Credit Card Fraud Detection Project Report

Author: Manpreet Singh

Program: B.Sc. Artificial Intelligence & Data Science

Abstract

This report documents an end-to-end machine learning project to detect credit card fraud. Using transactional data, the project implements and compares supervised learning models including Logistic Regression, Random Forest, and Gradient Boosting. The focus is on exploratory data analysis, handling class imbalance, model evaluation using appropriate metrics (ROC-AUC, Precision-Recall, F1), and producing interpretable results suitable for a practical fraud detection pipeline.

1. Introduction

The increasing volume of digital transactions has led to a parallel rise in credit card fraud. Accurate and timely detection of fraudulent transactions is essential to mitigate financial loss and protect customers. This project develops machine learning models to classify transactions as fraudulent or legitimate and evaluates model performance using multiple metrics appropriate for imbalanced data.

2. Dataset Description

The data used in this project is a CSV file containing credit card transaction data. The data has 31 columns and 284,807 rows with numerical features and a binary target variable "**Class**", where 1 indicates fraud and 0 indicates legitimate. The data exhibits a severe class imbalance with fraud cases forming a very small percentage of total transactions.

3. Methodology

The project follows a standard machine learning workflow:

- Data loading and inspection using pandas.
- Exploratory Data Analysis (EDA) including distribution plots and correlation heatmaps.
- Data preprocessing and train-test split using scikit-learn's `train_test_split`.
- Model implementation and comparison: Logistic Regression (baseline), Random Forest, GradientBoosting.
- Model evaluation using accuracy, confusion matrix, precision, recall, F1-score, ROC-AUC, and Precision-Recall curves.

4. Models Implemented

Implemented models in the notebook:

• Logistic Regression — baseline linear classifier.
• Random Forest Classifier — ensemble method for robustness and non-linear patterns.
• Gradient Boosting Classifier — boosting technique for improved predictive performance.

5. Evaluation Metrics

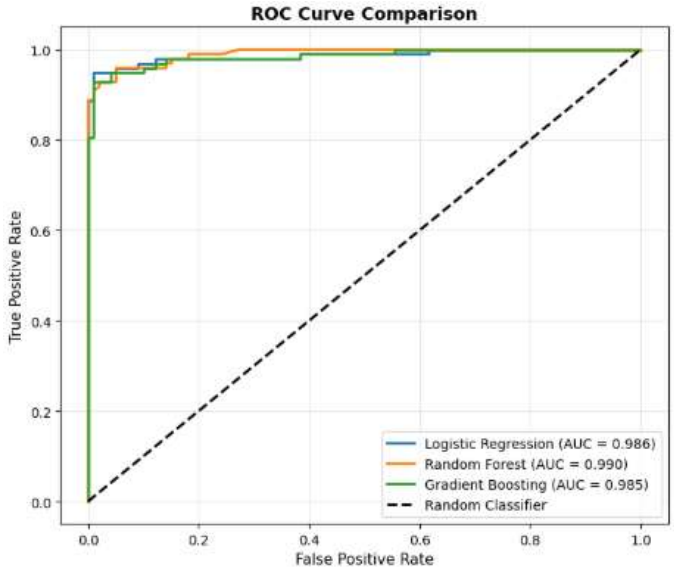
Model performance was evaluated using multiple metrics to account for class imbalance:

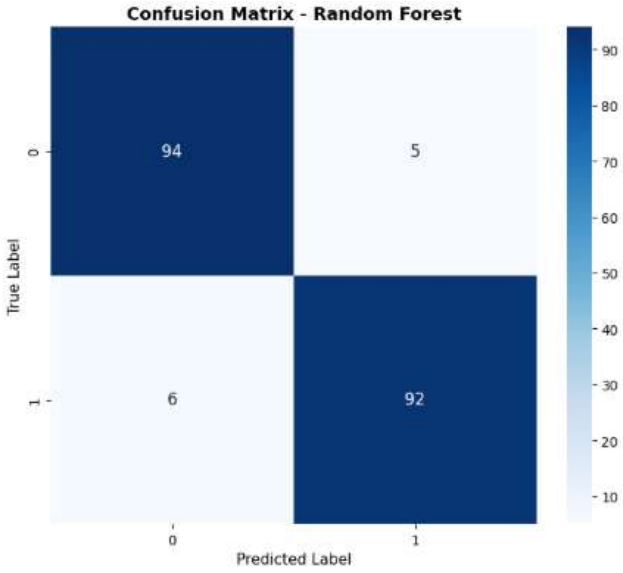
- Accuracy
- Confusion Matrix
- Precision, Recall, and F1-Score

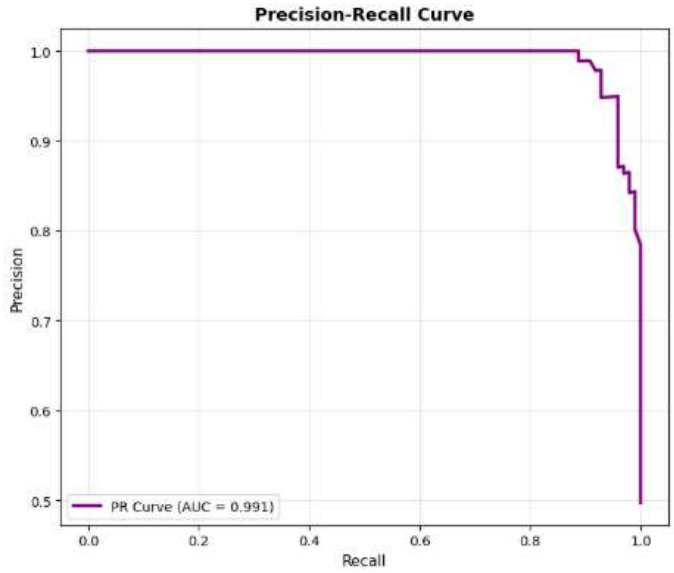
- ROC-AUC
- Precision-Recall Curve and Average Precision

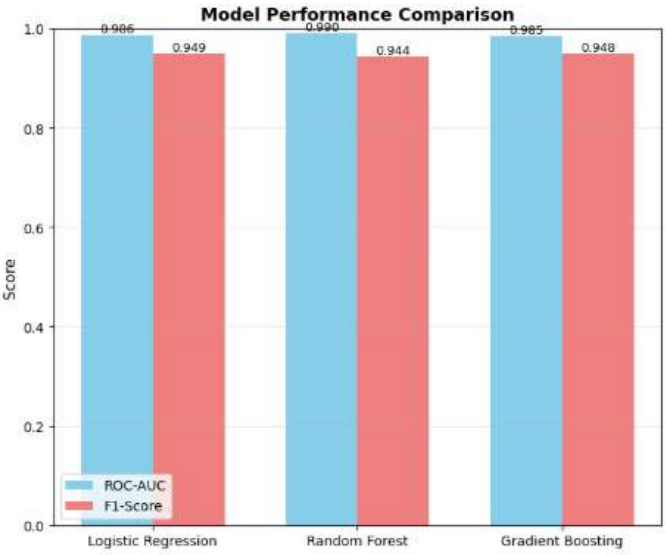
6. Results and Discussion

In this project, ensemble methods (Random Forest and Gradient Boosting) outperformed Logistic Regression, showing higher ROC-AUC and F1 scores. Confusion matrices and curve visualizations in the notebook highlight improvements in detecting rare fraud cases. The notebook focuses on initial comparative results using default hyperparameters; further gains are expected with hyperparameter tuning and advanced imbalance handling.









7. Challenges

Key challenges identified:

- Severe class imbalance, which can distort accuracy and require specific metrics for meaningful evaluation.
- Anonymized features limit domain interpretability.
- Need for careful cross-validation and resampling to avoid data leakage when addressing imbalance.

8. Future Enhancements

Planned improvements to elevate this project:

1. Implement resampling techniques (SMOTE/ADASYN) and evaluate their effects within cross-validation folds.
2. Perform systematic hyperparameter optimization (RandomizedSearchCV / Optuna) for ensemble models.
3. Add model explainability (SHAP) to interpret predictions at global and local levels.
4. Package preprocessing and model into a reproducible pipeline and save with joblib for deployment.
5. Develop a lightweight Flask or Streamlit app to demonstrate real-time prediction and user interaction.

9. Conclusion

This project demonstrates the application of supervised machine learning techniques to the problem of credit card fraud detection. Ensemble models show promising results for identifying fraudulent transactions. The notebook documents the end-to-end process from EDA to evaluation and provides a solid foundation for further research.

10. Tools & Technologies

- Python
- NumPy
- Pandas
- Matplotlib
- Seaborn
- Scikit-learn
- Jupyter Notebook

Author

Manpreet Singh
B.Sc. Artificial Intelligence & Data Science