Journal of Scientific Research and Reports

Volume 29, Issue 7, Page 84-89, 2023; Article no.JSRR.101581
ISSN: 2320-0227

Fraud Detection of AD Clicks Using Machine Learning Techniques

Neeraja [a], Anupam [a], Sriram [a], Subhani Shaik [a] and V. Kakulapati [a*]

[a] Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana-501301, India.

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JSRR/2023/v29i71762

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://www.sdiarticle5.com/review-history/101581

Original Research Article

Received: 12/04/2023
Accepted: 16/06/2023
Published: 20/06/2023

ABSTRACT

Although all businesses face the possibility of fraud, those that rely on internet advertising face an especially high risk of click fraud, which may lead to inaccurate click statistics and unnecessary expenditures. The cost per click for advertising channels might skyrocket if enough people click on the ads. Internet advertising is becoming a significant revenue source for many websites. Under this model, advertisers pay the publisher a flat rate for each click-through from the ad to the advertiser's site. Since spending much on Internet advertising requires significant resources, the term "click fraud" refers to an attack tactic in which the perpetrator repeatedly clicks on a single link for the sole purpose of generating illicit revenue. By clicking on a pay-per-click (PPC) ad many times using a script, fraudsters may trick online advertisers into paying for clicks that never happened. We may use a variety of methods to identify click fraud anytime a human or computer program clicks on a particular link, and then use the click-through rate to ascertain whether the clicker is legitimate. This work provides a machine-learning strategy for predicting user click fraud, which will enable us to distinguish between fraudulent and legitimate clicks and, therefore, identify fraudulent users from legitimate ones. We have used KNN, SVC, and Random Forest models for this purpose.

**Corresponding author: E-mail: vldms@yahoo.com;*

Keywords: Advertise; fraud click; machine learning; prediction; user; legitimate; link.

1. INTRODUCTION

Fraudulent activity is a constant threat, with click fraud being particularly rampant in the realm of online business advertising. With the rise of web technologies and media, advertising firms have pivoted from traditional newspapers and TV commercials to online and in-app promotions to draw in fresh customers. Internet behemoths like Google, Yahoo, and Facebook generate most of their revenue from online ads [1]. They act as intermediaries between advertisers and publishers, agreeing on a fee for every user action. Ad networks pay content publishers based on the number of visitors they drive to advertisers. However, this compensation model harbors a security threat known as Click Fraud. The result is compromised click data that leads to wasted financial resources. Ad channels can also drive-up costs by flooding ads with meaningless clicks. Supporting websites, online advertising has become a crucial mode of financing. Advertisers pay publishers each time their ad is clicked, leading to their website. Attackers who seek to illegally profit use "click fraud," repeatedly clicking on a specific link [2].

The high-stakes financial nature of Internet marketing is to blame. Click fraud is the most typical kind of ad fraud in performance marketing. Ad-click fraud happens when bad actors click on ads they shouldn't, resulting in an inflated bill for the advertiser from the ad network. In performance advertising, click fraud is by far the most common kind of fraud.

In this kind of trickery, bad actors try to trick the ad network into redirecting their ads by clicking on them when they aren't relevant. Criminals are also eager to reap the benefits of technological progress. possibilities to commit fraud against such people in the hopes of obtaining monetary reparation. They are capable of adopting a fictitious persona [3]. Click fraud occurs when one party (people or bots) generates clicks with the intent of earning money or using up a competitor's resources.

This is the term for fraudulent clicking. In the end, this threatens the foundations of Internet commerce and marketing, as sponsors cannot take part in auctions and poor-quality content can't be produced in any case. The longer advertising keeps the site running, the better off it will be. Combating click fraud is a time-consuming and difficult task. When click fraud bots are used regularly, they evolve and adapt in response to investments in resources and massive detection systems [4].

To combat click fraud, many prediction methods have been created. A statistical model that may identify the specific IP addresses used in these fraudulent actions is one such method [5]. Most existing approaches for identifying click fraud in advertisements are conducted offline.

The objective of this work is strategies for vetting potential clients, weeding out non-paying consumers, and identifying those most likely to repay their loans. Lenders and borrowers alike would benefit from the increased transparency that would result from the use of such prediction algorithms. Whether or not a client will miss their next payment is a yes/no decision in this binary classification issue. Since the dataset is skewed, prioritized precision and recall are above accuracy. In contrast to the precision-recall curve, the False Negative value of confusion metrics favors logistic regression.

Often, high-dimensional feature spaces and complex machine-learning algorithms are used together to identify fraud. Also, the data sets are often massive. For the largest advertising firms, daily ad clicks might reach hundreds of millions. As a result, the cost of both training and using such models for lookups might be prohibitive. Thus, most fraudulent clicks on advertisements are not committed online. Ad networks and marketers can only respond so much to this. If fraud could be detected in real-time, ad networks and advertisers would have a higher chance of investigating and taking action against it.

2. RELATIVE WORK

Brand advertising and performance advertising are the two biggest types of digital ads. Getting their name out there to as many people as possible is the primary objective of brand advertising. Before digital marketing came along, brand advertising was the most common type of marketing [6]. At that time, TV, print publications, billboards, and hoardings were the most common ways to advertise outdoors. Advertisers in this format care about how many people their ads reach relative to the money they have spent on promotion. Typically, pay-per-view pricing models are used in brand advertising, where advertisers get paid for each view of their web ad. Facebook advertising is a great example of a network that is made to promote a certain business.

Since advertising income is going down, content creators need to find other ways to make money. Ads on the site bring in the cash. By displaying relevant advertisements to their visitors and encouraging them to click on such ads, content creators may effectively monetize their traffic while simultaneously delivering useful services to their consumers. Advertisers, ad networks, and website publishers often make financial transactions based on page views, form submissions, and clicks. Paid placements and cost-per-click models are not exclusive to Google; Yahoo and Bing provide similar options. Some search engines pay for each click on an advertisement by taking on the duty of displaying ads that are relevant to the user's query [7,8].

Google isn't the only search engine that offers paid placements and pay-per-click models. Yahoo and Bing also have similar options. It is the job of certain search engines to display ads that are relevant to a user's query; in exchange for a fee, they collect data on how often users click on such ads. Ad fraud may affect both forms of digital marketing [9,10].

Ad fraud, especially impression fraud, is rampant in the industry. Advertising platforms like Google Display and Facebook's Ad Network (FAN) are particularly susceptible to this issue. Commercial placement networks are works with independent publishers. In addition, shady publishers look for imprints to use in fake news releases [11,12]. In advertising interpretation, the most common kind of fraud is announcement fraud. In this scam, dishonest participants fabricate clicks on advertisements in the hopes of tricking the ad network into overcharging them. To this end, fraudsters look for fresh opportunities to steer fraud against these parties, with the hope of rerouting some commercially illicit earnings into their own hands [13]. Click fraud occurs when someone, whether it be a person or a piece of software, generates many identically invalid clicks to earn a profit or drain an opponent's account. In the long run, this poses a danger to the ABC economics of online advertising, as it will likely result in the departure of advertisers from agreements and the inability of high-quality content to be funded by ads [14].

3. METHODOLOGY

Several ML methods, including K-nearest Neighbours (KNN) classification, Random Forest Classification, and Support Vector Machine (SVM) classification, were evaluated. With an accuracy of 88%, random forest performed effectively.

KNN: (K-nearest Neighbor): This technique is a nonparametric supervised learning learner that makes utilizes spatial closeness to either label or predict the grouping of data points.

Random Forest Classification (RVC): To get an accurate forecast or categorization, this method uses many models learned on similar data and then averages their findings.

Support Vector Machine (SVC): It is a method for machine learning that is utilized primarily for categorization issues and requires supervision. With SVC, data points are projected into a multifaceted space, and then the best hyperplane for categorizing the points into two groups is determined.

Forecasting has enhanced the rate of clicks, but more importantly, it is dynamically learned from data provided without any input from humans' knowledge of the field.

4. RESULTS AND DISCUSSION

4.1 Data Collection

The Kaggle dataset of 6582 rows × 7 columns Fig. 1 has features like Ip, app, device, OS, channel, and click _scaler and is attributed, however, we are primarily interested in the Ip address and fraud detection.

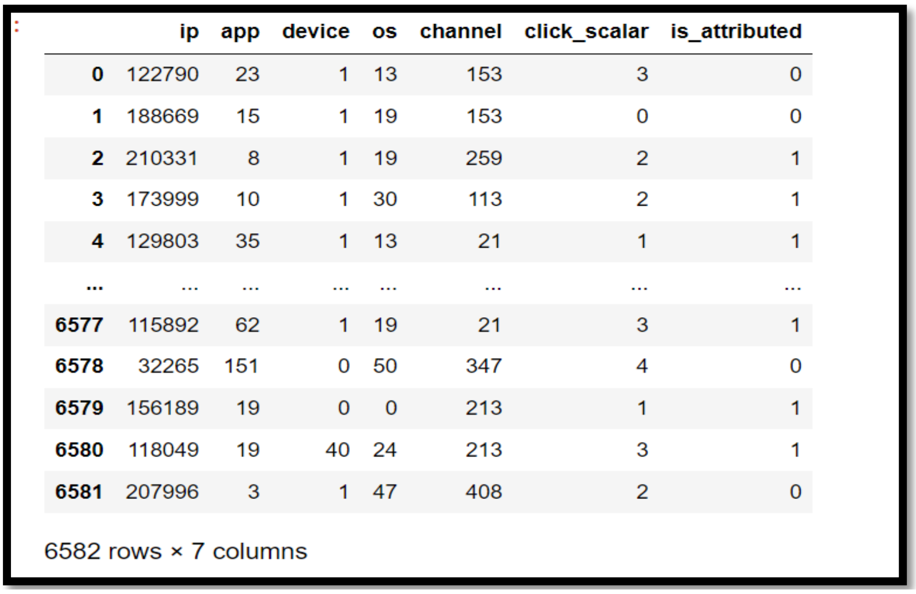
4.2 Data Preprocessing

The obtained Kaggle dataset has to be preprocessed and then separated the data into train and test data, with 80% of each being used for training and 20% for testing, then verify the association among the seven variables.

We converted the Ip addresses to machine-understandable code and then trained and tested data, and applied KNN, SVC, and Random Forest classifier models shown in Fig. 2 we related the frauds through is attributed likewise if it's 0 it’s a fraud and if it's 1 it's not a fraud.

Table 1. The comparative analysis of different algorithms

S. No.	Methodology and tools	Dataset	Accuracy	Results
1	KNN Classifier	Kaggle	Approximately 87%	It is good when clustering
2	Random Forest Classifier	Kaggle	88%	It performs well
3	SVC	Kaggle	87%	It also functions well to detect fraud



	ip	app	device	os	channel	click_scalar	is_attributed
0	122790	23	1	13	153	3	0
1	188669	15	1	19	153	0	0
2	210331	8	1	19	259	2	1
3	173999	10	1	30	113	2	1
4	129803	35	1	13	21	1	1
...	...	...	...	...	...	...	...
6577	115892	62	1	19	21	3	1
6578	32265	151	0	50	347	4	0
6579	156189	19	0	0	213	1	1
6580	118049	19	40	24	213	3	1
6581	207996	3	1	47	408	2	0

6582 rows × 7 columns

Fig. 1. Sample data in collected data

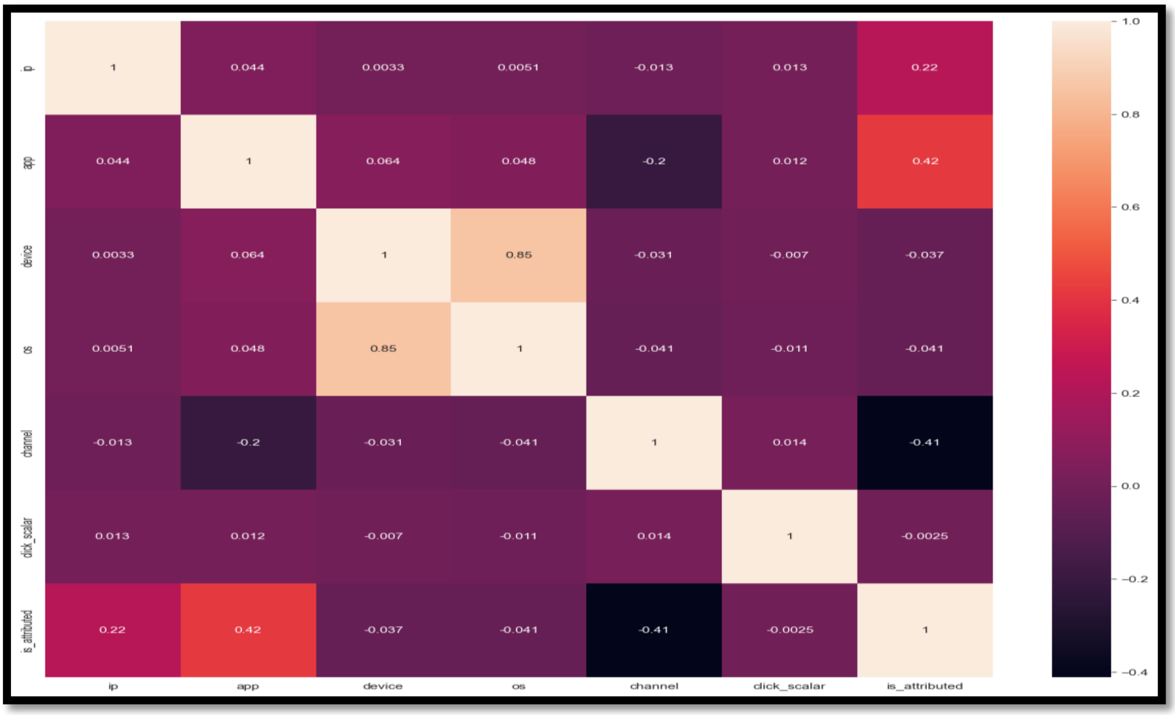
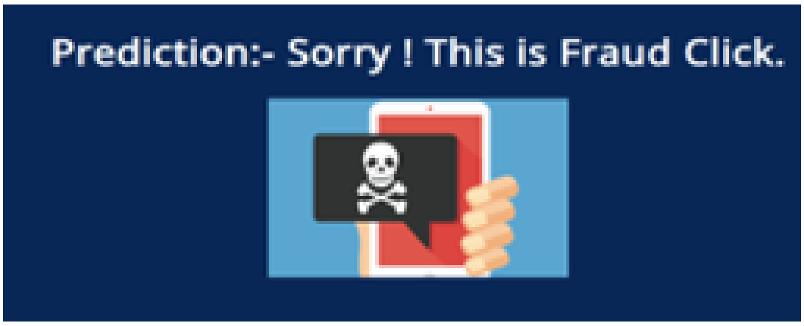


Fig. 2. Heat map of ad click data



Prediction:- Sorry ! This is Fraud Click.

Fig. 3. Predicting the result of fraud ad click

5. CONCLUSION

The above suggests that it is feasible to identify ad-click fraud in real-time using elementary classifiers such as this. Even with a skewed sample size. Classifiers based on machine learning, such as K-Nearest Neighbours (KNN), Random Forest, and Support Vector Machine, were put to the test. Our tests showed that Random Forest was the most effective method, with an accuracy rate of 88%. The adoption of prediction has enhanced the click-through rate, but more importantly, it automatically learns from the data without any human domain expertise.

Ad clicks fraud analysis suffers from a lack of publicly accessible and properly labeled datasets. As a result, there is a limit to how much can be tested and investigated. Ad-click fraud may be manufactured and interspersed with real visitors to a website. In the context of certain ad click fraud attacks, such as those utilizing botnets or human clickers, this might be used to evaluate the efficacy of ad click fraud detection methods. This may also lead to the accessibility of a wealth of other personally identifiable information, such as HTTP headers, referrer URLs and cookies.

6. FUTURE ENHANCEMENT

In the future, keeping up with click fraud is a never-ending game of cat and mouse because fraudsters who utilize bots are always innovating new techniques to evade detection. Due to the ever-evolving nature of bots, the process of developing tools to detect and prevent click fraud should be maintained. We need cutting-edge interventions to both prevent and deal with these issues. However, since it has only been tested in a small number of studies, care should be used when using feature integration as a combination of two or three characteristics to improve click fraud detection.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Available:https://www.kaggle.com/competitions/talkingdata-adtracking-fraud-detection/overview/evaluation
2. Available:https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=1916&context=etd_projects
3. Available:https://github.com/cliang1453/adfraud_detection_visualization
4. Available:https://www.googleadservices.com/pagead/aclk?sa=L&ai=DChcSEwiOyoTmj4b-AhUgk2YCHYuuAfsYABACGgJzbQ&ohost=www.google.com&cid=CAASJuRowGFx1zTOGwzKOe0UeS4rojn3HqeisS6Mi7qA3Uwm-Ml4Neug&sig=AOD64_39xolm4s4UGkZhdST8S6oVHLMi0Q&q&adurl&ved=2ahUKEwjZ__7lj4b-AhXhTmwGHZS0BVQQ0Qx6BAgLEAE
5. Available:https://scholar.google.co.in/scholar?q=ad+click+fraud+detection&hl=en&as_sdt=0&as_vis=1&oi=scholart
6. Available:https://support.google.com/google-ads/answer/42995?hl=en#:~:text=Clicks%20on%20ads%20that%20Google,them%20from%20your%20account%20data.
7. Available:https://www.researchgate.net/publication/336666758_Machine_Learning_Based_Ad-click_prediction_system
8. Available:https://www.ijeat.org/wp-content/uploads/papers/v9i3/C5518029320.pdf

9. Available:https://www.ijert.org/advertisement-click-fraud-detection-system-a-survey
10. Available:https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/
11. Available:http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html
12. Available:https://www.geeksforgeeks.org/support-vector-machine-algorithm/
13. Available:https://www.techtarget.com/searchsecurity/definition/click-fraud-pay-per-click-fraud
14. Available:https://towardsdatascience.com/everything-about-svm-classification-above-and-beyond-cc665bfd993e

Peer-review history:
The peer review history for this paper can be accessed here:
https://www.sdiarticle5.com/review-history/101581