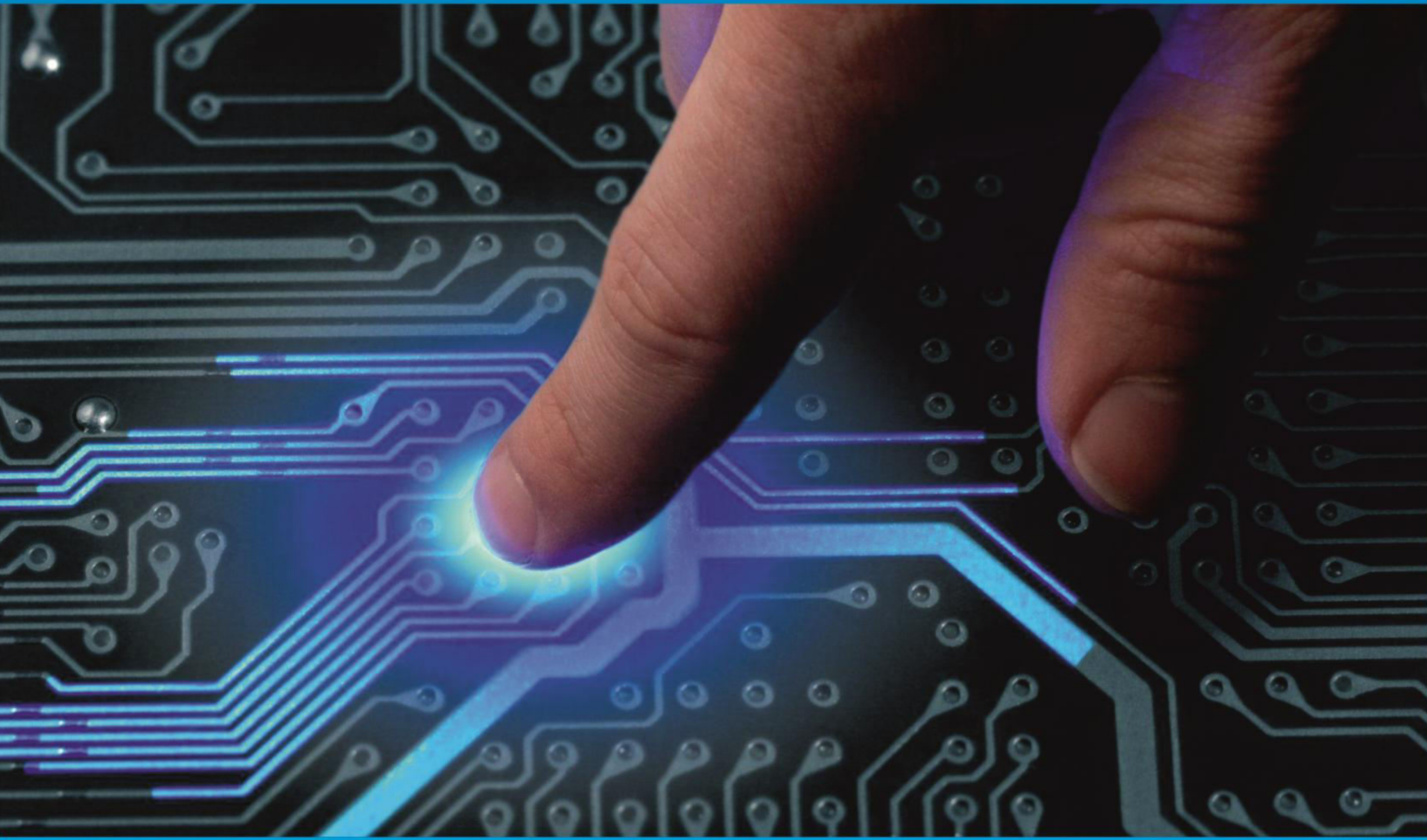
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 4, April 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.488

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904039 |

Detection of Fraud in Mobile Advertising by Using Machine Learning

Dr.I.Poonguzhali[1], Raja Keerthivasan[2*], T.Saidinesh[3], K.Manoj[4]

Associate Professor, Department of Electronics and Communication Engineering, Panimalar Institute of Technology, Chennai, India [1]

Student, Department of Electronics and Communication Engineering, Panimalar Institute of Technology, Chennai, India [2*, 3, 4]

ABSTRACT: With ongoing advancements in the field of technology, mobile advertising has emerged as a platform for publishers to earn profit from their free applications. An online attack commonly known as click fraud or ad fraud has added up to the issue of concerns surfacing mobile advertising. Click fraud is the act of generating illegitimate clicks or data events in order to earn illegal income. Generally, click frauds are generated by infusing the genuine code with some illegitimate bot, which clicks on the ad acting as a potential customer. These click frauds are usually planted by the advertisers or the advertising company so that the number of clicks on the ad increases which will give them the ability to charge the publishers with a hefty sum per number of clicks. A number of studies have determined the risks that click fraud poses to mobile advertising and a few solutions have been proposed to detect click frauds. The solution proposed in this paper comprises of a social network analysis model – to detect and categorize fraudulent clicks. This social network analysis model takes into consideration a wide range of parameters from a large group of users around the world. In this work, we study& analyse the fraud detection. In this work, we analyse the performance of machine learning classification methods, and classify as "is attributed" or "not attributed". A machine learning model like XGBoost, Lightlgm, multiple encoding methods are applied for the prediction process. The complete implementation is done through Google Colab (Python-Jupyter Notebook).

KEYWORDS: Mobile advertising, Machine learnimg, Google colab, Fraud

I. INTRODUCTION

Recently, mobile advertising has evolved expeditiously as it provides publishers a platform to expand their audience reach by putting their advertisements in mobile applications. Statistically, mobile in-app advertising is estimated to surpass a revenue limit of approximately

$17 billion by the end of 2020. The mobile advertising industry is bilateral as on one hand, it helps developers boost app monetization and on the other hand, it expands the install horizons to which an application can reach. It allows advertisers to take their product out to new audiences and app developers to escalate their application reach to new markets.

Another common name for mobile advertising is in-app advertising. This in-app advertising comprises of four major components, namely: 1) The advertiser, 2) The user, 3) The publisher, and

4) The ad network. A user, in mobile advertising is one who views the ad. The owner of the product which is being advertised is considered to be the advertiser whereas the person to whom the application, in which the advertisement is advertised, belongs is the publisher. Furthermore, the third party who links advertisers to publishers is called the ad network. These ad networks aim at generating as a percentage of publisher's profit.

Click fraud is a type of fraud, which puts the Cost per Click model in jeopardy. Certain fraudulent sources came up with the idea of generating illegitimate clicks on the advertisement, which encouraged unethical groups to hire these sources to increase the amount of user action on their advertisement and generate money from it. A click fraud occurs when a bot, a computer code, an automated script or a person, pretending to be genuine user, generates a random number of clicks on an advertisement without any legitimate interest in it.

II. EXISTING AND RELATED WORK

In this section, we firstly provide a description of CSBPNN architecture for click fraud detection. Secondly, we use ABC to optimize BPNN connection weights and feature selection synchronously. Thirdly, the error function is corrected by adding cost parameters to BPNN. Finally, the cost sensitive BPNN model based on ABC is employed to the problem of click fraud detection. To summarize, Fig. 1 shows our detection framework, it consists of four components: data preprocess, feature selection, CS-BPNN classification model training phase, prediction phase. For data preprocess, we calculated feature vector values from millions of click data by user behavior analysis. In the general process of feature selection, we adopt Artificial Bee Colony algorithm to search for the best feature subset. The data of click fraud is extremely imbalanced and the cost of incorrect classification for a normal publisher is far below a fraud publisher. We use SMOTE to generate synthetic fraud samples and convert BPNN into a costsensitive classifier. In order to improve the performance of the classifier, BP neural network weights are optimized at the same time with feature subset selection.

Back Propagation Neural Network The click data are mass and random, and the association between derived click features and click fraud tendentiousness is often complex and nonlinear, so it is not easy to establish the detection model. Neural network is a machine learning technology that simulates human brain neural systems to realize artificial intelligence , and establishes the detection model based on the existing data. Therefore, BPNN is a good choice for fraud detection to complete the classification of publishers. In fact, the essence of neural network is to fit the real functional relationship between feature and target through weights and activation function. The output of each neuron in the hidden layer is

$$H_j = f_i\left(\sum_{i=1}^{n} w_{ij}^{(l)} x_i + b_j^{(l)}\right) \qquad (1)$$

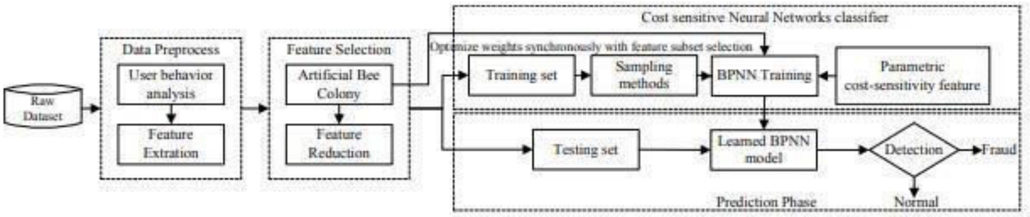


Figure 1. The overall framework of the proposed CSBPNN-ABC click fraud detection system

where j=1, 2,…, q, q is the number of neurons in this layer, l represent the neuron in the l layer, n is the total number of inputs to neuron j, i x is the i-th input of the neuron, ()l wij represents the connection weights between the neurons i and the neurons j, and ()l j b is the bias. f () is an activation function. The output of each neuron in the output layer is

$$O_k = f\left(\sum_{j=1}^{q} w_{jk}^{(l)} H_j + b_k^{(l)}\right)$$

where k=1, 2,…, p, and p is the number of outputs. f () can be a step, sign, sigmoid, Gaussian or linear function, which

is defined according to the different task. In this paper, we choose sigmoid function. The neural network utilizes the error function to carry out the back propagation to adjust the weights. The define of conventional error function is

$$E = \frac{1}{2}\sum_{k=1}^{p}(d_k - O_k)^2$$

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.488 |

|| Volume 9, Issue 4, April 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0904039 |

where dk is expected outputs. As in (3), adjusting the connection weights will change the error and BP neural network is prone to fall to local optimality. So the ABC algorithm is used to global optimize the weights to minimize error. The speed of convergence rate is improved at the same time. The inputs of BPNN is feature subset also searched by ABC.

Because of the different misclassification costs of fraud and normal class, we introduce the cost feature to correct the error function. That does not try to minimize global misclassification error, but misclassification costs. The process will be described in next subsection. The CSBPNN with two hidden layers created for click fraud detection is shown in Fig. 2

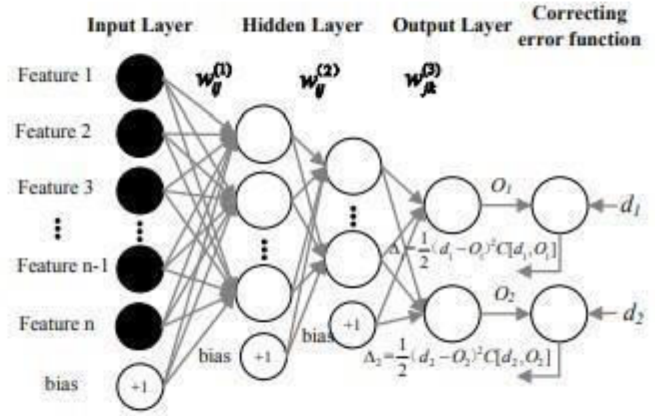


Figure 2. Architecture of the CS-BPNN for click fraud detection

III. LITERATURE SURVEY

A.“A Comparative Study of Ensemble Learning Methods for Classification in Bioinformatics “ byAayushi Verma, Shikha Mehta in 2017 7th International Conference on 2017 Jan 12 (pp. 155-158). IEEE.

A novel ensemble learning approach “BBS method” which stands for Bagging, Boosting and Stacking with appropriate base classifiers for the classification of the five UCI datasets taken from the field of Bioinformatics. Experiments are conducted using Weka and Java Eclipse and it has been observed empirically that our approach gives better accuracy with lower root mean square error rate using the technique of ensemble learning. Henceforth we conclude that our proposed ensemble learning method is more suitable in handling the classification problem in the bioinformatics domain. Such approaches can be efficiently used in related real-life scenarios of classification domain.

B. “Survey Paper on Crime Prediction using Ensemble Approach” by AyisheshimAlmaw, Kalyani Kadam in International Journal of Pure and Applied Mathematics 2018

Crime is a foremost problem where the top priority has been concerned by individual, the community and government. It investigates a number of data mining algorithms and ensemble learning which are applied on crime data mining Crime forecasting is a way of trying to mining out and decreasing the upcoming crimes by forecasting the future crime that will occur. Crime prediction practices historical data and after examining data, predict the upcoming crime with respect to location, time, day, season and year. In present crime cases rapidly increases so it is an inspiring task to foresee upcoming crimes closely with better accuracy. Data mining methods are too important to resolving crime problem with investigating hidden crime patterns.so the objective of this study could be analysing and discussing various methods which are applied on crime prediction and analysis.

C. "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone" by Davide Chicco and Giuseppe Jurman in Chicco and Jurman BMC Medical Informatics and Decision Making

Cardiovascular diseases kill approximately 17 million people globally every year, and they mainly exhibit as myocardial infarctions and heart failures. Heart failure (HF) occurs when the heart cannot pump enough blood to meet the needs of the body. Available electronic medical records of patients quantify symptoms, body features, and clinical laboratory test values, which can be used to perform biostatistics analysis aimed at highlighting patterns and correlations otherwise undetectable by medical doctors. Machine learning, in particular, can predict patients' survival from their data and can individuate the most important features among those included in their medical records.

D. "A machine learning approach to predict crime using time and location data" by Shama,NishatinBRAC University

In this work, they recognizing the criminal activity patterns of a place is paramount in order to prevent it. Law enforcement agencies can work effectively and respond faster if they have better knowledge about crime patterns in different geological points of a city. To use machine learning techniques to classify a criminal incident by type, depending on its occurrence at a given time and location. In that they use the dataset, which containing San Francisco's crime records from 2003 -2015. For this supervised classification problem, k-NN, Logistic Regression, Random Forest classification models were used. As crime categories in the dataset are imbalanced, oversampling methods, such as SMOTE and undersampling methods such as Edited NN, Neighborhood Cleaning Rule were used.

IV. PROPOSED WORK

Architecture Diagram:

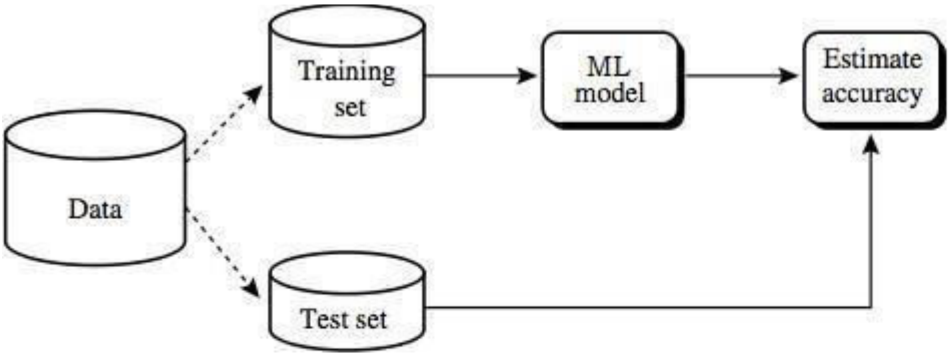
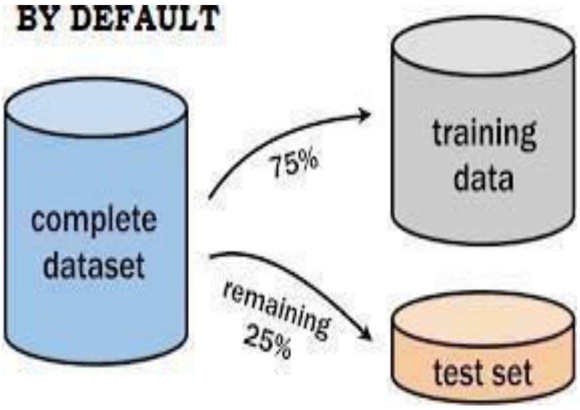


Figure3 :Architecture of machine lerning classification methods

In this work, we are going to study & analysing about the fraud detection. In this work, we are going to analyse the performance of machine learning classification methods, and classify as "is attributed" or "not attributed". A collection of machine learning model like XGBoost, Lightlgm, multiple encoding methods are applied for the prediction. A data is split into 3 parts like train, validation, test. To find the data is attributed or not. The complete implementation can be done through Google Colab (Python-Jupyter Notebook).

Splitting the Dataset:



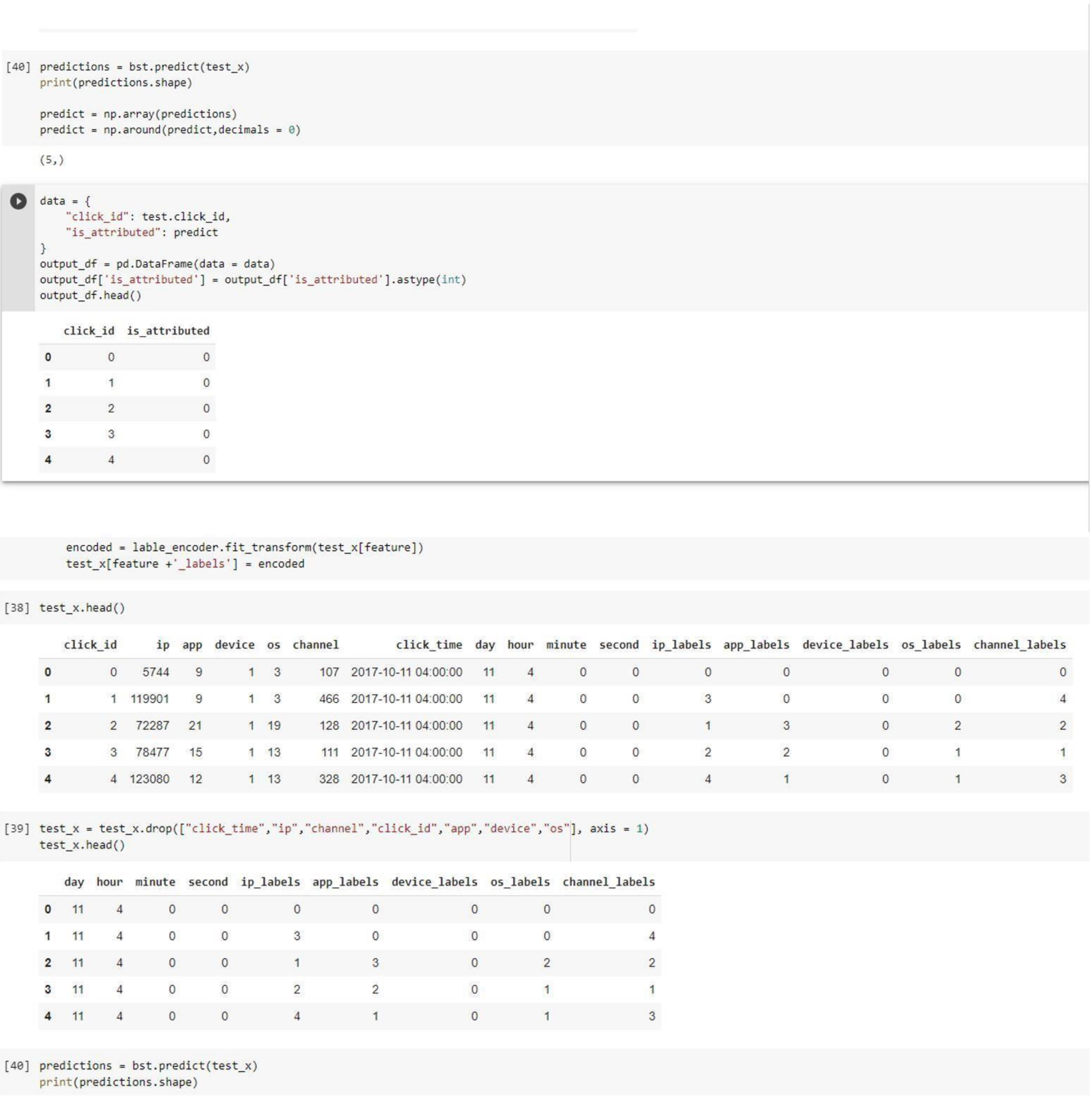
In General, on data splitting 75% of total data allocated for training & 25% o total data allocated for testing (By Default)

Libraries:

- OS
- Pandas
- Numpy
- Seaborn
- Matplotlib
- Scikit-Learn

| DOI: 10.15680/IJIRCCE.2021.0904039 |

Result



```
[40] predictions = bst.predict(test_x)
     print(predictions.shape)

     predict = np.array(predictions)
     predict = np.around(predict,decimals = 0)

     (5,)

     data = {
         "click_id": test.click_id,
         "is_attributed": predict
     }
     output_df = pd.DataFrame(data = data)
     output_df['is_attributed'] = output_df['is_attributed'].astype(int)
     output_df.head()
```

	click_id	is_attributed
0	0	0
1	1	0
2	2	0
3	3	0
4	4	0

```
        encoded = lable_encoder.fit_transform(test_x[feature])
        test_x[feature +'_labels'] = encoded

[38] test_x.head()
```

	click_id	ip	app	device	os	channel	click_time	day	hour	minute	second	ip_labels	app_labels	device_labels	os_labels	channel_labels
0	0	5744	9	1	3	107	2017-10-11 04:00:00	11	4	0	0	0	0	0	0	0
1	1	119901	9	1	3	466	2017-10-11 04:00:00	11	4	0	0	3	0	0	0	4
2	2	72287	21	1	19	128	2017-10-11 04:00:00	11	4	0	0	1	3	0	2	2
3	3	78477	15	1	13	111	2017-10-11 04:00:00	11	4	0	0	2	2	0	1	1
4	4	123080	12	1	13	328	2017-10-11 04:00:00	11	4	0	0	4	1	0	1	3

```
[39] test_x = test_x.drop(["click_time","ip","channel","click_id","app","device","os"], axis = 1)
     test_x.head()
```

	day	hour	minute	second	ip_labels	app_labels	device_labels	os_labels	channel_labels
0	11	4	0	0	0	0	0	0	0
1	11	4	0	0	3	0	0	0	4
2	11	4	0	0	1	3	0	2	2
3	11	4	0	0	2	2	0	1	1
4	11	4	0	0	4	1	0	1	3

```
[40] predictions = bst.predict(test_x)
     print(predictions.shape)
```

Figure4: Output for clicking fraud in encode and decode process

IV. CONCLUSION

Data in real world are complex, changeable and do not hold enough information for classification. In a problem like click fraud we have so many correlated variables, also high-entropy features together with low-entropy ones, which need to preprocess to get satisfied outcomes. The social network analysis model proposed is capable of detecting click frauds up to an accuracy of 91.23%. The analysis and comparison of different parameters helps put forward a clear and distinct vision towards the parameters, which impact the click fraud detection process. Furthermore, the model proposed can be made more efficient by including more parameters from the publisher's browsing data and a detailed study of the nature of the website

REFERENCES

1. I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.
2. F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” J. Mach. Learn. Res., vol. 12, no. Oct, pp. 2825–2830, 2011.
3. Benjamin EJ, Virani SS, Callaway CW, et al. Heart disease and stroke statistics—2018 update: a report from the American Heart Association. Circulation. 2018 Mar 20;137(12):e67–492.
4. Anderson JP, Parikh JR, Shenfeld DK, Ivanov V, Marks C, Church BW, et al. Reverse engineering and evaluation of prediction models for progression to Type 2 diabetes: an application of machine learning using electronic health records. J Diabetes Sci Technol. 2016 Jan;10(1):6–18. [5]. Montazeri M,Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. Technol Health Care. 2016 Jan 27;24(1):31–42.
5. Witten IH, Frank E, Hall MA. The WEKA workbench. Online appendix for “Data mining: practical machine learning tools and techniques.” 4th ed. Morgan Kaufmann; 2016.
6. Anand RS, SteyP,. Predicting mortality in diabetic ICU patients using machine learning and severity indices. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2018;2017:310– 9.
7. Welcome to Python.org [Internet]. Python.org. [cited 2018 Aug 5]. Available from:https://www.python.org/
8. “Total global mobile in-app advertising revenues 2015-2020”, [Online]. Available: https://www.statista.com/statistics/220149/total-wporldwide-mobile-app-advertising-revenues/ [Accessed: February 2020] .
9. “Report: Ad Fraud to hit $23 billion isn’t going down”,[Online]. Available: https://adage.com/article/digital/report-ad-fraud-hit-23-billion-isnt-go ing-down/2174721 [Accessed: February 2020].
10. H. Xu, D. Liu, A. Koehl, H. Wang, A. Stavrou((2014, Sepetember),”Click fraud detection on the advertiser Side”, in proceedings of the 19th European Symposium on Research in Computer Security(ESORICS), Poland, Europe. 11.F. Dong, H. Wang, L. Li, Y.Guo, T.F.Bissyande, T.Liu, G.Xu, J.Klein(2017, July),”FraudDroid: automated ad fraud detetcion for android apps”, in proceedings of ACM Symposium on Principles of Distributed Computing , Washington DC, US. 12.B. Liu, S. Nath, R. Govindam, J.Liu,(2014, April), “DECAF: Detetcing and characterizing ad fraud in mobile apps”, in proceedings of the 11th USENIX Symposium on Networked Systems Design and Implementation, Seattle, WA, US. 13.X. Zhang, X. Liu, H. Guo(2018, December), “A Click Fraud detection scheme based on cost sensitive BPNN and ABC in mobile advertising”, 4th IEEE International Conference on Computer and Communications(ICCC), Chengdu, China.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING