MOBILE ADVERTISEMENT FRAUD DETECTION USING MACHINE LEARNING

MR.S.Kaviarasan
Asst.Professor
Computer Science And Engineering
Panimalar Institute Of Technology
arasan.kavi@ @gmail.com

Nellore Sai Gokul
Student
Computer Science And Enginnering
Panimalar Institute Of Technology
nelloregokul @gmail.com

Chintagumpala Jaswanth
Student
Computer Science And Engineering
Panimalar Institute Of Technology
jashu7032425621@gmail.com

Madithati Nikhil Reddy
Student
Computer Science And Engineering
Panimalar Institute Of Technoloy
nikhilreddy1837@gmail.com

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. 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 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" 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 or "not attributed". A machine learning model like XGBoost, Lightlgm, multiple encoding methods are applied for the prediction process. The complete implementation can be done through Google Colab (Python-Jupyter Notebook).

I. INTRUDUCTION:

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

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.

Click frauds pose a huge risk to the advertising industry as businesses are estimated to lose $26 billion by 2020, $29 billion by 2021 and $32 billion by 2022(A study by cyber security company, Cheq), provided that these frauds remain unchecked. Till date, multiple methods have been proposed and various studies have been conducted for efficient detection and elimination of click fraud, for example, functionality test to characterize user engagement, UI state transition graphs to check against a set of heuristic based rules for detecting fraudulent behaviors, a DECAF method to control the execution of fraud detection policies.

we design a systematic approach to categorize fraudulent clicks. This approach's adoption of a set of custom-designed attributes enables the effective detection of click fraud. In addition, multiple correlation techniques including heat map using correlation have been used to successfully eliminate the detected frauds.

II. Literature Survey:

A. "A Comparative Study of Ensemble Learning Methods for Classification in Bioinformatics " by Aayushi Verma, Shikha Mehta in 2017 7t InternationalConference 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 Ayisheshim Almaw, 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

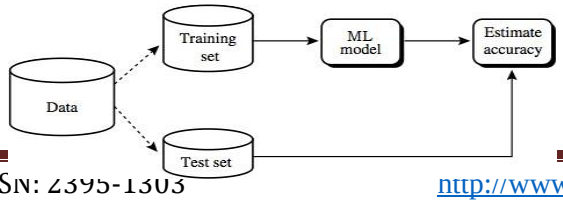
the most important features among those included intheir medical records.

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

E. "GUI BASED PREDICTION OFCRIME RATE USING MACHINELEARNING APPROACH" by Prithi S Aravindan S; Anusuya E; Ashok Kumar M in International Journal of Computer Science and Mobile Computing, Vol.9 Issue.3, March-2020, pg. 221-229

Crime rates are increases based on location and time. There is no specific reason for any criminal activity. To prevent this problem, Police sectors have to predict crime rate using machine learning. The aim is to investigate machine learning based techniques for crime rate by prediction results in best accuracy and explore in this work the applicability of data technique in the efforts of crime prediction with particular importance to the data set. The analysis of dataset is carried out by supervised machine learning technique (SMLT) to capture few vital information and to perform data validation, data cleaning and data visualization on the given dataset. The analysis does the prediction of accuracy by comparing the resultof different supervised machine learning algorithm **F." Detection and Analysis of Stress using Machine Learning Techniques " by Reshma Radheshamjee Baheti, Supriya Kinariwala. In International Journal of Engineering and Advanced Technology (IJEAT) 2019**



Every year tens of millions of people suffer from depression and few of them get proper treatment on time. So, it is crucial to detect human stress and relaxation automatically via social media on a timely basis. It is very important to detect and manage stressbefore it goes into a severe problem. A huge number of informal messages are posted every day in social networking sites, blogs and discussion forums. This paper describes an approach to detect the stress using the information from social media networking sites, like tweeter.

This paper presents a method to detect expressions of stress and relaxation on tweeter dataset i.e. working on sentiment analysis to find emotions or feelings about daily life. Sentiment analysis works the automatic extraction of sentiment related information from text. Here using Tensi Strength framework for sentiment strength detection on social networking sites to extract sentiment strength from the informal English text. TensiStrength is a system to detect the strength of stress and relaxation expressed in social media text messages.

TensiStrength uses a lexical approach and a set of rules to detect direct and indirect expressions of stress or relaxation. This classifies both positive and negative emotions based on the strength scale from -5 to +5 indications of sentiments. Stressed sentences from the conversation are considered &categorised into stress and relax.

G."Stress Detection in Speech Signal Using

Machine Learning and AI" by Dhole, S. N. Kale. In Machine Learning and Information Processing pp11-26

Individual person's speech is verbal way to have conversation with others. Speech many time probably becomes to know that individual person is in stressful condition or normal. These can lead with appropriate assessment of the speech signals into different stress types to evoke that the individual person is in a fit stateof mind.

In this work, stress identification and classification algorithms are developed with the aid of machine learning (ML) and artificial intelligence (AI) togetherwith MFCC feature extraction methods.

The machine learning and AI-based approaches use an intelligent combination of feature selection and neural optimization algorithms to train and to improve the classification and identification accurateness of the system. Comparison is done with approach of classical neural networks and fuzzy inference classifiers.

H."Stress Detection Using Classification Algorithm" by J. S. Kanchana, H. Thaqneem Fathima, R. Surya, R. Sandhiya. In International Journal of Engineering Research & Technology (IJERT) 2018 http://www.ijert.org ISSN: 2278-0181

Psychological problems are becoming a major threat to people's life. It is important to detect and manage stress before it turns into a severe health issue. Nowadays people share their feeling in social media regularly. It becomes easy to detect the stress of the users based on their social behaviour. Also, traditional stress detection methods are time consuming and costly. So the linguistic attributes in tweets can be leveraged to detect individual user stress. In this project, the stress states of users are classified using Naive Bayes classification algorithm and are categorized into stressed and non-stressed user

III. Proposed System:

Architecture Diagram:

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).

IV.Conclusion:

Data in real world are complex, changeable and do nothold enough information for classification. In a problem like click fraud we have so many correlated variables, also high-entropy features together withlow-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 thewebsite.

V.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. 2016Jan 27;24(1):31–42.

[6]. 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.

[7]. Anand RS, Stey P,. 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.

[8]. Welcome to Python.org [Internet]. Python.org. [cited 2018 Aug 5]. Available from: https://www.python.org/

[9] "Total global mobile in-app advertising revenues 2015-2020", [Online]. Available: https://www.statista.com/statistics/220149/total-wporldwide-mobile-a pp-advertising-revenues/ [Accessed: February 2020] .

[10]. "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].

[11]. 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.

[12]. 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.

[13]. 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.

[14]. 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.