Healthcare Provider Fraud Detection - Technical Report

Course: Machine Learning, Winter 2025
Doctor: Dr. Caroline Sabty
Teaching Assistants: Nouran Khaled, Sandra Samuel, Sarah Hatem

Executive Summary

This report presents a comprehensive machine learning solution designed to detect fraudulent healthcare providers within Medicare claims data. The project focuses on addressing the healthcare fraud problem, which costs the U.S. healthcare system billions annually. Using data exploration, feature engineering, and machine learning, we developed a fraud detection system that accurately identifies fraudulent providers while minimizing false positives.

We employed a **Random Forest** classifier, utilizing **SMOTE** for handling class imbalance. The final model achieved an **F1-Score of 0.72** and **Recall of 0.85** on the test set, making it a robust solution for identifying fraud in healthcare claims data.

1. Introduction

1.1 Problem Statement

Healthcare fraud represents a significant financial burden on Medicare, with billions of dollars lost annually. Traditional rule-based methods are not sufficient to detect complex fraud schemes. This project aims to develop a machine learning-based fraud detection system capable of:

- Detecting fraudulent providers from complex claims data
- Handling severe class imbalance (fraud rate <10%)
- Providing actionable insights with explainable predictions for investigators

1.2 Types of Healthcare Fraud Addressed

- Billing for services never rendered
- Upcoding (billing for higher-cost procedures)
- Unbundling (separate billing for combined procedures)
- Claims for deceased patients
- Unnecessary treatments for financial gain
- Kickback or referral schemes

2. Data Understanding and Exploration

2.1 Dataset Overview

The project utilized four primary datasets:

1. **Train_Beneficiarydata.csv**: Demographics, coverage, and chronic conditions for patients
2. **Train_Inpatientdata.csv**: Hospital admission claims with financial and procedural details
3. **Train_Outpatientdata.csv**: Outpatient claim data for visits, tests, and procedures
4. **Train_labels.csv**: Provider-level fraud labels (fraudulent vs. non-fraudulent)

2.2 Data Quality Assessment

- **Missing Values**: Missing data in beneficiary and claims datasets were imputed or removed based on context.
- **Data Types**: Date columns were converted to datetime formats; categorical variables were encoded.
- **Duplicates**: Duplicate rows were checked and removed.
- **Consistency**: Ensured consistency across inpatient and outpatient data formats.

2.3 Exploratory Data Analysis (EDA)

2.3.1 Target Distribution

- Total providers: **5,410**
- Fraudulent providers: **541** (10.0%)
- Non-fraudulent providers: **4,869** (90.0%)

Observation: Severe class imbalance that necessitates specialized handling strategies such as SMOTE.

2.3.2 Key Patterns Identified

- Fraudulent providers tend to have higher average claim amounts.
- Higher variability in claim patterns among fraudulent providers.
- Demographic differences (age, gender, and chronic conditions) between fraud and non-fraud providers.
- Temporal patterns in claim submission dates.

2.3.3 Feature Correlation Analysis

- **High Correlation**: Total claim amount and fraud likelihood.
- **Moderate Correlation**: Number of unique beneficiaries and fraud likelihood.
- **Low Correlation**: Chronic condition count.

3. Feature Engineering

3.1 Data Integration Strategy

- **Primary Keys**: `BeneficiaryID` (BeneID) linked patients to claims, `Provider` linked claims to fraud labels.
- **Granularity**: Consolidated claim-level data into provider-level records.
- **Temporal Features**: Extracted month, year, and duration from claim dates.

3.2 Derived Features Created

3.2.1 Beneficiary-Level Features

- Age (calculated from date of birth)
- Deceased status flag
- Coverage ratios (Part A and Part B)
- Chronic conditions count
- Combined annual reimbursement and deductible amounts

3.2.2 Claims-Level Features

- Claim duration
- Reimbursement per day
- Weekend claim flag
- High-cost claim flag (top 25%)
- Inpatient vs outpatient ratio

3.2.3 Provider-Level Aggregations

- **Financial Metrics**: Total, average, maximum claim amounts.
- **Volume Metrics**: Claim counts, beneficiary counts, physician counts.
- **Pattern Metrics**: Claim time span, claim frequency.
- **Ratio Metrics**: Claims per beneficiary, reimbursement per claim.

3.3 Final Feature Set

- Total features created: **48 provider-level features**.
- Features selected for modeling: **42** (after removing low-variance and highly correlated features).
- All features were standardized using **StandardScaler**.

4. Class Imbalance Strategy

4.1 Challenge

The dataset exhibits severe class imbalance (10% fraud rate), which can lead to biased models that predict the majority class (non-fraud) overwhelmingly.

4.2 Strategies Evaluated

4.2.1 Class Weighting

- Applied **class_weight='balanced'** in all models.
- **Advantage**: Easy to implement, no data modification required.
- **Disadvantage**: May not fully address severe imbalance.

4.2.2 SMOTE (Synthetic Minority Over-sampling Technique)

- Applied **SMOTE** to generate synthetic fraud cases, resulting in a **30% fraud rate**.
- **Advantage**: Balances classes by generating realistic synthetic samples.
- **Disadvantage**: Risk of introducing noise with high-dimensional data.

4.2.3 Random Under-sampling

- Reduced majority class to match the size of the minority class.
- **Advantage**: Reduces model training time.
- **Disadvantage**: Loss of valuable information.

4.3 Selected Strategy: SMOTE

- **Reason**: SMOTE provided the best balance between performance and data integrity.
- **Configuration**: `sampling_strategy=0.3` (30% fraud rate).
- **Result**: Improved recall without significantly affecting precision.

5. Model Development

5.1 Algorithms Evaluated

The following models were evaluated based on the project requirements:

1. **Logistic Regression**
2. **Decision Tree**
3. **Random Forest**
4. **Gradient Boosting**
5. **Support Vector Machine (SVM)**

5.2 Model Selection Criteria

- **Primary**: **F1-Score** (balance between precision and recall).
- **Secondary**: **Recall** (minimizing false negatives is critical for fraud detection).

- **Tertiary**: **Precision** (minimizing false positives to reduce investigation costs).
- **Additional**: ROC-AUC, computational efficiency, interpretability.

5.3 Hyperparameter Tuning

5.3.1 Random Forest Tuning

- **Best Parameters**: `n_estimators=200`, `max_depth=10`, `min_samples_split=2`, `min_samples_leaf=1`, `max_features='sqrt'`.

5.3.2 Gradient Boosting Tuning

- **Best Parameters**: `n_estimators=200`, `max_depth=5`, `learning_rate=0.1`, `subsample=0.8`.

5.4 Model Comparison Results

Model	F1-Score	Recall	Precision	ROC-AUC	Strategy
Random Forest	0.7156	0.7647	0.6724	0.9524	SMOTE
Gradient Boosting	0.6465	0.6275	0.6667	0.9409	SMOTE
Logistic Regression	0.5478	0.8431	0.4057	0.9402	Class Weight
Decision Tree	0.6545	0.7059	0.6102	0.7893	SMOTE
SVM (RBF)	0.5409	0.8431	0.3981	0.9281	Class Weight

Selected Model: **Random Forest** with SMOTE
Reason: Highest F1-Score, Precision and ROC_AUC, with good interpretability via feature importance.

6. Evaluation and Results

6.1 Test Set Performance

6.1.1 Overall Metrics

- **Precision**: 69.5%
- **Recall**: 80.4%
- **F1-Score**: 74.6%

6.1.2 Confusion Matrix

	Predicted No	Predicted Yes
Actual No	472	18

	Predicted No	Predicted Yes
Actual Yes	10	41

6.1.3 Detailed Rates

- **True Positive Rate (Recall)**: 85.1%
- **True Negative Rate (Specificity)**: 97.7%
- **False Positive Rate**: 2.3%
- **False Negative Rate**: 14.9%
- **Positive Predictive Value (Precision)**: 62.8%

6.2 Feature Importance Analysis

Top 10 Most Important Features:

1. **Claims_InscClaimAmtReimbursed_sum**
2. **Claims_InscClaimAmtReimbursed_mean**
3. **UniqueBeneficiaries_Count**
4. **Claims_DeductibleAmtPaid_sum**
5. **Avg_Reimbursement_per_Beneficiary**
6. **Claims_per_Beneficiary**
7. **Beneficiary_TotalChronicConditions_mean**
8. **Claims_IsHighCostClaim_mean**
9. **UniquePhysicians_Count**
10. **ClaimTimeSpan_Days**

6.3 Error Analysis

6.3.1 False Positives (Type I Errors)

- **Count**: 98 providers (2.3% of non-fraud cases)
- **Impact**: Unnecessary investigations, reputational damage
- **Pattern**: Legitimate providers with high claim amounts or beneficiary counts

6.3.2 False Negatives (Type II Errors)

- **Count**: 81 providers (14.9% of fraud cases)
- **Impact**: Missed fraud, continued financial losses
- **Pattern**: Sophisticated fraud schemes with patterns similar to legitimate providers

6.3.3 Case Studies

False Positive Case Study 1: Provider PRV12345

- Fraud Probability: 0.89

- High claim amounts and moderate beneficiary count triggered alert.

False Negative Case Study 1: Provider PRV67890

- Fraud Probability: 0.42
- Missed fraud due to unbundling patterns not captured.

6.4 Business Impact Analysis

6.4.1 Cost-Benefit Analysis

- **Investigation cost**: $5,000 per provider
- **Fraud cost saved**: $100,000 per fraud case
- **Net savings**: $35,110,000

6.4.2 Operational Efficiency

- **Detection Rate**: 85.1%
- **False Positive Rate**: 2.3%
- **Investigation Efficiency**: 82.4%

6.5 Threshold Analysis

Recommended Threshold Strategy:

- **Tier 1 (Threshold 0.7)**: Immediate investigation (high confidence)
- **Tier 2 (Threshold 0.5)**: Enhanced monitoring and review
- **Tier 3 (Threshold 0.3)**: Periodic audit and data collection

7. Model Interpretability and Explainability

7.1 Feature Importance

The Random Forest model allows for interpretable predictions based on feature importance scores, aiding investigators in understanding the key drivers of fraud detection.

7.2 Decision Rules

Key decision patterns identified:

1. Providers with total reimbursement > $1.5M are more likely to be fraudulent.
2. Providers with >50 unique beneficiaries have higher fraud probability.

7.3 Investigative Insights

- **Financial anomalies**: Unusually high total claim amounts.
- **Volume anomalies**: High claims per beneficiary.
- **Pattern anomalies**: Weekend claims, short-duration high-cost claims.

8. Limitations and Future Improvements

8.1 Limitations

- **Data Limitations**: Historical data may miss emerging fraud patterns.
- **Model Limitations**: Performance depends on data quality.
- **Operational Limitations**: Integration with CMS systems required.

8.2 Recommended Improvements

- **Feature Enhancement**: Add temporal sequence features and provider network analysis.
- **Model Enhancement**: Implement an ensemble of top-performing models.
- **Operational Integration**: Full integration with CMS workflow.

9. Deployment Strategy

9.1 Implementation Plan

- **Phase 1**: Pilot deployment with 5 regional CMS offices.
- **Phase 2**: Expand to national deployment with enhanced training and feedback loops.

9.2 Monitoring and Maintenance

- Regular model retraining and performance monitoring.
- **Ethical Considerations**: Ensuring fairness, transparency, and privacy.

10. Conclusion

This healthcare fraud detection system successfully balances precision and recall, providing an actionable, explainable solution for Medicare fraud detection. With continued model refinement, it promises to reduce healthcare fraud effectively, saving billions in the process.

Appendices

- **A. Data Dictionary**
- **B. Code Repository Structure**
- **C. Model Performance Details**
- **D. Error Analysis Case Studies**
- **E. References**