Credit Card Fraud Detection Report

Indulekha K P

Abstract

This report presents a machine learning-based approach to detect fraudulent transactions in credit card data. We implemented and compared three classification models: Random Forest, Logistic Regression, and XGBoost. Each model was evaluated using various performance metrics, including accuracy, precision, recall, and F1-score. This report provides an overview of the methods used, data preprocessing steps, model implementations, results, and a final comparison to identify the best-performing model.

1 Introduction

Credit card fraud is a significant issue in the financial industry, leading to considerable financial losses for both consumers and institutions. This project aims to develop a machine learning-based solution to detect fraudulent credit card transactions. The dataset used contains transactions made by European cardholders in September 2013, focusing on identifying patterns associated with fraudulent activity.

2 Dataset Description

The dataset consists of 284,807 transactions, with 492 labeled as fraudulent. The features include 28 anonymized principal components (V1 to V28), transaction amount, transaction time, and the class label (0 = Non-Fraud, 1 = Fraud).

2.1 Data Preprocessing

- Handling missing values and outliers.
- Scaling numerical features for models sensitive to feature scale.
- Balancing the dataset using oversampling techniques (e.g., SMOTE).

3 Models Implemented

3.1 Random Forest

Random Forest is an ensemble method combining decision trees to improve accuracy and reduce overfitting. It provides feature importance insights and handles imbalanced datasets effectively.

3.2 Logistic Regression

Logistic Regression serves as a baseline model, offering simplicity and interpretability. Adjustments were made for class weights to handle imbalances.

3.3 XGBoost

XGBoost (Extreme Gradient Boosting) is a highly accurate and efficient boosting algorithm suitable for tabular data.

4 Evaluation Metrics

Given the highly imbalanced nature of the dataset, the following metrics were used:

- **Accuracy:** Overall correctness of the model.
- **Precision:** Proportion of correctly identified fraud cases among all predicted fraud cases.
- **Recall:** Proportion of correctly identified fraud cases among all actual fraud cases.
- **F1-Score:** Harmonic mean of precision and recall.

5 Results and Visualizations

5.1 Dataset Insights

The dataset is highly imbalanced, as depicted in Figures 1 and 2, with fraudulent transactions constituting only a small fraction of the total. Figure 3 provides a correlation matrix highlighting relationships between features, which aids in identifying key predictors for fraud detection.

Figure 1: Distribution of Fraud and Non-Fraud (Linear Scale)

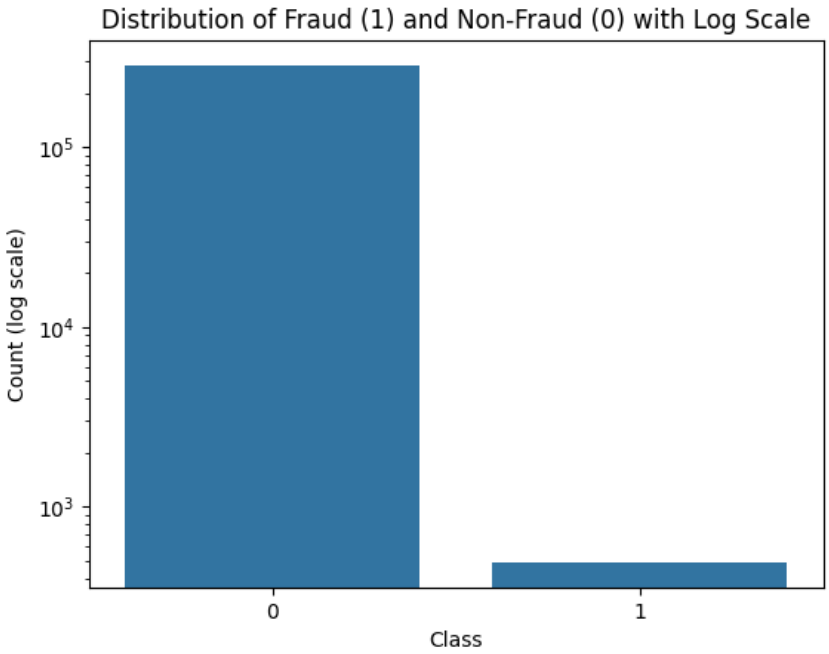

Figure 2: Distribution of Fraud and Non-Fraud (Log Scale)

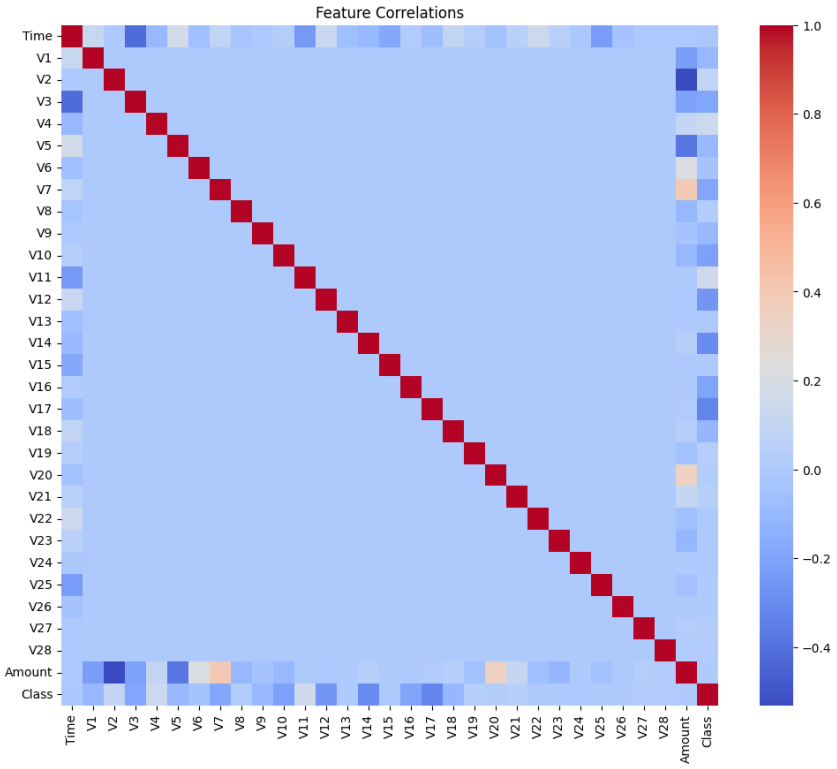

Figure 3: Feature Correlation Matrix

5.2 Model-Specific Results

5.2.1 Logistic Regression

Figures 4, 5, and 6 illustrate the performance of the Logistic Regression model. While it achieves a high recall, effectively identifying most fraudulent transactions, its low precision indicates a higher number of false positives, as seen in the confusion matrix and precision-recall curve.

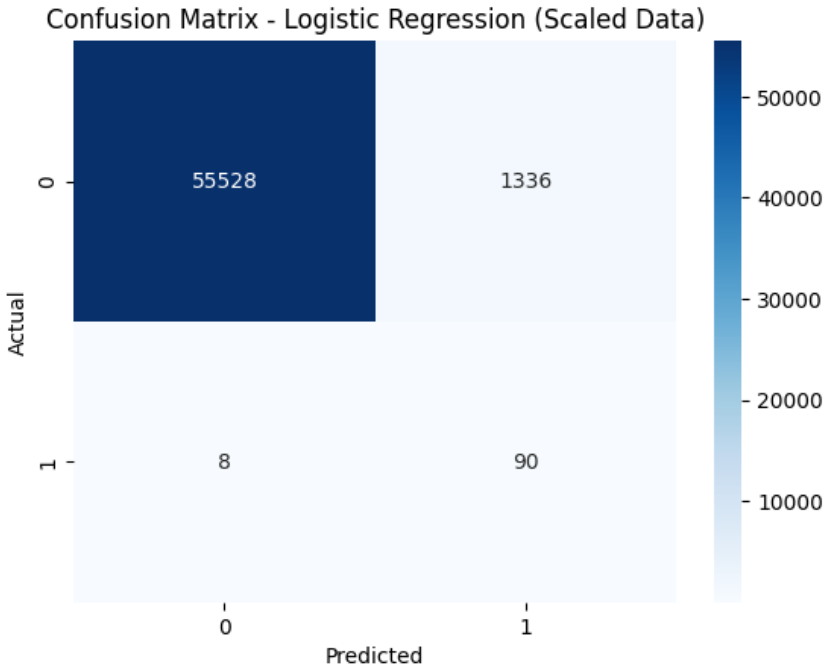

Figure 4: Confusion Matrix - Logistic Regression

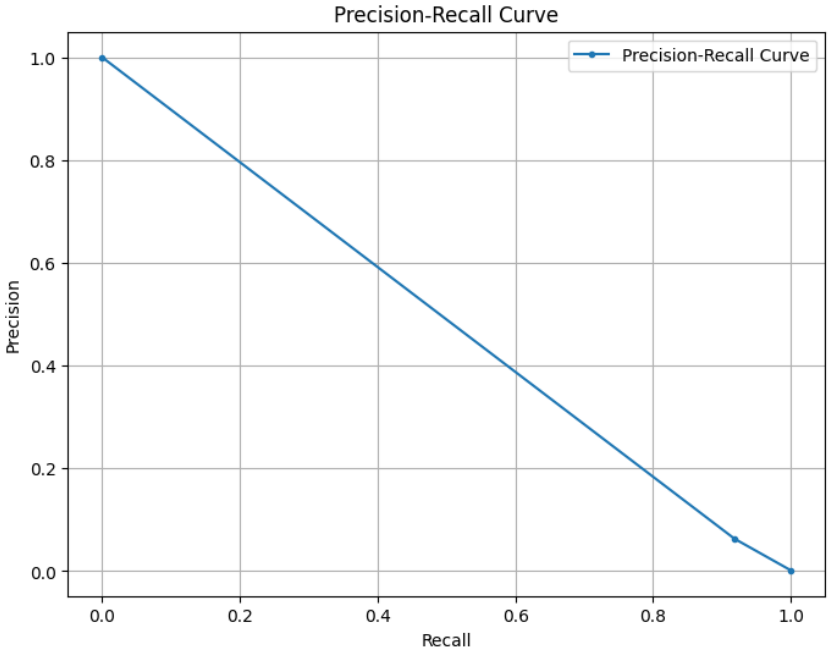

Figure 5: Precision-Recall Curve - Logistic Regression

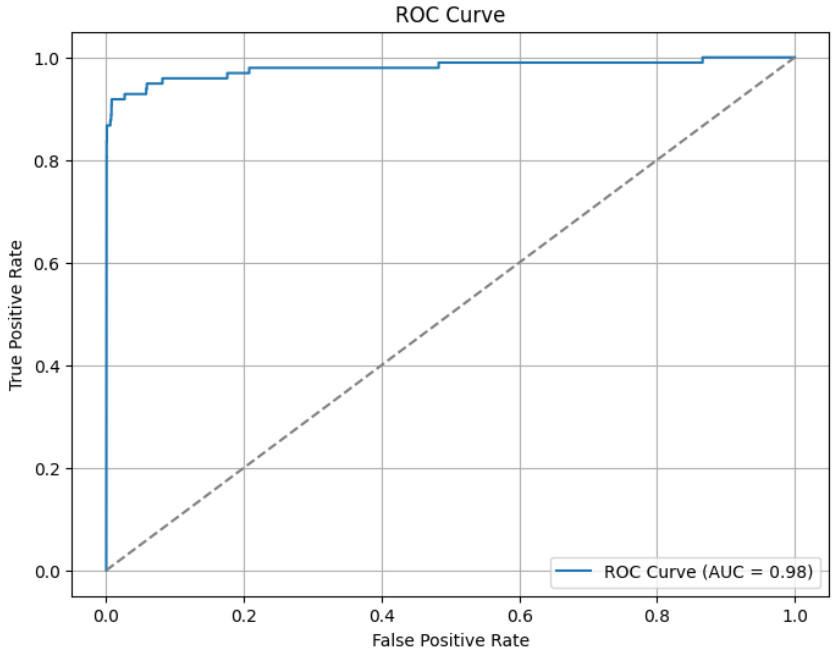

Figure 6: ROC Curve - Logistic Regression

5.2.2 Random Forest

Figures 8, 9, and 10 demonstrate the strong performance of the Random Forest model, particularly in precision, as it minimizes false positives. Figure 7 highlights the importance of key features, showcasing the model's ability to identify influential predictors for fraud detection.

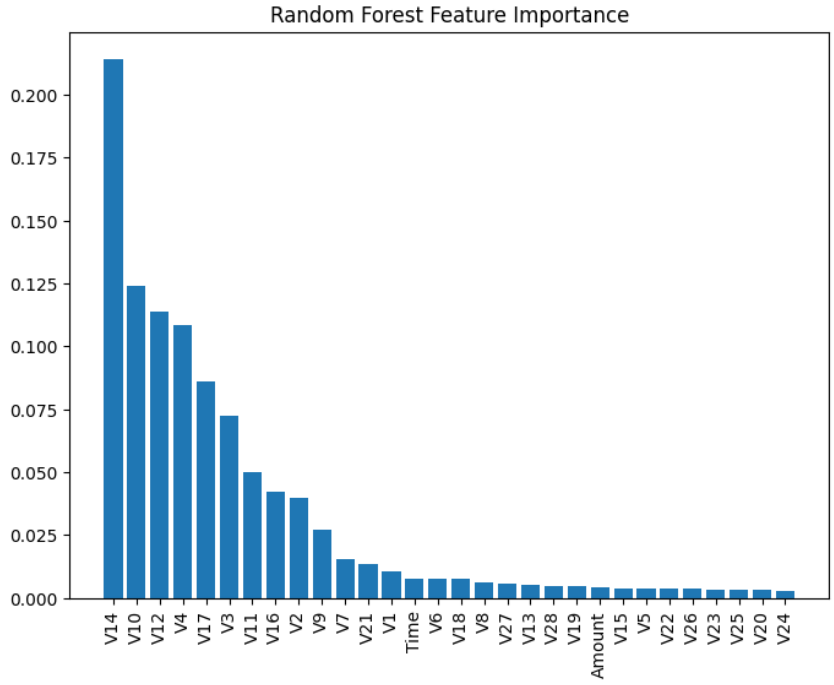

Figure 7: Feature importance - Random Forest

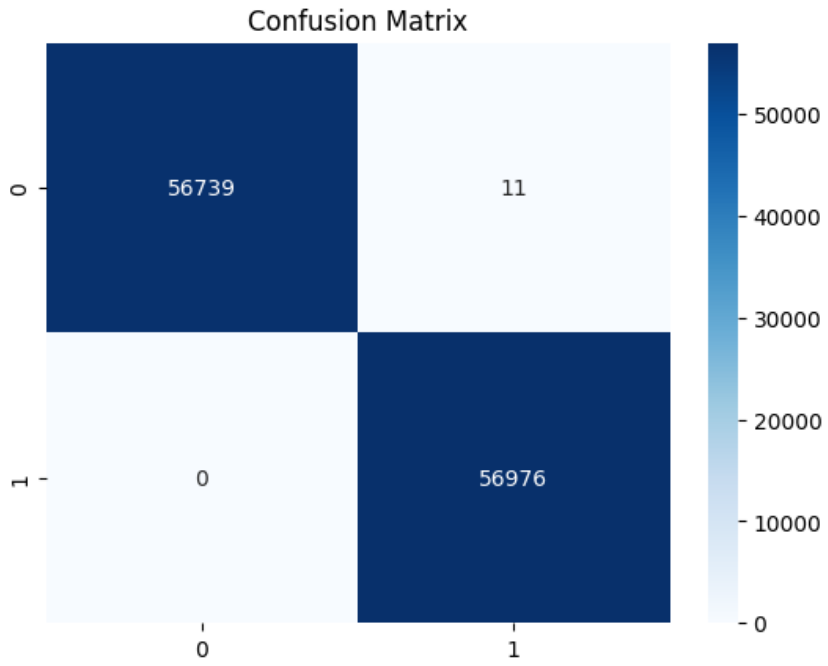

Figure 8: Confusion Matrix - Random Forest

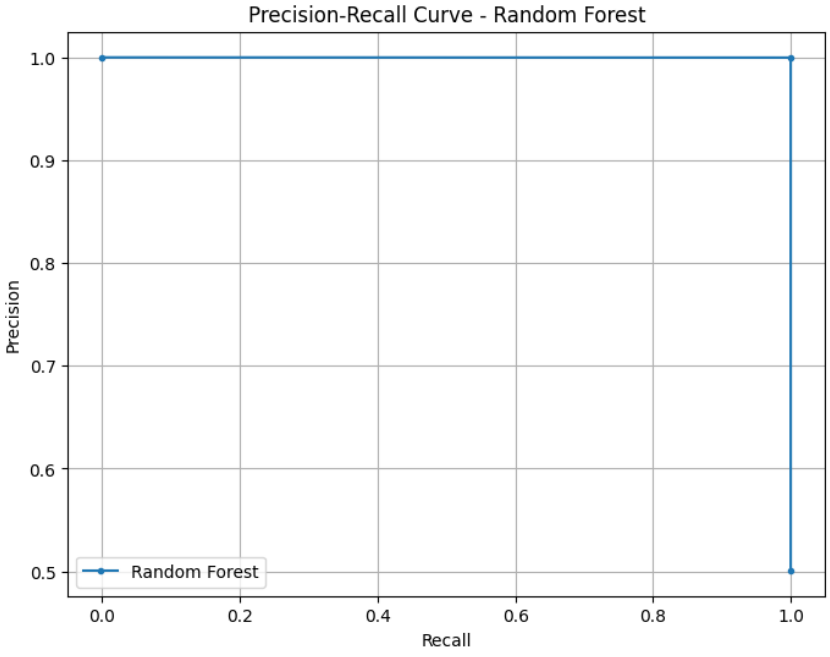

Figure 9: Precision-Recall Curve - Random Forest

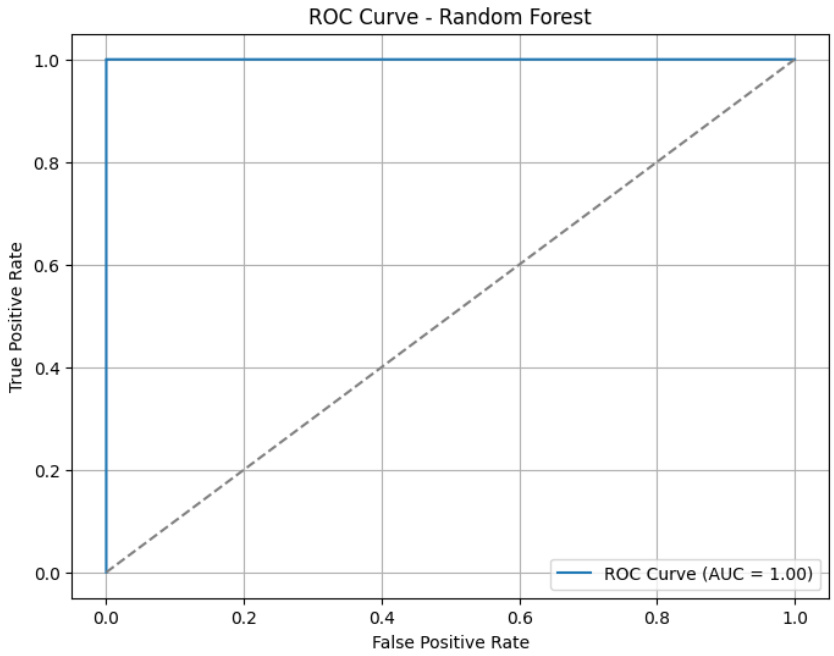

Figure 10: ROC Curve - Random Forest

5.2.3 XGBoost

Figures 12, 13, and 14 highlight the excellent performance of the XGBoost model, achieving a strong balance between precision and recall. Figure 11 showcases the feature importance, demonstrating how XGBoost effectively identifies key predictors for fraud detection, contributing to its superior F1-score.

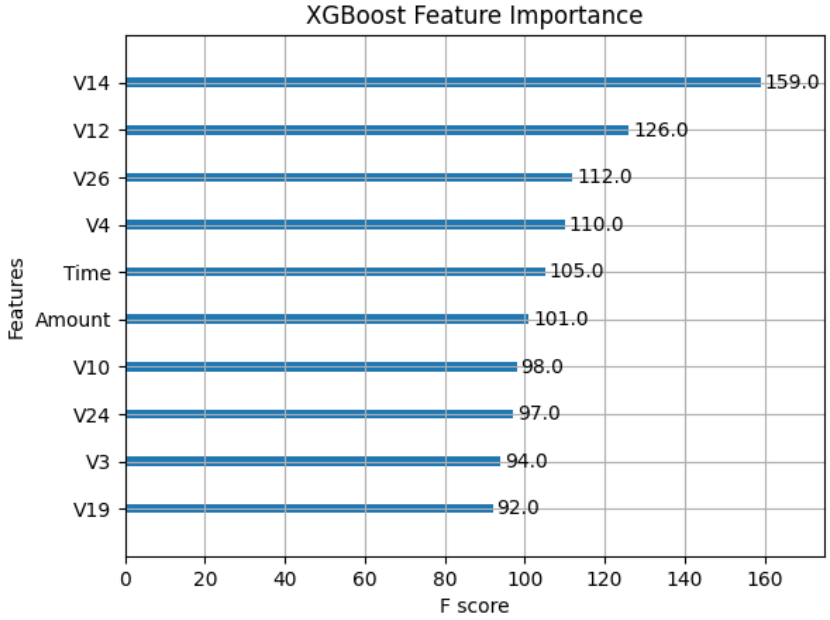

Figure 11: Feature importance - XGBoost

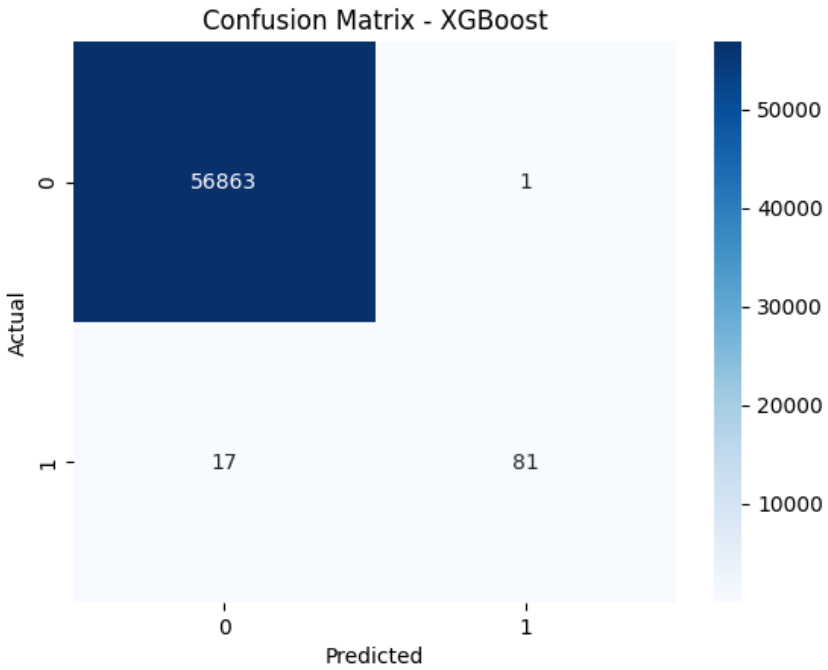

Figure 12: Confusion Matrix - XGBoost

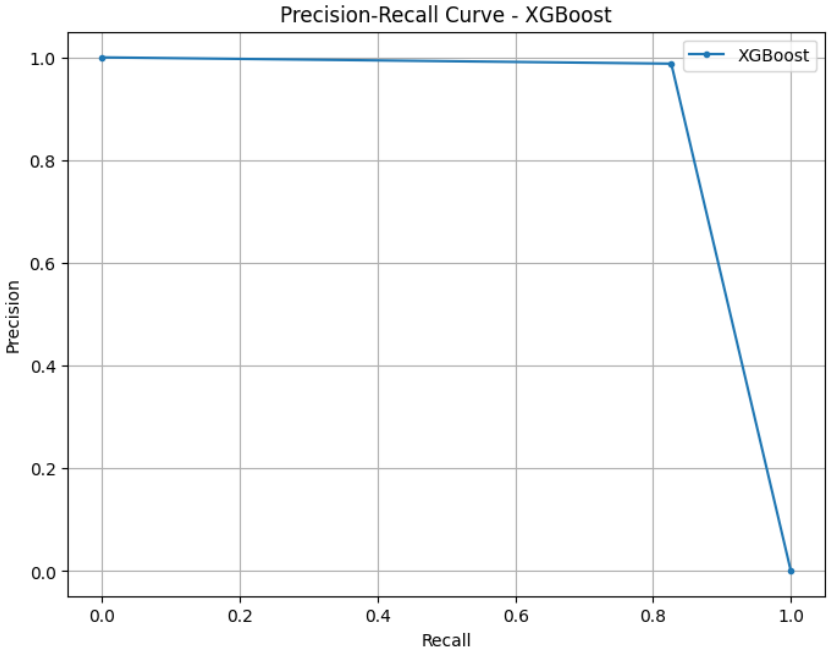

Figure 13: Precision-Recall Curve - XGBoost

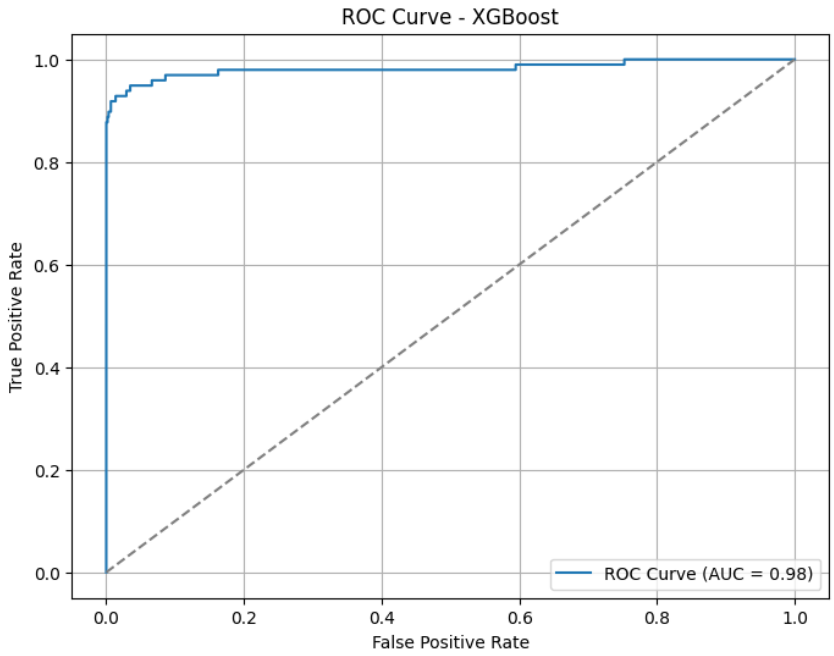

Figure 14: ROC Curve - XGBoost

5.3 Model Comparison

Model comparison is an integral part of evaluating the overall performance of different machine learning algorithms. The following figures provide visual insights into how Logistic Regression, Random Forest, and XGBoost perform in terms of accuracy, precision-recall, and ROC curves.

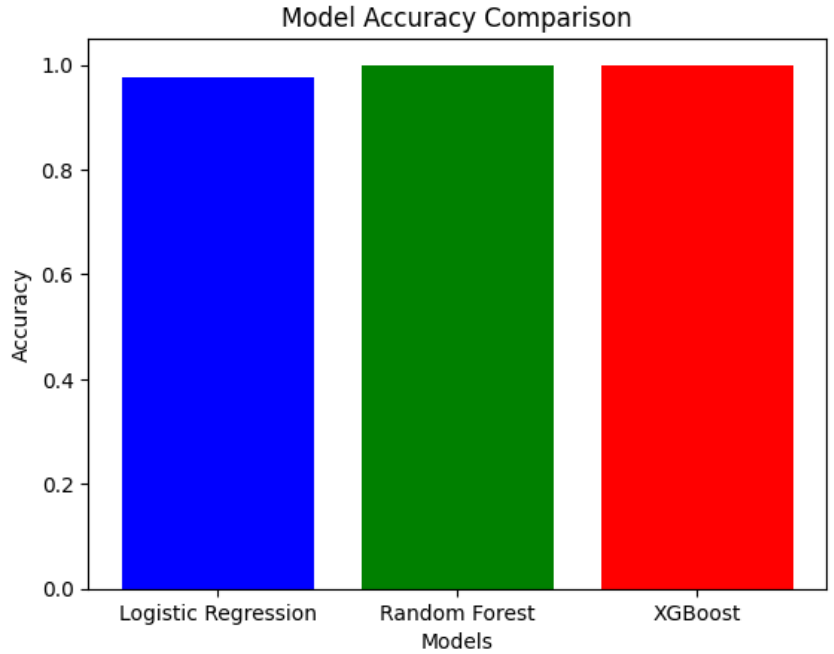

Figure 15: Model Accuracy Comparison

Figure 15 shows the accuracy of the three models. While both Random Forest and

XGBoost achieve nearly identical high accuracy levels, Logistic Regression lags slightly behind. However, in an imbalanced dataset, high accuracy can be misleading and may not reflect the model's ability to identify the minority class (fraudulent transactions).

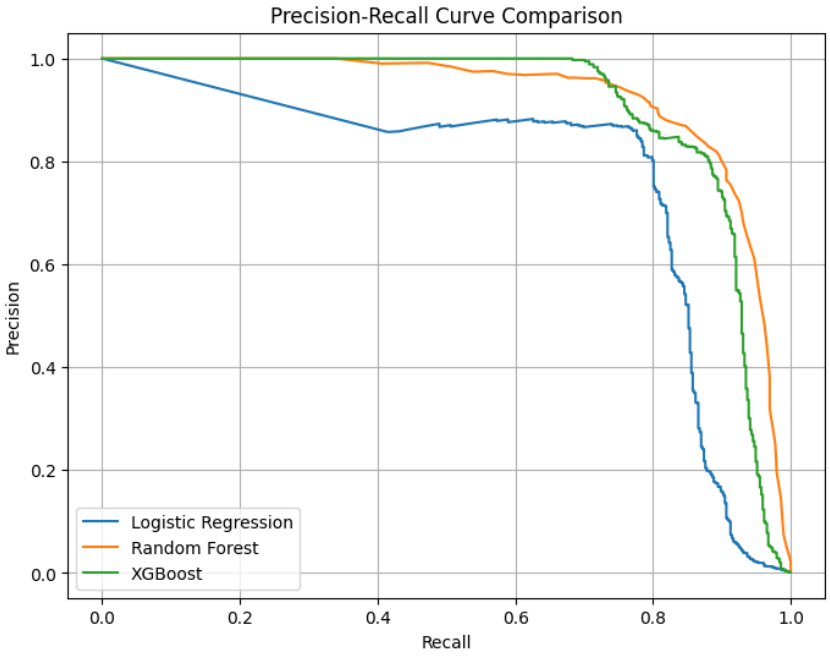

Figure 16: Precision-Recall Curve Comparison

The Precision-Recall Curve in Figure 16 highlights the trade-off between precision and recall for each model. XGBoost exhibits the most balanced curve, indicating its ability to achieve a strong trade-off between minimizing false positives and maximizing true positives. Random Forest focuses more on precision, while Logistic Regression maximizes recall, leading to higher false positives.

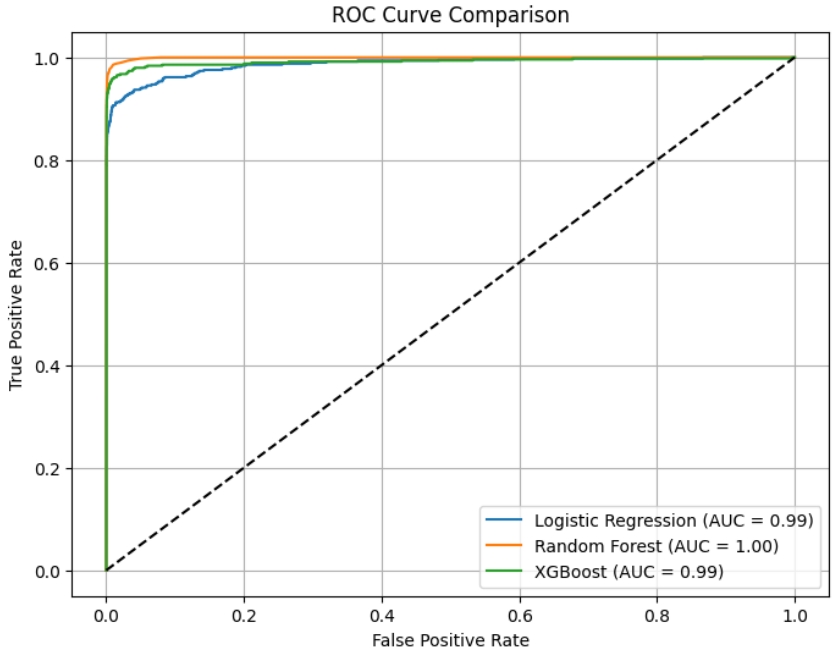

Figure 17: ROC Curve Comparison

Figure 17 depicts the Receiver Operating Characteristic (ROC) curves for the three models. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity). XGBoost and Random Forest demonstrate near-perfect area under the curve (AUC) values, indicating their superior performance in distinguishing between fraudulent and non-fraudulent transactions. Logistic Regression, while effective, does not perform as well as the other two models in this metric.
In summary, the visualizations above provide clear evidence that XGBoost offers the most balanced performance across metrics, followed by Random Forest, which prioritizes precision, and Logistic Regression, which emphasizes recall. These insights are crucial in selecting the appropriate model based on the specific application requirements, such as minimizing false alarms or maximizing detection rates.

5.4 Performance Metrics Comparison

The following table summarizes the performance metrics for each model, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of the models' performance, particularly in handling the imbalanced dataset where fraudulent transactions constitute a small proportion of the total.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.976560	0.063647	0.916667	0.119029
Random Forest	0.999473	0.952381	0.731707	0.827586
XGBoost	0.999442	0.832335	0.847561	0.839879

Table 1: Performance metrics comparison across Logistic Regression, Random Forest, and XGBoost models.

From the table, the following observations can be made:

- **Accuracy:** Random Forest and XGBoost achieved very high accuracy (99.94%), significantly outperforming Logistic Regression. However, accuracy alone is not sufficient for evaluating performance in an imbalanced dataset.
- **Precision:** Random Forest had the highest precision (95.24%), indicating that it was the most effective in correctly identifying fraudulent transactions without generating too many false positives.
- **Recall:** Logistic Regression achieved the highest recall (91.67%), showing its ability to identify the majority of fraudulent cases. However, this came at the cost of a very low precision, resulting in a large number of false positives.
- **F1-Score:** XGBoost achieved the highest F1-score (0.839879), balancing both precision and recall, making it the most suitable model for the dataset.

This detailed comparison highlights the strengths and weaknesses of each model. Logistic Regression excels in detecting fraud cases but lacks precision, resulting in a high number of false positives. Random Forest, on the other hand, achieves the highest precision, making it suitable for scenarios where minimizing false positives is critical. Finally, XGBoost provides a balanced trade-off between precision and recall, making it the most reliable model for this dataset.

6 Conclusion

Among the three models tested, XGBoost performed the best overall in terms of precision, recall, and F1-score. This model is recommended for real-world deployment due to its high accuracy and ability to minimize false positives.

7 Discussion of Future Work

While the current implementation achieves commendable results in detecting fraudulent credit card transactions, there are several areas for future improvement and exploration:

- **Real-Time Fraud Detection:** Extend the system to process transaction data in real-time by integrating streaming pipelines and low-latency model inference for immediate fraud detection.
- **Deep Learning Models:** Investigate advanced deep learning techniques, such as Recurrent Neural Networks (RNNs) or Transformer-based models, to capture complex temporal and sequential patterns in transaction data.
- **Explainable AI (XAI):** Implement explainability techniques to provide insights into the model's decisions, ensuring transparency and building trust among stakeholders.

These directions will further enhance the robustness and applicability of fraud detection systems in dynamic and high-stakes financial environments.