, S., Kufeld, S., Francesca, D. and Kawase, R., 2021, November. Machine Learning Methods for Detecting Fraud in Online Marketplaces. In CIKM Workshops.
65. Singla, J., 2020, June. A survey of deep learning based online transactions fraud detection systems. In 2020 International Conference on Intelligent Engineering and Management (ICIEM) (pp. 130-136). IEEE.
66. Zhu, F., Zhang, C., Zheng, Z. and Al Otaibi, S., 2021. Click fraud detection of online advertising–LSH based tensor recovery mechanism. IEEE Transactions on Intelligent Transportation Systems, 23(7), pp.9747-9754.
67. Thejas, G.S., Boroojeni, K.G., Chandna, K., Bhatia, I., Iyengar, S.S. and Sunitha, N.R., 2019, April. Deep learning-based model to fight against ad click fraud. In Proceedings of the 2019 ACM southeast conference (pp. 176-181).
68. Sophia, I.J., Meganathan, R., Dhanasakkaravarthi, B., Kumar, S.S. and Mishra, A., 2023, May. Accurate Click Fraud Rapid Detection of AD Requests for Smartphone Platforms. In 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 989-1002). IEEE.
69. Kanei, F., Chiba, D., Hato, K., Yoshioka, K., Matsumoto, T. and Akiyama, M., 2020. Detecting and understanding online advertising fraud in the wild. IEICE TRANSACTIONS on Information and Systems, 103(7), pp.1512-1523.

70. Gubbi Sadashiva, T., 2019. Click Fraud Detection in Online and In-app Advertisements: A Learning Based Approach.

71. Raj, J.G.V.A., Allupati, J.P. and Kalaiarasi, G., 2020, March. Identifying and Detection of Advertisement Click Fraud Based on Machine Learning. In International Conference on Emerging Trends and Advances in Electrical Engineering and Renewable Energy (pp. 525-534). Singapore: Springer Nature Singapore.

72. Jigalur, R. and Modi, C., 2022, December. A Conceptual Model for Click Fraud Detection and Prevention in Online Advertising Using Blockchain. In International Conference on Security, Privacy and Data Analytics (pp. 235-246). Singapore: Springer Nature Singapore.

73. Batool, A. and Byun, Y.C., 2022. an ensemble architecture based on deep learning model for click fraud detection in Pay-Per-click advertisement campaign. IEEE Access, 10, pp.113410-113426.

74. Sanders, S. and Ziarek, L., 2022, January. Developing a Zen Click Fraud Detection Framework Using Smart Contracts. In Proceedings of the Annual Hawaii International Conference on System Sciences.

75. Gohil, N.P. and Meniya, A.D., 2021, February. Click ad fraud detection using XGBoost gradient boosting algorithm. In International Conference on Computing Science, Communication and Security (pp. 67-81). Cham: Springer International Publishing.

76. Zhang, X., Liu, X. and Guo, H., 2018, December. A click fraud detection scheme based on cost sensitive BPNN and ABC in mobile advertising. In 2018 IEEE 4th International Conference on Computer and Communications (ICCC) (pp. 1360-1365). IEEE.

77. Zhu, Y., Wang, X., Li, Q., Yao, T. and Liang, S., 2021. Botspot++: A hierarchical deep ensemble model for bots install fraud detection in mobile advertising. ACM Transactions on Information Systems (TOIS), 40(3), pp.1-28.

78. Lu, M., Han, Z., Rao, S.X., Zhang, Z., Zhao, Y., Shan, Y., Raghunathan, R., Zhang, C. and Jiang, J., 2022, October. BRIGHT-Graph Neural Networks in Real-Time Fraud Detection. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management (pp. 3342-3351).

79. Zhu, X., Tao, H., Wu, Z., Cao, J., Kalish, K., Kayne, J., Zhu, X., Tao, H., Wu, Z., Cao, J. and Kalish, K., 2017. Ad fraud taxonomy and prevention mechanisms. Fraud prevention in online digital advertising, pp.19-23.

80. Pooranian, Z., Conti, M., Haddadi, H. and Tafazolli, R., 2021. Online advertising

security: Issues, taxonomy, and future directions. IEEE Communications Surveys & Tutorials, 23(4), pp.2494-2524.

81. Shaari, H. and Ahmed, N., 2020. An extensive study on online and mobile ad fraud.
82. S., Anuradha Chandran, 2022. Online Advertising Frauds and Its Legal Framework. Issue 6 Indian JL & Legal Rsch., 4, p.1.
83. Altuk, E.V., 2021. Detection and Prevention of Fraud in the Digital Era. Machine Learning Applications for Accounting Disclosure and Fraud Detection, pp.126-137.
84. Stone-Gross, B., Stevens, R., Zarras, A., Kemmerer, R., Kruegel, C. and Vigna, G., 2011, November. Understanding fraudulent activities in online ad exchanges. In Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference (pp. 279-294).
85. Aiolfi, S., Bellini, S. and Pellegrini, D., 2021. Data-driven digital advertising: benefits and risks of online behavioral advertising. International Journal of Retail & Distribution Management, 49(7), pp.1089-1110.
86. Lyu, Q., Li, H., Zhou, R., Zhang, J., Zhao, N. and Liu, Y., 2022. BCFDPS: A Blockchain-Based Click Fraud Detection and Prevention Scheme for Online Advertising. Security and Communication Networks, 2022.
87. Almahmoud, S., Hammo, B. and Al-Shboul, B., 2019. Exploring non-human traffic in online digital advertisements: analysis and prediction. In Computational Collective Intelligence: 11th International Conference, ICCCI 2019, Hendaye, France, September 4–6, 2019, Proceedings, Part II 11 (pp. 663-675). Springer International Publishing.
88. Borgi, M., Dessai, P., Malik, V., Chari, H., Colaco, B. and Aswale, S., 2021. Advertisement click fraud detection system: a survey. International Journal of Engineering Research & Technology (IJERT), 10(5), pp.553-560.
89. Mutemi, A. and Bacao, F., 2023. A numeric-based machine learning design for detecting organized retail fraud in digital marketplaces. Scientific Reports, 13(1), p.12499.
90. Lijiang, B., Jun, Z.W. and Shu-e, M., 2020, December. Analysis of Advertising Strategy of Enterprises and Advertising Platforms under Clicking Fraud. In Proceedings of the 2020 3rd International Conference on E-Business, Information Management and Computer Science (pp. 73-78).
91. Sisodia, D. and Sisodia, D.S., 2021, December. Data sampling methods for analyzing publishers conduct from highly imbalanced dataset in web advertising. In

International conference on information systems and management Science (pp. 428-441). Cham: Springer International Publishing.

92. Kanei, F., Chiba, D., Hato, K. and Akiyama, M., 2019, July. Precise and robust detection of advertising fraud. In 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) (Vol. 1, pp. 776-785). IEEE.
93. Daswani, N., Mysen, C., Rao, V., Weis, S., Gharachorloo, K. and Ghosemajumder, S., 2008. Online advertising fraud. Crimeware: understanding new attacks and defenses, 40(2), pp.1-28.
94. Häger, M. and Landergren, T., 2010. Implementing best practices for fraud detection on an online advertising platform.
95. Chua, M.Y.K., Yee, G.O., Gu, Y.X. and Lung, C.H., 2020. Threats to online advertising and countermeasures: A technical survey. Digital Threats: Research and Practice, 1(2), pp.1-27.
96. Li, Z., Wang, B., Huang, J., Jin, Y., Xu, Z., Zhang, J. and Gao, J., 2024. A graph-powered large-scale fraud detection system. International Journal of Machine Learning and Cybernetics, 15(1), pp.115-128.
97. Gabryel, M., 2018. Data analysis algorithm for click fraud recognition. In Information and Software Technologies: 24th International Conference, ICIST 2018, Vilnius, Lithuania, October 4–6, 2018, Proceedings 24 (pp. 437-446). Springer International Publishing.
98. Kanei, F., Chiba, D., Hato, K., Yoshioka, K., Matsumoto, T. and Akiyama, M., 2020. Detecting and understanding online advertising fraud in the wild. IEICE TRANSACTIONS on Information and Systems, 103(7), pp.1512-1523.
99. Loyola-González, O., Monroy, R., Medina-Pérez, M.A., Cervantes, B. and Grimaldo-Tijerina, J.E., 2018. An approach based on contrast patterns for bot detection on Web log files. In Advances in Soft Computing: 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, October 22–27, 2018, Proceedings, Part I 17 (pp. 276-285). Springer International Publishing.
100. Kwon, J., Lee, J. and Lee, H., 2011, May. Hidden bot detection by tracing non-human generated traffic at the zombie host. In International Conference on Information Security Practice and Experience (pp. 343-361). Berlin, Heidelberg: Springer Berlin Heidelberg.
101. Zhu, X., Tao, H., Wu, Z., Cao, J., Kalish, K., Kayne, J., Zhu, X., Tao, H., Wu, Z., Cao, J. and Kalish, K., 2017. Ad fraud categorization and detection methods. Fraud

Prevention in Online Digital Advertising, pp.25-38.