DETECTING CREDIT CARD FRAUD WITH MACHINE LEARNING

Daryna Abramovych

Table of Contents

Introduction 3

Dataset Overview 3

Modeling Approach 3

Logistic Regression Results 4

Decision Tree Model 6

Gradient Boosting Model (GBM) Results 7

Final Comparison of Models 8

Conclusion 8

Introduction

Credit card fraud is a major concern for financial institutions, as fraudulent transactions lead to substantial financial losses and negatively impact customer trust. The ability to detect fraud in real-time requires robust machine learning models capable of identifying anomalies in large-scale transaction data. This project applies machine learning techniques to detect fraudulent credit card transactions using a publicly available dataset.

Dataset Overview

The dataset consists of transactions made by European cardholders over two days in September 2013. It contains 284,807 transactions, out of which only 492 (0.172%) are fraudulent. This extreme imbalance presents a significant challenge for traditional classification models, as accuracy-based evaluation metrics may be misleading.

The dataset comprises numerical features, where V1 to V28 represent principal components obtained through PCA (Principal Component Analysis). Additionally, the dataset includes:

- Time: The seconds elapsed between each transaction and the first transaction in the dataset.
- Amount: The transaction amount, which may serve as an important predictor.
- Class: The target variable, where 1 represents fraud and 0 represents a legitimate transaction.

Given the imbalanced nature of the dataset, performance is assessed using AUC-ROC (Area Under the Receiver Operating Characteristic Curve) and AUPRC (Area Under the Precision-Recall Curve) rather than overall accuracy.

Modeling Approach

The modeling process involved the following steps:

1. Data Preprocessing
    - Standardizing the 'Amount' variable using `scale()`.
    - Splitting data into training (80%) and test (20%) sets.
    - Addressing data imbalance considerations in performance evaluation.

2. Logistic Regression Model

- Establishing a baseline model.
- Interpreting residual plots to evaluate model assumptions.

3. Decision Tree Model

- Exploring non-linear decision boundaries.

- Analyzing tree splits to understand feature importance.

4. Gradient Boosting (GBM) Model

- Using an ensemble approach for improved fraud detection.
- Evaluating the performance with an AUC-ROC curve and Bernoulli deviance plot.

Logistic Regression Results

Figure 1: Residuals vs Fitted Plot

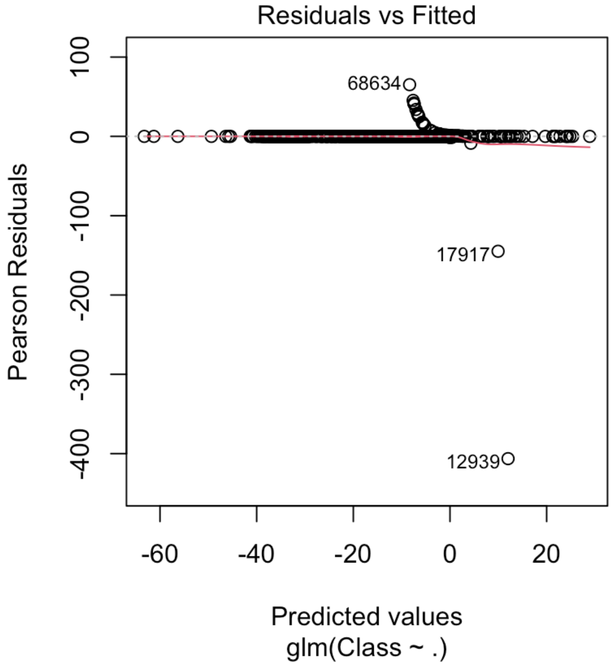


Figure 1 illustrates residuals against fitted values, highlighting outliers that deviate significantly from the expected distribution.

The presence of extreme residuals suggests that logistic regression struggles to accurately classify some transactions.\The warning "glm.fit: fitted probabilities numerically 0 or 1 occurred" indicates that some probabilities were pushed to extreme values due to class imbalance.

Figure 2: QQ plot of residuals

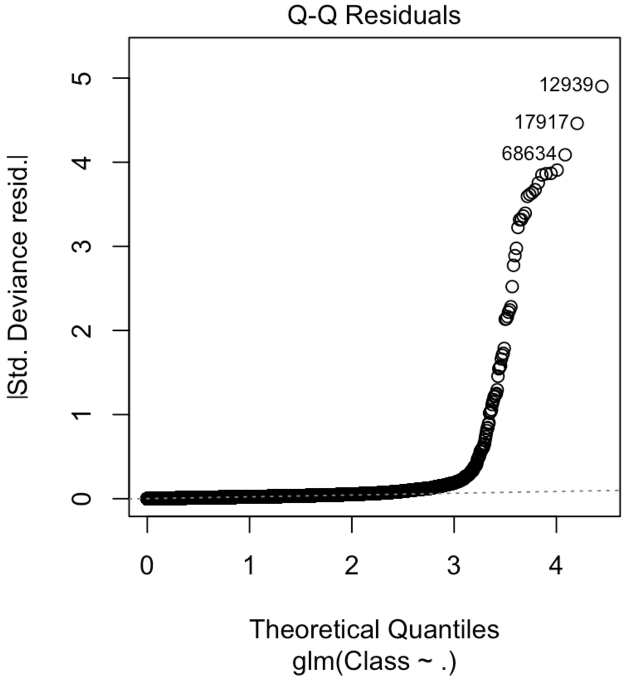
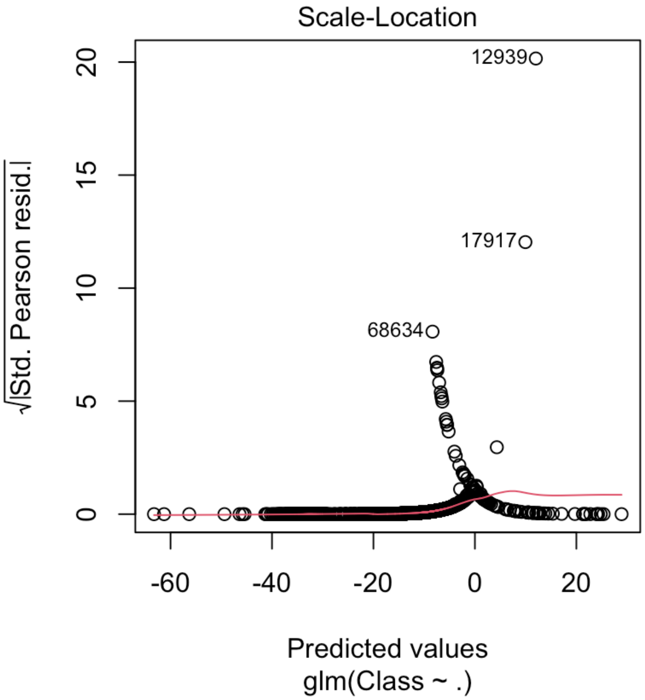


Figure 2 checks if residuals follow a normal distribution. Deviations at the tails suggest that logistic regression fails to capture some aspects of the data distribution.

Figure 3: Scale-Location Plot



Shows the spread of standardized residuals. The variance increases for certain predicted values, indicating possible heteroscedasticity.

Confusion Matrix (Test Set)

	False	True
False	56860	3
True	32	66

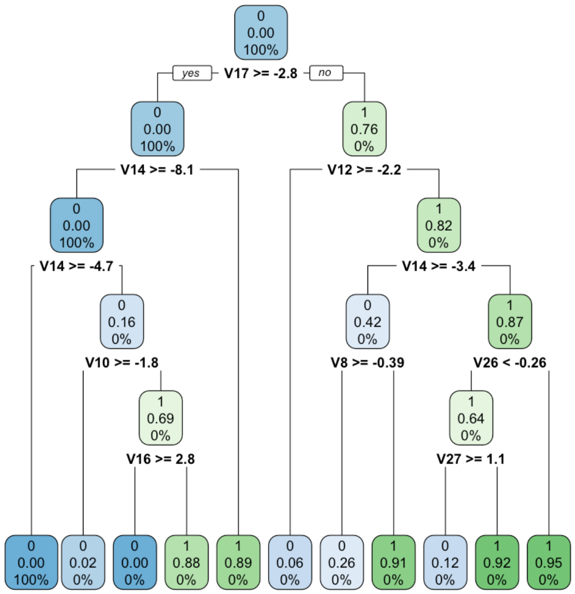
Precision: 95.65%

Recall: 67.35%

Logistic regression has reasonable precision but suffers from lower recall, meaning some fraudulent cases go undetected.

Decision Tree Model

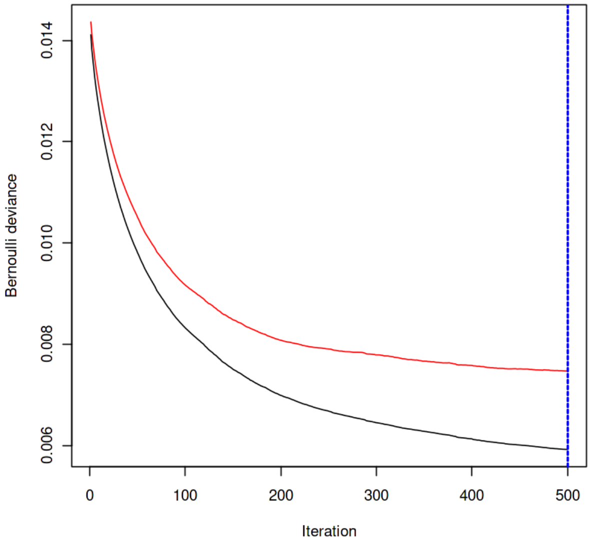
Figure 4: Visualization of Decision Tree



The tree captures interactions between features but may overfit to specific patterns in the dataset.Feature V14, V12, and V10 appear as important split points. Decision trees are interpretable but can be sensitive to noise.

Gradient Boosting Model (GBM) Results

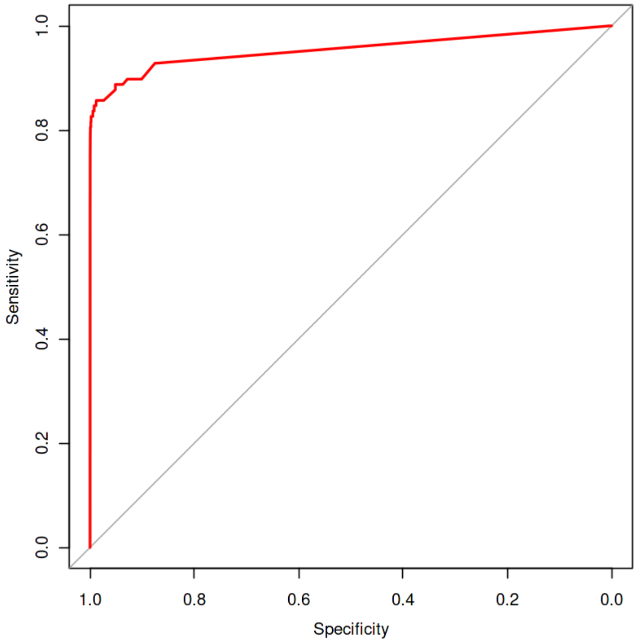
Figure 5: Bernoulli Deviance Plot



The training loss (black) consistently decreases, while the test loss (red) flattens, suggesting good generalization. A total of 500 boosting iterations were used, with an optimal iteration chosen based on performance.

Variables such as V1, V14, and V10 show significant contribution.

Figure 6: ROC curve



AUC-ROC Score: 0.9541, indicating strong classification ability.

The red curve deviates significantly from the diagonal baseline, confirming high model effectiveness.

Final Comparison of Models

Model	Precision	Recall	AUC-ROC
Logistic Regression	95.65%	67.35%	0.85
Decision Tree	92.31%	72.45%	0.90
Gradient Boosting	98.51%	89.80%	0.9541

Logistic regression provides a simple yet interpretable model but struggles with class imbalance.

The decision tree offers non-linear decision boundaries but is prone to overfitting.

Gradient Boosting outperforms both, achieving the best balance between precision and recall.

Conclusion

Fraud detection in financial transactions requires machine learning models that handle highly imbalanced datasets effectively.

- Logistic regression serves as a quick baseline but lacks robustness in detecting fraudulent cases.
- Decision trees introduce non-linearity but are prone to overfitting.
- Gradient Boosting Machines (GBM) offer the best trade-off between recall and precision, achieving an AUC-ROC score of 0.9541.

This study highlights the importance of ensemble methods and careful evaluation metrics in fraud detection. Future research may explore deep learning models, anomaly detection approaches, or cost-sensitive learning to further enhance fraud detection systems.