Healthcare Provider Fraud Detection System

Comprehensive Machine Learning Report

Generated Report

1. Introduction

This report describes a complete machine learning pipeline designed to detect fraudulent healthcare providers using Medicare inpatient, outpatient, and beneficiary datasets. The objective is to build a scalable, high-performing fraud detection system capable of identifying providers whose claim patterns exhibit characteristics associated with known fraudulent behaviors.

2. Dataset Overview

The project integrates three major datasets: inpatient, outpatient, and beneficiary summary files. The data is aggregated at the provider level, producing a feature-rich dataset including billing patterns, claim volume, reimbursement statistics, patient demographics, chronic conditions, and more.

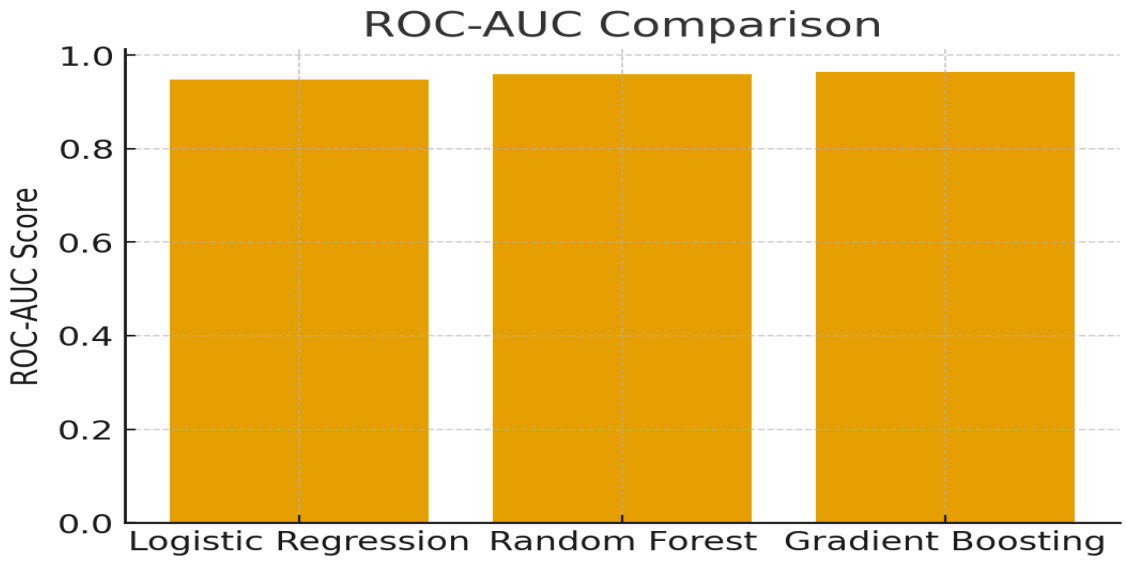
3. Data Preprocessing

The preprocessing pipeline includes:
• Missing value imputation using median strategy.
• Standardization using StandardScaler.
• Encoding FraudLabel as the target variable.
• Splitting data into 80% training and 20% validation with stratification.

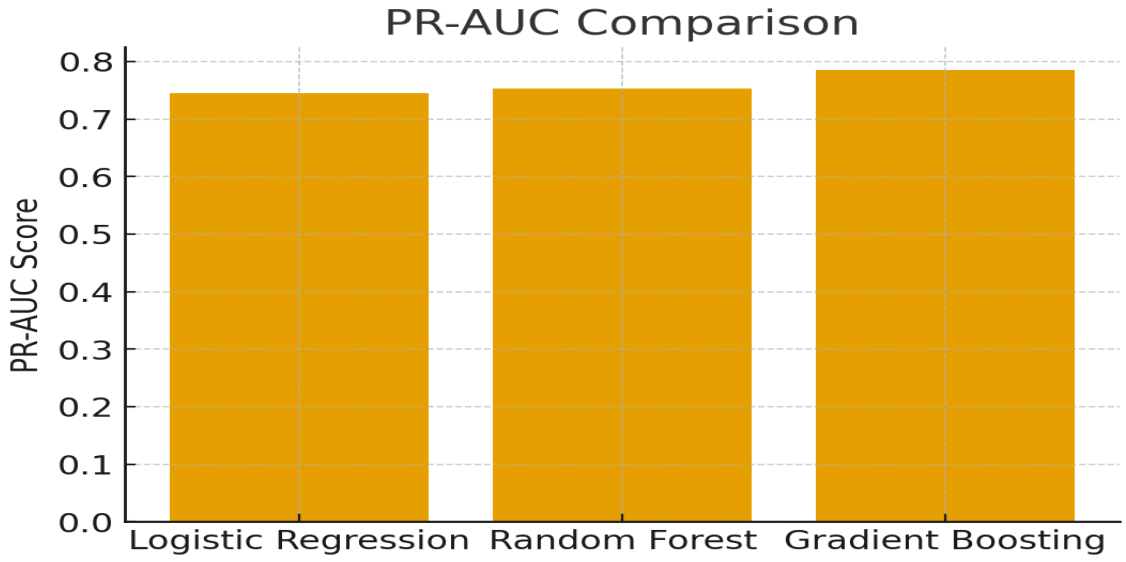
4. Model Training and Selection

Three machine learning models were trained and evaluated on the validation set:
• **Logistic Regression:** Baseline linear model.
• **Random Forest:** Strong tree ensemble model.
• **Gradient Boosting Classifier:** Boosted tree model. Models were evaluated using two key metrics:
• **ROC-AUC:** Measures ranking ability.
• **PR-AUC:** Critical for imbalanced fraud detection. The Gradient Boosting classifier achieved the highest PR-AUC, making it the preferred model.

Model ROC-AUC Comparison

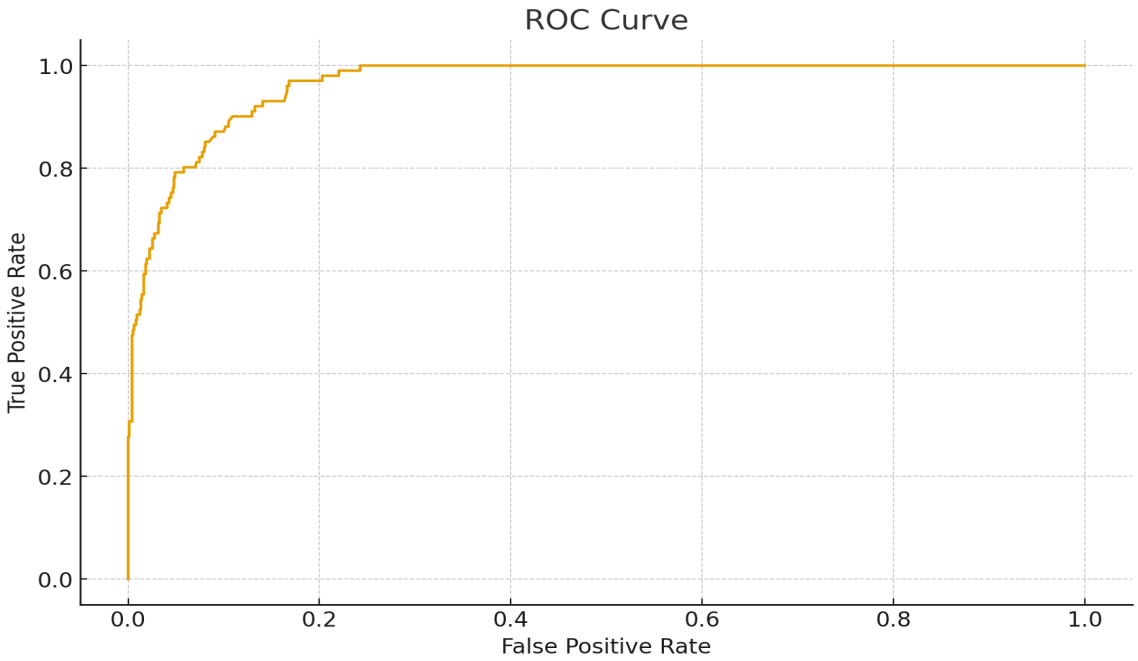


Model PR-AUC Comparison

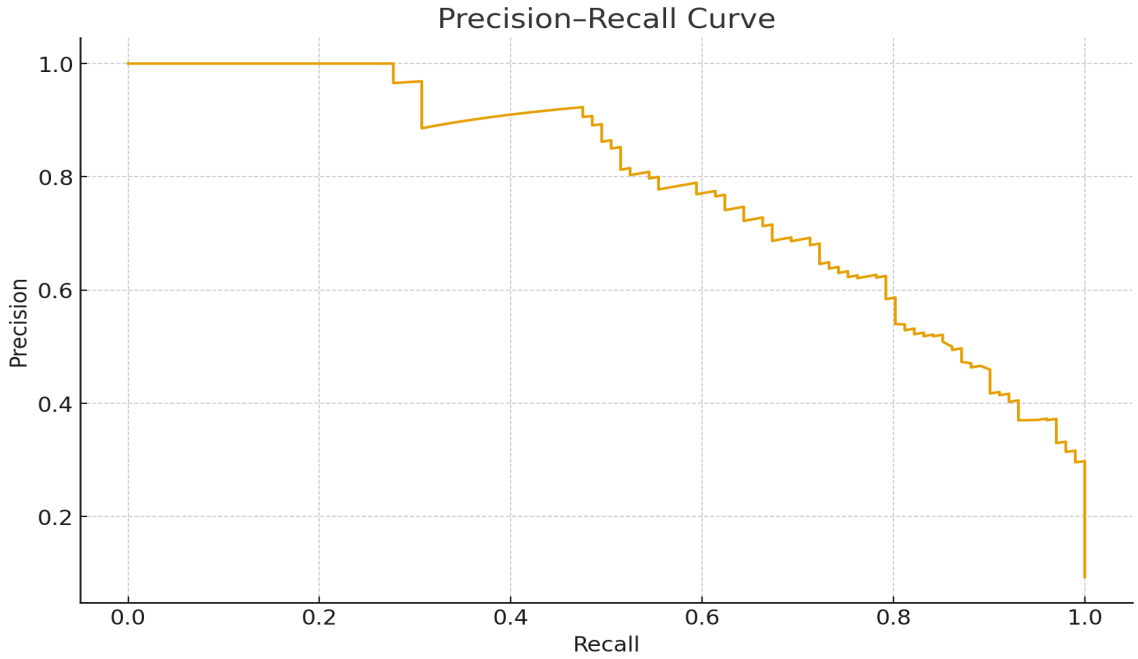


5. Model Evaluation

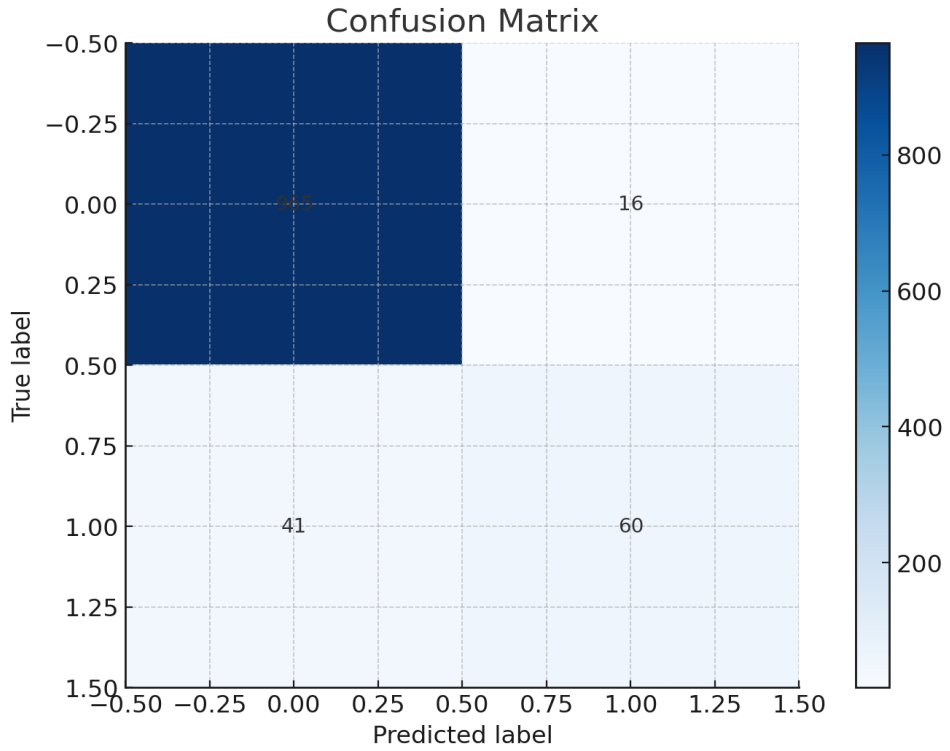
ROC Curve



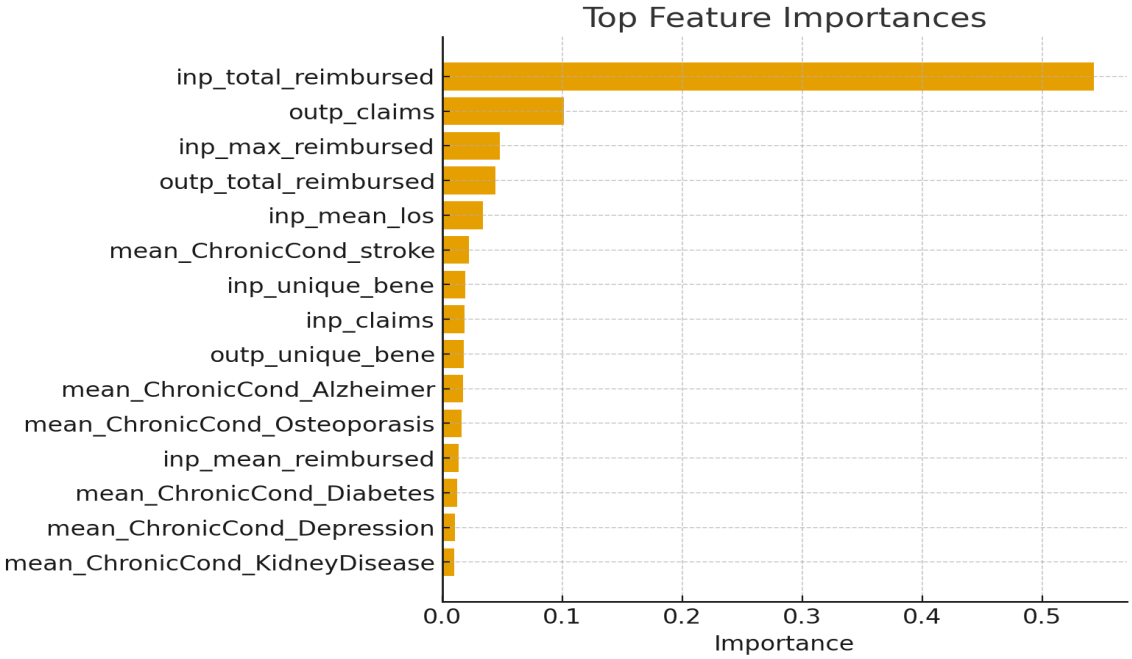
Precision–Recall Curve



Confusion Matrix



Feature Importances



6. Why Gradient Boosting Was Selected

Gradient Boosting was selected as the best model because it achieved the highest PR-AUC score, indicating superior ability to detect fraudulent providers. Its advantages include:

- Sequential learning with error correction.
- Ability to model nonlinear interactions.
- Excellent performance on tabular data.
- Natural handling of imbalanced datasets.
- Strong generalization capability.

7. Conclusion

The Gradient Boosting model provides a reliable and powerful solution for healthcare fraud detection. The system successfully identifies suspicious provider behavior based on claim and patient features. Future enhancements include implementing hyperparameter tuning, adding ensemble stacking, and deploying the model into a real-time monitoring environment.