International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 04 | Apr 2022 www.irjet.net p-ISSN: 2395-0072

Click Fraud Detection Of Advertisements using Machine Learning

Yameeni Mhaske[1], Ankita Gupta[2], Vaishnavi Bhosale & Prof. Krishnendu Nair

[1,2,3]Studens , Dept. of IT Engineering, Pillai College Of Engineering, Maharashtra, India
[4]Prof , Dept. of IT Engineering, Pillai College Of Engineering, Maharashtra, India

---

Abstract – *Ad networks charges advertisers for every click they get on their ads displayed on the website. So when a malicious user clicks on the ad without any intention of productive use the advertiser still has to pay the website owner and advertisers may face great loss (in millions) . There are many ways a click-fraud can occur such as botnets, competitors. In this project, we will present a method to detect such fraud by developing an effective click fraud detection algorithm essential for online advertising businesses. We tried various machine learning models and try to built a high-end algorithm having most accuracy and is effective throughout the datasets. Hence we used XGBoost and tried adding new features to dataset. The output file will contain the click-ip and the probability if the user actually downloaded the app after clicking on it.*

***Key Words*: Click Fraud, Ip-address, Number of clicks, click-time, is_attributed**

1. INTRODUCTION

World has come closer because of online marketing. Earlier small companies where struggling to achieve bigger audience but due to pay per click and other digital tools many companies are flourishing. According to census more than 4 billion people use internet on a daily basis and 2 billion shop online .Also according to google new survey more than 5 million clicks are generated on ads . But there are always more frauds in any busy marketplace. One such fraud is click fraud. Click fraud occurs when large number of useless clicks are produce intentionally and un-intentionally which cause huge loss of money to the advertisers. Mostly fraud occurs intentionally and Click fraud can be conducted in two main ways: (1) hiring a group of people to increase fraudulent traffic, and (2) deploying automatic clicks/click bots. Shreds of evidence show that using a click bot to launch a click fraud attack is much more effective and thus far more common than the manual type of attack involving humans.

2. Literature Survey

The 1st we referred to is A hybrid and effective learning approach for Click Fraud detection(2021)Authors: Thejas G.S., Surya Dheeshjith , S.S. Iyengar , N.R. Sunitha , Prajwal Badrinath.It uses the Cascaded Forest to concatenate the original dataset with additional columns and uses the XGBoost model for final classification.It uses the Cascaded Forest to concatenate the original dataset with additional columns and uses the XGBoost model for final classification. Using various click fraud detection models, we infer that CFXGB performs outstandingly well on performing comparative analysis. We also experimented with Parent node values and inferred that this parameter must be treated as a hyperparameter. Nevertheless, this model can be used as a generic model to solve other machine learning classification problems.

The 2nd paper referred to FCFraud: Fighting Click-Fraud from the User Side(2016)Authors:Shahrear Iqbal, Mohammad Zulkernine, Fehmi Jaafar, Yuan Gu. In this paper, we develop a technique, FCFraud, that protects innocent users by detecting the fraudulent processes that perform click-fraud silently. FCFraud executes as a part of the operating system's anti-malware service. It inspects and analyzes web requests and mouse events from all the user processes and applies RandomForest algorithm to automatically classify the ad requests. After that, it detects fraudulent ad clicks using a number of heuristics. In our experimental evaluation, FCFraud successfully detects all the processes running in the background and performing click fraud. It blocks them from further accessing the network thus removing the machine from the botnet. Most major operating system vendors own or operate ad networks and advertising is one of their major sources of income. As a result, we believe that adding FCFraud to the operating system's anti-malware service can greatly serve the interests of the operating system vendors and online advertisers and it can be a valuable addition to the server-based detection techniques.

The 3rd paper referred is Detection of Advertisement Click Fraud Using Machine Learning.(2020) Authors: B. Viruthika, Suman Sangeeta Das, E Manish Kumar, D PrabhuThe percentage of advertisement click fraud is found significant. Recent statistics have proven that the cause is major and would be increasing in the future. Thus, the proposed model has been developed to detect and minimize the malwares that monetises using click fraud. It is in need as criminals get profit out of it and the problem is on a large scale. Hence, the proposed system has overcome the problem to maximum extent and provides the result accurately.

The 4th paper is A Method for Detecting Fraud Advertisement (2020) Authors: Keesara Sravanthi, Batturi Pavankalyan, Sharath Chandra.We had done the advertising click fraud detection by using neural networks and we got a result of accuracy 91% with less false positives. We had used Gaussian naïve bayes classifiers before normalization. We got more false positives so then we had normalized the data and we tested the data so we got less number of false

positives. So we conclude that the neural network is the best technique to find the frauds in advertisements.

3. Proposed Work

We have gathered data from China's largest independent big data service platform, covering over 70% of active mobile devices nationwide. They handle 3 billion clicks per day, of which 90% are potentially fraudulent. According to the references we are planning to implement the techniques like Adaboost,, XGBoosts, with , feature encoding and engineering, and sequence modeling. We are implementing Jupyter Notebook scikit Learn and a confusion matrix to sort the data. We have discussed the model of our project using block diagrams and hardware software details. If this algorithm is applied into advertisement click fraud detection systems, the probability of fraud transactions can be predicted soon after click fraud occurs. Thereafter a series of anti-fraud strategies can be adopted to prevent advertisers from great losses and reduce risks.

4. System Architecture

The system architecture is given in Figure 1. Each block is described in this Section.

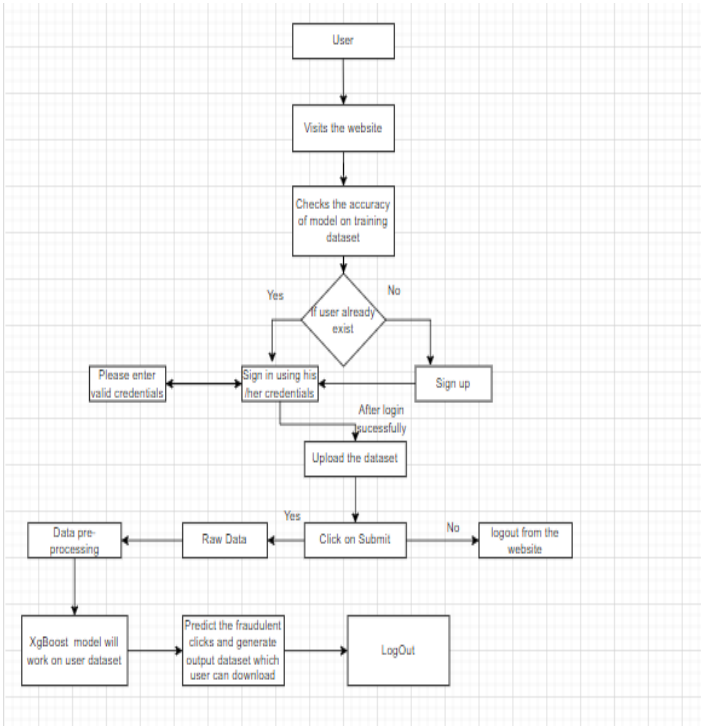


Fig. 1 Proposed system architecture

This is a block diagram for our system. It is the actual representation of the algorithm which we wish to implement.

The 1st step is to read the data set & then it is sent for sampling. Training & testing of the data set is done.

After the feature selection, the data will be sent to the algorithm which is the combination of the Adaboost, XGBoost.

The resultant data is stored in test sample data. The prediction of outcome is done Based on test sample data & the result of the combined algorithm.

Later the performance & accuracy results are plotted.

5 Requirement Analysis

The implementation detail is given in this section.

5.1 Software

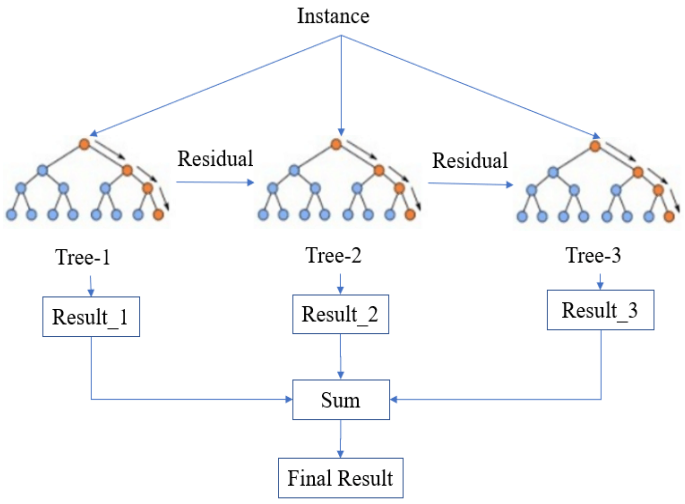
1. Operating system : Windows 8/10.
2. IDE Tool : Anaconda
3. Coding Language : Python 3.6 & up
4. APIs : Numpy, Pandas ,Seaborn, Matplotlib

5.2 Hardware

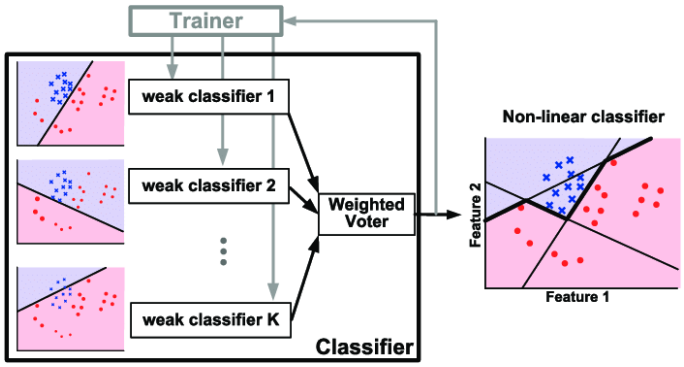
1. Processor : Pentium i3 or higher.
2. RAM : 4 GB or higher.
3. Hard Disk Drive : 20 GB (free).
4. Peripheral Devices : Monitor, Mouse and Keyboard

5.3 Algorithms

XGBoost: XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.XGBoost is an implementation of gradient boosted decision trees designed for speed and performance



AdaBoost:.AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.



COMPARISON

Performance of all learning algorithms used for Advertisement click fraud detection are compared in table The comparison is based on their accuracy, precision and specificity.

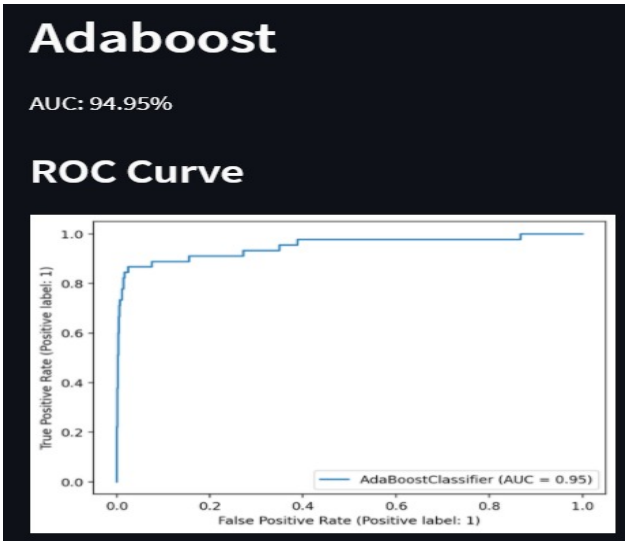
Classifier	Metrices		
	Accuracy	Decision	Specificity
XGBoost	0.962	0.997	0.987
Adaboost	0.949	0.996	0.979

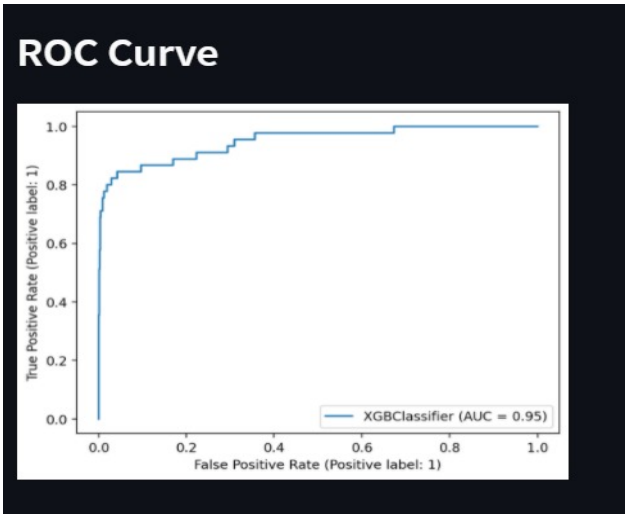
Result:

	click_id	is_attributed
0	0	0.000002
1	1	0.000013
2	2	0.000026
3	3	0.000470
4	4	0.000214

	click_id	is_attributed
0	False	False
1	False	False
2	False	False
3	False	False
4	False	False
...	...	...
18790464	False	True
18790465	False	True
18790466	False	True
18790467	False	True
18790468	False	True

18790469 rows × 2 columns





5.4 Data Preprocessing

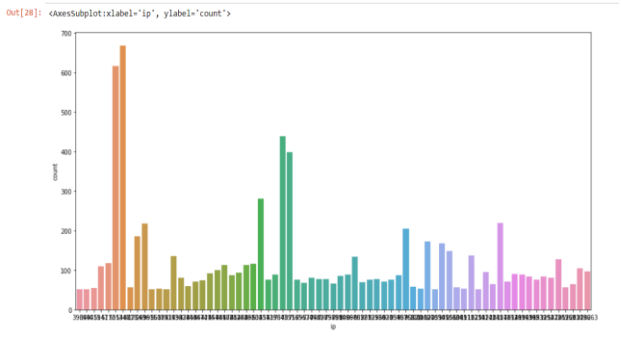
In this module selected data is formatted, cleaned and sampled. The dataset that we've considered is from talking data which is China's largest independent big data service platform, covering over 70% of active mobile devices nationwide. They handle 3 billion clicks per day, of which 90% are potentially fraudulent. The data preprocessing steps includes following:

a. Formatting: The data which has been selected may not be in a suitable format. The data may be in a file format and we may like it in relational databases or vice versa.

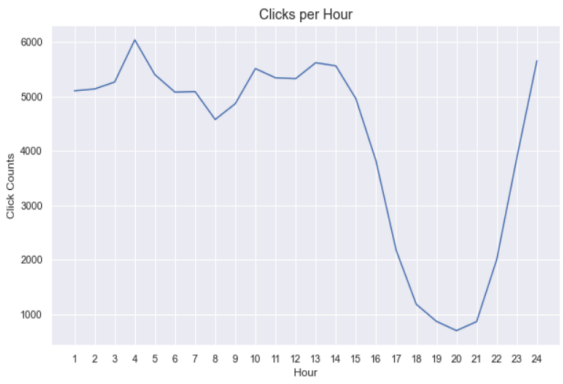
b. Cleaning: Removal or fixing of missing data is called cleaning. The dataset may contain records which may be incomplete or it may have null values. Such records need to be removed.

c. Sampling: As the number of frauds in the dataset is less than the overall transaction, class distribution is unbalanced in credit card transactions. Hence sampling method is used to solve this issue.

IP Count

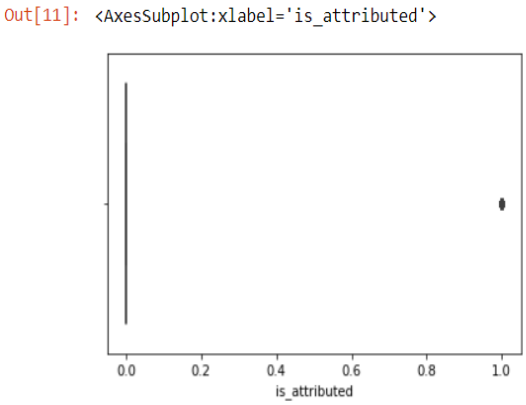


Clicks per hour

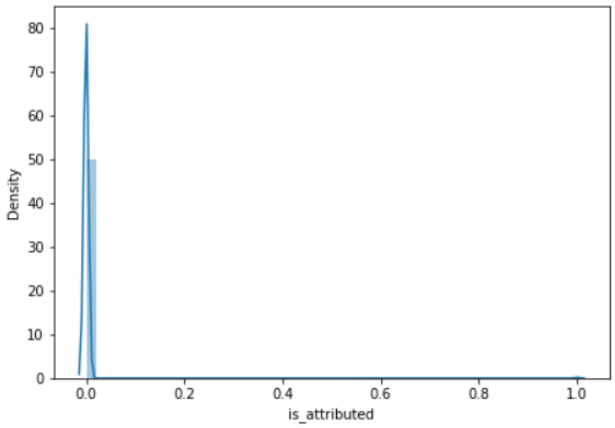


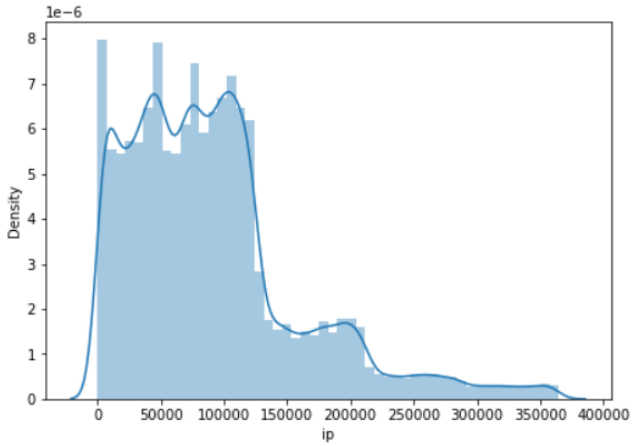
Outliers:

Our dataset contains approximate 99% of fraudulent clicks therefore we are getting a straight line at 0 on x axis. Whereas we have approximate 1% of valid clicks. Therefore the box-plot graph shows a dot at 1 on x-axis



Plot the Distribution plots for the features





6. CONCLUSION :

Due to evergrowing online market, companies are now shifting to selling their products or services online. They are advertising their products and services on internet through different websites. They therefore undertake the ppc(pay per click) campaign and pay the add network. But this also gave rise to Frauds and loss of money due to fraud. So we proposed a Machine learning model, XGBoost, that accurately detect fraud clicks and predict the ips causing fraud.

ACKNOWLEDGMENT

We would like to thank our guide and mentor for this project Prof. Krishnendu Nair who mentored us throughout our "Advertisement Click Fraud Detection using Machine Learning" project and cleared our concepts and helped us understand all the topics.

We would also like to thank the Head Of Department Of IT Dr. Satishkumar Varma for giving us an opportunity to understand and implement this project, which helped us a lot in understanding deep concepts of Software Designing, and how it works.

We thank our principal Dr. Sandeep Joshi for providing us with all the facilities and required equipment for this project. We would like to thank our guide and mentor for this SDL lab Prof. Smita Kadam who mentored us throughout our "Credit Card Fraud Detection" project and cleared our concepts and helped us understand all the topics.

We would also like to thank the Head of Department Of IT Dr. Satishkumar Verma for giving us an opportunity to understand and implement this project, which helped us a lot in understanding deep concepts of Software Designing, and how it works.

We thank our principal Dr. Sandeep Joshi for providing us with all the facilities and required equipment for this project.

REFERENCES

[1]. M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kotter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel. KNIME {The Konstanz information miner: Version 2.0 and beyond. SIGKDD Explorations Newsletter, 11(1):26{31, 2009.

[2]. G. E. P. Box. Non-normality and tests on variances. Bio metrika, 30(3/4):318{335, 1953.

[3]. L. Breiman. Bagging predictors. Machine Learning, 24: 123 {140, 1996.

[4].L.Breiman. XGBoosts. Machine Learning, 45(1):5{32, 2001.

[5]. C. Chambers. Is click fraud a ticking time bomb under Google? Forbes Magazine, 2012. URL http:// www. forbes. com/sites/investor/2012/06/18/is-click-fraud-a-ticking-time-bomb-under-google/.

[6]. C. C. Chang and C. J. Lin. LIBSVM: A library for supporting vector machines. ACM Transactions on Intelligent Systems Technology, 2(3):27:1{27:27, 2011.

[7].A. Chao and T. Shen. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. Environmental and Ecological Statistics, 10:429{443, 2003.

[8]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegel meyer. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321{357, 2002.

[9]. C. Chen, A. Liaw, and L. Breiman. Using XGBoosts to learn imbalanced data. Technical report, Technical Report No. 666, Department of Statistics, University of California, Berk eley, 2004.

[10]. W. Cohen. Fast effective rule induction. In Proceedings of the International Conference on Machine Learning, pages 115{123, Tahoe City, California, 1995.

[11].T. Cover. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1):21{27, 1967.

[12]. V. Dave, S. Guha, and Y. Zhang. Measuring and fingerprinting click-spam in ad networks. In ACM SIGCOMM Computer Communication Review, volume 42, pages 175{186, Helsinki, Finland, 2012.