Technical Report

Medicare Fraud Detection Project

1. Introduction

This project aims to detect healthcare providers who might be involved in fraudulent Medicare activities. Fraud causes large financial losses every year, so having a model that can flag suspicious providers helps investigators focus on the right cases. We used four datasets: beneficiary data, inpatient claims, outpatient claims, and provider labels.

2. Data Overview

We worked with four training datasets:

- Beneficiary data: patient details, chronic conditions, Medicare coverage.

- Inpatient data: hospital claims, reimbursements, patient and provider IDs.

- Outpatient data: non-hospital claims and reimbursements.

- Label data: provider and whether they are marked as fraud.

The main IDs linking data are BeneID and Provider. Fraud cases are rare, so the data is imbalanced.

3. Data Exploration

We checked dataset shapes, missing values, and column types. Fraud was only around 8–10 percent of the data. During EDA, fraudulent providers usually had more total claims, higher reimbursements, and more beneficiaries. These patterns helped guide feature engineering.

4. Feature Engineering

The fraud label is at the provider level, so we converted all claim-level data into one row per provider. We built features from inpatient claims, outpatient claims, and beneficiary data. These included claim counts, reimbursement statistics, Medicare coverage months, chronic condition rates, and combined metrics like total claims and inpatient-outpatient ratios. All features were saved in provider_features.csv.

5. Modeling

We split the dataset into training, validation, and testing. Three models were tested: Logistic Regression, Random Forest, and Gradient Boosting. Because fraud is rare, we used class weighting. Metrics such as precision, recall, ROC-AUC, and PR-AUC were used to compare models. Random Forest performed the best. We selected a threshold of 0.35, which improved the balance between catching fraud and avoiding false positives.

6. Final Results

The final test results were strong:

- Fraud precision: 0.66

- Fraud recall: 0.66

- Accuracy: 0.94

- ROC-AUC: 0.95

- PR-AUC: 0.73

The confusion matrix showed good detection rates with reasonable errors.

7. Error Analysis

False positives were providers predicted as fraud but actually normal. They often had high claims or reimbursements. False negatives were actual fraud cases that looked normal, which makes them harder to detect. These analyses helped explain the model's weaknesses.

8. Feature Importance

The most important features included total reimbursements, total claims, number of beneficiaries, and chronic condition rates. These matched patterns seen during EDA.

9. Conclusion

The Random Forest model performed well and can help identify suspicious provider behavior. Improvements could include using advanced boosting models, oversampling techniques, and adding time-based features.