Healthcare Provider Fraud Detection – Technical Report

1. Introduction

Healthcare fraud results in significant financial losses and reduces the effectiveness of medical systems. This project builds a data-driven machine learning system to detect potentially fraudulent healthcare providers using Medicare claims data...

2. Data Description

Train datasets: Inpatient, Outpatient, Beneficiary, Labels.

Test datasets: Inpatient, Outpatient, Beneficiary.

...

3. Data Quality & Preprocessing

Describes missing values, imputation strategy, column elimination, etc.

4. Feature Engineering

Provider-level aggregation: inpatient features, outpatient features, beneficiary demographics and chronic conditions, merging, final provider table.

5. Modeling

Models trained: Logistic Regression, Random Forest, Gradient Boosting.

Class imbalance handled via class_weight='balanced'.

Train-validation split is stratified.

6. Evaluation

ROC-AUC, PR-AUC, confusion matrix, threshold tuning.

Visualizations and metric interpretation included.

7. Error Analysis

False positives and false negatives examined.

Patterns discussed.

8. Conclusion

Project goals achieved: integration, modeling, prediction, error analysis.

9. Recommendations

Lower threshold suggested, additional feature engineering (ICD-9 codes, time-series, physician network analysis), retraining cycles.