IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11388

Credit card fraud detection using ML

Mr.Nitin Jagtap[1], Ms.Bhagyashri R. Patil[2], Ms.Khushbu Y. Sonawane[3],

Ms.Sanskruti D. Shinde[4], Mr.Parag N.Patil[5]

Professor Computer Engineering, SSBT COET, Bambhori , Jalgaon, Jalgaon, India[1]

Student of Computer Engineering, SSBT COET, Bambhori, Jalgaon, India[2-5]

Abstract: Credit Card Fraud detection is a challenging task for researchers as fraudsters are innovative, quick-moving individuals. The credit card fraud detection system is challenging as the dataset provided for fraud detection is very imbalanced. In today's economy, credit card (CC) plays a major role. It is an inevitable part of a household, business global business. While using CCs can of er huge advantages if used cautiously and safely, significant credit financial damage can be incurred by fraudulent activity. Several methods to deal with the rising credit card fraud (CCF) have been suggested. In this paper, an ensemble learningbased an intelligent approach for detecting fraud in credit card transactions using XGBoost classifier is used to detect credit card fraud, and it is a more regularized form of Gradient Boosting. XGBoost uses advanced regularization (L1 and L2), which increases model simplification abilities. Furthermore, XGBoost has an inherent ability to handle missing values. When XGBoost encounters node at lost value, it tries to split left right hands learn all ways to the highest loss.

Keywords: component, formatting, style, styling, insert

I. INTRODUCTION

Credit card fraud is a huge pain and comes with huge fees for banks and card provider companies. Financial companies try to prevent account abuse by using individual security responses. The more complex the security responses, the more fraudsters applying to obtain scammers, i.e. Change their strategies over time. Therefore, it is necessary to improve fraud detection strategies along with security units trying to prevent fraud. Most customers use credit card for buying things online. way, some of the customers can be the thief who has stolen the card of a person to make the online transactions. This is considered as the credit card fraud that must be detected. This fraud can also be in the form of any purchase by using the credit card in an unauthorized way. The cases of this kind of fraud are increasing. It is necessary to solve this challenging issue. Artificial intelligence is saving the time of humans in different fields. Especially machine learning, which is the branch of artificial intelligence is very helpful in performing the complex and difficult tasks.

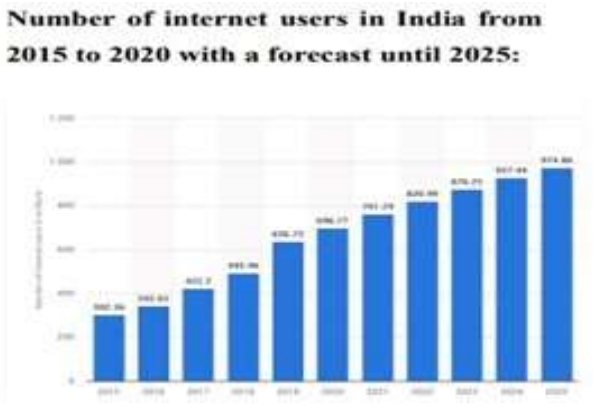


Fig.1 Number of internet users in India

II. SCOPE

Fraud act as the unlawful or criminal deception intended to result in financial or personal benefit. It is a deliberate act that is against the law, rule or policy with an aim to attain unauthorized financial benefit. Numerous literatures pertaining to anomaly or fraud detection in this domain have been published already and are available for public usage. A comprehensive survey conducted by Clifton Phua and his associates have revealed that techniques employed in this

DOI: 10.17148/IJARCCE.2022.11388

domain include data mining applications, automated fraud detection, adversarial detection. In another paper, Suman, Research Scholar, GJUST at Hisar HCE presented techniques like Supervised and Unsupervised Learning for credit card fraud detection. Even though these methods and algorithms fetched an unexpected success in some areas, they failed to provide a permanent and consistent solution to fraud detection. A similar research domain was presented by Wen-Fang YU and Na Wang where they used Outlier mining, Outlier detection mining and Distance sum algorithms to accurately predict fraudulent transaction in an emulation experiment of credit card transaction data set of one certain commercial bank. Outlier mining is a field of data mining which is basically used in monetary and internet fields. It deals with detecting objects that are detached from the main system i.e. the transactions that arent genuine. They have taken attributes of customers behaviour and based on the value of those attributes they've calculated that distance between the observed value of that attribute and its predetermined value. Unconventional techniques such as hybrid data mining/complex network classification algorithm is able to perceive illegal instances in an actual card transaction data set, based on network reconstruction algorithm that allows creating representations of the deviation of one instance from a reference group have proved efficient typically on medium sized online transaction. There have also been efforts to progress from a completely new aspect. Attempts have been made to improve the alert feedback interaction in case of fraudulent transaction. In case of fraudulent transaction, the authorised system would be alerted and a feedback would be sent to deny the ongoing transaction. Artificial Genetic Algorithm, one of the approaches that shed new light in this domain, countered fraud from a different direction. It proved accurate in finding out the fraudulent transactions and minimizing the number of false alerts. Even though, it was accompanied by classification problem with variable misclassification costs.

III. RESEARCH METHODOLOGY

Systematic literature reviews, for example, are a type of methodology, which conducts a literature review on a specific topic, could be used to detect fraud. A systematic review's primary goal in this context is to identify, evaluate, and Interpret the available studies in the literature that address the authors' research questions. A secondary goal is to identify research gaps and opportunities in the area of interest. First the credit card dataset is taken from the supply, and cleaning and approval is executed on the dataset which joins disposal of excess, filling void territories in sections, changing imperative variable into components or exercises then actualities is part into 2 sections, one is preparing dataset and another is check data set. Presently k crease move approval is done that is the special example is arbitrarily divided into k same and equivalent measured subsamples.

IV. CREDIT CARD FRAUD DETECTION TECHNIQUES

A. Logistic Regression An algorithm that can be used for both regression and classification tasks, but it is most commonly used for classification.' Logistic Regression is used to predict categorical variables using dependent variables. Consider two classes, and a new data point is to be checked to see which class it belongs to. The algorithms then compute probability values ranging between (0) and (1). Logistic Regression employs a more complex cost function, this cost function is known as the Sigmoid

B. Function or the Logistic Function.' [33]. LR also does not require independent variables to be linearly related, nor does it require equal variance within each group, making it a less stringent statistical analysis procedure. As a result, logistic regression was used to predict the likelihood of fraudulent credit cards [34]. Clarify the working of LR through the following scenario: The default variable for determining whether a tumor is malignant or not is y=1 (tumor= malignant); the x variable could be a measurement of the tumor, such as its size. The logistic function converts the x-values of the dataset's various instances into a range of 0 to 1. The tumor is classified as malignant if the probability exceeds 0.5. (As indicated by the horizontal line). B. K-Nearest Neighbours A simple, easy-to-implement supervised machine-learning technique that uses categorized input data to develop a function that gives a suitable output when given additional unlabelled data. Both classification and regression problems can be solved with the k-nearest neighbours (KNN) algorithm, which is quick and straightforward to apply. Uses label data to teach a function that generates an acceptable performance for new data. In the K-Nearest Neighbor algorithm, the resemblance between the new case and the cases that are already categorized is calculated. Once the new case is placed in a category that is most comparable to the available ones, it is applied to all remaining cases in that group. In an analogous fashion, KNN organizes all accessible data and categorizes new points depending on how similar they are. This describes anytime new data emerges, it is just a matter of fitting a K-N classification scheme to it. The algorithm is very straightforward and uncomplicated to put into practice. If a model does not need to be built, so some parameters and expectations may be tuned, it is unnecessary. The algorithm gets significantly slower as predictors/independent variables increase.

C. C. Random Forest Random Forest classifier finds decision trees in a subset of the data and then aggregates their information to that to get the full dataset's predictive power. Rather than relying on a single decision tree. The RF takes the predictions from each tree and forecasts the final output based on the majority votes of forecasts. Using a huge number

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022
DOI: 10.17148/IJARCCE.2022.11388

of trees in the forest improves precision and eliminates the issue of over fitting. It predicts output with high precision, and it runs efficiently even with large datasets. It can also keep accuracy when a large proportion of data is lost. Random Forest can handle both classification and regression tasks. It can handle large datasets with high dimensionality. It improves the model's accuracy and avoids the over fitting problem. We use twostep training techniques in the process of tree-based Random Forest: First, we generate the random forest by mixing N trees together, and then we estimate for each of the trees we generate in the first phase [31]. An ensemble algorithm employs the "random forest" artificial intelligence technique. Because it averts over-fitting by averaging the results, this approach outperforms single decision trees. Random Forest is an ensemble of diverse trees, similar to Gradient Boosted Trees, but unlike GBT, RF tree grow in parallel. Random Forests have a lot of uncorrelated trees. Because various trees are trained in parallel, the overall model diminishes a large number of variances. Random Forest treats each tree as a separate classifier that has been trained on resampled data. As a result of employing this this learn strategy and divide, the model's overall learning ability is increased

D. D. XGBoost Algorithm XGBoost has been widely used in many fields to achieve state-of-the-art results on some data challenges (e.g., Kaggle competitions), which is a high effective scalable machine learning system for tree boosting. XGBoost is optimized under the Gradient Boosting framework and developed by Chen and Guestrin [18], which is designed to be highly efficient, flexible and portable. The main idea of boosting is to combine a series of weak classifiers with low accuracy to build a strong classifier with better classification performance. If the weak learner for each step is based on the gradient direction of the loss function, it can be called the Gradient Boosting Machines. XGBoost is an efficient and scalable implementation of the Gradient Boosting Machine (GBM), which has been a competitive tool among artificial intelligence methods due to its features such as easy parallelism and high prediction accuracy.

V. DATASET

The credit card fraud detection related dataset used from the publicly available kaggle dataset. The dataset contains transactions made by credit cards in September 2019 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172percent of all transactions. The dataset divided into two groups of training set with 70percent and testing set with 30percent. It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, . . . V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset.

VI. WHY XGBOOST

The following advantages make it adaptable to deal with the transient stability prediction:

• In XGBoost model, multithreading parallel computing can be automatically called, which is faster than the traditional ensemble learning testability with large amounts of data in the actual power grid

• That the regularization term addition to XGBoost, makes its generalization ability be improved, which makes up for the shortcoming that the decision tree is easy to be over-fitted.

• XGBoost is the tree structure model, which doesn't need to normalize the data collected by PMU in the power system. Furthermore, it can effectively deal with the missing values, which is suitable for PMU-based transient stability prediction to discover the relationship between features and transient stability

. • When we look towards the time complexity XGBoost gives an accurate result in minimum time thus time complexity plays an important Role.

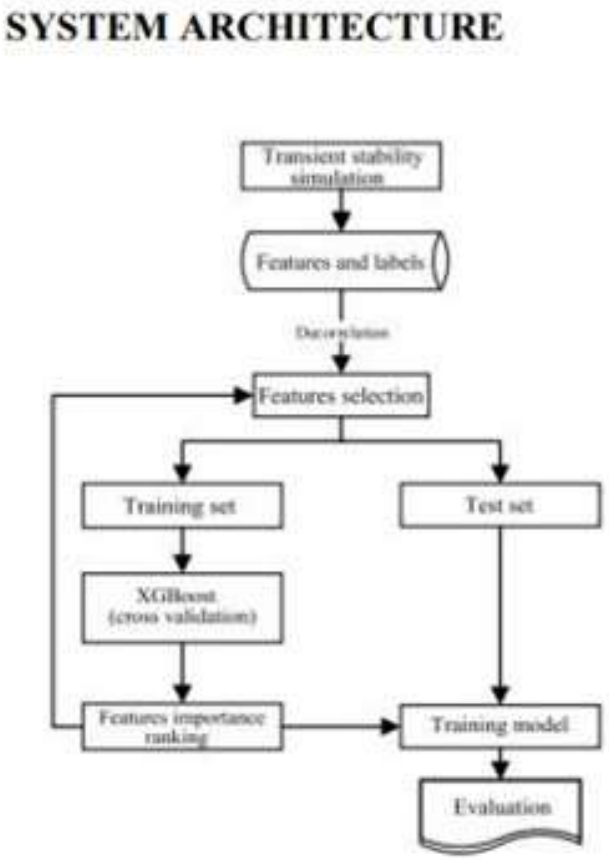
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11388

VII. SYSTEM ARCHITECTURE



VII. CONCLUSION

We can conclude from the above discussion that CCF is the major issue of financial sector that is increasing with the passage of time. More and more companies are moving towards the online mode that allows the customers to make online transactions. This is an opportunity for criminals to theft the information or cards of other persons to make online transactions. The most popular techniques that are used to theft credit card information are phishing and Trojan. So a fraud detection system is needed to detect such activities. Different machine learning algorithms are compared, including Logistic Regression, Random Forest, K-Nearest Neighbors, and XG Boost . Because not all scenarios are the same, a scenariobased algorithm can be used to determine which scenario is the best fit for that scenario. So on the time complexity basis we have choose the XG Boost a good algorithm to be performed to check detection.

REFERENCES

[1] S. H. Projects and W. Lovo, —JMU Scholarly Commons Detecting credit card fraud: An analysis of fraud detection techniques,|| 2020.
[2] S. G and J. R. R, —A Study on Credit Card Fraud Detection using Data Mining Techniques,|| Int. J. Data Min. Tech. Appl., vol. 7, no. 1, pp. 21–24, 2018, doi: 10.20894/ijdmta.102.007.001.004.
[3] —Credit Card Definition.|| https://www.investopedia.com/terms/c/creditcard.asp (accessed Apr. 03, 2021).
[4] —Credit Card Definition.|| https://www.investopedia.com/terms/c/creditcard.asp (accessed Apr. 03, 2021).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11388

BIOGRAPHY

Mr. Nitin Pundlik Jagtap, completed B.E (IT) and M.E. in Computer Science & Engineering. He is working as Assistant Professor in SSBT's College of Engineering and Technology since 2007. He is pursuing his PhD in Computer Science & Engineering in Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon. His areas of interest are Data mining, Machine Learning, Sentiment Analysis, and Data Analytics.

Ms. Bhagyashri R. Patil, UG candidate of Computer Engineering, Shrama Sadhana Bombay Trust"s College of Engineering and Technology, Jalgaon. Area of Interest: Machine Learning and Algorithm, Data Science.

Ms. Khushbu Y. Sonawane , UG candidate of Computer Engineering, Shrama Sadhana Bombay Trust"s College of Engineering and Technology, Jalgaon. Area of Interest: Machine Learning and Algorithm, Data Science.

Ms. Sanskruti D. Shinde, UG candidate of Computer Engineering, Shrama Sadhana Bombay Trust"s College of Engineering and Technology, Jalgaon. Area of Interest: Machine Learning and Algorithm, Data Science.

Mr. Parag N. Patil, UG candidate of Computer Engineering, Shrama Sadhana Bombay Trust"s College of Engineering and Technology, Jalgaon. Area of Interest: Machine Learning and Algorithm, Data Science.