International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.060 ISSN: 2581-4621

Credit Card Fraud Detection Using Random Forest Algorithm In Machine Learning

N.Pavan[1], J.Siddharth[2], S.Ajay[3], S.V.R.Manimala[4]
[1,2,3]UG-Electronics and communication Engineering , Maturi Venkata Subba Rao Engineering College, Nadergul, Hyderabad-501510
[4] Professor, Electronics and Communication Engineering, Maturi Venkata Subba Rao Nadergul, Hyderabad-501510

Abstract: This The goal of this project is to employ machine learning to create a very successful credit card fraud detection system using the Random Forest method. The goal is to develop a solid system that can instantly detect and stop fraudulent transactions. The Random Forest algorithm is selected for the project because of its capacity to manage intricate data linkages. The project has several phases, including data gathering, preprocessing, and model selection.

Metrics like precision and recall are used to train, refine, and assess the model. The rapid detection of suspicious behaviors is ensured by real-time deployment, and stakeholder contact is facilitated by an intuitive interface. The project is driven by the pressing need to improve customer trust, increase financial security, and support technological innovation in the financial industry.

Stronger data security, a considerable decrease in financial losses, and favorable effects on industry practices are among the expected advantages. In the end, this project aims to create a cutting-edge system for detecting credit card fraud, using cutting-edge machine learning techniques to tackle modern problems.

I. INTRODUCTION

In the European Union, 700 million payment cards were issued in 2011. The volume of non-transaction cash excited in those years exceeded 3000 billion euros. In order to improve the security of European consumers, the use of emv1 for in-person transactions and 3D secure2 for online transactions is rapidly growing. Significantly more credit card transactions are still fraudulent despite the use of 3D-SECURE and EMV security measures: The annual amount of credit card theft in Europe is 1.5 billion euros. It is stated that in the future, credit card detection will use a data-driven fraud detection system in order to reduce the amount of credit card fraud that results from the inclusion of authentication methods.

According to Europol, international organized crime groups control the black market for credit card fraud and have an impact on non-cash payments all over the world. For organized crime groups, accepting payments using fake cards is a profitable and low-risk activity. These profits are subsequently utilized to create new deceptive strategies, finance other crimes, or start legitimate businesses (money laundering).

Experts argue that the highly structured and global nature of criminal networks demands international police cooperation. But the 1EMV (Europay, MasterCard, Visa) standard is designed for payment cards that use chips.

2 3D secure: the cardholder's identity is doubled by a PIN sent by SMS. 3.EUROPOL: European Police Organization for Police Cooperation and Criminal Justice Repression.

We will be working with a dataset in this project that has 284.807 transactions total, 492 of which are fake. The dataset contains 31 characteristics, the 31st of which is a binary variable with 1 fraudulent transaction and 0 valid transactions. Additionally, the random forest, svm, and naïve bayes algorithms will be put into practice. Additionally, the confusion matrix is used to gauge the classification model's performance. This confusion matrix computes the model's performance based on recall, accuracy, precision, and f1-score c.

II.LITERATURE REVIEW

• Dal Pozzolo et al. (2015) used an unbalanced dataset to tackle fraud detection as a binary classification problem. They emphasized that Random Forest's resilience to overfitting and capacity to manage unbalanced data contributed to its efficacy. In dynamic contexts, their unique adaptive learning approach enhanced classifier performance.

The importance of feature engineering in improving fraud detection systems is highlighted by Bahnsen et al. (2016). Their investigation revealed that transaction time and amount-based features greatly increased detection rates. The great interpretability and accuracy of Random Forest were praised. The significance of well-engineered features in anticipating fraudulent activities is highlighted by these findings.

• Carcillo et al. (2019) presented SCARFF, a scalable system utilizing Apache Spark for real-time fraud detection. This system maintained good detection accuracy while processing large-scale datasets in an effective manner. It was noted that Random Forest can adapt to streaming data, which makes it appropriate for situations involving real-time fraud detection.

• Abdallah et al. (2016) conducted a thorough analysis of several fraud detection systems, with a particular emphasis on Random Forest. Issues including unbalanced datasets and the requirement for real-time processing were covered in the survey. As one of the best options for detecting credit card fraud, Random Forest was acknowledged for its exceptional performance and capacity to manage huge datasets

III. EXISTING SYSTEM

More techniques than just the Random Forest algorithm are used in machine learning for credit card fraud detection. Modeling the likelihood that a transaction is fraudulent based on input information is one popular method called logistic regression. Understanding the variables contributing to fraud can be facilitated by its interpretability and simplicity.

Support vector machines (SVM) are an additional technique that facilitates transaction classification by identifying the hyperplane that most effectively distinguishes between fraudulent and non-fraudulent data points. Non-linear correlations between features can be handled with SVM, which performs well in high-dimensional domains.

Furthermore used, especially in deep learning designs, are artificial neural networks (ANN). complicated patterns and correlations in data can be automatically learned by ANNs, which makes them useful for identifying complicated fraud schemes. Sequential data analysis tasks, such finding patterns in a string of transactions, frequently make use of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks.

A strong classifier is produced by combining several weak learners using ensemble techniques like Gradient Boosting Machines (GBM) and AdaBoost. By emphasizing the cases that were incorrectly classified in earlier iterations, these techniques iteratively boost the accuracy of the model. Identification of anomalies or peculiar patterns in transaction data can be facilitated by the use of anomaly detection techniques like Isolation Forest and One-Class SVM. When transactions greatly depart from the norm, these techniques flag them as possibly being fraudulent. To further prepare data for machine learning algorithms, sophisticated data preparation methods including feature scaling, dimensionality reduction, and outlier detection are crucial. Feature engineering is also essential for the selection and development of pertinent features that record significant transactional data.

Several approaches are available for detecting credit card fraud besides the Random Forest algorithm. These include ensemble methods like Gradient Boosting Machines (GBM) and AdaBoost, as well as individual methods like Logistic Regression, Support Vector Machines (SVM), and Artificial Neural Networks (ANN). Still, every approach has drawbacks of its own. SVMs can be computationally demanding and require careful parameter adjustment, but the restricted complexity of logistic regression may make it difficult to detect sophisticated fraud patterns.

The interpretation of ANNs can be difficult, as they are prone to overfitting, especially deep architectures like RNNs and LSTMs. Ensemble approaches are more complex and have a higher chance of overfitting, notwithstanding their effectiveness. While they are confined to unsupervised learning, anomaly detection methods like Isolation Forest and One-Class SVM may have trouble determining the right thresholds. Even though they can be interpreted, rule-based systems are less flexible when it comes to changing fraud tendencies and necessitate specialized knowledge when creating rules. Further data preparation methods might also cause information loss and complexity even when they are useful. In order to create an effective credit card fraud detection system, organizations need to consider these drawbacks in addition to the benefits of each technique.

IV. RANDOM FOREST ALGORITHM

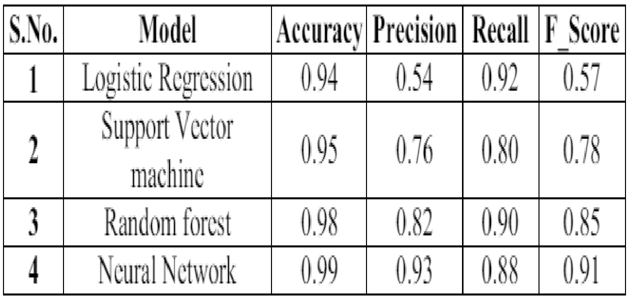
For machine learning problems including regression and classification, Random Forest is an ensemble learning method that is employed. The Random Forest algorithm's key details are as follows:

- Ensemble Approach: To increase overall accuracy and robustness, Random Forest gathers predictions from several decision trees it has built during training.
- Decision Trees: Individual decision trees that base their conclusions on characteristics in the input data are the base learners of a Random Forest.
Using a random subset of features during training lowers the likelihood of overfitting and increases tree variety.
- Bootstrap Aggregating (Bagging): All trees in Random Forest are trained using a bootstrapped sample (randomly sampled with replacement) from the training data using the bagging approach.
- Voting Mechanism: A majority vote amongst the individual trees determines the final forecast in categorization tasks. The mean of the predictions is what matters for regression problems.
- Good Generalization and High Accuracy: Random Forest has a propensity for good generalization to new, unknown data, big dataset handling, and high accuracy.
Compared to individual decision trees, Random Forest is less prone to overfitting because it aggregates the predictions of numerous trees.
- Feature Importance: One way to determine which features have the greatest influence on prediction performance is to use the algorithm's feature importance metric.
Random Forest exhibits versatility as it may be employed for a range of machine learning issues, owing to its ability to handle both numerical and categorical features in data.
- Parallelization: Utilizing contemporary computing architectures, training individual trees in a Random Forest can be expedited through parallelization.
- As a result, credit card fraud detection with the Random Forest algorithm is a thorough procedure that makes use of cutting-edge machine learning methods to provide a reliable, real-time system. Prioritizing precision, comprehensibility, expandability, and ongoing enhancement guarantees the system's efficacy in handling the dynamic obstacles presented by credit card fraudulent activities. making sure the models are not equivalent. So, how does random forest ensure that the behavior of any tree in the model isn't overly influenced by the behavior of the other trees? It is traversed by the two following ways.
- Decisions about baggage Assembling Bootstrap Data Diverse tree designs can be obtained by making small adjustments to the training set. Utilizing replacement sampling, random forest generates unique trees by letting each tree take a random sample from the dataset. The term "bagging" refers to this process.
- Instead of training each tree on a different piece using bagging, we are not breaking up the training data into smaller chunks. If the sample size is N, then every tree (unless otherwise noted) will have a training set of size N. Our method makes use of a random sample of size N with replacement in place of the original training data.

- We could provide [1, 2, 2, 3, 6, 6] to one of our trees if we have [1, 2, 3, 4, 5, 6] as our training set. As you can see, both lists include six items each, and because of our sampling strategy, the training data we provide our tree is randomly selected, therefore the words "2" and "6" appear twice
- The algorithm's performance is assessed using a certain table. It is employed to provide a performance summary of the classification algorithm. The reason it is named an error matrix is that it uses a matrix to show the algorithm performance error. The matrix is made using expected and actual parameters.
- True positive (TP): The actual value (YES) supported the model's forecast (YES). False positive (FP): When YES was predicted by the model, the actual result was NO. Type-I error is another term for it.
- Mistake Type-II: A false negative (FN) occurs when the model predicts a result of YES, but the actual value is NO.
- True negative (TN): The outcome was indeed NO as well as what the model had anticipated. Numerous computations, including accuracy and precision, can be performed using this matrix. The accuracy of a classification refers to the proportion of true positives and negatives out of all cases classified correctly. It shows the frequency of accurate outcome predictions made by the model. The formula for accuracy is shown below.
Precision equals (TP+TN)/(TP+FP+TN+FN).
- PRECISION: This number represents the proportion of cases correctly identified as positive (true positives) as opposed to cases that are assumed to be positive. It can be calculated using the formula below.
(TP+FP)/TP = Precision
- RECALL: The proportion of cases that are accurately categorized as positive (true positives) is this. It can be calculated using this formula.
= TP/(TP+FN) is the recall.
- F1-SCORE: The F1-SCORE is the harmonic mean of precision and recall. The F1 score is greatest when the recall and precision numbers are identical. It can be calculated using this formula. (2(Precision*Recall))/(Precison+Recall) is the F1-score.

V. RESULT AND CONCLUSION

The robustness and precision of the Random Forest algorithm are highlighted in this study, which shows how good it is at detecting credit card fraud. Effective preprocessing, feature engineering, and model optimization are important components. We made sure that real-time implementation and effective data processing were achieved by utilizing cloud platforms, Jupyter Notebook, Python, and Apache Spark. Adding more techniques to the Random Forest can improve detection even more. All things considered, Random Forest works well as a tool to lower financial losses and increase transaction securit



S.No.	Model	Accuracy	Precision	Recall	F_Score
1	Logistic Regression	0.94	0.54	0.92	0.57
2	Support Vector machine	0.95	0.76	0.80	0.78
3	Random forest	0.98	0.82	0.90	0.85
4	Neural Network	0.99	0.93	0.88	0.91

Table1: Comparisons of Other Methods

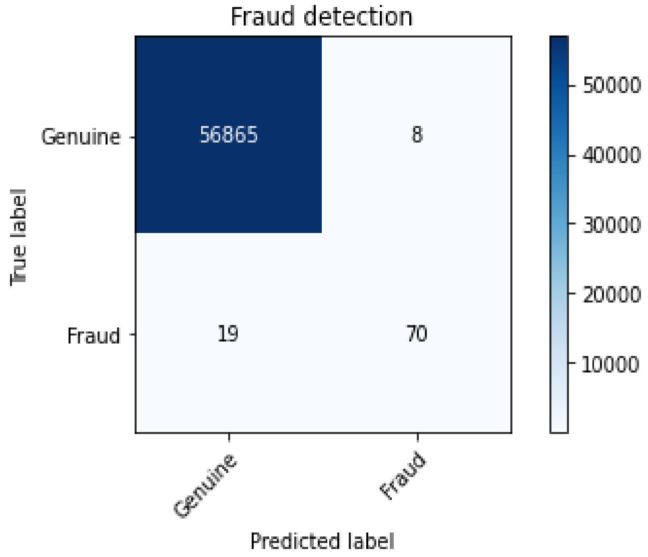


Figure 1: Confusion Matrix of Testing Data (Random Forest)

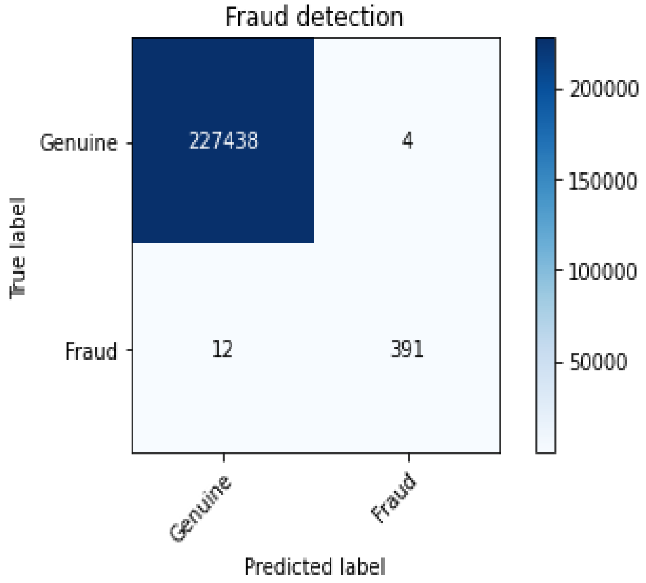


Figure 2: Confusion Matrix of Training Data (Random Forest)

References

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

2. Clifton phua1, vincent lee1, kate smith1 & ross gayler2 “ A Comprehensive Survey of Data Mining-based Fraud Detection Research” published by School of Business Systems, Faculty of Information Technology, Monash University,Wellington Road, Clayton, Victoria 3800, Australia

3. “Survey Paper on Credit Card Fraud Detection by Suman” , Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014
4. “Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen- Fang YU and Na Wang” published by 2009 International Joint Conference on Artificial Intelligence
5. “Credit Card Fraud Detection through Parenclitic Network Analysis By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral” published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages
6. “Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy” published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018 “Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi” published by International Journal of
7. Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016
8. David J.Wetson, David J.Hand, M Adams, Whitrow and Piotr Jusczak “Plastic Card Fraud Detection using Peer Group Analysis” Springer, Issue 2008.