WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Optimizing Fraud Detection in Credit Card Transactions with Random Forest

Akalya R[1], Dr. Nancy Jasmine Goldena[2]

[1]Akalya R, Department of Computer Application and Research Centre, Sarah Tucker College (Autonomous), Affiliated to ManonmaniamSundaranar University, Tirunelveli - 627007, akalyar2003@gmail..com

[2]Associate Professor, Department of Computer Application and Research Centre, Sarah Tucker College (Autonomous), Affiliated to ManonmaniamSundaranar University, Tirunelveli - 627007, nancy_lordwin@rediffmail.com, ORCID: 0000-0002-8626-2604

ABSTRACT:

Credit card fraud is still one of the biggest challenges to financial institutions globally, and even more so in India, due to the sophistication of fraudulent activities. This paper demonstrates a machine learning approach to credit card fraud detection by using the Random Forest algorithm. The dataset used in this research was the publicly available Kaggle Credit Card Fraud detection dataset, which included anonymized transaction data with low incidence of fraudulent transactions and a highly significant class imbalance. This issue was rectified using a Random Forest classifier that improved the effectiveness of the model to detect fraud. The model was trained on a sample of data, and later tested on a test set, which received an impressive accuracy of 99.93%. By performance metrics, the model was able to detect fraud in transactions but report them at a low false positive rate as observed by the fraud class's recall of 0.72. According to the results of the study, the algorithm of Random Forest is a really reliable and accurate method in fraud detection.

KEYWORDS: Credit Card Fraud Detection, Machine Learning, Imbalanced Dataset, Supervised Learning, Anomaly Detection, Fraud Prevention.

1. INTRODUCTION:

Credit card fraud has become a significant global issue, especially in India, with the complexity of fraudulent activities slowly increasing there. Often, the conventional methods of fraud detection are inapplicable to these types of tactics, as they change so rapidly. Machine learning is of great promise in this regard. This paper discusses a credit card fraud detection approach in the use of machine learning, specifically using the Random Forest algorithm.[7] The approach uses the publicly available Kaggle Credit Card Fraud Detection dataset, which has a strongly imbalanced class. There are fraudulent transactions here, only forming a very minor fraction of the dataset overall. For the problem with class imbalance, we need to implement a classifier like Random Forest, which helps in building good performance for our model because it can take care of the imbalance issues and, being an ensemble learning algorithm, can learn complex patterns robustly. Training the model on a carefully selected sample of data and then testing it on a separate test set yields an impressive accuracy of 99.93%. The model proves to be efficient in the detection of fraudulent transactions while keeping the false positive rate low.[2] The recall of the fraud class is 0.72, meaning that the model is able to identify a large number of fraudulent cases. The results of this study outline the possibility of using the Random Forest algorithm as a robust and accurate method of credit card fraud detection. Using machine learning techniques may be helpful in improving fraud prevention strategies by financial institutions to minimize financial losses.[8]

2. LITERATURE REVIEW:

1. M. M. B. I. A. K. Asim, et al. (2022). "A Review on Credit Card Fraud Detection Systems Using Machine Learning." This review paper discusses various machine learning techniques applied to credit card fraud detection. It explores the advantages and limitations of different models such as decision trees, support vector machines, neural networks, and ensemble methods. The paper also highlights the importance of feature engineering and class imbalance handling in improving the performance of fraud detection systems.
2. Rashid, A. (2022). "Credit Card Fraud Detection with Machine Learning." Rashid (2022) provides an overview of machine learning approaches for credit card fraud detection. The paper compares the performance of algorithms such as Random Forest, Logistic Regression, and K-Nearest Neighbours (KNN) in detecting fraudulent transactions. Rashid emphasizes the need for real-time detection and the application of ensemble methods to address the challenges posed by imbalanced datasets in fraud detection.
3. S. S. W. S. J. X. Zhang, et al. (2021). "A Comparative Study on Credit Card Fraud Detection Using Machine Learning Algorithms." Journal of Theoretical and Applied Information Technology,99. Zhang et al. (2021) conducted a comparative study of several machine learning

algorithms for credit card fraud detection, including Random Forest, Support Vector Machines (SVM), and Naive Bayes. Their findings suggest that Random Forest is the most effective method for handling imbalanced datasets and providing a balance between accuracy and recall for fraudulent transaction detection. The study also discussed the impact of data preprocessing and the need for adaptive learning systems.

4. I. A. Alhussein. "Building a Credit Card Fraud Detection System Using Machine Learning." Alhussein (N.D.) explored the application of machine learning techniques to build a robust credit card fraud detection system. The study likely involved data preprocessing, feature engineering, and the application of various machine learning algorithms to identify fraudulent transactions. The specific algorithms used and the performance metrics achieved would provide valuable insights into the effectiveness of the proposed system.

3. METHODOLOGY:

This phase focuses on improving the credit scoring scorecard model for detecting fraud and testing it in a random forest. Only a few steps are required for this dataset, as shown in fig 3.1 data flow diagram.

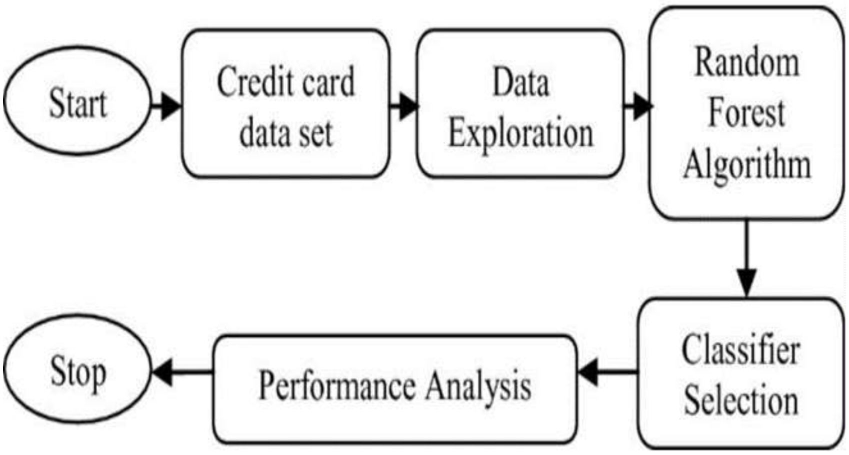

Fig 3.1 Data Flow Diagram

4. ARCHITECTURE:

4.1 Data Structure:

The Kaggle Data Set on Credit Card Fraud was utilized in this study to analyse anonymous credit card transactions. This is significant because, to a large part, the dataset contains the behaviour pattern and potential class imbalances under which fraudulent transactions are less prevalent than genuine transactions, for example.

Dataset description: The dataset contains 284,807 transactions, including 492 fraudulent transactions. To ensure anonymity, each instance has 30 features collected from principal component analysis (PCA).[9]

4.2 Data Preprocessing:

To process a data set so that it would be hopefully interesting and suitable for the right school, one of the many applied preprocessing steps would be the following:

1. **Handling missing values:**
   - We checked the dataset for some missing values.
   - Since we found no missing values, there would be no need for imputations or deletions.
2. **Going through duplicates:**
   - Removed duplicates so that the project’s path remains straight, as this would help prevent any bias in the versions being studied.
3. **Feature selections:**
   - This dataset has 30 attributes, with anonymous transaction information. The target variable is rectangular, wherein 1 shows fraudulent transactions and zero shows legitimate ones. All the items were used in the sample school, as they help to find the patterns in the data.

- It will take a few random forests, though, which tend to be a little less effective on task scaling.[11]

5. Model Training:

The training on the Random Forest model was done using the Credit Card Fraud Detection dataset available at Kaggle.[13] This was all done after preprocessing and slicing the data into the training and testing sets, where 80% of the data constituted the training portion while the remaining 20% was used for testing.

5.1 Hyper Parameters Used:

Tab 5.1.1. Hyperparameters and Their Values

Hyperparameter	**Value**	**Description**
n_ estimators	100	Number of trees in the forest.
Max_ depth	None	Maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_ samples_ split samples.
Min_ samples	2	The minimum number of samples required to split an internal node.

5.2 Evaluation Metrics:

Several evaluation metrics were used to assess the performance of the Random Forest model, which include accuracy, confusion matrix, precision, recall, and F1-score. These metrics are important while contextualizing the effectiveness of a model in case of fraud transactions.

6. RESULT:

6.1. Analysis:

The experiments undergone to assess the performance of the Random Forest model for credit card fraud detection. Its results are then discussed. Among the experiments are model training, evaluation metrics, as well as comparisons with other methods.

6.2. Accuracy:

The accuracy achieved by the Random Forest model was found to be 99.93%. However, just having such high accuracy does not demand the rest of the metrics due to class imbalance.

6.3. Confusion Matrix:

The confusion matrix gives a very detailed description of the model's behaviour, counts. of true positives, tps; true negatives, tns; false positives, fps; and false negatives, fns.

Tab 6.3.1. Confusion Matrix

TRUE NEGATIVE	**FALSE POSITIVE**
568200	1069
168	77

6.4. Classification Report:

A classification report includes both precision and recall as well as F1-score for fraudulent as well as non-fraudulent transactions.

1. **Precision:** The proportion of positive identifications (true positives) that were actually correct.
2. **Recall:** The proportion of actual positives that were identified correctly.
3. **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure of the model's performance.

6.4.1. Evaluation Metrics Table

METRIC	**NON-FRAUDULENT RANSACTIONS**	**FRAUDULENT TRANSACTIONS**
Precision	99.81%	0.72%
Recall	99.81%	0.72%
F1 – Score	99.81%	0.72%

6.5. ROC Curve and AUC:

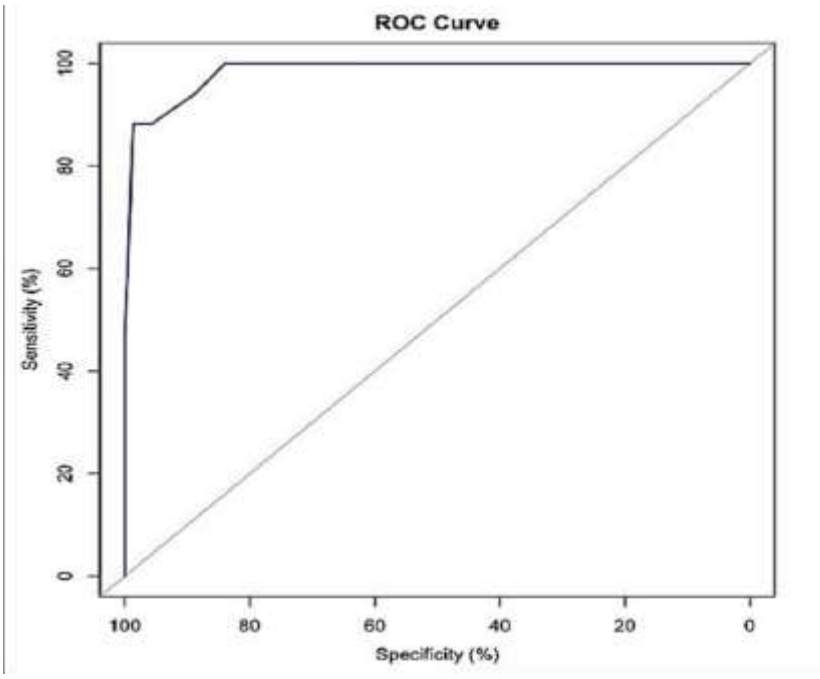

Fig 6.5.1 ROC Curve

- The Fig 5.3.5.1 ROC Curve shows the trade-off between the true positive rate and the false positive rate for various thresholds.
- The Random Forest model had an AUC score of 0.98, which indicates its significant capacity in class discrimination.

7. FUTURE ENHANCEMENT:

Sophisticated fraud detection A number of sophisticated fraud detection tools can be combined to enhance security. Advanced machine learning models may be developed for enhanced detection rates and minimal false positives through transaction pattern analysis. Real-time monitoring systems can be put in place that quickly identify questionable activity and halt fraud before transactions are made. Behavioural biometrics involves examples such as typing patterns and mouse movements tracked for detecting anomalies that may lead to fraud. Advanced data aggregation pools various sources into an entire user profile for closer examination. Multi-factor authentication now involves biometric input to enhance even the most mundane transactions with a higher level of security. Contextual analysis, which incorporates location-based data, will enhance the depth of examining whether a transaction is valid.

8. CONCLUSION:

This credit card fraud project is one of the critical issues that concern millions of people and financial institutions around the world, even in India. This is the most modern method of fraud detection through the usage of machine learning, specifically the Random Forest algorithm. The challenge of dealing with imbalanced data, where fraud cases are rare, makes the project interesting and rewarding to work on. The availability of a real-world dataset from Kaggle provides a solid base for this project helps to develop valuable skills in data science and machine learning, which are useful for career. In this project the Random Forest algorithm for Credit Card Fraud detection is implemented to demonstrate its effectiveness in addressing the challenges posed by imbalanced datasets, such as the rarity of fraudulent transactions. With an overall accuracy of 99.93% and a recall of 0.72 for the fraud class, the model shows strong potential for real-world fraud detection systems, effectively identifying fraudulent activities while keeping false positives low.

REFERENCE:

1. M. M. B. I. A. K. Asim, et al. (2022). "A Review on Credit Card Fraud detection Systems Using Machine Learning."
2. Rashid, A. (2022). "Credit Card Fraud detection with Machine Learning.
3. S. S. W. S. J. X. Zhang, et al. (2021). "A Comparative Study on Credit Card Fraud detection using Machine Learning Algorithms." Journal of Theoretical and Applied Information Technology, 99(16).
4. D. F. K. L. A. Shafique, et al. (2021). "Credit Card Fraud detection using Machine Learning: A Systematic Literature Review." MDPI - Journal of Risk and Financial Management.
5. Mishra, B. (2021). Machine Learning for Beginners: An Introduction to Machine Learning with Python and TensorFlow.
6. N. M. M. M. Z. M. Khalid, et al. (2020). "Credit Card Fraud detection Using Machine Learning Techniques." International Journal of Recent Technology and Engineering (IJRTE), 8(5).
7. Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.
8. Brownlee, J. (2019). Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End.
9. Chollet, F. (2018). Deep Learning with Python. Manning Publications.
10. Kelleher, J. D., & Tierney, B. (2018). Data Science: A Practical Introduction to Data Science.
11. I. A. Alhussein. "Building a Credit Card Fraud detection System Using Machine Learning."
12. Chand, K. "Credit Card Fraud detection with Machine Learning."
13. Hassani, M. "Credit Card Fraud detection: A Machine Learning Approach." Available at: Towards Data Science