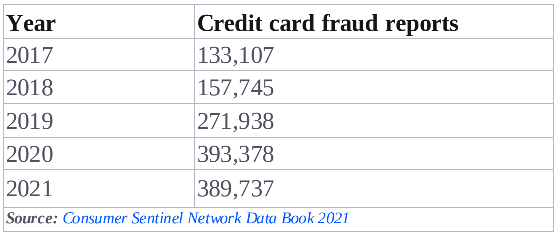
Credit Card fraud Detection

Akshay Sutar[1]

Data Analyst

Aim and Objectives

Credit card fraud is a significant issue that can result in substantial financial losses for individuals and businesses. According to a study credit card fraud has become more predominant due to the increasing popularity of online transactions and e-commerce platforms. Between 2020 and 2021, card fraud losses worldwide grew by over 10%, marking the biggest increase since 2018. An estimated 30 billion dollars were lost by retailers and cardholders, with the United States accounting for about 12 billion of those losses, according to the source. The payments industry watchdog Nilson Report published a prediction in December 2022 that states that over the following ten years, card fraud losses in the United States will hit $165.1 billion. Card-not-present fraud, which includes online, over-the-phone, and mail-order transactions, is the only kind of credit card theft that is thought to have caused $5.72 billion in losses for the United States in 2022.



Year	Credit card fraud reports
2017	133,107
2018	157,745
2019	271,938
2020	393,378
2021	389,737

Source: *Consumer Sentinel Network Data Book 2021*

Table 1

A key challenge with payments fraud data is class imbalance. In the Kaggle dataset, roughly 99.8 percent of the transactions are labeled as legitimate and 0.2 percent as fraudulent. Class imbalance can make it difficult for standard models to learn to distinguish between the majority and minority classes. In response to this issue, this study aims to develop an efficient model for detecting credit card fraud using machine learning techniques.

1. To review the literature on credit card fraud detection techniques and identify the most effective methods for detecting fraudulent transactions.

2. To preprocess the dataset obtained from Kaggle for credit card fraud detection, which contains 284,807 transactions, of which only 0.172% is fraudulent.

3. To implement and compare the performance of three machine learning algorithms, namely logistic regression, decision tree classifier, and random forest classifier, for detecting credit card fraud.

4. To evaluate the performance of the implemented models using accuracy score, F1 score, precision score, and recall score.

Literature Review

According to a study by Joseph and Manikandan (2022), credit card fraud has become more prevalent due to the increasing popularity of online transactions and e-commerce platforms. The authors reviewed various techniques for detecting credit card fraud, including machine learning and deep learning algorithms. They found that machine learning algorithms such as logistic regression, decision trees, and random forests have shown promising results in detecting credit card fraud.

Machine learning algorithms can analyze large volumes of transaction data, allowing them to identify Patterns and anomalies that could be indicative of fraud. Supervised learning techniques such as decision trees, logistic regression, and neural networks have been used to build models that classify transactions as either fraudulent or non-fraudulent.

Similarly, Randhawa et al. (2018) proposed a technique for detecting credit card fraud using machine learning algorithms. They used AdaBoost and majority voting methods to improve the accuracy of their model. The authors achieved a precision rate of 99.7% and an accuracy rate of 99.9% using their proposed method.

On the other hand, Huang et al. (2020) used variational auto-encoding (VAE) to detect credit card fraud. They found that the VAE model achieved a precision rate of 99.5% and an accuracy rate of 99.9%. However, the authors noted that the recall rate of this model was not increased while increasing the precision and F-measure.

In a study by Esenogho et al. (2022), the authors proposed a neural network ensemble with feature engineering for improved credit card fraud detection. They achieved a precision rate of 99.9% and an accuracy rate of 99.9% using their proposed method.

Methodology

The dataset contains transactions made by a cardholder in duration in 2 days i.e., two days in the month of September 2013. Where there are total 284,807 transactions among which there are 492 i.e., 0.172% transactions are fraudulent transactions. This dataset is highly unbalanced. Since providing transaction details of a customer is considered to issue related to confidentiality, therefore most of the features in the dataset are transformed using principal component analyses (PCA). V1, V2, V3,..., V28 are PCA applied features and rest i.e., 'time', 'amount' and 'class' are non-PCA applied features, as shown in table 2.

Sr. No.	Attribute	Details
1	Time	Time in seconds to specify the elapses between the current transaction and first transaction.
2	V1 : V28	Non-descriptive variables resulting from a PCA dimensionality reduction to protect sensitive data
3	Amount	Transaction amount
4	class	0 – Normal 1 - Fraud

Table 2

Type of Transaction	Count	Percent
Fraud	492	0.173
Normal	284,135	99.827
Total	284,807	100

Table 3

In this study, we preprocessed the dataset obtained from Kaggle for credit card fraud detection. We used techniques such as data cleaning, normalization, and feature selection to prepare the dataset for machine learning algorithms.

We implemented and compared the performance of three machine learning algorithms, namely logistic regression, decision tree classifier, and random forest classifier, for detecting credit card fraud. We used the scikit-learn library in Python to implement these algorithms.

We evaluated the performance of the implemented models using accuracy score, F1 score, precision score, and recall score. We used a ratio of 70:30 to split the dataset into training and testing sets.

Conclusion

The study found that the random forest classifier achieved the highest accuracy score of 99.99%, followed by the logistic regression model with an accuracy score of 94.55% and the decision tree classifier with an accuracy score of 99.82%. The random forest classifier also achieved the highest F1 score, precision score, and recall score.

ML Model	Accuracy	Precision	Recall Score	F1 Score
Logistic Regression	94.55	97.34	91.59	94.38
Decision Tree Classifier	99.82	99.74	99.89	99.82
Random Forest Classifier	99.99	99.98	100	99.99

The study demonstrated that machine learning algorithms, particularly the random forest classifier, can be effective in detecting credit card fraud. However, the study also highlighted the need for further research to improve the precision rate of the models.

References

Esenogho, E., Mienye, B. D., Swart, T. G., Aruleba, K., & Obaido, G. (2022). A neural network ensemble with feature engineering for improved credit card fraud detection. IEEE Access, 10, 1598-1608.

Huang, T., Cheng, G., & Huang, K. (2020). Using variational auto encoding in credit card fraud detection. IEEE Access, 8, 195402-195411.

Joseph, R. E., & Manikandan, L. C. (2022). A review on credit card fraud detection techniques. International Journal of Engineering and Technology, 14(04), 56-63.

Kaggle. (2018). Credit Card Fraud Detection. Retrieved from <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

Randhawa, K., Loo, C. K., Seera, M., Lim, C. P., & Nandi, A. K. (2018). Credit card fraud detection using AdaBoost and majority voting. IEEE Access, 6, 14277-14284.

[1] https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

[2] https://www.ijert.org/a-review-on-credit-card-fraud-detection-techniques-2

[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10280638/

[4] https://ijcsmc.com/docs/papers/August2019/V8I8201911.pdf

[5] https://ijcrt.org/papers/IJCRT23A5164.pdf

[6]https://nilsonreport.com/articles/card-fraud-losses-worldwide-2/

[7] Consumer Sentinel Network Data Book 2021 (ftc.gov)