International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 06 | June - 2022 Impact Factor: 7.185 ISSN: 2582-3930

MACHINE LEARNING APPROACH FOR CLICK FRAUD DETECTION

Dr. G. Madhukar, Associate Professor, Dept of CSE, CMR Technical Campus

K Teja Niranjan, Dept of CSE, CMR Technical Campus

N Madhuri, Dept of CSE, CMR Technical Campus

A Vamshidhar Reddy, Dept of CSE, CMR Technical Campus

ABSTRACT - Mobile advertising has gained popularity in recent years as a means for publishers to monetize their free applications due to the increase of Internet usage. Click fraud is one of the main concerns in the in-app advertising industry. Click fraud involves online advertisements that have been clicked on. Pay-per-click fraud involves online advertisements that have been clicked on. Advertisements that pay per click typically target potential customers by charging a fee per click. With machine learning as a solution, we designed the system to detect click fraud using naive bayes, xgboost classifier, random forest, decision tree with gradient boosting, extra tree classifier with gradient boosting, and we observed decision tree with gradient boosting outperformed other algorithms with 96.07% accuracy.

Key Words: click fraud detection, online advertisement, Real time fraud detection

1. INTRODUCTION

Fraudulent clicks on pay-per-click ads are designed to divert the budgets of advertisers. There are several parties who are engaging in click fraud. Consider the top three criminals, competitors, webmasters, and fraud circles to understand who is clicking on your ad fraudulently. They serve ads to users and agree on a price per action. According to the frequency of visitors to the advertiser, the ad network pays the content publisher. With this payment model, however, there are security risks, such as click fraud. The number of fraudulent clicks for smartphones doubled in four months (ppc, 2019) from 1 in 5 in 2017. There is a significant portion of web traffic that is fraudulent based on these click fraud statistics. Click fraud always results in financial losses for the advertiser, regardless of its form. Most ad click fraud is committed by competitors. Make yourself more competitive by wasting your competitor's click billing budget. When webmasters commit click fraud, they display ads on their sites to generate fraudulent revenue. To increase sales, they choose to click on these ads instead of creating and developing their website. Click farms are a way to trick people into clicking on ads all day long to make money on click fraud. Compared to automated scripts, we find it more beneficial to use real people, as compelling click performers can lead to clicks on your advertisement.

2. METHODOLOGY

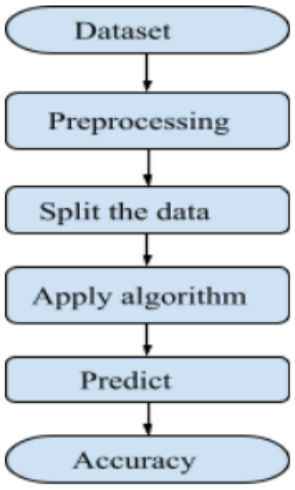

Figure- 2.1 Data Flow Diagram

The diagram basically describes a program control overflow. The first step in your project is to fetch the dataset and remove all kinds of errors, missing values and noisy data. This is sometimes referred to as data pre-processing. After the data has been processed, it will try to split the data. Training and test datasets that try to apply the decision tree algorithm individually. After applying these algorithms, you will get two types of results for both the test and training datasets, and these results will be compared in the next step. These steps of applying the algorithm to get the values continue until you have the accuracy you need for your project.

Dataset

Talking Data, China's largest independent big data service platform, covers more than 70% of active mobile devices nationwide. It processes 3 billion clicks per day, 90% of which are potentially fraudulent. The current approach to prevent click fraud by app developers is to measure user click behaviour across the portfolio and flag IP addresses that generate a lot of clicks but don't install the app. I used this information to create an IP blacklist and a device blacklist.

The dataset contains 100001 records, column are 8 and label is 0

Attribute information:

1)Ip

2)App

3)Device

4)Os

5)channel

6) click time

7) is attribute

3. MODELING AND ANALYSIS

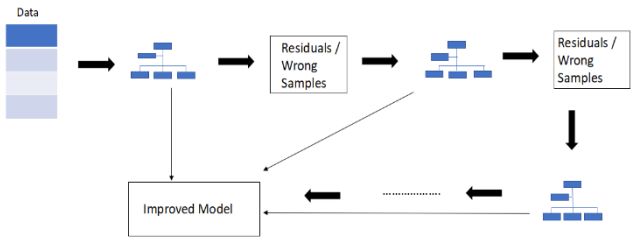

FIGURE 3.1: Architecture Diagram

The project architecture represents the full functionality of the click fraud detection project program. First, we collect data from various sources such as websites and Kaggle. Then remove the noisy data and try to pre-process the data. After the pre-processing is complete, it tries to apply the decision tree algorithm to the dataset. Therefore, after application, you will get two results. For example, if you get the correct results, try applying a decision tree algorithm to this data. They are added to the improved result collection, incorrect samples are reprocessed, and the process continues until reasonable accuracy is found.

4. RESULTS AND DISCUSSION

Evaluation metrics

True Positive: That is when we anticipate Jesus and the actual result is also Yes.

True Negative: In this case, we are predicting "no" and the actual output is also "no".

False positives: If you predicted "yes", it was actually "no".

False Negatives: If I expected it to be no, it wasn't.

accuracy=TP+TN/TP+FP+TN+FN

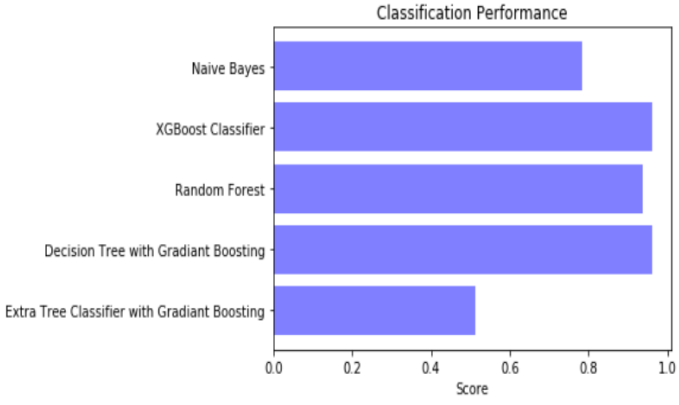

Figure 4.1: Bar Graph

We have trained 5 machine learning algorithms and the above bar graph accuracy comparison is given below

sno	Algorithm names	Accuracy
1	Naive Bayes	78.21 %
2	XGBoost Classifier	96.06 %
3	Random forest	93.62 %
4	Decision tree with gradient boosting	96.07 %
5	Extra tree classifier with Gradient boosting	51.10 %

Table 4.2: Accuracy Comparison of Algorithms

We have observed that decision tree algorithm has performed better than other algorithms so we finalized decision tree.

5. CONCLUSION

We have developed a click fraud detection mechanism that can be used in the real world. You used a dataset with different attributes. We have used many click fraud detection algorithms such as Naive Bayes, xgboost classifier, decision tree gradient boosting, additional tree classifier with gradient boosting, and random forest. Of all these algorithms, xgBoost works very well with a project accuracy of 0.9606. This machine learning template can be used to identify real and fake users.

6. FUTURE SCOPE

If many resources are available, you can increase the number of decision trees to get accurate results. You can also apply multi-grain scans to improve data preprocessing. You can also add the consumer's geographic location as an attribute to analyze and customize the results. Also, if you use this geographic location to see if a person or bot is trying to click from the new location, you'll see a warning flag telling you that the new user is clicking. I think these ideas need further discussion as they are input attributes that are useful for classification systems and projects.

7. ACKNOWLEDGEMENT

The success of this paper includes help from our guide as well. We are grateful to our guide, Dr. G. Madhukar, Associate professor, CMR Technical Campus, for his expertise that assisted us in our research.

8. REFERENCES

[1]. Antoniou, D., Paschou, M., Sakkopoulos, E., Sourla, E., Tzimas, G., Tsakalidis, A., & Viennas, E. (2011). Exposing click-fraud using a burst detection algorithm. In 2011 IEEE

[2]. symposium on computers and communications (ISCC) (pp. 1111–1116). IEEE.

[3]. Arisoy, E., Sainath, T. N., Kingsbury, B., & Ramabhadran, B. (2012). Deep neural network language models. In Proceedings of the NAACL-HLT 2012 workshop: will we ever really replace the N-gram model? on the future of language modeling for HLT (pp. 20–28). Association for Computational Linguistics.

[4]. Breiman, L. (1996). Stacked regressions. Machine Learning, 24(1), 49–64. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[5]. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd Acm Sigkdd international conference on knowledge discovery and data mining (pp. 785–794). ACM.

[6]. Chenoweth, T., ObradoviĆ, Z., & Lee, S. S. (2017). Embedding technical analysis into neural network-based trading systems. In Artificial intelligence applications on wall street (pp. 523–541). Routledge.

[7]. Click fraud statistics: The click fraud blog. (2019). Click Cease. URL: https://www. lickcease.com/blog/click-fraud-statistics/.

[8]. Click fraud - The 5 most common forms of click fraud. (2019). URL: https://fruition. net/about/blog/types-click-fraud-detect/.