

Research Paper

Hybrid cnn attention model for accurate ad click fraud detection

Dr. U. Thirupalu, Professor, Depttment of CSE Audisankara College of Engineering and Technology, Gudur, ummadi70@gmail.com

G.Sreelekha, PG Scholar, Depttment of CSE Audisankara College of Engineering and Technology, Gudur, sreelekhagadi98@gmail.com

Abstract— The internet scam of clicking is a bane to online advertisers whose campaigns data is being fabricated. Attention- based convolutional neural network will form the basis of the paper in order to enhance the deep learning models. It enables this model to nullify the majority of most noteworthy time and behavior features of sequential data of the click streams. They are geared towards identifications accuracy improvements. In cases where the sequential dependency learning and contextual representation are upgraded, the attention-enhanced CNNs would be effective in fraud classification as compared to the normal CNNs. The reason is that this is done to achieve real-time prediction in a web application developed with the Flask based on the proposed model. Dynamic online advertising ecosystems are successful, useful and effective in nature.

Keywords— *Online advertising systems, attention mechanism, convolutional neural network, deep learning, clickstream analysis, real-time prediction, fraud classification, ad click fraud detection*

I. INTRODUCTION

The campaign performance, as well as communication with the users is an important aspect of the online economy, which can be monitored by online advertising in the form of pay-per-click (PPC) approach. The click ad fraud has since taken place due to the increased popularity of PPC advertisement. The bots or malicious people contribute to the creation of fake clicks to raise the cost and performance of advertisement. Ad fraud causes an estimated billions of dollars loss to advertisers and thus the necessity to find it in a more efficient way [1], [5].

Click fraud hurts the revenues, analytics and trust of publishers, advertisers and ad networks. Click fraud reduces ROI and conversion that is translated into bad business decision. There should be fraud deterrence within the online advertising or a web advertising system as fraud eradicates integrity within the business and confidence [2]. The regulatory investigations have revealed that Internet advertisement fraud is increasing at the pace that is too high to track and control [3].

The early measurement and research on online ad transactions had indicated that the fraudulent attempts to make a click were dynamic, spread and hard to differentiate between the user activity. The algorithm of thresholds/rule-based-detection is not applicable in large scale, as the fraud-offenders are capable of breaking the system [4]. Fraud detection solutions based on data are proposed to be inefficient compared to the static ones.

The learning of both complex behavioral patterns in large volumes of clickstream data has been performed using deep learning and machine learning. The algorithms argue the accuracy of identification despite the fact that not all of them can store sequential correlations and give rate considerations to behavioral considerations. The architecture of attention based CNN is utilized to complement the contextual click behavior in order to detect online ad click fraud.

Ensemble, independent deep learning system, taking into account all the input variables and simple machine learning model system are some of the methods that are popular in detecting ad click fraud. This paper introduces a model of attention-based CNN, which is designed in a particular way to pay attention to the contextual awareness and resistance to the frauds in terms of clickstream items. The difference lies in the implementation of Flask based real-time web application in comparison with the other previous works where the monitoring of frauds is not real-time.

II. LITERATURE SURVEY

Communications technology in the future will push digital advertising to 646B with half of the digital advertising being driven by mobile internet. Since click fraud manipulates the content on the advert and results in enormous losses, it impede the progress. The machine learning model employed in this paper is the Random Forest, LightGBM, XGBoost, AdaBoost, Decision Tree, Gradient Boosting and MLP to identify the fraudulent activities linked to online advertisements with a view to clicks. The hard model decision-making has been brilliantly figured out by LIME as transparency. It reveals that the Random Forest model is one of the strong and versatile

models as it is discovered that the accuracy of the model is so high even after the data preparation and the verification of the model through accuracy, precision, recall and F1-score. This approach of the study is desirable to the platform designers and marketers because the interpretability and reliability of such a study is positive. The reason why it remains relevant and methodologically efficient is that it is not inferior compared to the recent publications[6].

The probe is also used to ascertain false clicks in commissioner click statistics. Ad clicks will be in anticipation of bad labeling of the user. Data is used in training systems that are implemented in classifying click fraud. The first one is the stage of generating and retrieving the attributes to crude log data. Thereafter, we will evaluate the classifiers with the view to selecting the models and generate an ensemble model. final evaluation of our model using the data of human judges [7].

Click fraud empties the failure cost of publisher advertising. The trend of clicks based systems of fraud detection is excellent. The properties are used to formulate the new features activated by our click fraud detector algorithm in BuzzCity data. The recall and model correctness are measured. The last type of ensemble model having six learning algorithms was trustworthy on all three measures of performance. The generalizability concept may be indicated by the improved performance of the remaining model during training, test and validation data sets [8].

It needs PR in the marketing of the mobile apps. The killing this ecosystem is the practice of click fraud because the process of clicking the adverts is repeated by malware or bots. The existing methods used to detect click frauds are centered on the server requests. False negative is reduced by distributing the clicks all over the world or by using proxies. AdSherlock is a mobile application company whose premium is named AdSherlock and an application based on customer-side and based on click fraud detector. AdSherlock AdSherlock program is mathematically difficult to determine the clicks request on-line and off-line. AdSherlock is based on URL tokenization which generates off-line reliable and probable patterns. The online click requests are known through the use of ad request tree models to identify the click fraud. AdSherlock prototypes are prototypes that are put into test. Binary instrumentation is used to make the placement of online detector in the program fast. The click fraud detection systems have more false positives compared to AdSherlock. In this article [9], mobile advertising fraud detection and request identification are provided.

Online advertising fraud increases the risks in the marketing, advertisement and e-business. The Internet advertising is vulnerable to the click fraud. Online marketers have better-filtered traffic that requires the best click fraud protection. The Internet advertisement should have a strong system of screening fraud. It is another machine learning model on click fraud that is being tested. And with 200 million click trends we analyzed 4 days. Through the analysis of the portfolio click path of an individual, we may determine the IP addresses that registered significant numbers of click sessions but none of the apps were installed. LightGBM research is carried out by utilizing Gradient Boosting Decision Trees. In this process, the level of accuracy is 98 percent. To justify our findings, the literature review was adopted [10].

The studies in ad click fraud detection have established that the machine learning algorithms of Decision Trees, Random Forest, Gradient Boosting, Support Vectors machine learning and Naive Bayes have been employed to identify the fraudulent click activity. Lightgbm ensemble models and XGBoost models were used to deal with feature interactions and imbalance of data in order to work on the detection accuracy. This is because such strategies are time limited and are based on manual properties.

In the recent past, CNNs, RNNs and hybrid architecture have been able to learn and replicate complex and sequential click behavior. These models are better than the old fashioned classifiers, which do not entail attention strategies to select useful features or to administer real time deployment. This gap led to the attention based CNN methodology that enhances contextual learning process and provides a useful framework to detect ad clicks frauds in the real world.

III. METHODOLOGY

Ad click fraud is required to be detected in terms of model training, feature selection and data preparation. Supervised learning is a type of annotated clickstream data, which includes time, device, browser and user activity records. The identification of category attributes, the normalization of numerical variables and elimination of inconsistent and missing data make the model more stable. Recursive Feature Elimination (RFE) is used to pick out important features to aid in the improvement of identification and dimensional reduction. Eighty percent of the data in processing is received by training and testing sets. The examples of the machine learning models that are taken into consideration based on their accuracy, precision, recall, and F1 are Logistic Regression, Support Vector machine, Decision tree, random Forest, K-Nearest Neighbors, Naive bayes, gradient Boosting, XGBoost, and LightGBM. The information on the online ads are collected by the usage of the web-based click tracking in a real world ad click data. Clicks were analyzed in embedded tracking machines by the saves made. During the process of browsing on the browser and clicking adverts, the information is copied. In this work, the recordings of clicks were 5000 although 20 attributes of the input were used after preprocessing and feature selection. All of them are comprised of OS, IP, device type, browser, temporal, user behavioral and interaction data. On the real life, web advertising is 75 percent and 25 percent real and fake respectively.

Besides machine learning, deep learning establishes complex click patterns. When ANN and DNN are trained in nonlinear features interaction, RNN models are trained in time correlation of click sequences. CNNs have the ability to produce hierarchical features and emphasize on meaningful and contextual input of sequence. CNNs that focus on attention filter noise and base their focus on the patterns to detect frauds. The Flask-based model is used to determine the fraud through prediction. Dynamic online advertising can also be demonstrated to be scalable, user-friendly, and effective.

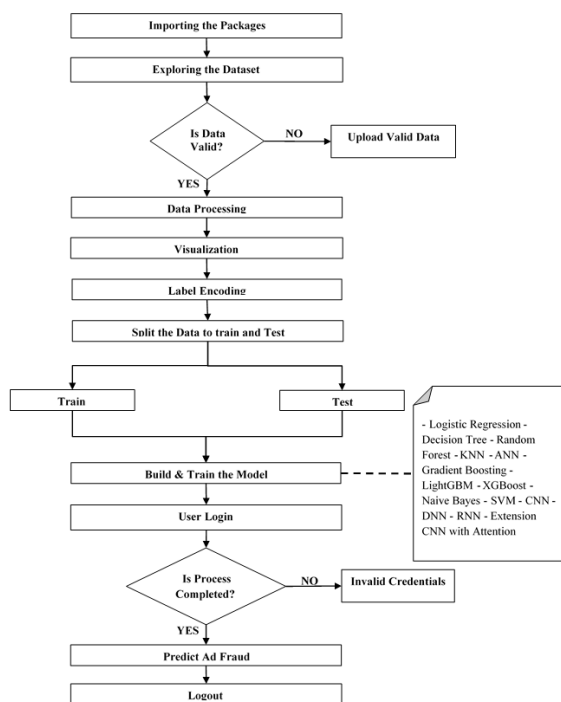


Fig1 work flow

A) Proposed System

The solution to be proposed incorporates machine learning and deep learning with regards to detecting the fraud ad clicks. It processes tagged click stream data comprising of temporal, device, browser, OS and user behavior. Categorical features labeling, dropping values that are missing, numerical data normalization, and Recursive Feature Elimination (RFE) to select the most important features are all part of data preparation. Machine learning classifiers such as the Logistic regression, the Support Vector Machine, the Decision Tree, the Random Forest, the K-Nearest Neighbors, the Naive Bayes, the Gradient Boosting, the XGBoost and the LightGBM classify the clicks as genuine or fake based on the persistence of the behavior patterns.

The system encourages the interactions of the complex and nonlinear, sequential click interactions with ANN, DNN, CNN, and the RNN deep learning models to promote their detection. The CNN is also contextualistically oriented to other practicalities that will assist it in establishing a better categorization of frauds. To train the model to identify fraud in real time, click information of the user can be sent to a Flask web app that trains the model. This is a flexible and expansive structure applied in the online ad in reality to counter developments in fraudulent methods.

The news article on CNN with attention is the major contribution of the work to detect ad click frauds. The suggested CNN structure interprets and dynamically locates time and behavioral features of the clickstream information in perspective as opposed to the conventional CNN models. It has boosted the sequential pattern detection and fraud detection, contextual learning.

B) System Architecture

The architecture shown below includes browsing and clicking of a user on a web site with the help of a piece of

JavaScript. It is a script that will be used to record clicks, time stamp, browser communication and sequence of navigation. Veracity Trust Network is the one where this information is not lost as it is encrypted. The raw data shall be analyzed as encrypted nested click data which will be in form of JSON files. The json files are initially compressed then converted to structured CSV so as to apply machine learning and preprocessing.

The module relies on random undersamples as a method to counteract imbalance and missing values as well as encode categories after data preparation. The behavioral and temporal user interaction information is extracted with the help of the feature engineering. Continuous data is discretized into discrete intervals which enhance model and learning interpretation. The trained features guide the deep learning and machine learning approaches of identifying the click fraud. Internet advertisement that is characterized by fraudulent clicks of advertisement are properly and suitably identified through integration of real time data collection and feature appraisals.

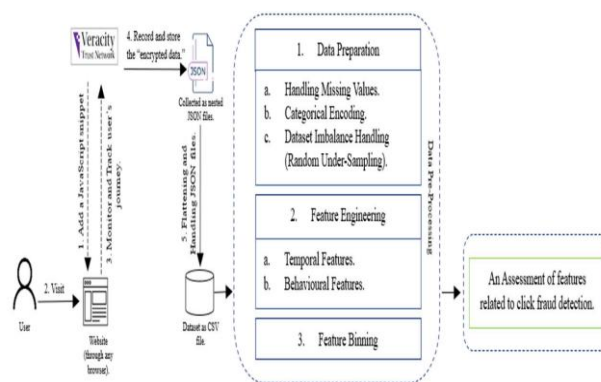


Fig 2. Architecture

C) MODULES

1. Importing the Packages:

It is a collection of packages required to import needed python modules to evaluate, visualize, machine, deep learning, preprocess and manage data. It realizes the computing environment and is an assurance that all the requirements frameworks and features will be present in the later stages of the systems.

2. Exploring the Dataset:

This module assists in loading and performing a simple initial research to become familiar with the format, characteristics of features and data distribution of the ad click fraud. It helps in the detection of trends, irregularities and abnormalities before the pre-processing and model development.

3. Data Processing:

The Data Processing will entail the handling of the missing values, correcting inconsistencies and normalizing of the numerical variables on top of preparing the information prior to the modeling. In the case of machine learning and deep learning algorithms, it ensures that the information that is being gathered is of quality.

4. Visualization:

This module will graphically represent the trend of clicks, the nature of browsers, operating system, and device to use and clicking tendency. The visual clues may be applied in the detection of click fraud, pattern recognition and examine user activity.

5. Label Encoding:

Here in this module, information that is stored in one category is transformed to a similar value using label encoding. This enables machine learning models and deep learning models to be in a position to handle the form of browser, form of operating system and device.

6. Data Splitting (Train-Test):

This module data operations are performed aiming at splitting the data in general, 80/20 to training and tests. The testing set is utilized to test the performance and the training set is utilized to produce models which will provide reasonable and good checking.

7. Model Training:

The module trains numerous machine learning models like Logistic Regression, Decision Tree, random forest, KNN, Naive Bayes, SVM, gradient boosting, LightGBM, XGBoost, ANN, CNN, DNN and RNN. The two models are learnt to learn the trends of clicking behavior in order to differentiate on legitimate and invalid clicks.

8. Model Evaluation:

The module 8 uses the performance measures of accuracy, precision, recall and F1-score, to test trained models. The measures can be utilized to compare models and choice algorithms in fake ad click identification.

9. Flask Server:

It is a Flask based server on which the developed fraud detection model will be applied. Any person can make predictions on the real time web-based interface.

10. User Login:

User Login takes care of the user authentication security. To achieve this, only the authorized users will be allowed to access the system that will enable them to read the prediction results and post data to ensure security.

11. Predict Ad Fraud:

Click data files can be inputted into the internet interface. The system compares the information in order to determine the authenticity of the clicks and implements the acquired model.

12. Logout:

User sessions are safe shut down. It eliminates post-session access that is not wanted.

D) ALGORITHMS

a) Deep Neural Network (DNN):

DNN is a multilayer feedforward neural network that is capable of representing complex, nonlinear interactions. It examines the time when the user clicks, user and device information behavior in attempts to identify fraud. DNNs can detect abnormalities with the assistance of hierarchical

representations and can be effectively used with extensive data sets.

b) Recurrent Neural Network (RNN):

It is a type of neural network that recalls the temporal relationship. The sequential click activity is modeled and the pattern abnormalities are detected on time basis. RNNs can be taken advantage of in fraud detection based on time and sequence of clicks.

c) CNN with Attention Mechanism:

The CNN with attention is an attention-enhanced CNN that involves the use of the attention layer to add convolutional feature extraction with a layer of attention to obtain the true features and time periods. The method enhances sequential pattern matching and click stream contextual understanding. Attention-based CNN can effectively be used in detecting fraud as compared to CNN and machine learning models due to the concentration on the relevant features and noise reduction.

IV. EXPERIMENTAL RESULTS

Different machine learning and deep learning algorithms of identifying fraudulent ad clicks used a labeled dataset of ad clicks to test. The data has been preprocessed to fill the gaps in the data, code the categorical variables, standardize the numeric ones, as well as to select the features according to Recursive Feature Elimination. The processed data were separated into 80:20 training and testing parts. The ensemble based models proved to be superior to the ordinary classifiers that have the capability of modeling the intricate interactions amongst the characteristics and that are more effective in dealing with the data unbalance. Logistic Regression, Decision Tree, Random Forest, KNN, Naive Bayes, Support Vector machine, Gradient Boosting, XGBoost and LightGBM were all the models that were trained and assessed using the most popular performance measures such as accuracy, precision, recall and F1.

Random Forest, Decision Tree, Gradient Boosting, and XGBoost have above 98% in terms of detection of fraudulent click pattern. Neural Network that was recreated repeatedly made 97 percent hits on the next action of clicking compared to deep learning models. The Convolutional Neural Network based on attention was the most accurate with a figure of 99.73 percent out of the rest which is indicative of how sequential pattern recognition and emphasis on contextual features depend on the attention processes. These results support the fact that feature selection and attention-based deep learning are useful in improving fraud detection, false positive and convenient and scalable online advertisement tool.

The CNN based on attention has an accuracy of 99.73 due to sufficient data preprocessing, attention selection and attention mechanism whereby attention is given to discriminative behavioral and temporal features. The CNN based hierarchical feature extraction is better than generalization and fraud detection although the attention layer is focused on trendy click patterns to remove noise.

Accuracy: To find out the reliability of the test, compare the positive and negative results. A mathematical explanation:

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: Precision gives an indication of positive instances or categorization perfection. The precision is calculated through the following:

$$Precision = \frac{TP}{(TP + FP)}$$

Recall: The proportion of the correct predictions made in the positive cases divided by the total positive cases is the ability of a model that identifies all instances of machine learning class.

$$Recall = \frac{TP}{(FN + TP)}$$

F1-Score: High F1 value indicates that there is a correct machine learning model. Precision and recall are to be used together to increase the accuracy of models. Accuracy measures the prediction efficiency of the model on a collection of data.

$$F1 = 2 \cdot \frac{(Recall \cdot Precision)}{(Recall + Precision)}$$

Algorithm Name	Accuracy (%)	Precision (%)	Recall (%)	F-Score (%)
Logistic Regression	97.267	96.573	96.343	96.457
Decision Tree	98.933	98.044	99.277	98.639
Random Forest	98.333	97.010	98.888	97.939
KNN	73.200	48.649	49.902	43.215
ANN	67.267	53.533	52.713	52.569
Gradient Boosting	98.000	96.454	98.645	97.475
LightGBM	98.533	97.349	99.006	98.137
XGBoost	98.667	97.579	99.097	98.304
Naïve Bayes	89.333	85.492	92.527	87.606
SVM	73.800	36.900	50.000	42.463
CNN	95.667	93.212	96.326	94.599
DNN	73.800	36.900	50.000	42.463
RNN	98.733	97.625	98.978	98.382
CNN with Attention (Proposed)	99.7	99.61	99.87	99.74

Table 1: Performance Comparison table

The attention-based CNN represents the proposed model and achieves the highest performance among all evaluated algorithms.

V. CONCLUSION

This paper has been developed to present a deep learning-based and machine learning model of detecting ad clicks fraud. To compare a set of classifiers, a labelled set of click stream data was processed and features were selected with recursive feature elimination. Machine learning models in Ensemble that comprised of Decision Tree, Random Forest, Gradient Boosting, LightGBM and XGBoost were found to recognize with a score of more than 98 percent following the tests but deep learning models had developed complex and consecutive patterns of click actions.

The introduction of attention mechanism in the Convolutional Neural Network boosted the accuracy of the contextual feature learning and sequential pattern recognition to 99.73 percent. The trained model was also described as usable and scalable since it was implemented as a Flask-based web application to make prediction of real-time fraud with the use of the trained model. The results portray that attention-based deep learning which is preprocessed by strict preprocessing could be useful in identifying fake clicks on advertisements in dynamic contexts of online advertisement.

REFERENCES

- [1] Juniper Research, Hampshire, U.K. Quantifying the Cost of Ad Fraud: 2023–2028. Accessed: Jul. 12, 2024. [Online]. Available: https://fraudblocker.com/wp-content/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.

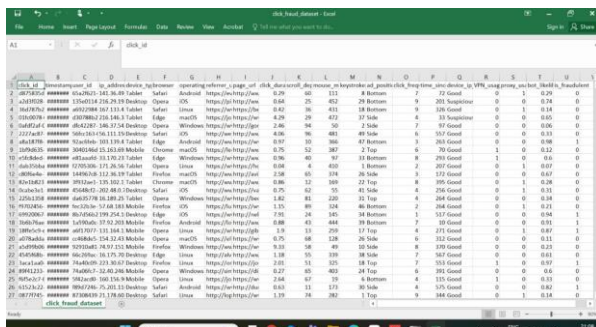


Fig3 dataset

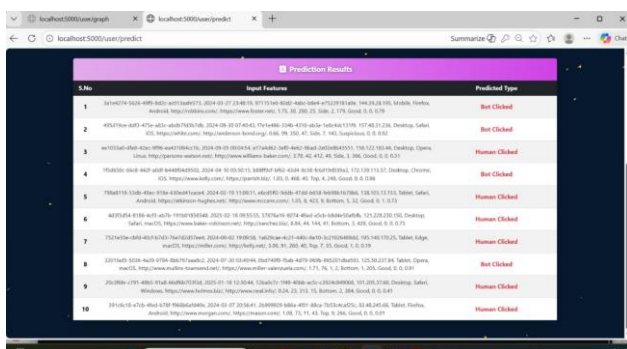


Figure 4: Prediction Results

The figure shows the results of predicting the click data that were submitted. The fact that the system can determine the fraudulent clicks can be evidenced by the fact that there is a label to each record which shows whether it was Human Clicked or Bot Clicked.

- [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/wp-content/uploads/2023/09/Lunio_Wasted_Ad_Spend_Report_2024_V2.pdf
- [6] D. Berrar, "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
- [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.
- [9] C. Phua, E.-Y. Cheu, G.-E. Yap, K. Sim, and M.-N. Nguyen, "Feature engineering for click fraud detection," in *Proc. Work. Fraud Detect. Mob. Advert.*, 2012, pp. 1–10. [Online]. Available: <http://palanteer.sis.smu.edu.sg/fdma2012/doc/FirstWinner-Starrystarrynight-Paper.pdf%5Cnpapers2://publication/uuid/9290A6CF-A861-4058-99F4-D39706B0619A>
- [10] E.-A. Minastireanu and G. Mesnita, "Light GBM machine learning algorithm to online click fraud detection," *J. Inf. Assurance Cybersec.*, vol. 2019, pp. 1–12, Apr. 2019, doi: 10.5171/2019.263928.