IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.066 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 10, Issue 11, November 2023

DOI: 10.17148/IARJSET.2023.101110

Using a Machine Learning Model to Predict Fraudulent Credit Card Transactions

Dr. Dinesh D. Patil[1], Dr. Priti Subramanium[2], Pooja Balu Wankhede[3]

Head of Dept. & Associate Professor, Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal, India[1]

Associate Professor, Shri Sant Gadge Baba College of Engineering and Technology Bhusawal, India[2]

Student’ Shri Sant Gadge Baba College of Engineering and Technology, Bhusawal, India[3]

Abstract: The project's goal is to use machine learning models to predict fraudulent credit card transactions. From both the bank's and the customers' perspectives, this is essential. The banks are unable to bear the loss of consumer funds to dishonest individuals. Since the bank is in charge of the fraudulent transactions, every fraud results in a loss for the bank. This document can be used as a template to type your own text into or as a set of instructions. When developing the model, we must address the imbalance in the data and test different algorithms until we find the optimal model. When developing the model, we must address the imbalance in the data and test different algorithms until we find the optimal model. Examining the valuable transaction and comparing it to a fresh, current transaction will reveal whether not the transaction is fraudulent. The purpose of this paper is to provide a comprehensive review of various techniques for detecting fraud. We suggested a system that uses a decision tree, random forest, and logistic regression to identify fraud in the processing of credit card transactions. After then, we use XGBoost to fit the dataset and categorize the transactions as either fraudulent or not. Because we were working with a heavily skewed dataset, we used the F1 and ROC AUC scores to evaluate our model's performance.

Keywords: Fraud Detection Techniques, Logistic Regression, Random Forest, Decision Tree, XGBoost Classifier.

I. INTRODUCTION

The individual utilizing the credit card has no relationship whatsoever with the cardholder and has no intention of paying back the amount they have made. The problem of credit card fraud detection, has been extensively studied. Because of this, we were able to examine widely used datasets and algorithms for the same goal and come up with a method that would work better than them. In detecting various fraudulent credit card transactions, many modern techniques based on Artificial Intelligence and Data Warehousing have evolved. In terms of finding data and identifying and stopping fraudulent transactions, machine-learning-based credit card fraud detection creates a model that yields the best results. For historical data, user’s pattern and behaviour are used to check and verify that the transaction is fraud or not [1]. This review paper presents an analysis of different machine learning and data mining techniques based on CC fraud detection methods [2][3]. Several learning algorithms, such as Random Forest [4][5], decision trees [6][7], and logistic regression [8], have been proposed for credit card fraud detection. Random Forest's generalization performance is superior. Even though it performed incredibly well, it required a long time to train and didn't deliver great results when working with a large dataset. A step up from Random Forest, XGBoost dramatically decreased training times and improved memory usage efficiency [9][10]. By using machine learning to detect credit card fraud, a model that yields optimal results in terms of data discovery, fraud detection, and prevention is developed. Various fundamental issues are at play, such as the system's fast reaction time, cost sensitivity, and feature pre-processing. Machine Learning (ML) is a branch of artificial intelligence that builds predictions on past data trends using a computer [11].

II. LITERATURE SURVEY

Suryanarayana, Venkata S. et al. [12] Because credit cards are so widely used, fraud appears to be a big issue for the credit card business. Since banks and businesses are hesitant to reveal the exact number of losses brought on by fraud, it is very difficult to find statistics on the impact of fraud. None the less, there are still many unanswered concerns regarding the optimal course of action because public data has little to do with privacy-related issues. The inability to quantify the amount of fraud we have observed, as well as the amount of fraud that goes unreported or unrecognized, presents another challenge to estimating the loss from credit card fraud. Dornadula and Vaishnavi Nath et al. [13] Fraud involving credit cards is an easy and accessible objective. Online payment options have expanded due to e-commerce and numerous other websites, raising the possibility of online fraud. Researchers are now using a variety of machine learning techniques to identify and analyse fraud in online transactions as fraud rates rise.

Chen et al. [14] Suggest a technique for gathering user questionnaire-responded transaction (QRT) data via an online survey. In addition, it employs the QRT models and a support vector machine (SVM) trained on the data to forecast brand-new transactions. An overview published by Phua et al. [15] Summarizes the difficulties that exist and various computational methods for detecting fraud across these fields. Brause et al. [16] Investigated the possibility of combining advanced data mining techniques and neural networks to achieve high fraud coverage while minimizing false alarms.

III. RESEARCH METHODOLOGY

For research purposes, the credit card dataset is accessible. By creating various, uncorrelated variables that eventually max, it obtains this. The dataset's detail, which consists of 31 columns with time, V1, V2, V3, ... V28 PCA applied features, amount, class labels.

- Time: The lapses between the current transaction and the initial transaction are measured in seconds.
- V1, V2, V3, ... V28 Attributes: The outcome of PCA's dimensional reduction to safeguard sensitive features and user identity is displayed in these 28 columns.
- Amount: Transaction Amount.
- Class Label: Binary class 1 and 0, Binary class 1 is non-fraudulent and binary class 0 is fraudulent transaction.

A. Performance Measures for Categorization:

i. Accuracy: Percentage of positive and negative values that were correctly predicted relative to the total values.

ii. Precision: Percentage of exact positive prediction among all positive estimates.

iii. F1-Score: Percentage of all actual positive values that were correctly predicted as positive.

iv. Recall: Integrates recall and precision into a single metric. The harmonic means of recall and precision, or the simple average of recall and precision.

v. ROC AUC Score ROC: The ROC-AUC curve can be used when it's necessary to see how well the classification model performs on charts. It is a well-liked and significant metric for assessing how well the classification model is performing. True positive rates rise when actual positive values do in a given dataset. In moreover, the true negative value might increase if real negative values rise in tandem with the dataset's overall values. Therefore, the recall is a useful metric for imbalanced datasets because it can change dramatically in response to changes in TP and TN values. Since ROC AUC does not skew the amount of test or evaluation data, it is a more accurate indicator of classifier performance than accuracy.

B. Techniques are Applied in ML:

i. Logistic Regression: A simple algorithm known as logistic regression computes the likelihood of an event occurring by estimating the relationship between one dependent binary variable and independent variables. The regulation parameter C manages the trade-off between maintaining the model's simplicity and increasing complexity. When the model gets more complicated and the power of regulation decreases for large values of C, overfitting of the data takes place. Going to follow the methodology, each dataset parameter "C" is determined before the logistic regression model is fitted to the training data.

ii. Decision Tree: Consequently, starting with the decision tree, the decision tree classifier is used to create the model. The algorithm's ‘criterion,’ which determines when to stop splitting the tree and is comparable to ‘max depth,’ is set to ‘entropy.’ This indicates that the tree can split four times.

iii. Random Forest: Two utilizes of RF, an ensemble method that is thought of as group learning, are regression and element classification. Irregular patterns are learned through deep trees. This method can lower the average of the variation in the value of RF when deep trees learn the same portion of the training sample.

iv. XGBoost**:** The XGB Classifier is the name of the XG boost model for classification. XGBoost is an ensemble machine learning algorithm based on decision trees that employs a gradient boosting framework. Therefore, artificial neural networks typically outperform all other algorithms or frameworks when used with unstructured data and prediction problems.

IV. RESULT AND DISCUSSION

Given that the dataset we were working with was highly skewed, we used ROC AUC and F1 to evaluate our model's performance. It is evident that the percentage of fraud is 0.17%. The imbalance of classes will be addressed later.

DOI: 10.17148/IARJSET.2023.101110

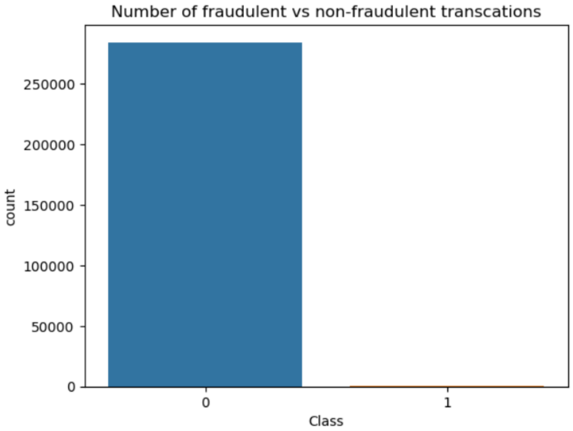

Fig. 1 Number of Fraudulent vs Non-Fraudulent Transactions

Outliers' Treatment:

For this specific dataset, we are not treating any outliers. Since every column has already undergone PCA transformation, it is assumed that the outlier values will be taken into account during the data transformation process.

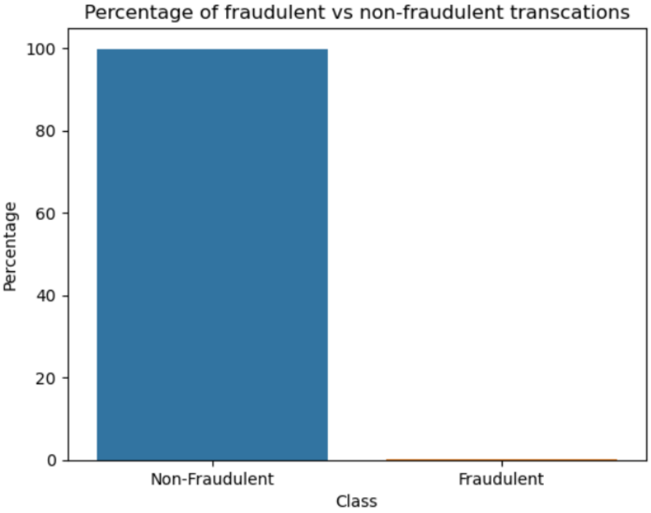

Fig. 2 Percentage of Fraudulent vs Non-Fraudulent Transactions

Observe the Distribution of Classes with Time:

Analysis**:** With regard to time, we see no detectable pattern for fraudulent and non-fraudulent transactions. As a result, we can remove the time column.

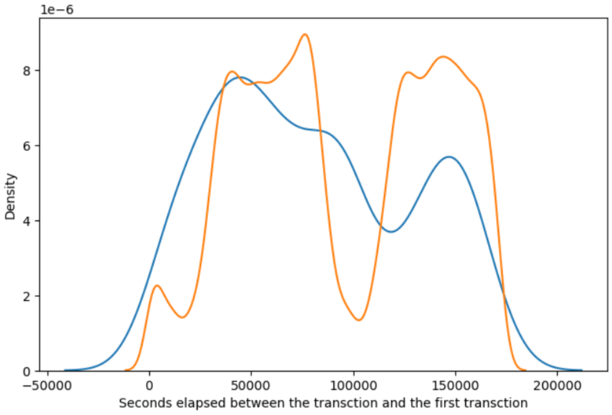

Fig. 3 Observe the Distribution of Classes with Time

DOI: 10.17148/IARJSET.2023.101110

Observe the Distribution of Classes with Amount:

Analysis: We can see that fraudulent transactions are mostly concentrated in the lower range of amounts, whereas non-fraudulent transactions are distributed evenly across the low to high range of amounts.

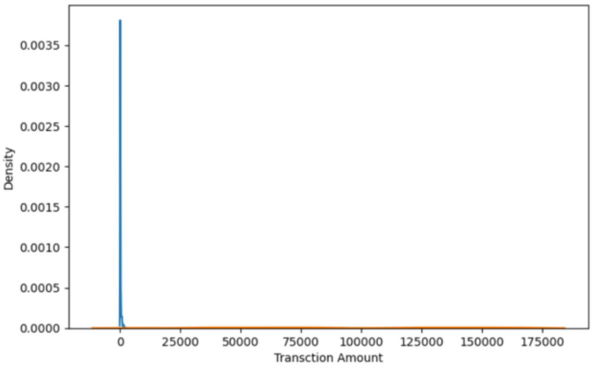

Fig. 4 Observe the Distribution of Classes with Amount

Tuning Hyperparameter C:

In Logistic Regression, C is the inverse of regularization strength. Higher C values indicate less regularization.

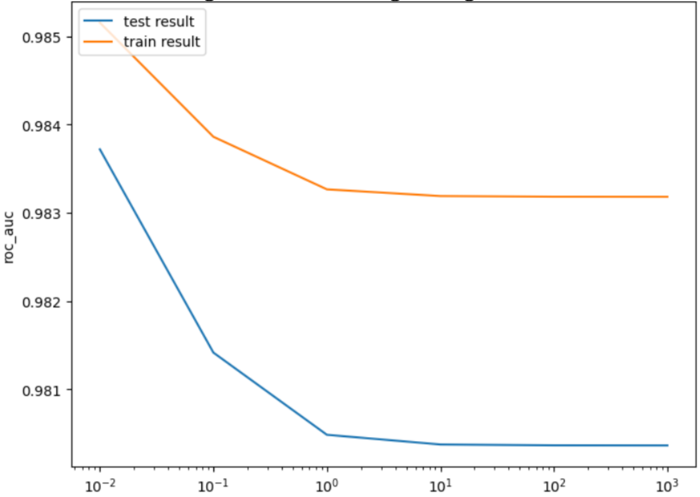

Fig. 5 Test Result & Train Result ROC AUC Curve

At C = 0.01, the highest test roc_auc is 0.9837192853831933.

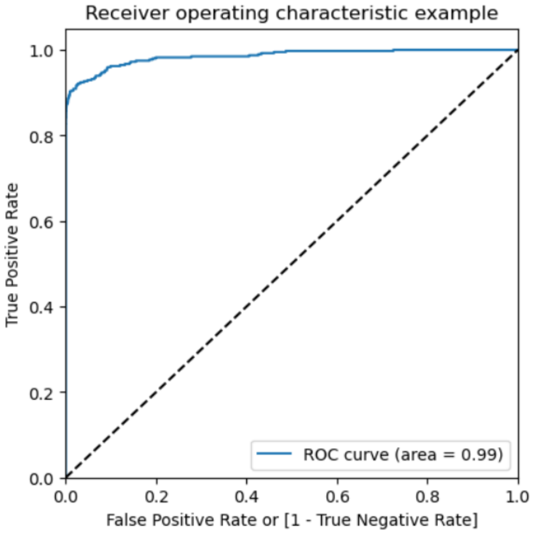

Fig. 6 ROC Curve Area Train Set

DOI: 10.17148/IARJSET.2023.101110

An Excellent ROC 0.99 was attained on the train set.

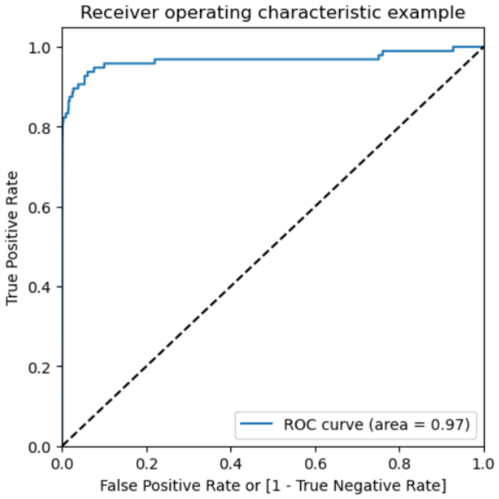

Fig. 7 ROC Curve Area Test Set

As we can see, the test set's ROC is very good at 0.97, which is very near to 1.

Model summary

- **Train set**
- Accuracy = 0.99
- Sensitivity = 0.70
- Specificity = 0.99
- F1-Score = 0.76
- ROC = 0.99
- **Test set**
- Accuracy = 0.99
- Sensitivity = 0.77
- Specificity = 0.99
- F1-Score = 0.65
- ROC = 0.97

After learning from the train set, the model does generally do well in the test set.

V. CONCLUSION

A strong classifier is able to adapt to the evolving nature of fraud. A fraud detection system's top priorities are minimizing false-positive cases and accurately predicting fraud cases. Every business case is different when it comes to how well machine learning techniques work. This project has looked into the dataset's imbalance or skewed nature. The effectiveness of this model has been greatly influenced by data pre-processing, especially feature extraction, as is the case with any machine learning workflow. With a ROC-AUC score, this model has demonstrated effective handling of credit card fraud cases that are imbalanced. This model exceeded the XGBoost Model on the same Dataset and demonstrated effective handling of imbalanced cases of credit card fraud, displaying a ROC-AUC score, F1 score, and recall.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Shri Sant Gadge Baba College of Engineering and Technology in Bhusawal, India, for supporting this project's work.

REFERENCES

[1]. Y. Sachin, E. Duman, “Detecting Credit Card Fraud by Decision Tree and Support Vector Machine”, In Proceedings of the international multi-Conference of Engineers and Computer Scientists, Hong Kong, 2011, pp. 1-6.

[2]. Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim, and Asoke k. Nandi, " Credit Card Fraud Detection Using AdaBoost and Majority Voting”, IEEE, vol. 6,2018, pp. 14277-14285.

DOI: 10.17148/IARJSET.2023.101110

[3]. B. Pushpalatha, C. Willson Joseph, " Credit Card Fraud Detection Based on the Transaction by Using Data mining Techniques”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue2, February 2017, PP 1785-1794.

[4]. Jalinus, N., Nabawi, R. A., & Mardin, A. (2017). The Seven Steps of Project-Based Learning Model to Enhance Productive Competences of Vocational Students. In 1st International Conference on Technology and Vocational Teacher (ICTVT 2017). Atlantis Press. Advances in Social Science, Education and Humanities research (Vol. 102, pp. 251-256).

[5]. Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi and Gianluca Botempi,” Credit card Fraud Detection: A realistic Modeling and a Novel Learning Strategy”, IEEE Trans. on Neural Network and Learning system, vol.29, No.8, August 2018.

[6]. Y. Sahin, and Duman, E., (2011),” Detecting credit card fraud by ANN and logistic regression.”, In Innovations in Intelligent Systems and Applications (INISTA), 2011 international Symposium on (pp.315-319). IEEE.

[7]. Behera, Tanmay Kumar, and Suvasini Panigrahi. "Credit card fraud detection: a hybrid approach using fuzzy clustering & neural network." In 2015 second international conference on advances in computing and communication engineering, pp. 494-499. IEEE, 2015.

[8]. S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang and C. Jiang, "Random Forest for credit card fraud detection," 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Zhuhai, China, 2018, pp. 1-6, doi: 10.1109/ICNSC.2018.8361343,

[9]. V. Jain, M. Agrawal and A. Kumar, " Performance Analysis of Machine Learning Algorithms in Credit Cards Fraud Detection", 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020, pp. 86-88, Doi: 10.1109/ICRITO48877.2020.9197762.

[10]. F. Wan," XGBoost Based Supply Chain Fraud Detection Model", 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), 2021, pp. 355-358, doi: 10.1109/ICBAIE52039.2021.9390041.

[11]. Y. Abakarim, M. Lahby, and A. Attioui,” An efficient real time model for credit card fraud detection based on deep learning”, in Proc. 12th Int. Conf. Intell. Systems: Theories Appl., Oct. 2018, pp. 1–7, doi: 10.1145/3289402.3289530.

[12]. Suryanarayana, S. Venkata, G. N. Balaji, and G. Venkateswara Rao, "Machine learning approaches for credit card fraud detection." Int. J. Eng. Technol 7.2 (2018), pp. 917-920.

[13]. Dornadula, Vaishnavi Nath, and Sa Geetha. "Credit card fraud detection using machine learning algorithms." Procedia computer science 165 (2019): 631-641.

[14]. R.C. Chen, M.L. Chiu, Y.L. Huang, L.T. Chen, “Detecting credit card fraud by using questionnaireresponded transaction model based on support vector machines”, Proceedings of the Fifth International Conference on Intelligent Data Engineering and Automated Learning, vol. 3177, October 2004, pp. 800– 806.

[15]. C. Phua, V. Lee, K. Smith, and R. Gayler, “A comprehensive survey of data mining-based fraud detection research,” arXiv preprint arXiv:1009.6119, 2010.

[16]. R. Brause, T. Langsdorf, M. Hepp,"Neural data mining for credit card fraud detection ", Proceedings of the International Conference on Tools with Artificial Intelligence, 1999, pp. 103– 106.