International Core Journal of Engineering Volume 7 Issue 11, 2021
ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0043

An Advertising Fraud Detection Method based on Automatic Feature Mining of Time Series Data

Shuo Wang, Kang Yang and Tongle Wang

Mininglamp Technology, Shanghai 200000, China.

Abstract

With the continuous development of the mobile Internet, the digital advertising marketing market is also showing a trend of vigorous development. The market scale is increasing year by year, and the subsequent advertising fraud problem has gradually become a stubborn disease in the digital advertising marketing market. The traditional rule-based method does not consider the update speed, complexity and randomness of advertising fraud, while the method based on machine learning often ignores the importance of data form and the timing dependence of fraudulent user behavior. In order to solve the above problems, this paper proposes an advertising fraud detection method based on time series data automatic feature mining. First of all, considering the importance of the data form, we constructed the user behavior time series data based on the original data, and then used the improved LSTM-based autoencoder to extract high-level features for the automatic mining of user fraud characteristics, taking into account the timing dependence of fraudulent user behavior It uses multi-scale convolutional neural network to extract local salient features of different scales. In order to avoid information loss, the residual network idea is used to combine high-level features and local salient features as extracted user fraud features. Finally, a fully connected layer is used to merge Use softmax to map user fraud characteristics to whether a user is fraudulent. This article conducts experiments on real advertising traffic data. The experimental results show that the method in this article has effectively improved the F1 value.

Keywords

Advertising anti-fraud; Automatic Feature Mining; Stacked Autoencoder; Multi-CNN.

1. Introduction

With the continuous development of the mobile Internet, the digital advertising marketing market is also showing a trend of vigorous development. The market scale is increasing year by year, and the subsequent advertising fraud problem has gradually become a stubborn disease in the digital advertising marketing market. Advertising fraud refers to a series of illegal acts carried out by advertisers, advertising agencies, advertising operators and advertising publishers in order to profit themselves or harm others in the process of digital advertising. The advertising traffic generated by these behaviors cannot bring expected benefits to advertisers, but will consume advertisers' advertising costs, bring huge economic losses to advertisers, affect the image of advertising media platforms, and destroy the digital advertising marketing market. Therefore, how to conduct advertising anti-fraud is very important for advertisers and the digital advertising marketing market. In response to the problem of advertising fraud, some scholars have conducted related research. Traditional methods are often based on expert rules, using expert experience to find out unreasonable invalid traffic behaviors, and define them as filtering rules, so as to achieve real-time monitoring and filtering of traffic. For example: Metwally et al. (2005) proposed a network advertising anomaly

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0043

mining system based on association rules, which uses association rules to mine the relationship between the features in anomalous data to detect fraud; Metwally et al. (2007) in order to further explore fraudulent gangs , Proposed a similarity-based search algorithm, by finding all the largest cliques in the graph, the algorithm was extended to detect fraudulent alliances of any size. Aiming at anomalous clicks on ads, Zhang et al. (2008) and Antoniou et al. (2011) analyze the number of user visits within a specific time interval to detect repetitive click anomalies; Kitts et al. (2015) use multiple rules Filters and filtering strategies to detect click fraud; Faou et al. (2016) use social network analysis to detect fraudulent malicious adware; In recent years, with the continuous development of artificial intelligence technology and its application in various fields , More and more researchers are focusing on the research of artificial intelligence-based abnormal traffic detection methods, and begin to use machine learning algorithms to fit more complex relationships to build advertising anti-fraud models to improve fraudulent traffic detection Performance. Mouawi et al. (2018) et al. constructed five types of abnormal traffic characteristics from a business perspective, such as click duration, click frequency, the ratio of unique IP in the number of clicks, the download ratio of each application, and the time difference in the number of ad clicks. , By constructing the above features to improve the recognition performance of the machine learning model; in response to the problem of unbalanced data distribution in traffic anomaly detection, Thejas et al. (2019a) proposed a hybrid deep learning model, which is composed of an autoencoder, It is composed of neural network and semi-supervised generative confrontation network, which is used to realize the anomaly detection of ad cheating traffic in the predictive imbalanced data set. Taking into account the possible periodic characteristics of abnormal traffic data, Thejas et al. (2019b) established a multi-time-scale periodic model to predict advertising click behavior in minutes and hours, and adopted the Akaike Information Criterion (Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) fine-tune the model. Taking into account the time series of user advertisement clicks, Thejas et al. (2021) constructed an anomaly detection framework CFXGB for advertising cheating traffic through fusion rules and supervised learning algorithms, and extracted mixed features under multi-granular time windows, and trained user classification The filter indirectly filters cheaters and complex invalid traffic. Gohil and Meniya (2020) constructed the AdDetect platform, which uses semantic analysis and machine learning to automatically detect the semantics of the in-app advertising library, and recognizes and restores the main and non-main modules of the application based on the hierarchical clustering module. To form a semantic feature vector and input it into the support vector machine for anomaly detection.

The above methods can solve the problem of ad fraud to a certain extent, but there are still the following problems that need to be solved urgently. First, how to construct a traffic data format is essential for fraud detection, but existing methods often focus on model construction. It ignores the importance of data-side construction. The second is automatic feature mining. The accuracy of user fraud features is very important for the detection of fraudulent users. How to automatically mine user fraud features is also one of the issues that need to be studied at present. The third is that fraudulent users have timing dependence in their behaviors, and judging whether a user is a fraudulent user depends on the user's behavior over a period of time.

In order to solve the above-mentioned challenges, this paper proposes an advertising fraud detection method based on automatic feature mining of time series data to identify fraudulent users in mobile advertising, and shows the performance of fraudulent user detection in real mobile advertising data sets. This method first constructs time series data based on user behavior from the original traffic data, and then proposes an autoencoder based on the time series data of user behavior to automatically extract high-level features, and continuously improves the extracted user fraud features through self-supervised learning. Then, a Multiscale Convolutional Neural Network (Multi-CNN) is used to capture the local prominent features of user behavior in different time intervals. In order to prevent the loss of user fraud information caused by Multi-CNN, we use the idea of residual network to splice high-level features and local prominent features as the final extracted user fraud features, and finally

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0043

use a fully connected layer and use softmax to map user fraud features to Whether the user is a fraudulent user.

The contributions in the method proposed in this paper are summarized as follows:

1) Considering the importance of the form of advertising data, this article constructs time series data based on user behavior. The time series data is based on advertising information, media information, user equipment information, user location information, user equipment operating system information, etc. Improve the data representation on the data side to improve the performance of fraudulent user detection.

2) In order to automatically mine user fraud characteristics, this paper proposes the use of autoencoders to automatically construct high-level features, continuously improve the representation of user fraud characteristics through self-supervised learning, and automatically mine features useful for fraudulent user detection.

3) Considering the timing dependence of fraudulent user behavior, a method using Multi-CNN to capture the local prominent features of traffic in different time intervals is proposed.

We evaluated our method on a dataset of real scenarios. Experimental results show that this method can effectively detect fraudulent users, and is better than the current popular machine learning methods in F1 value.

2. Proprosed Approach

The model framework proposed in this paper is shown in Figure 1. First, construct the time series data based on user behavior, and then input the time series data based on user behavior into the autoencoder layer to automatically mine the high-level features of user fraud. The hierarchical features are input to the Multi-CNN layer, and the local prominent features in different time intervals are extracted, and the timing dependence of user fraud in different time intervals is considered. Finally, in the output layer, in order to prevent the loss of user fraud information caused by the Multi-CNN layer, drawing on the idea of residual network, the high-level features and local prominent features output by the autoencoder layer and the Multi-CNN layer are spliced together as user fraud features Indicates that after a fully connected layer with a Softmax layer, the user's fraud characteristics are mapped to whether it is a fraudulent user.

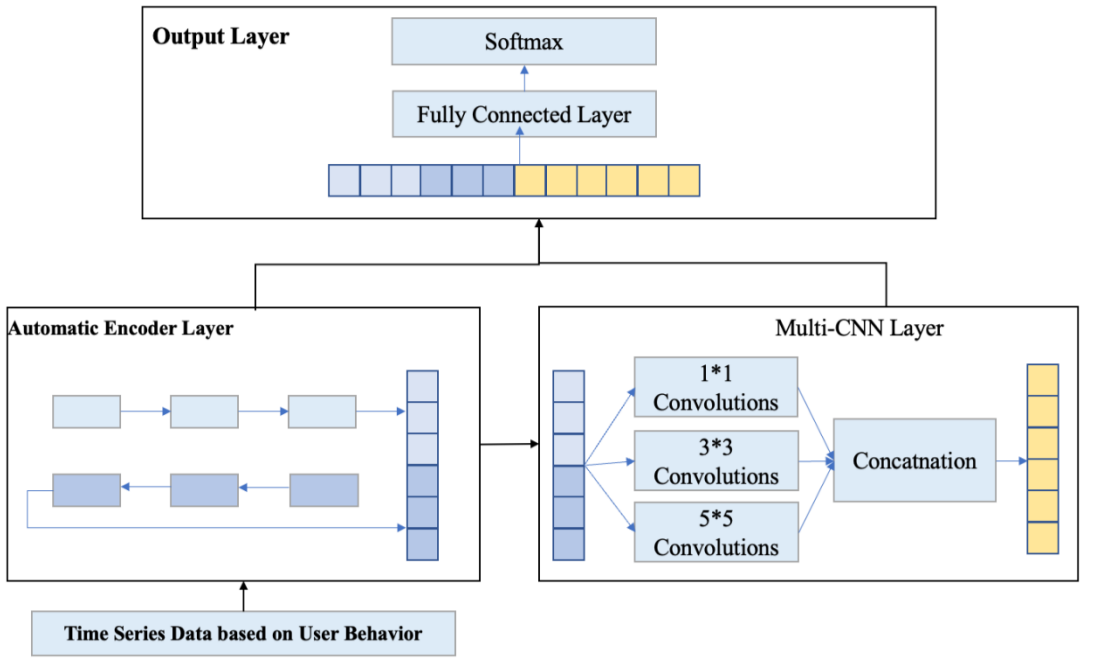

Figure 1 Model Framework

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0043

2.1 Time Series Data Construction based on User Behavior

We use 24 hours as a unit to detect fraudulent users. When constructing time series data, we sample once every hour, and there are 24 sampling points for each statistical feature of each user. The statistical characteristics involve each user's exposure, clicks, IP address, activities, equipment, operating system, media, region, advertiser, agency, brand, etc.

2.2 Autoencoder Layer

The quality of feature construction is critical to the detection of fraudulent users. In order to automatically mine user fraud characteristics, this paper proposes to use an autoencoder to construct high-level features, and the autoencoder uses a two-way improved LSTM to improve the ability of feature extraction. The improved LSTM integrates the Transformer for feature extraction. First, the sequence information is extracted through the transformer for multi-dimensional feature extraction, and then the LSTM gate processing mechanism is used to fit and establish long-term dependencies, thereby realizing long-sequence information modeling. The LSTM structure of the fusion transformer is shown in Figure 2.

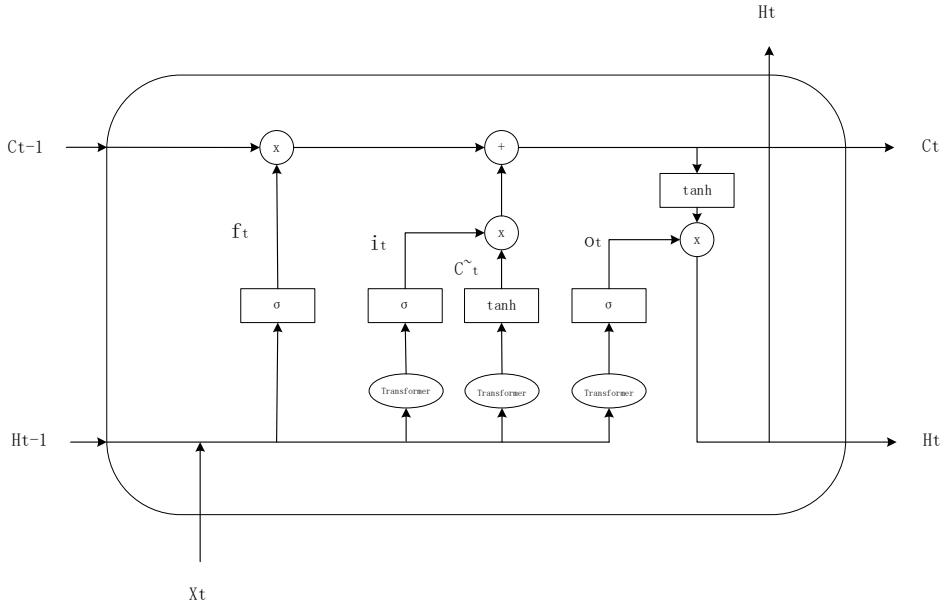

Figure 2 Improved LSTM with transformer

The output information Ht-1 at the previous time, the cell state information Ct-1, and the current input Xt are used as the input of the current unit. And through the forget gate to selectively discard information, the process is shown in formula (1).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$

Input the data into the input gate, use the transformer for feature extraction, and then input the extracted features into Sigmoid to determine which values to update. At the same time, use tanh to create a candidate vector, multiply the two outputs, and finally merge the information into the unit state information , The process is as formula (2).

$$\begin{aligned} i_t &= \sigma(W_i \cdot Transformer([h_{t-1}, x_t] + b_i)) \\ \tilde{C}_t &= \tanh(W_c \cdot Transformer([h_{t-1}, x_t] + b_c)) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \end{aligned} \tag{2}$$

Input the data to the output gate of LSTM, extract the features through the transformer, and input it to Sigmoid for non-linear fitting. At the same time, the unit state information is input to the tanh layer

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0043

for fitting, and finally the two outputs are multiplied as the output of the unit. The process is as formula (3).

$$o_t = \sigma(W_o \cdot Transformer([h_{t-1}, x_t]) + b_o)$$
$$h_t = o_t * \tanh(C_t) \quad (3)$$

Utilizing the powerful feature extraction capabilities of transformer to overcome LSTM's shortcomings in data feature extraction, the model can not only obtain the long-term dependence of sequence information, but also extract the features of the sequence well, so that the performance of the model can be further improved.

2.3 Multi-CNN Layer

In the task of advertising fraud detection, there is a local dependence on the time series data of user behavior in different time intervals. In order to improve the performance of fraudulent user detection, we uses Multi-CNN with a scale of 1*1, 3*3, and 5*5 convolution kernels. Extract local prominent features.

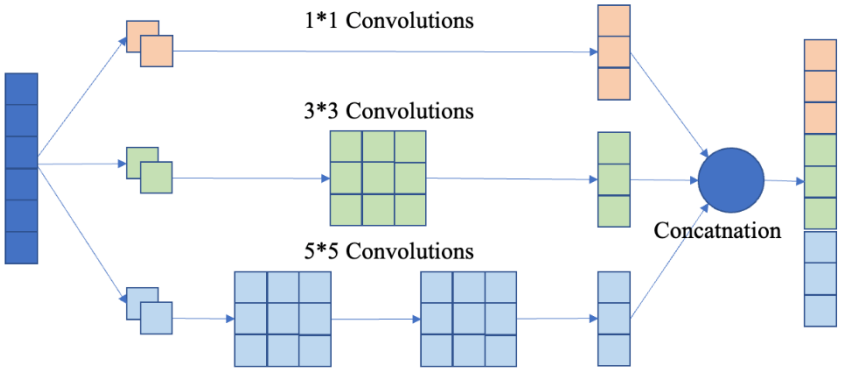

Figure 3 Multi-CNN Structure

In this paper, multiple filters are used for feature extraction under different scale convolution kernels to capture more abundant local features at different scales. The convolution operation can be expressed as formula (4).

$$c_{ij} = f\left(w_j \bullet x_{i:i+h-1} + b_j\right) \quad (4)$$

Among them, j, b $\in$ R, wj represents the j-th filter, b is a bias term, f represents a non-linear function, and h represents the size of the convolution kernel.

In our approach, ReLU is selected as the activation function in the convolution operation. The calculation speed of ReLU is very superior compared to activation functions such as sigmoid and tanh. The form of ReLU is as formula (5).

$$g(x) = \max(0, x) \quad (5)$$

2.4 Output Layer

In order to prevent the loss of high-level features by the Multi-CNN layer, the idea of residual network is borrowed from the output layer and the high-level features and local prominent features output by the autoencoder layer and the Multi-CNN layer are spliced together as the user fraud feature representation, and then A fully connected layer with a sigmoid function is used to map user fraud characteristics to the probability of fraudulent users. In the training process, the cross-entropy loss function is used to update. If the user fraud feature is represented as x, the calculation formula and loss function are as formula (6-7):

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \quad (6)$$

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0043

$$loss = -y_i \log(\ h_\theta(x)) - (1 - y_i)\log(1 - h_\theta(x)) \quad (7)$$

3. Experiments

3.1 Dataset

This article uses real-world advertising facts and returned digital features as the original advertising traffic data. The data set includes one month of real data, and about 3.5 billion traffic is generated every day, of which abnormal traffic is about 0.1%-5%. We randomly select 7 days of full data, apply the above time series data construction method to the daily raw data, generate time series data level (Real Scene Series Dataset, RSSD), and randomly select 5 million pieces of data for the training and training of this method. Verification, the ratio of training set to test set is 7:3.

3.2 Evaluation criteria

In the performance evaluation and comparison experiment of fraudulent traffic detection in this paper, precision, recall, and F1 value are used as evaluation indicators to perform performance evaluation among different models. Assuming that TP, FP, TN, and FN are the number of normal samples detected correctly, the number of normal samples detected incorrectly, the number of abnormal samples detected correctly, and the number of abnormal samples detected incorrectly, the relevant evaluation indicators are calculated as follows:

$$Precision = \frac{TP}{TP+FP}$$
$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (8)$$
$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

3.3 Experimental Result

In order to demonstrate the effectiveness of the method in this article, the following method is used as a baseline, and comparative experiments are carried out on the RSSD data set:

1) LightGBM model, Minastireanu et al. (2019) proposed to detect click fraud based on LightGBM.

2) The CFXGB model, the CFXGB detection framework proposed by Thejas et al. (2021), integrates rules and supervised learning algorithms, and trains a classifier to perform fraud detection through mixed features extracted under a multi-granular time window.

3) XGBoost, XGBoost is an improvement of the gradient boosting algorithm, and it has excellent performance in various competitions and on various tasks.

Table 1 Comparison table of model experiment results

Model	Precision(%)	Recall(%)	F1(%)
LightGBM	92.12	91.86	91.99
CFXGB	95.82	93.44	94.62
XgBoost	91.15	90.98	91.06
Ours	97.83	95.91	96.86
Improve(%)	**2.01**	**2.47**	**2.24**

The results of the comparative experiment are shown in Table 1. The accuracy, recall, and F1 values of this method are 97.83%, 95.91%, and 96.86%, respectively. They perform best on the RSSD data set, and compare with LightGBM, CFXGB, and XgBoost models. In comparison, F1 value gains of 4.87%, 2.24%, and 5.80% were obtained respectively, reaching the current best performance. The

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0043

values of Precision, Recall, and F1 are respectively 2.01%, 2.47%, and 2.24% higher than the current optimal model. It can be seen that the advertising fraud detection method based on time series data automatic feature mining has better performance.

In addition, in order to measure the effectiveness of the improved parts of each module of the method in this paper, a comparative experiment of the improved parts of each module is carried out. Autoencoder (LSTM) is an experiment that uses basic LSTM for automatic encoding, and other settings remain unchanged. CNN (With 1*1 Convolutions) replaced the model's Multi-CNN with a 1*1 convolution kernel CNN. Through the above two methods to verify the performance of the autoencoder layer and Multi-CNN layer.

Table 2 Comparison experiment of each module improvement part

Model settings	Precision(%)	Recall(%)	F1(%)
Autoencoder(LSTM)	96.25	95.00	95.62
CNN (With 1*1 Convolutions)	95.72	94.02	94.86
Complete Model(Ours)	97.83	95.91	96.86

The comparative experimental results of the improved parts of each module are shown in Table 2. It is not difficult to find that using the LSTM integrated with the transformer as the encoder unit increases the accuracy, recall, and F1 values by 1.58%, 0.91%, and 1.24, respectively. %. Using Multi-CNN to capture n-gram information in different time intervals improves accuracy, recall, and F1 by 2.11%, 1.89%, and 2.00%, respectively, compared with a single-scale CNN using a 1*1 convolution kernel. The experimental results show that the improvement effect in the Autoencoder layer and Multi-CNN layer is very significant.

4. Conclusion

This paper mainly studies the problem of ad fraud in mobile advertising, and detects fraudulent users by designing an algorithm model. We propose an advertising fraud detection method based on automatic feature mining of time series data to identify fraudulent users in mobile advertising fraud. First, build time series data based on user behavior based on the original data of mobile advertising, and then input the time series data to the improved autoencoder layer to automatically extract high-level features, and then use the Multi-CNN layer to capture local prominent features at different scales, and build Model the dependence characteristics of user fraud characteristics at different time intervals. In the output layer, the idea of residual network is borrowed, and high-level features and local prominent features are spliced together, and then input to the fully connected layer to detect fraudulent users. Finally, a comparative experiment was carried out on mobile advertising data in real scenes. The experimental results show that the performance of this method to detect fraudulent users is better than the existing fraud detection methods.

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 Symposium on Computers and Communications (ISCC) (pp. 1111-1116). IEEE.

[2] Faou, M., Lemay, A., Décary-Hétu, D., Calvet, J., Labrèche, F., Jean, M., Dupont, B. & Fernande, J. M. (2016). Follow the traffic: Stopping click fraud by disrupting the value chain. In 2016 14th Annual Conference on Privacy, Security and Trust (PST) (pp. 464-476). IEEE.

[3] Gohil, N., & Meniya, A. D. (2020). A survey on online advertising and click fraud detection. In 2nd National Conference On Research Trends in Information and Communication Technology.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202111_7(11).0043

[4] Kitts, B., Zhang, J., Wu, G., Brandi, W., Beasley, J., Morrill, K., Ettedgui, J. Siddhartha, S., Yuan, H., Gao, F., Azo, P., & Mahato, R. (2015). Click fraud detection: Adversarial pattern recognition over 5 years at Microsoft. In Real World Data Mining Applications (pp. 181-201). Springer, Cham.

[5] Kitts, B. J., Najm, T., & Burdick, B. (2008). Identifying automated click fraud programs. U.S. Patent Application No. 11/745, 264.

[6] Mouawi, R., Awad, M., Chehab, A., Hajj, I. H. E., & Kayssi, A. (2018). Towards a machine learning approach for detecting click fraud in mobile advertizing. In 2018 International Conference on Innovations in Information Technology (IIT) (pp. 88-92). IEEE.

[7] Metwally, A., Agrawal, D., and Abbadi, A. E. (2005). Using association rules for fraud detection in web advertising networks. In Prceedings of the 31st International Conference on Very Large Data Bases (pp. 169–180). VLDB Endowment.

[8] Metwally, A., Agrawal, D., El Abbadi, A. (2007). Detectives: Detecting coalition hit inflation attacks in advertising networks streams. In Proceedings of the 16th International Conference on World Wide Web, (pp. 241-250). ACM.

[9] Minastireanu, E. A., & Mesnita, G. (2019). Light gbm machine learning algorithm to online click fraud detection. J. Inform. Assur. Cybersecur, 2019.

[10] Thejas, G. S., Boroojeni, K. G., Chandna, K., Bhatia, I., Iyengar, S. S., & Sunitha, N. R. (2019a). Deep learning-based model to fight against ad click fraud. In Proceedings of the 2019 ACM Southeast Conference (ACM SE '19) (pp. 176-181). Association for Computing Machinery.

[11] Thejas, G. S., Dheeshjith, S., Iyengar, S. S., Sunitha, N. R., & Badrinath, P. (2021). A hybrid and effective learning approach for click fraud detection. Machine Learning with Applications, 3, 100016.

[12] Thejas, G. S., Soni, J., Boroojeni, K. G., Iyengar, S. S., Srivastava, K., Badrinath, P., Sunitha, N. R., Prabakar, N., & Upadhyay, H. (2019b). A multi-time-scale time series analysis for click fraud forecasting using binary labeled imbalanced dataset. In 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS) (Vol. 4, pp. 1-8). IEEE.

[13] Zhang, L., & Guan, Y. (2008). Detecting click fraud in pay-per-click streams of online advertising networks. In 2008 The 28th International Conference on Distributed Computing Systems (pp. 77-84). IEEE.