International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

Advertisement Click Fraud Detection System: A Survey

Mahantesh Borgi
Dept. Of Computer Engineering
Shree Rayeshwar Institute Of Engineering & Information Technology Shiroda-India

Viraj Malik
Dept. Of Computer Engineering
Shree Rayeshwar Institute Of Engineering & Information Technology Shiroda-India

Breznew Colaco
Dept. Of Computer Engineering
Shree Rayeshwar Institute Of Engineering & Information Technology Shiroda-India

Pratik Dessai
Dept. Of Computer Engineering
Shree Rayeshwar Institute Of Engineering & Information Technology Shiroda-India

Harsha Chari
Dept. Of Computer Engineering
Shree Rayeshwar Institute Of Engineering & Information Technology Shiroda-India

Shailendra Aswale
Dept. Of Computer Engineering
Shree Rayeshwar Institute Of Engineering & Information Technology Shiroda-India

Abstract:- **Web services have become an integral part of our day-to-day life including the various advertisements viewed on websites. Revenue is generated by companies through advertisement by selling clicks (known as Pay-Per-Click model). The company is paid for each click performed by the user on link published on the webpage. The clicked link redirects the user to sponsoring company's content. Invalid clicks that are generated either by humans or through a software as a malpractice for earning money is known as click-fraud. Several different time features are combined into a time print. Machine learning is performed for understanding up to what extent the unusual time prints occur in a data for distinguishing invalid clicks and identifying click fraud. The results generated shows that time prints are useful tool for the improvement of the quality of click fraud analysis and increases the overall accuracy of the observations.**

Keywords –Fraud detection, Logistic Regression, Support Vector Machine, Traffic Analysis, JavaScript

I. INTRODUCTION

According to [1], in the online advertising market, the advertiser takes decision about the investments he would make for the advertising, the advertiser then chooses a rate-plan created by the publisher which is suitable for his fixed budget and submits the advertisement for publishing. The ad network is responsible for contracts and legal matters with a publisher that displays the advertisement and receive a part of payment as a commission depending on the impressions made on the advertisement displayed.

A variety of revenue models are used by the users and the publishers which can attract more customers to them and can help them earn profit as well. One such revenue model which is most commonly seen in the industry is the Pay-Per-Click(PPC) model. According to [2], a fixed amount is paid by advertiser to the publisher for each click made by the user. Click-fraud indulges in malpractices for deceiving the revenue model such as PPC by creating records for false clicking activities on online advertisement. The deceiver may have intentions of creating losses to the advertiser by increasing the false impression so that the advertiser pays more or he may be the agent of the publisher to increase the profits.

Pay Per Click (PPC) model has been implemented in [17] which shows that auctions and bidding play a vital role in the advertising economy. This has created opportunities for various IT companies and have boosted their wealth at a very high rate. The model exhibits properties such as a greater relevance and effective conversion rates where keywords are entered by a user for searching a specific advertised product.

According to [8] the Advertiser is the one who aims to make a business with a profit through product selling by means of online advertising. Promotions can be shown on webpage directly as a part of website or can be added as a pop-up related to an action of a user. The ad-network generates economy by taking payments from the advertiser for the promotion on a website or an application. The creation, development, maintenance and the deployment of the webpage or the application is the responsibility of the publisher. Ad Network behaves like a middle-man between the advertiser and publisher and streamlines the process.

Click fraud should not only be controlled from the technical point of view, but also involve the issue of integrity. If you want to cheat, no technology can be completely overcome and directly impact its market future. To solve click fraud well, or develop an effective strategy to prevent click fraud, or replace the current bidding ranking with another new business model, the current legislation is not perfect, consideration to make use of laws to strengthen the constraint on click fraud is required. The purpose of the study in [21] is to prevent the click fraud, reduce the harm of click fraud to the Internet, and make efforts to build a healthy and harmonious network.

This paper has been categorized as follows: Section II gives the overview about the different existing research works and also explains the methodologies used previously. Section III provides an overview to the ideas and concepts

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

in the proposed model. Section IV concludes the paper with the observations and scope for future improvements.

II. LITERATURE REVIEW

The advertiser may encounter the problem of Online Click Fraud in various ways. With the high increase in the internet usage the rate of click frauds has seen a steep climb over past few years. Multiple approaches for tackling the issues have been put forth by various researchers. The following literature gives a basic descriptive idea about various existing computational methodologies available.

A. REGRESSION

Machine Learning methods such as time-print and traditional classification model are used in [3]. Click data associated with time to check the behavior of users at particular time is analyzed. In [7] Traffic module, Analysis module, and Calculation module are used to monitor the behavior of user for click fraud detection.

Pixel cluster used in [8] for transit monitoring cursor movements uses applet or java script to keep track of transit of the cursor, it analysis the cursor movements and determines the confidence level to check whether cursor movements are indicative of fraud. Client-side scripting and server-side scripting is performed in [10]. It has click verification website to identify valid and invalid clicks and prevents advertising from getting billed for fraudulent activity.

Support vector machine (SVM) model are implanted in [17] where AUC (Area Under Curve) value of proposed algorithm is significantly improved in comparison with ELM and the weighted ELM based method. The investigation on Bot signature, Distribution tests, Scale families and reference curves were used to design Microsoft click filtration system. The design in [21] includes real-time and near real-time components and it also has consistence rapid update capability in fraud detection.

In [22] JavaScript support and mouse event test were carried out along with browser functionality test and browser behavior examination that collected information about real world click traffic. To perform effective identification of human and bot clicks a vital element is the testing throughout the client's functionality at the advertiser's side, which is implemented at both the server as well as client sides incurring negligible overhead. Filtering on Rule based, Classification based and Clustering based are carried over data in [30].

TABLE I: REGRESSION

Sr No.	Title	Methodology And Tools	Techniques	Dataset	Merits	Demerits	Results	Accuracy
[3]	Identification of Click Fraud and Review of Existing Detection Algorithm. (2019)	1.Machine learning 2.Naïve-Bayes	Mix Adjustment Algorithm	NA	A loop is created to monitor and optimize upon testing on many systems which are dynamic in nature.	The modification of the algorithm as well as its extension still needs to be accurately worked upon.	Testing can be done using the algorithm as well as comparison can be done to enhance its performance further	99.6%
[7]	Real-Time Ad Click Fraud Detection (2020)	1.Machine Learning, 2.Naïve-Bayes 3.Heuristics	1.Logged Event based ,Client Based 2. Sessionization, Heuristics 3.Explored, Conversion Sparsity	Kaggle Records: 184,903,890	1.Achieves high precision on unbalance dataset. 2. Compute costs below 1μs per click.	1.There is a very limited cost of utility for classification in a real time scenario 2.Datasets required for detection of click fraud is sparse in availability	On the basis of precision the methods of machine learning performed using neural network have best precision	MLP = 95.43%, Naïve-bayes =89.97% Gradient Boosting =71%
[8]	Deep Learning-based Model to Fight Against Ad Click Fraud (2020)	1.Machine Learning. 2.Multinomial Naïve bayes 3.Logistic regression, 4. Random Forest	1.Auto Encoder 2. Semi Supervised GAN	Kaggle Records: 184,903,890	1. High accuracy is achieved even on small dataset. 2.Probability of fake and real clicks is computed.	1.Require prior classification to anomalies. 2. Loss error will be high for bots.	Supervised GAN gives good accuracy.	89.7%
[10]	Click Fraud Detection and Prevention System for Ad Network. (2018)	1.Machine Learning 2.Naïve-Bayes 3.Python Language	Rues: 1.Online Methods 2.Offline methods	NA	Good performance against different types of attack	1.Low frequency attacks are undetected. 2.Lack of public data sets about click related subjects,	Click Fraud Detection using Online and Offline rules gives good result.	NA

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

[17]	Click Fraud Detection: Adversarial Pattern Recognition over 5 Years at Microsoft (2017)	1.Machine learning. 2.Naïve-Bayes 3.Data mining	1.Bot Signatures 2.Machine induced decision trees	Kaggle	1.The design containing components based on real-time as well as near-real-time are consistent 2. The rapid update capabilities are very consistent in detection of frauds.	Clickbots can be configured to blend in to avoid detection.	Adversarial Pattern Recognition has been steadily improved, enhanced and is very consistent in detection of frauds.	NA
[21]	The study on preventing click fraud in internet advertising. (2020)	1.Machine Learning 2.Naïve-Bayes	Random forest classification algorithm	Records: 5772649	1.High accuracy is obtained in this method. 2.It can efficiently unbalance data.	1.Encountered problems in data analysis. 2.Real dataset and deep neural network for analysis is difficult to find.	Third Party Detection technologies to prevent click fraud based on technical measures is efficient than singular algorithms.	93 %
[22]	Machine Learning based Ad-click detection system. (2019)	1.Machine Leaning 2.Naïve Bayes. 3.Logistic regression. 3.Naïve Bayes. 4.Decision Tree.	1.Time stamp 2.Scikit Learning Technique	NA	1.Uses machine learning techniques to give best precision. 2.As the data keeps on growing, the accuracy increases.	Dataset required for training analysis is difficult to get.	Logistic regression effectively predicts weather the person clicked on ads or not	96%
[30]	Malicious Click Detection in Web Advertising Data. (2018)	1.Machine Learning 2.Naïve Bayes 3.Decision table,	1.Random Forest 2.Classifiers 3.Filtering: →Rule-based →Classification-based →Clustering-based	Records: 35 million	1.Highly efficient 2. Result Verification achieves high prediction using Filtering.	1.Coverage of a blacklist is limited. 2.For the limitation of supervised method, false positive is inevitable.	Malicious clicks are detected by performing an in-depth analysis using framework	97.60%

B. SUPPORT VECTOR MACHINE (SVM)

The analysis in [9] concludes that performance of classification schemes is improved when data is pre-processed by performing feature selection and resampling. Resampling is done using SMOTE and then performing classification using Adaboost with random forest as base classifier which in turn produced the most improved results. Weights, ranks and compared rank are assigned in [13] along with the threshold. The algorithm can be implemented in a system that involves the processing of multiple entities which are required to test and compare the results in multiple environments.

Python Programming language has been used in [16] where online rules and offline rules are implemented to check the IP address, the number of clicks on the IP address and movement of cursor before clicking on the ad. This data is used to determine the normal person behavior and fraudulent behavior. Probability-based model approach is implemented in [25] which is better than learning based probabilistic estimator model. It has distorted towards particular time indexed label that hampers the accuracy, it is possible to even predict a behavior on the seconds scale of time or even a smaller timescale which gives better results.

TABLE II: SUPPORT VENTOR MACHINE(SVM)

Sr No.	Title	Methodology And Tools	Techniques	Dataset	Merits	Demerits	Results	Accuracy
[9]	A Click Fraud Detection Scheme based on Cost Sensitive BPNN and ABC in Mobile Advertising. (2018)	1.Support Vector Machine(SVM) 2.Neural Network 3. Artificial Bee Colony	1.CS-BPNN Classification 2. Synthetic Minority Over-Sampling Technique (SMOTE)	Buzz City Training: 5,862,839 Testing: 2,598,815	1.Gives best result for almost all cost ratio. 2.Optimizes the feature selection and weights simultaneously.	For Cost ratio 7 it does not give best results.	Click Fraud detection using BPNN and ABC gives good result as per this approach	NA
[13]	User click fraud detection method based on Top rank k frequent pattern mining (2019)	1.Support Vector Machine(SVM) 2.Top Rank k algorithm 3.Linear Regression	1.Click frequency 2.Time complexity (time spent on ad, arrival time)	NA	1. It is highly accurate. 2. It is better than traditional frequent pattern mining.	Dynamic clicks cannot be dealt well when the algorithm of frequent pattern mining is used without the optimization of Top-Rank-k.	The outcome proves that the method used is quite efficient in its correctness and checks patterns using graphs.	98%
[16]	Advertisement Click-Through Rate Prediction Based on the Weighted-ELM and	1.Support Vector Machine(SVM) 2.WELM-Adaboost	1.Data preprocessing 2.Area Under Curve	NA	The value of AUC (Area Under Curve) of the algorithm	The accuracy in prediction is quite low because of the imbalance in the	WELM-Adaboost algorithm is significantly	NA

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

	Adaboost Algorithm (2017)	algorithm 3.Logistic regression 4.Prediction Model based on Deep Neural Network			proposed has a significant improvement when it is compared to the methods based on ELM and the Weighted -ELM.	advertising data being distributed	better than ELM and Weighted ELM method	
[25]	Click Stream Data Analysis (2016)	1.Support Vector Machine(SVM) 2.MARS 3. Decision tree 4.Naive Bayes	1.WEKA: →Clustering →SVC classifier	Train: 3,173,834 Test: 2,598,815	Decision tree is consistent in fraud detection.	1.Access to user click information is limited. 2.Lower accuracy using more attributes.	Can be used for the pre-processing when online processing is required.	59.38%

C. TRAFFIC ANALYSIS

The algorithm in [5] helps to determine the frequently occurring patterns without using top-k optimization with real time clicks. The method extracts information about click process events, it also describes the frequency of click of the sample events. Using the weightage of click stream it is able to calculate evaluation score, the data is used to generate click stream density function. However, the selection of k value in this method is very important and is optimized in subsequent research. [11] refers to the Internet advertising service platform using defense system that have been established using the third party to detect the click fraud which uses random forest classification algorithm producing highly accurate results of positive class 93% and negative class 91%.

The proposed fraud detection in [14] uses CS-BPNN classification and SMOTE. BPNN is implemented in the variable click-fraud detection algorithm environment which avoids the local optimization of neural network and feature redundancy using Artificial Bee Colony. Data mining, honeypot, traffic analysis is used in [23] to verify whether the advertisement is normal, so that if any ad is being detected then honeypot sticks to it and exposes the bot leading to blocking of its IP. In [29] Content match advertising, social engineering attacks, Phishing and Blacklist evasion bidding style in which Bing search process is used which consists of policies. If any of the policy have not been satisfied by the ad then that ad will be detected as fraudulent.

TABLE III: TRAFFIC ANALYSIS

Sr No.	Title	Methodology And Tools	Techniques	Dataset	Merits	Demerits	Results	Accuracy
[5]	Click Fraud Detection using Traffic Analysis (2019)	Traffic Analysis	1.Traffic matrix construction 2.Traffic partitioning 3. Pooling	Records:217, 334,190	It helps to separate out clicks made by malware that are present in legitimate click streams	Can be defeated by reducing the network loads.	Organic click fraud attacks are detected by efficiently searching for patterns which are repetitive in nature coming from clickstreams of an ad network.	NA
[11]	Behavioral Verification: Preventing Report Fraud in Decentralized Advert Distribution System. (2017)	1. Traffic Analysis 2.Ad-Report method. 3.Python Language	1.Cost-Per-Click 2.Cost-Per-Impression 3.Cost-Per-Action.	NA	1.Multiple reports generated in short time. 2. It gets better result because it checks target users.	1.User privacy decreases 2. Digital signature were not completely detected.	Filtering Ad-Reports on basis of honest and dishonest users.	NA
[14]	Click fraud monitoring based on advertising traffic. (2017)	Traffic Analysis	1.Traffic Module 2.Analysis module 3.Fraud module 4.Calculation Module	NA	It Includes techniques for analyzing multiple aspects of advertising traffic.	Client identification may be inefficient due to single aspect of advertisement traffic	The technology is directed towards analyzing aspects of advertising traffic and monitoring click fraud.	NA
[23]	A survey on Online Advertising and Click Fraud detection. (2020)	1.Traffic analysis 2.k-Nearest Neighbours.	1.Data mining 2.Machine learning 3.Honeypot	Records: 200 million	1.High performance 2. Reliability 3.Availability over traditional Models.	The bluff ads may not be clicked if they are irrelevant.	Machine learning and deep learning provide accurate solution to the problem of click fraud.	98%

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

[29]	Search Advertiser Fraud (2017)	1.Traffic Analysis 2.Bing's search engine platform.	Anomaly detection strategies	Real Time Data (Input Data for Search Engine)	1. Accepts manual reporting. 2.Payment fraud detection is high.	Detection of the Web crawler can be avoided by attackers using 'Cloaking' on advertised web page.	Bing's policies have successfully contained fraud that may be costly.	NA

D. JAVASCRIPT

Light-GBM (Light Gradient Boosting Machine) algorithm is implemented in [2]. When there is a data which contains multiple attributes,feature parallelism can be implemented concurrently. When a large amount of data needs to be processed concurrently the data parallelism is brought into use. Voting parallelism focuses on data in which attribute has multiple features and votes for decision making. The prime advantage is the high performance along with the reliability and availability over traditional logistic regression models. Feature engineering of the data along with appropriate selection allows to improve the detection performance. In [4] the methodologies used are Logistic Regression, Support Vector, Machine Random Forest and Multinomial Naïve bayes, where 100000 random inputs were given and the discriminator classified 63000 as valid clicks and remaining as invalid clicks and it also provided good accuracy in determining the valid and invalid click.

SDK library (real ad network) click generation attacks are detected in [15]. Potential risk of automated occurrence of an online fraud attacks in mobile advertising is evaluated and it also showed that 75% of the network was highly vulnerable to click fraud attacks. Pixel Clustering and methods like cost-per-click, cost-per-impression and cost-per-action are implemented in [15] where ads were filtered on the basis of user clicks and fraudulent users (whether Botnets or click frame) for great amount of data in minimal amount of time.

TABLE IV: JAVASCRIPT

Sr No.	Title	Methodology And Tools	Techniques	Dataset	Merits	Demerits	Results	Accuracy
[2]	Fighting Click-Fraud from the User Side. (2016)	1.Fight Click-Fraud (FCFraud) 2.JavaScript	1. HTTP Requests(GET Function) 2. JavaScript support and mouse event test. 3.Blacklist.	Records:165, 426	1.It detects the clickbots in a busy time period. 2.It works on different displays	1.Inability to detect complex JavaScript.	FCFraud acts as addition to the techniques of server-based detection	85.6%
[4]	Click Fraud Detection on the Advertiser Side. (2016)	1.JavaScript 2.Proactive functionality testing. 3.Passive browsing behavior examination.	1)JavaScript support and mouse event test. 2)Browser functionality test. 3)Browsing behavior examination	Records: 9900	1.It is not easy for the clickbots to pass the functionality. 2 Functionality of client can be properly tested.	Superior clickbots can spoof the information of application layer.	Effective in identifying while incurring negligible overhead.	99.1%
[15]	Method for performing real-time click fraud detection, prevention and reporting for online advertising (2016)	1.Client-side Scripting code 2.Server-side tracking code	1. HTTP Requests(GET Function) 2)JavaScript support and mouse event test. 3.Browser functionality test.	NA	If the rules set are not satisfied, the user will be redirected to a non-paying advertisement.	It is prone to additional risks such as parasite ware.	Helps in Identification of clicks that are valid and invalid to prevent from being billed for fraudulent activity.	NA
[20]	Pixel Cluster Transit Monitoring for detecting Click Fraud. (2016)	1.Pixel Clustering. 2.JavaScript	1.Cluster Analysis 2.Applet 3.Javascript.	NA	1.Real time analysis is done using JavaScript on Webpage.	JavaScript may crash during execution.	Uses applet or java script to keep track of transit of the cursor to detect false clicks	NA

E. MISCELLANEOUS

In [1], Machine learning, Heuristic search and Naïve-Bayes method were used for fraud detection on very high dataset using Logged Event Based, Client Base, Heuristic search, Click IP's and Explored Conversion Sparsity which achieves very high precision on unbalanced dataset. In [6] Data Analysis Algorithm such as k-Nearest neighbor for recognizing click fraud are used. Special java script on advertiser's website is included which collects the important information about used behavior.

Java Script validation is performed in [2] to check click on HTTP request to detect the malware whether it executes the attack in a time-period when there is a high traffic(busy period). FCFraud also works in case a user has a touchscreen monitor. Adding FCFraud to the OS can serve the online advertising. Attention Mechanism, Stacked Autoencoder and Sparse Data

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

Prediction is implemented in [18] which has improved the click through rate and most importantly it automatically learns from the data with any human domain knowledge.

In [19] techniques like Logistics regression, Naïve Bayes, Support vector machine and decision tree have been used. Frequency of time-stamp of user on website is checked and the time user spend on the website is analyzed. Traffic partitioning, pooling and isolating click spam are used in [24] for traffic analysis. The network clickstream is analyzed thoroughly for recurring patterns that may show up due to anomalous activities thus increasing the efficiency .

Automated click fraud based on clickable captcha's is used in [26] which requires simple operation and less storage space. Click fraud is identified based on valid users. In [27] an android application has been designed which provides high true positive rates and low false negative rates, it accesses the phone application to check for frauds on the web ads. Exploratory data analysis, data pre-processing and data prediction is performed in [28] for click fraud detection, the proposed model has been developed to detect and minimize the malwares.

TABLE V: MISCELLANEOUS

Sr No.	Title	Methodology And Tools	Techniques	Dataset	Merits	Demerits	Results	Accuracy
[1]	Light GBM Machine Learning Algorithm to Online Click Fraud Detection. (2019)	LightGBM (Light Gradient Boosting Machine) algorithm.	1.Feature parallelism. 2. Data parallelism 3.Voting parallel	Kaggle (Records:200 million).	1.High performance 2. Reliability 3.Availability over traditional Models	Insufficient resources for training.	Improves the detection performance step by step.	98%
[6]	Detection of Advertisement Click Fraud Using Machine Learning (2020)	XGBoost algorithm	1.Exploratory data analysis 2.Data pre-processing 3.Data prediction	NA	The percentage of advertisement click fraud is found significant compared to regression model.	Cost of prediction utility for classification in is very limited.	Detect and minimize the malwares that monetizes using click fraud.	NA
[12]	Crowdsourcing for Click Fraud Detection (2019)	1.Click Fraud Crowd-sourcing 2.Android Studio	Android Application.	CFCA Records: 500,000	1.Provides high true positive rate. 2.It has low false negative rate.	Performs additional step to merge its library.	Click Fraud Crowd-sourcing gives good result.	93%
[18]	A Multi-time-scale Time Series Analysis for Click Fraud Forecasting using Binary Labeled Imbalanced Dataset (2019)	1.Learning Based Probabilistic Estimator 2.Probabilistic Based (PB) Modelling 3.AR (Auto Regressive modelling)	1.Data Preprocessing 2.Data Smoothing	Kaggle Record: 184,903,890	This approach allows behavior to be forecasted in seconds and time scales .	It is difficult to detect Botnet which can click an ad by impersonating itself as a genuine user.	Probability model is found to be good than the learning based probabilistic estimator model	96%
[19]	Data Analysis Algorithm for Click Fraud Recognition. (2018)	1.k-Nearest Neighbours. 2.Differential Evolution.	1.Clustering 2.Classification 3.Data collection	Records: 30,000	1.The operator can evaluate succeeding clicks by his own. 2.It can be used for various sizes of data.	Real algorithms are hard to get which deals with user classifications of organic versus non organic.	k-NN classifier application for clustering algorithm is very good.	97.11%
[24]	A survey on online click fraud execution and analysis. (2018)	1. Group blooms filter algorithm. 2.Time blooms filter algorithm.	CAPTCHA Generation and Validation	NA	1.Require simpler operations. 2.Verifies Input from user which is difficult for a bot to generate.	1.Loading captcha's require time and space. 2.Critical issue is to identify copy clicks.	Click fraud is identified based on valid users using CAPTCHA validation.	64.07%.
[26]	Click Fraud Detection (2019)	1.DATA MINING: →SMOTE → ADASYN → NCL 2.ROC Plots	1.AdaBoost 2.Bagging 3.LogitBoost 4.Rotation Forest	NA	Precision is high for large dataset compared to regression model.	Re-sampling resulted in reduction in performance measures.	Classification improved when data was pre-processed by performing feature selection.	92.60%
[27]	Combating online fraud attacks in mobile-based advertising (2016)	1.Implementation of android app 2.SDK library	1.Honey advertisement. 2.Detecting anomalous behaviours.	NA	Analyzed potential risks of services of mobile advertisement.	Current results are not sufficient because there are very less	Evaluation of potential risk in mobile advertising related to	75 %

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

						number of ad networks that are tested.	automated online fraud attacks.	
[28]	Click-through rate (CTR) prediction (2018)	1.Sparse data prediction Methods 2. ASAE Model	1.AUC (area under ROC) 2.Logloss (cross entropy)	Training: 149,639,105 Test: 20,257,594	1. Pre-training is not required. 2.High- and low-order feature interactions are learnt.	The pre-training stage introduces an overhead which reduces efficiency.	Model mines the relationship between features, which improves the CTR	79.81%

III. METHODOLOGY

Data collection: Dataset is collected from Kaggle named TalkingData AdTracking Fraud Detection Challenge which contains attributes such as click_time, ip, app, device, OS, channel, attributed_type and is_attribute but we are focused on ip address and click_time to detect the fraud.

Data Preprocessing: Collected dataset from Kaggle needs to be preprocessed in order to eliminate the null and unwanted values by using WEKA/python tool. WEKA has application such as Preprocess, Classify, Cluster, Select Attribute and Visualize.

The system needs to implement the rules which are called as Offline Rules. Offline rules will appear such as Blacklist-Rule and Time-Period Rule, where Blacklist-Rule will maintain the Blacklist IPs which occur frequently. When the user visits the website it's IP address will be matched with the Blacklist IP's to check if it a fraudulent IP. Time-Period Rule will detect the fraud based on the time that the normal person and fraud agent/botnet will utilize to click on the ad.

The model uses Logistic regression method. This method basically implements the Sigmoid Activation Function where the decisions can be made as following:

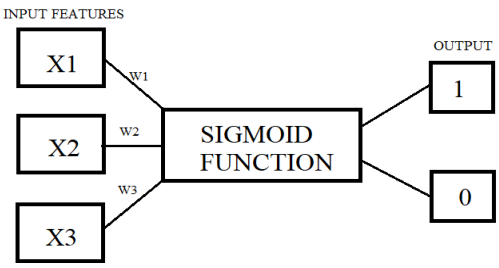

Fig1.: Working of Sigmoid Function

The result 1 denotes that the resulting click is genuine.
The result 0 denotes that the resulting click is malicious.

Sigmoid Activation Function determine the independent variable based on the dependent variable in the form of 0 and 1.
A system architecture is proposed as shown below:

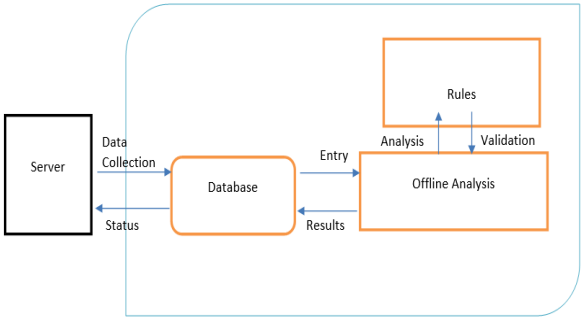

Fig.2: Proposed Model System

A request for data from server/google server/system is made from the database side. Only required data (Valid IP address and time of click) will be requested into database so that the memory consumption is efficient. The data will be further send for offline analysis which will check for rules such Blacklist rule and time period rule. Further the results will be validated and send to database. From database transaction of data will be analyzed by server for detection of click fraud activities.

IV. CONCLUSION

The paper provides a brief overview and comparison between the different techniques used in advertisement click fraud detection. The traditional techniques used earlier for click detection and click validation are used for the purpose. Every technique mentioned has its own merits and demerits. A few common approaches like Logistic Regression, Traffic Analysis and JavaScript, Support Vector Machine are summarized and represented in the form of tables. The implementation of the traditional methods is cumbersome. The traditional methods are less accurate.

It has been observed that techniques under the JavaScript enabled system have an average accuracy of 90%. In system approach the advertisers can independently detect click fraud activities performed by clickbots and human clickers.

From the above tables most of the time logistic regression and offline rules based on user interface system gives a good accuracy in detecting advertisement click fraud. The Regression model gives higher accuracy for higher number of attributes and the offline rule(Blacklist Rule) helps in

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 05, May-2021

minimizing overhead processes of similar data. The user interface system deals with the fetching of advertisement and usually also monitors the activity of a user with respect to the input functions mentioned. It also tracks the time spent by the users before clicking an advertisement to which is the most important validation property to verify if the click can be a genuine or a fraudulent.

REFERENCES

[1]. Elena-Adriana and Gabriela, Light GBM Machine Learning Algorithm to Online Click Fraud Detection, researchgate.net, 2019.
[2]. Md Shahrear Iqbal, Mohammad Zulkernine, Fehmi Jaafar,Yuan Gu, FCFraud: Fighting Click-Fraud from the User Side, IEEE, 2016.
[3]. Mukesh Patel School of Technology, Identification of Click Fraud and Review of Existing Detection Algorithms, IEEE, 2019.
[4]. Haitao Xu, Daiping Liu, Aaron Koehl, Click Fraud Detection on the Advertiser Side, eecs.northwestern.edu, 2016.
[5]. Shishir Nagaraja, Ryan Shah,University of Strathclyde, Clicktok: Click Fraud Detection using Traffic Analysis, WiSec '19: Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks, 2019.
[6]. B. Viruthika, Suman Sangeeta Das, E Manish Kumar, D Prabhu, Detection of Advertisement Click Fraud Using Machine Learning, International Journal of Advanced Science and Technology, 2020.
[7]. Apoorva Srivastava, San Jose State University, REAL-TIME AD CLICK FRAUD DETECTION, scholarworks.sjsu.edu, 2020.
[8]. G. S. Thejas, Kianoosh G. Boroojeni, Kshitij Chandna, Isha Bhatia, S. S. Lyenger, N. R. Sunitha, Deep Learning-based Model to Fight Against Ad Click Fraud, Florida International University, ACM SE '20: Proceedings of the 2020 ACM Southeast Conference, 2020.
[9]. Xin Zhang, A Click Fraud Detection Scheme based on Cost sensitive BPNN and ABC in Mobile Advertising, Xuejun Liu,Han Guo School of Computer Science and Technology Nanjing Tech University Nanjing, China, IEEE, 2018.
[10]. Paulo S. Almeida, João J. C. Gondim, Click Fraud Detection and Prevention System for Ad Networks, researchgate.net, 2018.
[11]. Qi Wang, Linzhi Li, Yadong Xu, LiLi Yang, Advertisement Click-through Rate Prediction, cse.scu.edu, 2017.
[12]. Riwa Mouawi , Imad H. Elhajj, Ali Chehab and Ayman Kayssi, Crowdsourcing for click fraud detection, researchgate.net, 2019.
[13]. Lijiao Pan, User click fraud detection method based on Top-Rank-k frequent pattern mining, Shibiao Mu and Yingyan Wang Yiwu Industrial & Commercial College, worldscientific.com, 2019.
[14]. Kam Kouladjie, Seattle, CLICK FRAUD MONITORING BASED ON ADVERTISING TRAFFIC, US PATENT, 2017.
[15]. Method For Performing Real-Time Click Fraud Detection, Prevention And Reporting For Online Advertising, VALIDCLICK, INC., Conway, US PATENT, 2016.
[16]. Sen Zhang,Qiang Fu,Wendong Xiao, Advertisement Click-Through Rate Prediction Based on the Weighted-ELM and Adaboost Algorithm, hindawi.com/journals/sp/2017/2938369, 2017.
[17]. Jing Ying Zhang, Click Fraud Detection: Adversarial Pattern Recognition over 5 Years at Microsoft ,Brendan Kitts, ouyangchen.com, 2017.
[18]. Thejas G. S., Jayesh Soni , Kianoosh G. Boroojeni, A Multi-time-scale Time Series Analysis for Click Fraud Forecasting using Binary Labeled Imbalanced Dataset, people.cis.fiu.edu, 2019.
[19]. Marcin Gabryel, Data Analysis Algorithm for Click Fraud Recognition, Institute of Computational Intelligence, Czestochowa University of Technology, researchgate.net, 2018.
[20]. Patrick O'Sullivan, Ballsbridge (IE), Pixel cluster transit monitoring for detecting click fraud, US PATENT, 2016.
[21]. Zhi Li, Weichen Jia,School of Media and Law, The Study on Preventing Click Fraud in Internet Advertising, Zhejiang University Ningbo Institute of Technology, Ningbo, China, csroc.org.tw/journal/JOC31-3/JOC3103-20, 2020.
[22]. S. Saraswathi, Vallidevi Krishnamurthy, Machine Learning Based Ad-click Prediction System, .researchgate.net, 2019.
[23]. Nayanaba Gohil, Arvind D Meniya, A Survey on Online Advertising and Click fraud detection, Nayanaba Gohil Department of Information Technology Shantilal Shah Engineering College Bhavnagar, Gujarat, India, .researchgate.net, 2020.
[24]. Dr.R.Kayalvizhi,Kapil Khattar,Piya Mishra, A Survey on Online Click Fraud Execution and Analysis, ripublication.com/ijaer18/ijaerv13n18_59, 2018.
[25]. Ladislav Beránek, Václav Nýdl, Radim Remes, Click Stream Data Analysis for Online Fraud Detection in E-Commerce, semanticscholar.org, 2016.
[26]. Kaveri Kar, A Study on Machine Learning Approach using Ensemble Learners for Click Fraud Detection in Online Advertisement, gujaratresearchsociety.in/index.php /JGRS/article, 2019.
[27]. Geumhwan , Junsung Cho, Youngbae Song, Combating online fraud attacks in mobile-based advertising, jis-eurasipjournals.springeropen.com, 2016.
[28]. Qianqian Wang,Fang'ai Liu,Shuning Xing,Xiaohui Zhao, Click-through rate (CTR) prediction, hindawi.com/journals, 2018.
[29]. Joe DeBlasio, Saikat Guha, Geoffrey M. Voelker, Alex C. Snoeren, Search Advertiser Fraud, 2017.
[30]. Leyi Song , Xueqing Gong , Xiaofeng He , Rong Zhang , Aoying Zhou, Malicious Click Detection in Web Advertising Data, East China Normal University 3663 North Zhongshan Road, Shanghai, China, citeseerx.ist.psu.edu, 2018.