

# AD CLICK FRAUD DETECTION USING A NOVEL ENSEMBLE MODEL BASED ON USER BEHAVIOR PATTERNS

K.Naga Vishnu<sup>1</sup>, Miss.C.Yamini<sup>2</sup>

<sup>1</sup>Student, Mca 2<sup>nd</sup> Year Kmmips, Tirupati, Affiliated To S.V. University, Tirupati, A.P, India

<sup>2</sup>Associate Professor, Dept Of Mca, Kmmips, Tirupati, Affiliated To S.V. University, Tirupati, A.P, INDIA

\*\*\*

**ABSTRACT-** In online advertising, click fraud poses significant challenge, draining budgets and threatening the industry's integrity by redirecting funds away from legitimate advertisers. Despite ongoing efforts to combat these fraudulent practices, recent data emphasizes their widespread and persistent nature. Toward detecting click fraud effectively, this study employed a comprehensive feature engineering and extraction approach to identify subtle differences in click behaviour that could be used to distinguish fraudulent from legitimate clicks. Subsequently, a thorough evaluation was conducted involving nine diverse machine learning (ML) and Deep Learning (DL) models. After Recursive Feature Elimination (RFE), the ML models consistently demonstrated robust performance. DT and RF surpassed 98.99% accuracy, while GB, LightGBM, and XGBoost achieved 98.90% or higher. Precision scores, measuring accurate identification of fraudulent clicks, exceeded 98% for models like ANN.

**Key Words:** Click fraud, machine learning, deep learning, online-advertising, bot detection, pay-per click, fraud.

## 1.INTRODUCTION

Today's world increasingly depends on web services, including marketing activities conducted on websites and smart phones. Among the most essential of these services are marketing or advertising campaigns, which are visible across a spectrum of websites and applications and take the form of advertisements to attract visitors and potential customers to draw attention to the promoted service or product. Through advertising campaigns, advertisements will be displayed on relevant web pages to increase profits. In these campaigns, advertisers pay fees for each click they receive on the ad, which is referred to as pay-per-click (PPC). Clicks on these advertisements may originate from legitimate, unsuspecting web users. Still, they may also occur due to malicious clicks conducted by individual software developed by rivals with illegal intentions. In these cases, the motivations may include maximizing the benefits of an organization or extracting excessive fees from advertisers. Juniper Research's, the most recent ad fraud report estimates that by the end of 2023, the anticipated cost to advertisers will reach \$84 billion, representing more than a fifth (22%) of all online advertising expenditures, based on the analysis of over

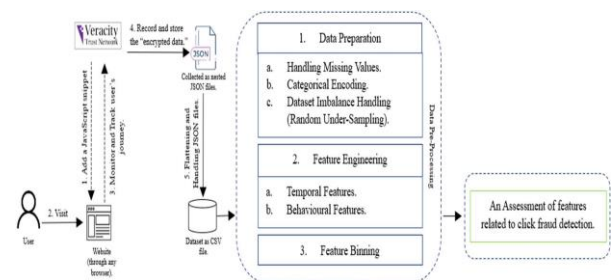
78,000 data sets of ad activity spanning 45 countries and eight major global regions. Moreover, by the end of 2023, it's estimated that 17% of PCs and desktop clickthrough's will be invalid, not delivering a return on ad spend (ROAS). The forecast indicates that although the number of valid clickthrough's will rise from 160 billion in 2023 to nearly 235 billion by 2028, the number of fraudulent clickthrough's will increase from 37 billion in 2023 to more than 65 billion by 2028. Since malicious bots are growing in number and diversity, extensive research has been conducted to understand what causes these misleading clicks and how they can be detected and predicted. To detect instances of click fraud, a range of artificial intelligence (AI) models are used to analyse when an advertisement is clicked on by either a human or a computer program. By assessing the authenticity of a click, these models aim to distinguish between legitimate and fraudulent interactions. Click fraud is frequently committed by automated means, such as bots or similar tools that resemble genuine human activity on websites. By repeatedly clicking on advertised items, these bots mislead platforms into believing that an actual human being is engaged in the advertised product or service. Many clicks originating from a single device can be detected more quickly when an advertising network or advertiser becomes suspicious of click fraud activity. Cyber criminals, however, can bypass this mechanism by using virtual private networks (VPNs) to route bot communication through a wide array of frequently changing internet protocol (IP) addresses. Furthermore, they can commit click fraud by using multiple computers in different geographical locations, thus allowing them to diversify the source of clicks, which can be high, medium, or low in volume. Despite marketers' efforts to fight click fraud, current statistics indicate that the issue is widespread and is expected to worsen in the future. According to the latest available data, marketers incurred about \$71.37 billion in 2024 due to fraudulent clicks. Due to the emergence of botnets based on fraudulent clicks, it has become increasingly essential to investigate this issue in detail to determine effective solutions. Furthermore, due to recent advances in artificial intelligence (AI) technologies and their wide application in cybersecurity, various defensive systems designed for advertising networks have been developed to detect click fraud activity. Thus, attackers have become more proficient at committing click fraud,

adopting tactics that mimic everyday user behaviour to avoid detection. In light of this evolution, there is a growing need for more robust and cutting-edge solutions to detect click fraud. Our objective in this study is to develop multiple machine learning (ML) and deep learning (DL) models capable of distinguishing between humans and bot visitors to a website. In this study, we worked on a real-time dataset containing the browsing behaviour of internet users as they interact with web sites. Following the preprocessing of the dataset, we derived a set of novel features and subsequently identified the most influential ones. Based on these selected features, ML models were applied to the dataset, and then, a comprehensive validation and test sets, showcasing its generalizability with an accuracy of 59.39%. Gradient boosting models (GBM) and their variations have been extensively used in past research due to their effectiveness in extracting features and classifying clicks. These models work by sequentially creating many decision trees. A version of GBM known as extreme gradient boosting (Boost) offers a more controlled version of Gradient Boosting. For instance, Phua et al. employed generalized boosted regression models and identified distinct spatiotemporal patterns in fraudulent clicks. They adopted simple statistical techniques to extract features related to click behaviour, frequency, and high-risk actions, significantly improving model performance and preventing overfitting during training. In another study, Ministering and Masonite utilized a state-of-the-art ML algorithm called light gradient boosting (Light) to analyse the actions of visitors who frequently click on ads but do not complete the desired action, like downloading an app. By engineering features and extracting time-related information from click times, they achieved an impressive 98% accuracy. Addressing the complexities of click fraud in the advertising sector, Singh and Sidonia emphasized the need for careful algorithms. They chose the gradient tree boosting (GTB) model, which outperformed 11 other ML algorithms in detecting fraudulent clicks. Dash and Pal constructed a click fraud detection model employing different ML methods, including SVM, KNN, DT, RF, and GBDT, to understand better the behaviour of individuals who regularly click on advertisements without accomplishing the desired action.

**2. RELATED WORKS**

Previous approaches have focused on using AI technologies such as machine learning (ML) and deep learning (DL), in addition to several successful methods already used for detecting click fraud. The primary purpose here is to protect advertisers from suffering significant expenditure due to misleading clicks, which can significantly influence the success of their marketing campaigns. Earlier research has demonstrated that employing AI methods to differentiate between valid and false ad clicks has proven quite effective. Most solutions for tracking the origin of a click have depended on ML techniques. Various tree-based

models, such as Decision Trees (DT), Random Forests (RF), and Extremely Randomized Trees (ERT), have shown impressive performance. Decision Trees build individual tree-node models, while Random Forests and Extremely Randomized Trees create multiple decision trees simultaneously using different strategies. Furthermore, other variations of tree-based methods have been employed. For instance, Berrari utilized Random Forests (RFs) with skewed bootstrap sampling to determine whether a publisher's clicks were legitimate or fraudulent. They included the click profile (time gaps between clicks) for analysis. Two tests were conducted using different subsets of included features. The first model exhibited the best performance, with an average accuracy of 49.99% in the validation and 42.01% in the test set. Yan and Jiang trained several classifiers with numerical features like IP addresses and the count of clicks at different time intervals during the day, along with statistical features. Classifier modellers, Bayesian networks (BNs), decision tables, REP Tree, and naïve Bayes (NB) were employed. Their findings highlighted that tree-based methods outperformed Bayes's approaches due to the imbalance between fraudulent and valid clicks, with fraudulent clicks being the majority. Perera et al. introduced a novel ensemble model based on user behaviour patterns from click data. They derived new valuable features from raw data that couldn't be used in their original forms to detect click fraud. They experimented with various classifiers, ultimately creating an ensemble model that integrated the six most effective classifiers: bagging with J48, bagging with Repartee, bagging with RF, Metaset with J48, Logbooks with J48, and random subspace with J48. This ensemble approach demonstrated its effectiveness on and Random Forest models. All models performed well, with an accuracy of up to 87%. As part of the new framework, Sisodia proposes two new stack generalization structures: one for resampling and the other for classification. The proposed structure's performance was compared with that of the previous literature based on the FDMA 2012 dataset. Stack generalization with gradient tree boosting (GTB) achieved a 66% average precision (AP). In another work, according to Sisodia et al., feature importance was tested using GTB, which proven to be an effective model design strategy that improves classification performance.



**Figure 1 : Research Methodology**

all of these features have been eliminated. As a result, we were able to obtain around 225 features. Notably, about 200 out of the 225 raw features are explicitly related to mouse movements action; these features are named c1, c2, c3, ..., and c200. An advertiser's page may contain a variety of actions the user takes (for example, scroll down, scroll up, click a button, move a mouse, play a video, etc.). If the action is "mouse movements," then the record corresponds to additional features that explain the behaviour of one mouse movement at a given moment. Therefore, if the user moves the mouse five times, this will correspond to five additional features for that specific click (features from c1 to c5), each describing one mouse movement (such as time, coordinates (x, y) of mouse movement, etc.). Accordingly, if there are 77 mouse movements, then there will be 77 additional features (features from c1 to c77), which will describe detailed information regarding each mouse movement and so on. B. DATASET PRE-PROCESSING

#### 4. HANDLING MISSING VALUES

It is common for real-time datasets to contain missing values, one of the most common problems that threaten data quality. In our dataset, approximately 30% of values are missing. Missing values on several features reached around 97% of the values; imputing such huge quantities will almost certainly result in bias, compromising the validity and accuracy of the model's predictions. Thus, if the missing data percentage reaches 80% or higher, the feature will be eliminated since the quantity of information captured in that feature is insufficient and will not contribute to the prediction model. In this study, the Miss Forest approach was used to impute missing data for the remaining features, given that most of the features in our dataset are categorical and imputing missing data using simple statistical methods such as mode (most frequent values) or mean and median (for numerical features) may affect the quality of the data and impose bias. Miss Forest outperformed and was more efficient than other imputation methods, such as traditional statistical approaches or K-NN based imputation.

#### 3. ENCODING CATEGORICAL FEATURES

Given that some raw and extracted features, such as the 'action' and 'user agent,' contain categorical values, they are converted into numerical features. Label Encoding techniques were employed for various features such as 'host', 'action', 'os' and so on. This approach allocates a unique integer to each distinct category or label within a categorical variable. This transformation simplifies the data, making it better suited for ML algorithms that require numerical input. Initially, our dataset consisted of 73,303 records, in which a total of 41,954 human clicks were recorded, while 32,362 bot clicks were recorded. As can be seen, there is an imbalance between the distribution of clicks on the two classes. Dataset imbalance occurs when

classes are distributed unevenly in a dataset, with some classes having significantly fewer instances than others. Thus, in this paper, we applied a random under-sampling technique to address this issue. We end up with 64,724 records and 32362 clicks in each class, benign or fraudulent. All datasets have imperfections, especially those collected in real-world scenarios. As mentioned earlier, we encounter termed challenges such as missing data and imbalanced class distributions. In addition, some raw features were highly correlated with the target class, which might have produced overfitting. As a solution to this issue, we eliminated these strongly correlated features to enhance the model performance and generalization.

#### 5. RESULT

In the context to click fraud detection, it is essential to reduce false negatives to prevent real cases of fraudulent activity from going undetected. False negatives happen when fraudulent clicks are mistakenly categorized as genuine. This can cause advertisers to suffer financial losses as well as damage their reputations. Never the less, it's crucial to find a balance between this goal and the possible effects of a false positive on business. False positives happen when legitimate clicks are mistakenly reported as fraud, which can cost advertisers real chances to engage with customers while generating revenue. In this regard, we should focus on performance metrics that prioritize both minimizing false negatives and managing the impact of false positives. Key metrics include Precision, Recall, and F1 score.

#### 7. CONCLUSION

In the landscape of online advertising, the persistent threat of click fraud looms large, undermining the industry's integrity and siphoning funds away from legitimate advertisers. Our study, aimed at addressing this challenge, embarked on a comprehensive exploration of feature engineering and extraction techniques to discern subtle nuances in click behaviour, crucial for distinguishing between fraudulent and genuine clicks. We meticulously evaluated nine ML and three DL models to identify the most effective approaches in combating click fraud. Our findings underscore the robust performance of ML models, particularly after RFE. Notably, DT and RF models surpassed 98.99% accuracy, while GB, Light, and Boost achieved accuracy rates of 98.90% or higher. Impressively, models such as ANN exhibited precision scores exceeding 98%, indicating their adeptness at accurately identifying fraudulent clicks.

#### 6. REFERENCES

[1] Juniper Research, Hampshire, U.K. Quantifying the Cost of Ad Fraud: 2023–2028. Accessed: Jul. 12, 2024. [Online]. Available: [https://fraudblocker.com/wpcontent/uploads/2023/09/Ad-Fraud-Whitepaper\\_Juniper-Research.pdf](https://fraudblocker.com/wpcontent/uploads/2023/09/Ad-Fraud-Whitepaper_Juniper-Research.pdf)

[2] X. Zhu, H. Tao, Z. Wu, J. Cao, K. Kalish, and J. Kayne, *Fraud Prevention in Online Digital Advertising*. Cham, Switzerland: Springer, 2017.

[3] A. K. Wood and A. M. Ravel, "Fool me once: Regulating fake news and other online advertising," *S. Cal. L. Rev.*, vol. 91, p. 1223, Jan. 2017.

[4] B. Stone-Gross, R. Stevens, A. Zarras, R. Kemmerer, C. Kruegel, and G. Vigna, "Understanding fraudulent activities in online ad exchanges," in *Proc. ACM SIGCOMM Conf. Internet Meas. Conf.*, Nov. 2011, pp. 279–294.

[5] (2024). *Wasted Ad Spend Report 2024*. [Online]. Available:[https://lp.lunio.ai/wpcontent/uploads/2023/09/Lunio\\_Wasted\\_Ad\\_Spend\\_Report\\_2024\\_V2.pdf](https://lp.lunio.ai/wpcontent/uploads/2023/09/Lunio_Wasted_Ad_Spend_Report_2024_V2.pdf)

[6] D. Berar, "Random forests for the detection of click fraud in online mobile advertising," in *Proc. Int. Work. Fraud Detect. Mob. Advert. (FDMA)*, Singapore, 2012, pp. 1–10. [Online]. Available:[http://berrar.com/resources/Berrar\\_FDMA2012.pdf](http://berrar.com/resources/Berrar_FDMA2012.pdf)

[7] J. H. Yan and W. R. Jiang, "Research on information technology with detecting the fraudulent clicks using classification method," *Adv. Mater. Res.*, vol. 859, pp. 586–590, Dec. 2013, doi:10.4028/www.scientific.net/amr.859.586.

[8] K. S. Perera, B. Neupane, M. A. Faisal, Z. Aung, and W. L. Woon, "A novel ensemble learning-based approach for click fraud detection in mobile advertising," in *Mining Intelligence and Knowledge Exploration (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8284. Berlin, Germany: Springer, 2013, pp. 370–382, doi: 10.1007/978-3-319-03844-5\_38.