FindDefault (Prediction of Credit Card fraud)

1. Introduction

Fraud detection in financial transactions is a critical task for financial institutions to prevent financial losses and protect customer assets. This project aims to build and evaluate various machine learning models to detect fraudulent transactions using a dataset of financial transactions. The project involves data preprocessing, model training, evaluation, and comparison of different models on both imbalanced and balanced datasets.

2. Data Description

The dataset used in this project contains financial transaction records, including both fraudulent and non-fraudulent transactions. Key features include transaction amounts and various anonymized features. The dataset is heavily imbalanced, with a much smaller number of fraudulent transactions compared to non-fraudulent ones.

3. Methodology

a. Data Preprocessing

1. Loading the Dataset: The dataset is loaded using pandas.

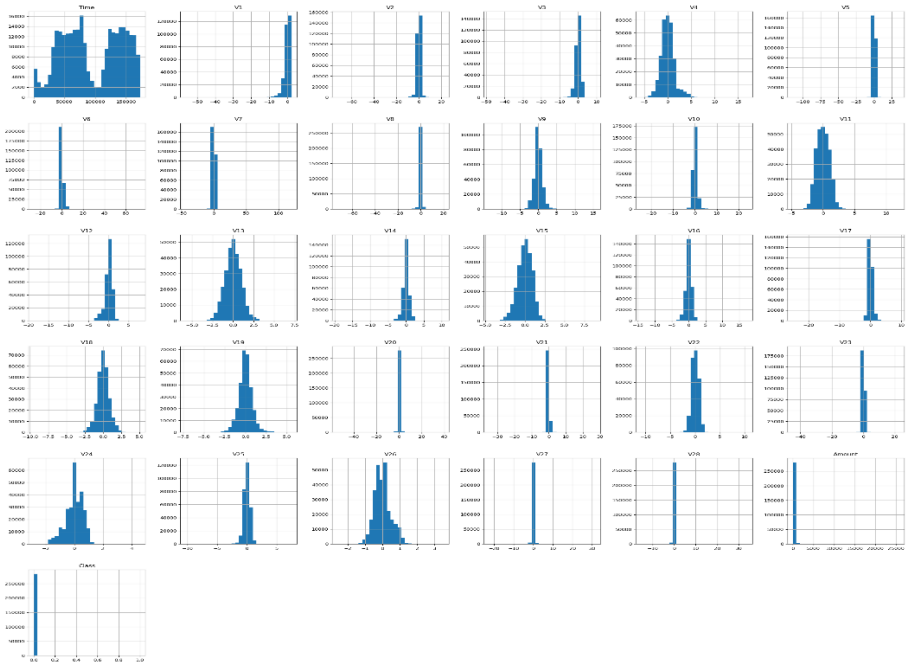
2. Splitting the Dataset: The data is split into training and validation sets using `train_test_split`.

3. Standardization: Features are standardized using `StandardScaler` to improve model performance.

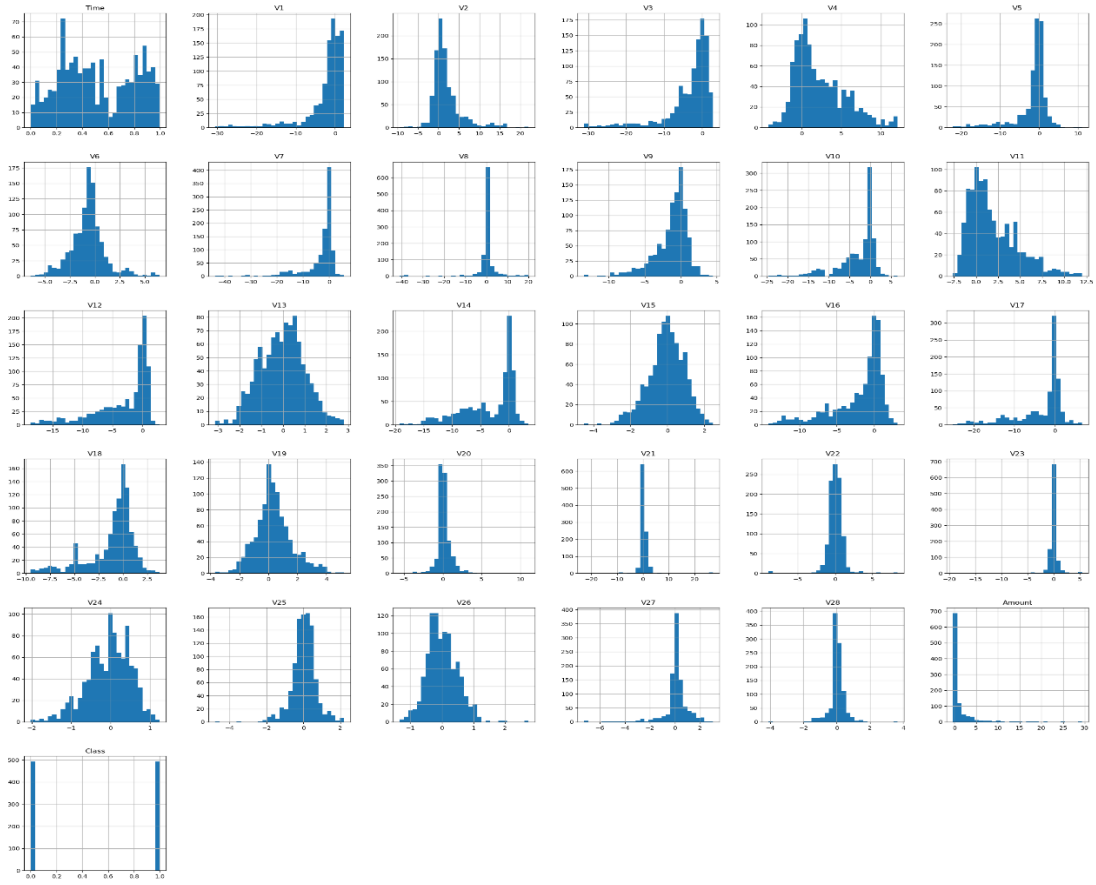
b. Handling Imbalance

Two versions of the dataset are prepared:

1. **Imbalanced Dataset:** The original dataset with its inherent imbalance.



2. **Balanced Dataset:** The balanced dataset was created using under-sampling of the majority class (non-fraudulent transactions) to match the size of the minority class (fraudulent transactions). This technique randomly selects a subset of non-fraudulent transactions equal to the number of fraudulent transactions, resulting in a balanced dataset.



c. Model Training

Five different machine learning models are trained and evaluated on both the imbalanced and balanced datasets:

1. Logistic Regression
2. Shallow Neural Network
3. Random Forest Classifier
4. Gradient Boosting Classifier
5. Linear Support Vector Classifier (SVC)

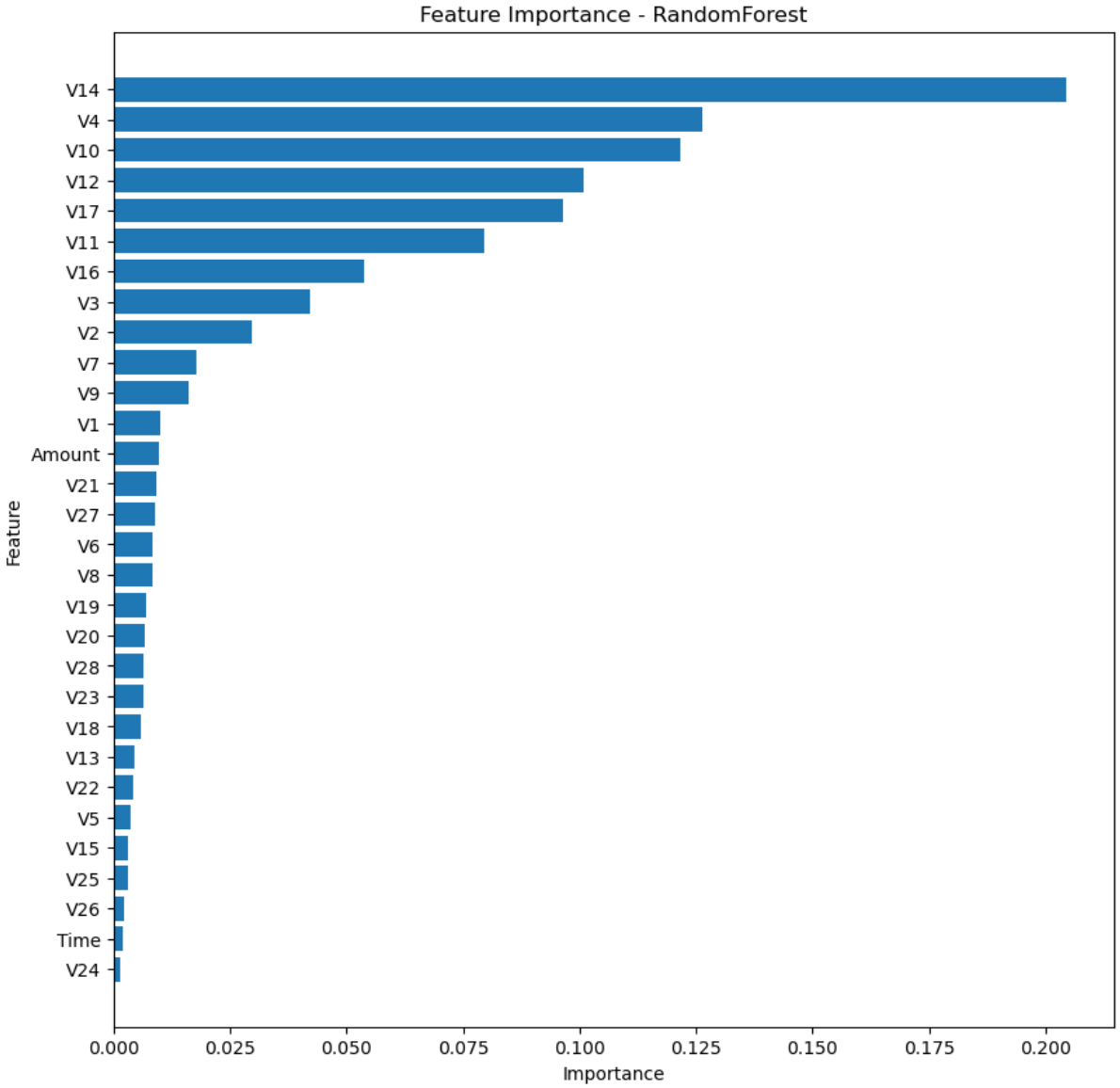
Additionally, the Random Forest model is further optimized using `RandomizedSearchCV`.

d. Model Evaluation

Models are evaluated based on precision, recall, F1-score, and ROC-AUC score. These metrics provide insights into the performance of the models, especially in terms of their ability to correctly identify fraudulent transactions.

e. Feature importance

Understanding which features contribute most to the model's predictions is crucial for both interpretability and improving model performance. In this project, feature importance was evaluated using the RandomForestClassifier, as it provides an inherent measure of feature importance. The RandomForest model's feature importances were extracted and visualized to identify the most significant predictors of fraudulent transactions. This analysis helps in understanding the factors that drive fraud detection and can guide further feature engineering or selection processes.



Here we can see the most contributing feature to the least.

4. Results

a. Result Matrix:

Model	Precision (Not Fraud)	Recall (Not Fraud)	F1-Score (Not Fraud)	Precision (Fraud)	Recall (Fraud)	F1-Score (Fraud)	Accuracy	AUC-ROC
Logistic Regression	1	1	1	0.91	0.68	0.78	0.999	0.93
Shallow Neural Network	1	1	1	0.84	0.82	0.83	0.999	0.99
Random Forest	1	1	1	0.89	0.57	0.69	0.999	0.95
Gradient Boosting	1	1	1	0.76	0.7	0.73	0.999	0.98
LinearSVC	1	1	1	0.77	0.82	0.79	0.999	N/A
Logistic Regression (Balanced)	0.96	1	0.98	1	0.95	0.98	0.98	0.99
Shallow Neural Network (Balanced)	0.96	1	0.98	1	0.95	0.98	0.98	0.99
Random Forest (Balanced)	0.9	1	0.95	1	0.89	0.94	0.94	0.98
Gradient Boosting (Balanced)	0.96	0.96	0.96	0.95	0.95	0.95	0.96	0.99
LinearSVC (Balanced)	0.96	1	0.98	1	0.95	0.98	0.98	N/A
Random Forest (Tuned)	0.94	1	0.97	1	0.93	0.96	0.97	0.99

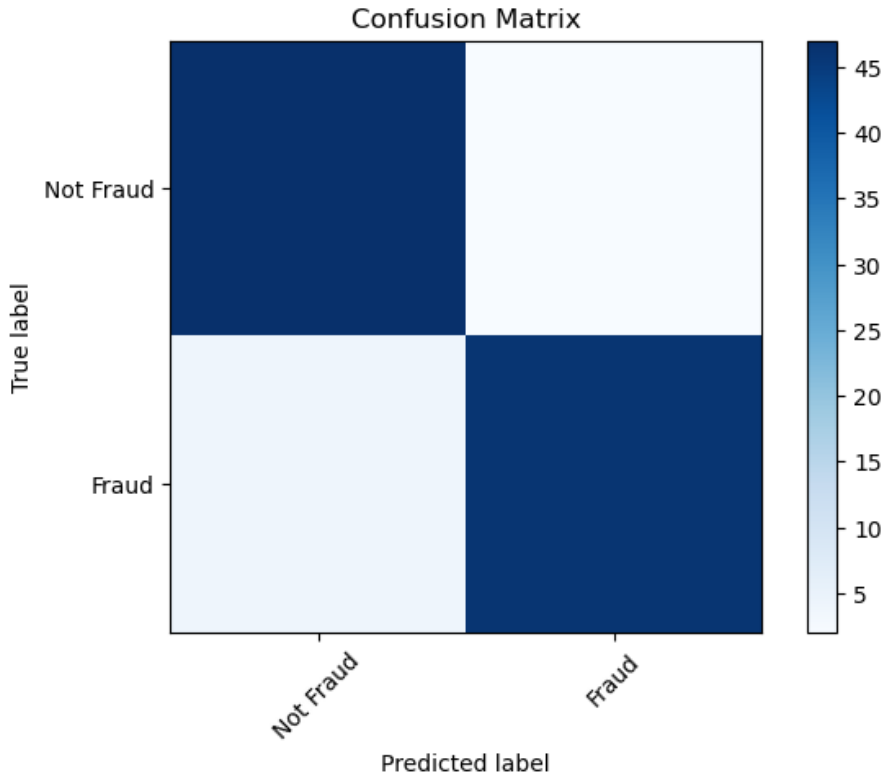
b. Model selection:

Among the various models tested, Logistic Regression was selected as the best model for the following reasons:

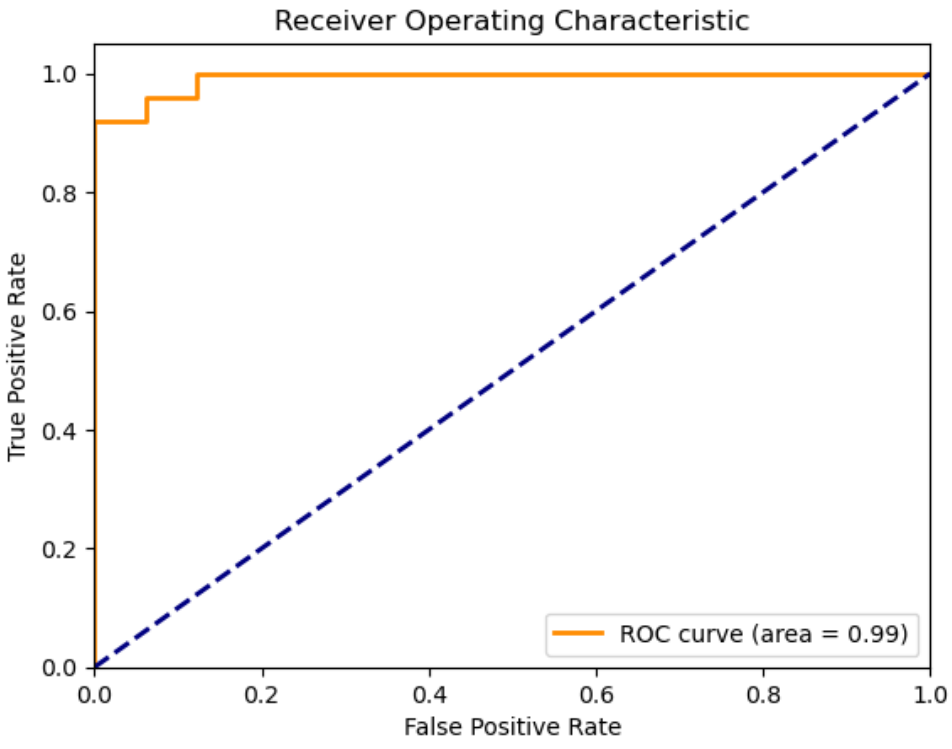
- **Interpretability**: Logistic Regression provides clear insights into which features are contributing to the prediction, making it easier to understand and trust the model.
- **Performance:** It demonstrated high accuracy, precision, recall, and AUC-ROC scores, indicating robust performance in detecting fraudulent transactions.
- **Efficiency:** Logistic Regression is computationally efficient, allowing for quicker training and predictions, which is beneficial for real-time fraud detection scenarios.
- **Scalability:** The simplicity and efficiency of Logistic Regression make it highly scalable for large datasets, ensuring it can handle the volume of transactions typically encountered in a real-world setting.

This decision balances performance with practicality, ensuring that the model is not only effective but also suitable for deployment in a production environment

Confusion Matrix of the chosen model:



ROC curve of the chosen model:



5. Conclusion

The project successfully demonstrates the application of various machine learning models for fraud detection in financial transactions. Key findings include:

- The shallow neural network and Linear SVC show strong performance on the imbalanced dataset, with high recall rates for detecting fraudulent transactions.
- On the balanced dataset, Logistic Regression and the shallow neural network achieve the highest F1-scores, making them the best models for this task.
- The tuned Random Forest classifier also performs well, indicating that hyperparameter optimization can further enhance model performance.

Future work could explore more advanced techniques such as deep learning models, ensemble methods, and real-time fraud detection systems.