Credit Card Fraud Detection Using Machine Learning

M.KoteswaraRao
Student
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, India.
mahankalikoteswararao78@gmail.com

J.SuryaNarayana
Student
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, India.
suryajakkula999@gmail.com

M.V.Aditya Kumar
Student
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, India.
adityakumarmudra@gmail.com

Y. Chandana
Asst. Professor
Department of Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet, India.
chandana.nrtec@gmail.com

Abstract—Nowadays credit card became one of the essential parts of the people. Sudden increase in E-commerce, customer started using credit card for online purchasing therefore risk of fraud also increases. Instead of carrying a huge amount in hand it is easier to keep credit cards. But nowadays that too becomes unsafe. Now a days we are facing a big problem on credit card fraud which is increasing in a good percentage. The main purpose is the survey on the various methods applied to detect credit card frauds. From the abnormalities, in the transaction, the fraudulent one is identified. We address this issue in order to implement some machine learning algorithm like Isolation Random Forest Algorithm in order to detect this kind of fraud. In this paper we increase the efficiency in finding the fraud. However, we discussed and evaluated employee criteria. Currently, the issues of credit card fraud detection have become a big problem for new researchers. We implement an intelligent algorithm which will detect all kind of fraud in a credit card transaction. We handled the problem by finding a pattern of each customer in between fraud and legal transaction. Random Forest Algorithm and Decision Tree Algorithm are used to predict the pattern of transaction for each customer and a decision is made according to them. In order to prevent data from mismatching, all attribute are marked equally.

Keywords— CreditCard, Criminal Transactions.

I. INTRODUCTION

At Present Situations as we can see that there is a huge increase online payment and the payment is mostly done with the help of credit cards. It becomes a big problem for marketing company to overcome with the credit card fraudulent activities. Fraudulent can be done in many ways such as tax return in any other account, taking loans with wrong information etc. Therefore, we need an efficient fraudulent detection model to minimize fraudulent activity and to minimize their losses. There are a huge number of new techniques which provide different algorithms which help in detecting number of credit card fraudulent activity. Basic understanding of these algorithms will help us in making a significant credit card fraudulent detection model. This paper helps us in finding doubtful credit card transaction by proposing a machine learning algorithms. Credit Card Fraudulent detection comes under machine learning, and the objective is to reduce such type of fraudulent activity[6].

This type of fraud is happening from past, and till now not much research has done here in this particular area. The types of credit fraud in transactions are bankruptcy fraud, behavioral fraud, counterfeit fraud, application fraud[3]. There are experiments done before on credit card fraudulent activity on basis of meta-learning. There is certain limit of meta-learning. There are two features which is introduced here in our report is True Positive and False alarm. Both these features play an important role in catching fraudulent because the rate of determining fraudulent behavior is quick[7].

II. DATASET DESCRIPTION

The dataset holds information about credit card transactions which has been made in a span of two days. The number of frauds have been calculated as 492 out of 284,807 transactions[2]. The details have been given in form of positive and non-positive numerical values. The dataset contains 31 features which has been labelled as V1-V28 due to confidential reasons. The feature which has been revealed are Time and Amount of transaction. Here time denotes the number of seconds elapsed from the first transaction of Day 1. Amount of transaction consists of positive value denoting deposit and non-positive value denoting withdrawal[12].

III. DATA PREPROCESSING

Data preprocessing is a way of making raw data more suitable for analysis. It involves cleaning, transforming, and integrating data to make it more complete, consistent, and understandable. Data preprocessing helps to improve the accuracy and quality of data mining or machine learning results.

In this project we have performed various preprocessing techniques which have helped clean the dataset into useful format.

IV. DATA VISUALIZATION

Data visualization is a way of showing data using graphics, such as charts, plots, infographics, and animations[8],[15].

Fig.3:Number of occurrences of each class label.

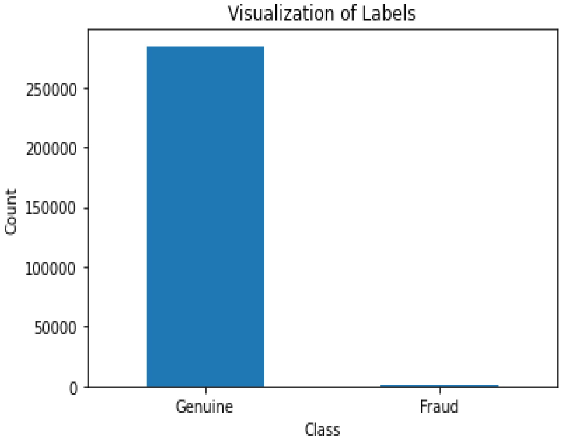


Fig.5:Approximate costing of online orders

V. MODEL BUILDING

The model is built in such way that the dataset we have is split into 70-30 using train_test_split().The model is built using Random Forest algorithm and Decision Tree Algorithm[9].

a)Isolation Random Forest (Existing system)

The previous detecting technique takes a long time to catch fraud which is basically depend on the database, not that much accurate and not give the result in-time. After that algorithm which is used for the detection of credit card fraudulent is generally on basis of analysis, fraudulent detection based on credit card transaction made by cardholder and the credit rate for cardholders.

There are certain limits of meta-learning. There are two features which is introduced here in our report is True Positive and False alarm. Both these features play an important role in catching fraudulent because the rate of determining fraudulent behavior is quick. For the better performance of model, we need a better classifier. Different classifier can be combined together with help of meta-learning.

Previously attempts have been made to work out Credit Card Fraud Detection system using SVM (Select Vector Machine). SVM makes use of hyperplane to classify the data points in a collection. A good hyperplane associates greater number of data points within its margin[5].

Processing a large amount of data sets can be inefficient due to the possibility of redundant data, which increases processing time. Therefore, it usually delayed in calculating the fraud or there might be probability to not calculate in time.

b) Random Forest(Proposed System)

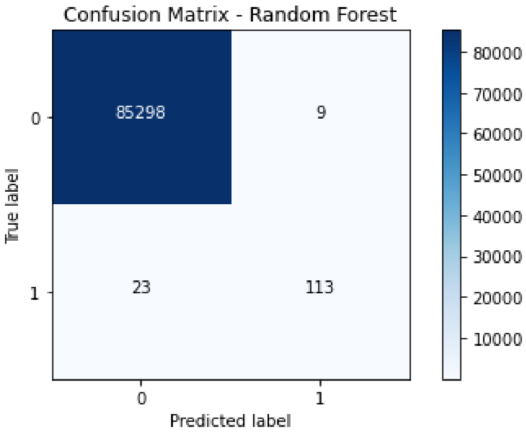
Belonging to the family of ensemble methods, Random Forest is a widely used machine learning algorithm. Ensemble methods combine multiple models to improve the accuracy and robustness of predictions. By utilizing decision trees- simple models that segment the feature space into smaller regions based on input feature values, Random Forest combines multiple models to form a powerful machine learning algorithm.

Multiple decision trees are created by the Random Forest algorithm, where each tree is trained on a random subset of input data and input features. The algorithm then aggregates the predictions of these individual trees to produce a final prediction. In order to aggregate results, two methods can be used depending on the problem type: majority vote for classification problems or averaging for regression problems.

The idea behind Random Forest is that by combining multiple decision trees, the model can capture a wider range of relationships between the input features and the target variable. Additionally, by using random subsets of the input data and features, By preventing models from becoming overly complex and overfitting to the training data, the algorithm mitigates the risk of poor performance on new and unseen data.

Random Forest outperforms other machine learning algorithms in multiple aspects, such as its capability to handle a large number of input features, its resilience to noisy data, and its ability to generate feature importance rankings that help uncover relationships between input features and target variables.

Random Forest is a versatile and powerful algorithm that can produce highly accurate predictions with relatively little tuning. Using this Random Forest We got 99.6% better Accuracy than Decision tree. And we are also performed train and evaluation of Dataset over Confusion Matrix-Random Forest.



c)Decision Tree (Proposed System)

A decision tree is a type of machine learning algorithm that is used for classification and regression analysis. It is a tree-like model where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label or a numerical value.

Decision trees are commonly employed in problems involving classification, where the objective is to anticipate a categorical output variable. The algorithm constructs a tree by dividing the data into smaller subsets repeatedly, based on the feature values. The splits are chosen to maximize the separation of the classes, usually based on metrics like information gain or Gini impurity.

Decision trees possess several benefits, including their simplicity in interpretation and visualization, which facilitates human comprehension of the model's prediction process. Additionally, decision trees can accommodate both categorical and numerical data, as well as manage missing values.

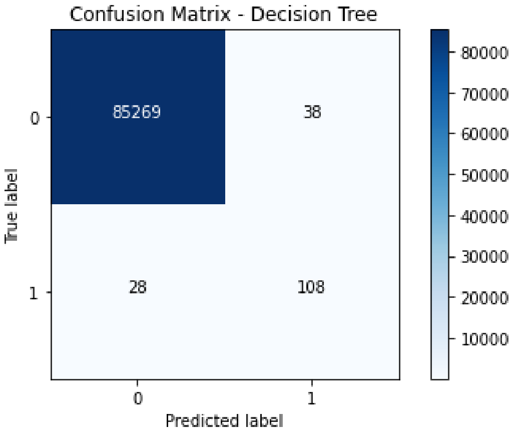
Nevertheless, decision trees have a tendency to overfit, particularly if the tree's depth is excessive or the feature count is substantial. Approaches like pruning and capping the tree's depth can alleviate overfitting. Furthermore, decision trees may not be the optimal selection for issues with interdependent features or situations with intricate, nonlinear decision boundaries.

As prediction of score is much important task according to our model therefore we are predicting the score on the basis of the given formula:

Score = 0.5 * TP + 0.5 * Deviation

Where, TP is True Positive value and Deviation is the deviation of outlier data from the standard data point.

On the basis of these score we made two classes 0 and 1. If the score is 1 it will move to class 1 and termed as legal transaction and if the score is 0 it will move to class 0 and termed as fraudulent transaction. At last, the accuracy is calculated on the basis of how many fraud transactions are there in our dataset and how many we predicted with the help of our model one without any replacement which results in creation of dataset for each tree with samples that are unique in nature.



VI. RESULTS AND DISCUSSION

We can see in the below table that Random Forest has the best performance in terms of accuracy and when compared to the Decision Tree and Isolation Random Forest which is proven to the best algorithm for the project until we have established the Credit Card Fraud Detection Using Machine Learning[13].

Algorithm	Accuracy
Isolation Forest	0.96
Random Forest	0.99
Decision Tree	0.98

Table1:Comparison of Performance of Regression Algorithms

VII. FUTURE SCOPE

There is a very strong possibility of the system being adopted as a norm for the major banking and financial services applications as fraud detection and prevention is the major checkpoint in financial and banking sector. The above system is also likely to be embedded in other applications based, modified as per platform-specific/application specific environment. The banks, financial and retail institutes have faced huge losses owing to cause of a robust and accurate system to predict and prevent the fraudulent transactions going on in an institution. This in-turn affects the business capabilities and consumer trust of the company. Thus, the organizations have moved their focus onto implementing a system which can depict inconsistent transactions, providing banks a privilege to act upon it take necessary measures.

VI. CONCLUSION

In this model, we detected the fraudulent transactions and recognized which illustrates the robustness of the proposed system. This proposed model took the trained dataset and performed classification on basis of them, if the transaction was legal then it moved to class 0 and if the transaction was fraud then it moved to class 1, and significantly improve the detection accuracy. The proposed method works efficiently in various platform, vivid environment and is a full_fledged cross platform application. The system has depicted robust, scalable and accurate performance to the degree that efficiency is taken into consideration in the Credit Card Fraud Detection System. The system takes into consideration various factors and has been fulfilling or meeting all the project specifications documented.

References

[1] R. R. Subramanian, R. Ramar, “Design of Offline and Online Writer Inference Technique”, International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 2S2, Dec. 2019, ISSN: 2278-3075.

[2]https://www.kaggle.com/mlg-ulb/creditcardfraud database of cards

[3] Delamaire. L. Abdou, HAH and Pointon. J,”Credit card fraud and detection techniques”, Banks and Bank Systems, Volume 4, Issue 2, 2009,2014.

[4] R. R. Subramanian, B. R. Babu, K. Mamta and K. Manogna, "Design and Evaluation of a Hybrid Feature Descriptor based Handwritten Character Inference Technique," 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Tamilnadu, India, 2019, pp. 1-5.

[5] Şahin, Y. G. and Duman, E. 2011. Detecting credit card fraud by decision trees and support vector machines.

[6] John Richard D. Kho, Larry A. Vea “Credit Card Fraud Detection Based on Transaction Behaviour” published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017.

[7] Yashvi Jain, Namrata Tiwari, ShripriyaDubey, Sarika Jain, “A Comparative Analysis of Various Credit Card Fraud Detection Techniques, Blue Eyes Intelligence Engineering and Sciences Publications 2019” .

[8] Learning Robert A. Sowah, Moses A. Agebure, Godfrey A. Mills, Koudjo M. Kaumudi, “New Cluster Undersampling Technique for Class Imbalance “of 2016 IJMLC.

[9] Baraneetharan, E. "Role of Machine Learning Algorithms Intrusion Detection in WSNs: A Survey." Journal of Information Technology 2, no. 03 (2020): 161-173.

[10] Mitra, Ayushi. "Sentiment Analysis Using Machine Learning Approaches (Lexicon based on movie review dataset)." Journal of Ubiquitous Computing and Communication Technologies (UCCT) 2, no. 03 (2020): 145-152.

[11] Mohamed Jaward Bah, Mohamed Hammad “Progress in Outlier Detection Techniques: A Survey” Hongzhi Wang, of the 2019 IEE. Aibus, S. et al. 2007. Application of Classification

Models on Credit Card Fraud Detection. IEEE

[12] Al Daoud, E. J. 1. 1. a. C. and Engineering, 1. 2019. Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset. 13(1), pp. 6-10. Alghamdi, M. et al. 2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. 12(7), p. 0179805.

[13] Awoyemi, J. O. et al, eds. 2017. Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI) IEEE.

[14] Bahnsen, A. C. et al, eds. 2014. Improving credit card fraud detection with calibrated probabilities. Proceedings of the 2014 SIAM international conference on data mining SIAM.

[15] Barandela, R. et al, eds. 2004. The imbalanced training sample problem: Under or aver sampling? Joint IAPR international workshops on statistical techniques in pattern recognition (SPR) and structural and syntactic pattern recognition (SSPR). Springer.