International Journal of Scientific Research in Computer Science, Engineering and Information Technology

ISSN : 2456-3307 OPEN ACCESS

Available Online at : www.ijsrcseit.com

doi : https://doi.org/10.32628/CSEIT2410612403

Smart Guard: A Comprehensive Approach to Ad Click Fraud Detection

Akash Vir*, Dr. Shivam Upadhyay

*Department of Computer Science & Engineering, Krishna School of Technology, Drs. Kiran & Pallavi Patel Global University, Vadodara, Gujarat, India

ARTICLEINFO

Article History:

Accepted : 28 Nov 2024

Published: 22 Dec 2024

Publication Issue

Volume 10, Issue 6

November-December-2024

Page Number

2018-2024

ABSTRACT

Digital advertising is plagued by ad click fraud, which can result in large financial losses and skewed statistics. Current research on ad click fraud detection is summarized in this systematic review, which also assesses different methods and approaches. The review addresses difficulties, outlines the efficacy of various approaches, and makes recommendations for future lines of inquiry. Ad click fraud is the practice of creating phony clicks on internet ads, which can be done by malevolent humans, click farms, or automated bots. Click metrics are inflated by this fake activity, resulting in false performance statistics and squandered advertising budgets. Ad click fraud must be identified and stopped in order to preserve the efficacy and integrity of digital advertising campaigns. The objectives of this review are to list popular methods for detecting ad click fraud, assess how well they work in practical situations, talk about the drawbacks and restrictions of the approaches currently in use, and recommend future lines of inquiry to improve ad click fraud detection. The analysis concluded that when it comes to identifying ad click fraud, machine learning and artificial intelligence (AI) algorithms typically perform better than rule-based approaches. However, the caliber and variety of the training data determine how effective these methods are. The results emphasize how crucial it is to fight ad click fraud by utilizing sophisticated detection methods. To increase detection accuracy and lower financial losses, advertisers should spend money on AI and machine learning-based solutions. To keep up with changing fraud strategies, future research should concentrate on hybrid techniques, real-time detection, and cross-platform analysis.

Keywords: Click Fraud Detection, Machine Learning, Artificial Intelligence, Deep Learning

Introduction

The advent of the Internet in the 20th century, coupled with ongoing advancements in networking and communication technologies and the globalization of cell phones, brought about a significant shift in Internet-based industries, including the advertising sector, which saw previously unheard-of levels of rapid expansion in the form of the digital advertising sector. The internet's ability to provide the online advertising sector with precise information on users' interests and details, which may be tracked by site cookies or shared on social networks, is one of its most noteworthy advantages. This facilitates real-time user personalization and targeting through web ads [1].

The unlawful practice of clicking on pay-per-click ads in order to boost publishers' profits or drain advertisers' budgets is known as click fraud. Artificial intelligence methods have been used more and more to address complex problems in various fields of study, including cybersecurity, with surprising results [1]. The unlawful practice of clicking on pay-per-click ads in order to boost publishers' profits or drain advertisers' budgets is known as click fraud. Artificial intelligence methods have been used more and more to address complex problems in various fields of study, including cybersecurity, with surprising results [2]. Click fraud can take many different forms; however, there are mostly two kinds: deception The program contains malicious code for phony ad clicks. Bot-driven frauds, such as fraudulent apps, use bot programs to click on ads automatically. The possibility of fraudulent behavior never goes away, and click fraud is especially common in the online commercial advertising space. Advertising companies have switched from traditional newspapers and TV commercials to online and in-app marketing in an effort to attract new clients as a result of the growth of digital technologies and media. Online advertisements are the main source of income for internet giants like Google, Yahoo, and Facebook [2-3].

They serve as go-betweens for publishers and advertisers, settling on a price for each user interaction. Publishers of content are compensated by ad networks according to the quantity of visitors they bring to advertisers. Nevertheless, a security risk called Click Fraud is present in this compensation mechanism.The outcome is compromised click data, which results in financial resources being wasted. Additionally, ad channels might increase expenses by overflowing advertising with pointless clicks. Online advertising has emerged as a vital source of funding for supporting websites. Every time a click on an advertisement takes a user to the publisher's website, advertisers pay the publisher.Attackers who want to make money unlawfully employ "click fraud," which involves repeatedly clicking on a certain link [3]. The field of click fraud detection is fundamentally ambiguous. Broadly speaking, it entails determining the purpose of the clicks that are received based solely on technical information (like the IP address and other details supplied by HTTP requests) and contextual data (like prior accesses from the same IP, for instance). Malicious click detection thus entails comparing each access activity with what is typical of normal users; however, this is challenging to codify, as context-dependent behavior is both context-dependent and nondeterministic [4]. A whole industry has been established to deceive customers and web advertisers. Some are malicious, like hackers; others were made fraudulently for the benefit of another organization; still others were intentionally spiteful and intended to steal advertisements from specific networks. By default, click fraud costs "tens of thousands" of internet advertisers "hundreds of millions of dollars" annually instead of generating profits for advertising. Malicious apps and malware typically generate click fraud and are responsible for around "30% of click traffic in ad networks." With mobile malware, click frauds have become much more common. Clearly, scammers make genuine apps or purchase reputable men. [5].

Literature Survey

In addition to other noteworthy solutions already in use, prior attempts have particularly employed AI-based methods like ML and DL to detect click fraud. These are meant to shield advertisers from being billed for fictitious clicks, which may be very costly and degrade the effectiveness of their advertising efforts. AI algorithms are used on both the server and user sides to detect and prevent fraudulent clicks because prior research has demonstrated that applying them to distinguish between genuine and illegitimate ad clicks produces positive effects.

This section examines earlier research done to identify and stop click fraud. The following standards form the basis of our review:

- case studies that made use of ML and DL, two AI approaches;
- Research that looked into how to identify click fraud in the advertising sector (impressions, publishers, and click fraud are some of the forms of fraud that hurt the advertising sector). Our primary focus will be on research on click fraud.

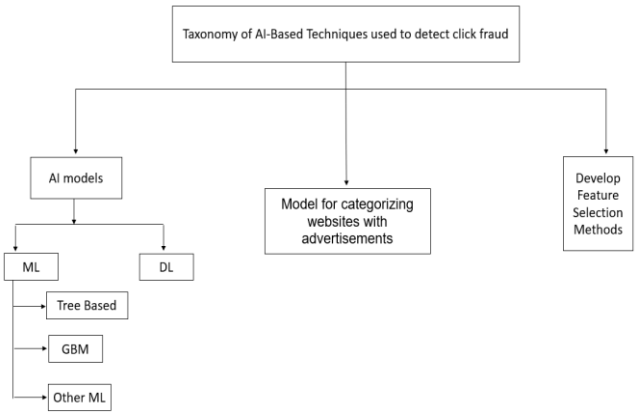

Figure: Taxonomy of AI-Based Techniques

Machine Learning-based: Comprising Random Forest, Neural Network (NN), Ripper, PART, Naïve Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and (RF). The justification for taking these algorithms into account is that they use a variety of methods to create intelligent models. Furthermore, these techniques are commonly employed to create supervised classification models in which the training dataset specifically contains the class value. Additionally, they are frequently employed for constructing categorization models across many fields [1]. A number of machine learning techniques were assessed, such as Support Vector Machine (SVM) classification, Random Forest classification, and K-nearest Neighbors (KNN) classification. With an 88% accuracy rate, random forest worked well. The K-nearest Neighbor (KNN) technique is a nonparametric supervised learning learner that uses geographic proximity to either predict or classify how data points should be grouped. Random Forest Classification (RVC): This technique averages the results of numerous models that have been trained on comparable data in order to get an appropriate prediction or classification. Support Vector Machine (SVC): This machine learning technique is used mostly for classification problems and necessitates supervision. Data points are projected into a multidimensional space using SVC, and the optimal hyperplane for grouping the points into two categories is then identified [3].

Gradient Tree Boosting: The goal of GTB is to identify an estimating function F(x) that, given an input variable x, can accurately predict the output variable y. The GBM approach iteratively adds weak learners hi(x) from a preset collection of weak learners H to improve the predictions of the previous iterations.

The first step is to assign the model's starting value to the constant function F0(x), which best fits the training set in terms of the loss function. In every iteration, the model adjusts the parameters by determining the gradient of the loss function that corresponds to the parameters and adjusting the parameters to reduce the loss function. The settings are adjusted by figuring out the optimal step size γ that minimizes the loss function.

The final estimation function F(x) is composed of the weights of the weak learners hi(x), which are determined by the optimal step sizes found in each iteration. It has been shown that this method

increases forecast accuracy while successfully reducing empirical risk [6].

Deep learning methods, which also greatly improved our accuracy. TensorFlow bots, a multi-layered neural network with an autoencoder linked and Keras backends. Due to an extremely uneven dataset, we employed a semi-supervised GAN to generate fake data in order to both serve as an attack and increase the accuracy of our neural network [7].

Multimodal Learning, Contrastive Learning: Click on Multimodal and Contrastive Learning Network (MCCF) fraud detection. Based on observations of the differences in demographic information, behavior patterns, and media relationships between fraudsters and authentic users on e-commerce platforms, MCCF specifically combines wide and deep features, behavior sequence, and heterogeneous networks to extract click representations. Furthermore, all three modules are integrated into contrastive learning, which combines them to produce the final predictions. Using a real-world dataset of 3.29 million clicks on the Alibaba platform, we investigate the effectiveness of MCCF. The experimental results show that the recommended approach can raise the F1 score by 15.6% and the AUC by 7.2% when compared to the state-of-the-art methods [8].

XGBoost: The XGBoost algorithm is used to categorize clicks that are fake. This approach is mostly used for multi-class classification, feature extraction, and selection. It is used in the fields of medicine, text classification, and business intelligence [9]. Boosting is a process in which models are added one after the other, improving the earlier models by fixing their mistakes. A method known as gradient boosting involves building models that forecast the residuals of earlier models before making a final prediction [10]. XGBoost, a machine learning technique that falls within the gradient boosting decision tree family, is frequently characterized as being extremely effective. It has a solid reputation for processing data in big databases quickly and reliably [11].

A common indicator of exploitative information is Benford's law. Benford's apportionment is not consistent; the smaller numbers are more certain than the larger ones. Using double-dealing Benford's law, you will examine specific numbers and focuses and identify those that show up both now and later when they are claimed to, and then they are the suspect shortly after [12].

The recent techniques that use machine learning. This advanced system can make predictions, evaluate data patterns, and adapt to new situations even without explicit programming. It becomes better with time thanks to experience, which makes it incredibly helpful for activities involving fraud detection. By using past data to find fraudulent patterns and anomalies, machine learning algorithms can increase their ability to recognize and halt fraudulent activity. Finding patterns or occurrences that substantially depart from the norm or expected behavior is accomplished using the approach of anomaly detection. It entails examining data to find anomalies or odd occurrences that point to fraud or questionable activity. Finding irregularities makes it easier to spot possible fraud and permits preventative actions to lessen risks. Several strategies are used in bot detection approaches to find and stop automated clickbots. The technique uses JavaScript challenges to evaluate user interactions, browser fingerprinting to examine distinct browser characteristics, and CAPTCHA challenges to confirm the user's humanity.These methods facilitate efficient bot detection and avoidance by differentiating between human users and bots. In order to find trends, collaborative filtering entails exchanging click fraud data with ad networks, fraud detection platforms, and industry partners. To improve click fraud detection, it makes use of collective intelligence. The accuracy and efficacy of click fraud prevention are increased by cooperation and knowledge exchange [13].

The algorithm known as LightGBM (Light Gradient Boosting Machine). The approach is a kind of gradient boosting decision tree, or GBDT, which is typically

used for regression, sorting, and classification. It also facilitates effective parallel training. At each stage, the algorithm approximates loss functions using the second-order Taylor approximation and piecewise constant trees. The second-order approximation is then minimized by training a decision tree, which is comparable to Newton's approach [14].

AdSherlock is a deployable and efficient method for client click detection. AdSherlock is a client-side method that is distinct from current server-side methods. The significant computation needed to identify click requests is divided between an offline and an online approach. AdSherlock's offline approach creates exact and probabilistic patterns for ad designs using URL tokenization. These patterns are used throughout the online process to detect click requests and click fraud patterns in conjunction with an ad request tree model for click fraud detection. [15].

Research Directions

As digital advertising continues to dominate the internet ecosystem, click fraud is one of the most common and financially damaging issues that publishers, ad networks, and advertisers deal with. The dynamic subject of machine learning (ML)-based ad click fraud detection and mitigation still has a lot of room for more research, despite the notable progress that has been done. With an emphasis on recent advancements and opportunities for innovation, this section explores the possible future uses of machine learning for web-based ad click fraud detection.

A. Real-Time Fraud Detection in Web-Based Environments

The requirement for real-time analysis is among the most urgent issues in web-based ad click fraud detection. Click farms, bot-driven traffic, and fraudulent user habits are examples of fraudulent operations that frequently happen quickly, necessitating the deployment of real-time systems that can handle and analyze enormous volumes of data. Optimizing algorithms to make conclusions in milliseconds is the key to the future of ML-based ad click fraud detection. This will allow networks and advertisers to act quickly to prevent or lessen fraudulent clicks. In order to accomplish this, methods like real-time reinforcement learning and streaming data analytics may be essential. These approaches enable models to continuously learn from fresh data and adjust instantly.

B. Integration of Behavioral and Interaction-Based Features

Traditional machine learning algorithms usually use device identifiers, IP addresses, and click patterns to identify fraud. However, scammers are getting better at hiding these glaring characteristics. A more detailed examination of user behavior and interaction patterns is probably what fraud detection will entail in the future. Features that can reveal whether a user is a person or a bot include click intervals, keyboard input speed, scrolling behavior, and mouse movement trajectories. By adding interaction-based features and behavioral biometrics to machine learning models, systems will be able to identify more subtle and difficult-to-copy patterns, improving detection accuracy.

C. Adversarial Machine Learning for Robustness

The methods used by scammers to evade detection are evolving along with fraud detection technologies. The goal of adversarial machine learning (AML) is to teach models to resist deceptive tactics. Adversarial approaches can be used to strengthen fraud detection systems' resilience against click fraudsters who purposefully provide inputs that trick conventional machine learning models in the context of web-based ad click fraud. Systems can become more resilient to changing fraudulent strategies by learning to identify and react to misleading inputs through the use of generative adversarial networks (GANs) and adversarial training.

D. Advanced Anomaly Detection Using Unsupervised Learning

Although supervised learning approaches are frequently employed in the detection of ad click fraud, they mostly depend on labeled datasets, which are either hard to come by or difficult to acquire. Clustering, autoencoders, and outlier identification is an example of an unsupervised learning technique that presents a possible avenue for fraud detection in the future without the need for labeled instances. Even in situations when there is no prior knowledge of fraud practices, these algorithms are able to detect unusual patterns in click activity. When it comes to identifying new or undiscovered types of fraud that might not be picked up by past data, unsupervised anomaly detection can be especially helpful. These unsupervised methods will probably be used more widely in web-based ad click fraud detection in the future to find novel fraud tactics that supervised models might overlook.

E. Multi-Modal Data Integration

Numerous types of heterogeneous data, such as clickstream data, device data, geographic data, and ad impressions, are produced by web-based advertising networks. Multi-modal machine learning models that integrate these various data sources will offer more thorough insights into user behavior and aid in the more precise identification of fraudulent conduct. Future machine learning models will be able to develop increasingly complex fraud detection systems that take into account a variety of behavioral indicators by merging data from web analytics, ad success measurements, user demographics, and even external sources like social media activity or network traffic. The development of more comprehensive and efficient fraud detection systems will require this all-encompassing approach to data.

Conclusion

Future developments in the field of ad click fraud detection are anticipated to be greatly aided by the application of cutting-edge machine learning techniques. As fraudsters continue to adapt their strategies, it will be essential to develop detection technologies that are more adaptable, transparent, and scalable. By embracing deep learning, explainable AI, federated learning, and collaborative efforts throughout the ad tech industry, machine learning-based fraud detection systems will become more dependable, efficient, and prepared to manage the increasingly complex world of ad click fraud. The next generation of fraud detection systems will focus on both identifying and preventing fraud in order to provide a more resilient and transparent digital advertising environment.

References

[1]. Malak Aljabri, Rami Mustafa A. Mohammad, 2023, “Click fraud detection for online advertising using machine learning,” Egyptian Informatics Journal, ISSN. 1110-8665, DOI. 10.1016/j.eij.2023.05.006.

[2]. Mahantesh Borgi, Viraj Malik, Breznew Colaco, Pratik Dessai, Harsha Chari, Shailendra Aswale, 2021, “Advertisement Click Fraud Detection System : A Survey,” International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181

[3]. Neeraja, Anupam, Sriram, Subhani Shaik, V. Kakulapati, 2023, “Fraud Detection of AD Clicks Using Machine Learning Techniques,” Journal of Scientific Research and Reports Volume, ISSN: 2320-0227, DOI:10.9734/JSRR/2023/v29i71762

[4]. Paulo S. Almeida, João J. C. Gondim, 2018 “Click Fraud Detection and Prevention System for Ad Networks,” ENIGMA - JOURNAL OF INFORMATION SECURITY AND CRYPTOGRAPHY

[5]. Anshuman Dash, Satyajit Pal, 2020, “Auto-Detection of Click-Frauds using Machine Learning,” IJESC - Vol. 10.

[6]. Lokesh Singh, Deepti Sisodia, N. L. Taranath, 2023, "Gradient Boosting-Based Predictive Click Fraud Detection Using Manifold Criterion Variable Elimination Gradient Boosting-Based Predictive Click Fraud Detection Using Manifold Criterion Variable," IFIP International Federation for Information Processing 2023 Published by Springer Nature, ISBN: 9783031382963, DOI: 10.1007/978-3-031-38296-3

[7]. Thejas G S, Kianoosh Boroojeni, Isha Bhatia, Sunitha N R, 2019, "Deep Learning-based Model to Fight Against Ad Click Fraud Deep Learning-based Model to Fight Against Ad Click Fraud", ACM Southeast Conference—ACMSE 2019—Session 2: Short Papers—ISBN: 978-1-4503-6251-1, DOI: 10.1145/3299815.3314453

[8]. Weibin Li, Qiwei Zhong, 2021 "Multimodal and Contrastive Learning for Click Fraud Detection" License Attribution 4.0 International.

[9]. B. Viruthika, Suman Sangeeta Das, E.Manish Kumar, D Prabhu, 2020 "Detection of Advertisement Click Fraud Using Machine Learning" International Journal of Advanced Science and Technology Vol. 29, No. 5, (2020), pp. 3238 - 3245.

[10]. Thejas G.S., Surya Dheeshjith, S.S. Iyengar, N.R. Sunitha, Prajwal Badrinath, 2021 "A hybrid and effective learning approach for Click Fraud detection" Machine Learning with Applications 3 (2021) 100016.

[11]. Tanvir Rahman Akash, Md Shakil Islam, and Md Sultanul Arefin Sourav, 2024 "Enhancing business security through fraud detection in financial transactions" Global Journal of Engineering and Technology Advances, 2024, 21(02), 079–087.

[12]. Ananya Smirti, 2021, "Detecting of Fraud Click on Advertisement" INTERNATIONAL JOURNAL OF RESEARCH IN SCIENCE & TECHNOLOGY e-ISSN:2249-0604; p-ISSN: 2454-180X.

[13]. Ranjeet Vishwakarma, Rajesh Dhakad, 2024 "Online Advertising and Fraud Click in Online Advertisement: A Survey" International Journal of Computer Applications (0975 - 8887).

[14]. Elena-Adriana MINASTIREANU and Gabriela MESNITA, 2019 "Light GBM Machine Learning Algorithm to Online Click Fraud Detection" Journal of Information Assurance & Cybersecurity DOI: 10.5171/2019.263928.

[15]. Mahesh Bathula, Rama Chaithanya Tanguturi, Srinivasa Rao Madala, 2021 "Click Fraud Detection Approaches to Analyze the Ad Clicks Performed by Malicious Code" IOP Publishing doi:10.1088/1742-6596/2089/1/012077