TECHNICAL REPORT

Medicare Provider Fraud Detection Using Machine Learning

GIU – Winter 2025
Team Members:

- Member 1 – Data Loading & Exploration
- Member 2 – Feature Engineering & Dataset Creation
- Member 3 – Modeling & Optimization
- Member 4 – Evaluation, Error Analysis, Documentation

1. Introduction

Medicare fraud is a major challenge for healthcare systems, costing billions in unnecessary reimbursements every year. Fraudulent providers often submit inflated, duplicated, or false claims, making early detection crucial for maintaining financial sustainability and patient safety.

This project aims to develop a machine learning pipeline to identify fraudulent healthcare providers using Medicare claim-level data. The dataset includes inpatient, outpatient, beneficiary, and fraud labels, which were aggregated into a provider-level modeling dataset.

The final output is a classification model that predicts whether a provider is potentially fraudulent based on their billing patterns.

2. Data Overview

We used the following raw datasets (CSV files provided):

- **Train_Inpatientdata.csv**
- **Train_Outpatientdata.csv**
- **Train_Beneficiarydata.csv**
- **Train.csv** (provider-level fraud labels)

The raw claim-level data contains thousands of rows per provider, with fields such as:

- Reimbursement amounts
- Deductibles
- Claim durations
- Procedure & diagnosis codes
- Claim counts per provider

The **fraud label (`PotentialFraud`)** has two values:

- `"Yes"` → Fraudulent provider
- `"No"` → Non-fraudulent provider

Fraud cases represent **~8%** of providers — a highly imbalanced classification task.

3. Data Preparation (Member 1)

Member 1 performed initial data cleaning and exploration:

3.1 Data Loading

- All raw CSV files were loaded and inspected.
- Shape, column names, and missing values were checked.
- Dataset consistency (shared Provider IDs) was validated.

3.2 Missing & Anomalous Values

- Some numeric fields contained missing values, zero values, or medical anomalies.
- Member 1 identified:
    - Rare extreme reimbursement amounts
    - Providers with ≥ 200 outpatient claims per month
    - Large variations in inpatient claim counts

3.3 Fraud Distribution

- Fraud vs non-fraud imbalance was confirmed (~8% fraud).

4. Feature Engineering (Member 2)

Member 2 aggregated claim-level data to build a **provider-level feature dataset** — one row per provider.

4.1 Aggregated Claim-Based Features

Counts

- `num_inpatient_claims`
- `num_outpatient_claims`
- `total_claims`

Reimbursement Statistics

- `ip_reimb_mean`, `ip_reimb_sum`
- `op_reimb_mean`, `op_reimb_sum`

Deductible Statistics

- `ip_deduct_mean`, `ip_deduct_sum`
- `op_deduct_mean`, `op_deduct_sum`

These features capture providers' billing behavior patterns.

4.2 Label Encoding

`PotentialFraud` values were converted to a binary variable:

```
fraud_flag = 1 if PotentialFraud in ["Yes", "Y"] else 0
```

4.3 Final Dataset

- Shape: **(n_providers, ~14 features)**

- Saved as **provider_level_features.csv**

5. Modeling (Member 3)

Member 3 split the data into **train/test (80/20, stratified)** and tested several models.

5.1 Models Trained

- **Logistic Regression**
- **Decision Tree**
- **Gradient Boosting (GBClassifier)** ← *best*
- (Optional: Random Forest, XGBoost – depending on member 3's work)

5.2 Handling Imbalance

Fraud detection suffers from class imbalance. Member 3 tried:

- **SMOTE**
- **Class weight balancing**
- **Threshold tuning**

5.3 Model Selection Criteria

- ROC-AUC
- PR-AUC
- Recall (fraud)
- F1-score (fraud)
- Confusion matrix

5.4 Saved Best Model

Chosen model:

Gradient Boosting Classifier

Saved as:

```
models/primary_model.pkl
```

6. Model Evaluation (Member 4)

Member 4 evaluated all models using test data and produced:

- Confusion matrices
- ROC curves
- Precision–Recall curves
- False positive & false negative analysis
- Business interpretation

6.1 Evaluation Metrics

Model	ROC-AUC	PR-AUC	Recall (fraud)	F1 (fraud)
Gradient Boosting	**0.953**	**0.736**	0.554	**0.619**
Logistic Regression	0.949	0.733	**0.881**	0.555
Decision Tree	0.772	0.367	0.594	0.574

6.2 Confusion Matrix (GB Model)

	Predicted No Fraud	Predicted Fraud
Actual No Fraud	957	24
Actual Fraud	45	56

Interpretation:

- Good at detecting non-fraud (TN = 957)
- Reasonable fraud detection (TP = 56)
- Some missed fraud cases (FN = 45)

7. Error Analysis (Member 4)

7.1 False Positives (FP)

Providers flagged as fraud but are not fraudulent typically had:

- Very high total reimbursement
- Extremely high deductible totals
- Many outpatient claims

These providers have **extreme billing behavior**, similar to real fraud patterns.

7.2 False Negatives (FN)

Fraud providers that were missed typically had:

- Moderate or low claim counts
- Under-the-radar reimbursements
- Less obvious financial anomalies

The fraud signal was weaker → model could not distinguish them from normal providers.

7.3 Conclusion of Error Analysis

- FP: Model is **too strict** with large claim volumes.
- FN: Model needs **more nuanced features** to capture subtle fraud behavior.

8. Business Interpretation

Fraud detection is critical in healthcare:

- Prevents **financial loss**
- Reduces **unnecessary medical procedures**
- Supports **government oversight**
- Prioritizes **high-risk providers** for auditing
- Saves time and resources through **automated screening**

A model with high ROC-AUC and PR-AUC greatly enhances the ability to identify suspicious providers early.

9. Recommendations for Future Work

To improve accuracy:

9.1 Add Richer Features

- Diagnosis codes (ICD)
- Procedure codes
- DRG codes
- Provider specialty
- Geographic region

9.2 Advanced Modeling

- XGBoost or LightGBM
- Stacking ensembles
- Isolation Forest for anomaly detection
- Deep autoencoders for pattern extraction

9.3 Better Imbalance Solutions

- SMOTE variants (Borderline-SMOTE, SMOTEENN)
- Cost-sensitive learning
- Threshold optimization for higher fraud recall

9.4 Temporal Features

- Claims per month
- Sudden spikes in activity
- Seasonal patterns

These would help capture sophisticated fraud behavior.

10. Conclusion

This project successfully built a full end-to-end fraud detection system including:

- Data exploration
- Feature engineering
- Model training
- Model evaluation
- Error analysis
- Business interpretation
- Reporting & presentation

The best-performing model, **Gradient Boosting**, achieved **ROC-AUC = 0.953** and showed strong ability to distinguish fraudulent vs non-fraudulent providers.

The findings provide valuable insights into suspicious billing patterns and pave the way for more advanced fraud analytics systems.

11. References

- Scikit-learn documentation
- Imbalanced-learn documentation
- Medicare Fraud Research Publications
- GIU Course Materials (Machine Learning, Database Programming)