Technical Report: Healthcare Provider Fraud Detection

Team DataOrbit

December 3, 2025

Abstract

This report details the development of an intelligent fraud detection system for the Centers for Medicare & Medicaid Services (CMS). Addressing the $68 billion annual cost of healthcare fraud, this project leverages machine learning to identify high-risk providers. The methodology moves beyond traditional rule-based systems by implementing advanced feature engineering, clinical code grouping, and ensemble modeling techniques to detect sophisticated fraud schemes while maintaining interpretability.

Contents

1 Executive Summary

Data Orbit was contracted to assist Medicare in detecting fraudulent healthcare providers. Existing systems rely heavily on rule-based methods that capture obvious patterns but often fail to identify complex schemes such as upcoding, unbundling, and services not rendered.

The objective of this project is to build an end-to-end machine learning pipeline that:

- Identifies potentially fraudulent providers from multi-table claims data.
- Handles severe class imbalance (approx. 9.3% fraud rate).
- Minimizes false positives to prevent unnecessary investigations and reputational damage.
- Provides interpretable predictions to aid investigators.

This report documents the data processing, feature engineering strategies, and modeling methodologies employed to achieve these goals.

2 Data Understanding and Preparation

The analysis utilizes the Healthcare Provider Fraud Detection dataset, comprising four primary files: Beneficiary Data, Inpatient Claims, Outpatient Claims, and Provider Labels.

2.1 Data Integration and Quality Assessment

The raw data consisted of 558,211 claims across 5,410 providers.

- **Merging Strategy:** The Inpatient and Outpatient datasets were merged with Beneficiary data using `BeneID`. Subsequently, these were joined with the target labels using `Provider`.
- **Missing Values:** High missingness was observed in `ClmProcedureCode_6` (100% missing), which was dropped.
- **Date Harmonization:** A discrepancy was discovered between `ClaimStartDt` and `AdmissionDt` in 32 inpatient cases. The analysis revealed that admission often precedes the billing period. To capture the true length of stay, `ClaimStartDt` was standardized using admission dates.
- **Schema Alignment:** To enable a unified analysis, the Outpatient dataset (which lacks Diagnosis Groups) was harmonized with Inpatient data by filling missing `DiagnosisGroupCode` values with ”Not Applicable”.

2.2 Exploratory Data Analysis (EDA)

Key patterns identified during exploration included:

- **Post-Discharge Billing:** A calculated flag `IsPostDischargeBilling` revealed that 100% of claims where the billing end date succeeded the discharge date were associated with fraudulent providers.
- **Claim Duration:** Fraudulent providers exhibited distinct distributions in claim duration compared to legitimate providers.
- **Deductible Standardization:** Missing deductible amounts were imputed based on service type logic ($1,068 for Inpatient, $0 for Outpatient) to maintain financial data integrity.

3 Feature Engineering

To transform raw transactional data into provider-level predictive features, extensive engineering was performed.

3.1 Clinical Code Grouping (Dimensionality Reduction)

The raw dataset contained thousands of sparse ICD-9 codes. Using these directly would result in high dimensionality and noise. We implemented a clinical grouping strategy using CMS dictionaries:

- **Method:** Diagnostic and Procedural codes were mapped to 17 clinically significant "Super Groups" (e.g., *Circulatory System*, *Trauma*, *Respiratory*, *Neoplasms*).
- **Metric:** Instead of simple binary flags, we calculated the frequency (count) of these groups per claim, capturing the complexity and severity of the patient's condition.

3.2 Demographic and Financial Features

- **Age Calculation:** Exact age was derived relative to a reference date (2010-01-01).
- **Renal Disease:** The `RenalDiseaseIndicator` was converted to a binary flag, as chronic kidney disease is a high-cost co-morbidity often targeted in fraud schemes.
- **State/County:** Geographic codes were mapped to their respective abbreviations and names to analyze regional fraud concentrations.

3.3 Provider-Level Aggregation Strategy

The fundamental unit of analysis is the **Provider**, not the Claim. We aggregated 558,211 claims into 5,410 provider records using a domain-driven aggregation dictionary:

- **Skewed Financials:** `mean`, `std`, and `99.5th percentile` were used for reimbursement amounts to capture outliers typical of fraud.
- **Clinical Counts:** `max` and `sum` were used for disease groups to identify providers specializing in high-risk procedures.
- **ID Cardinality:** `nunique` counts for `BeneID` and `AttendingPhysician` were calculated to detect network anomalies (e.g., a provider seeing an impossible number of unique patients).

3.4 Multicollinearity Reduction

Post-aggregation, the dataset contained 134 features. A Spearman correlation analysis identified highly redundant features (correlation > 0.9).

- **Selection Logic:** For every correlated pair, the feature with the lower correlation to the target variable (`PotentialFraud`) was dropped.
- **Result:** 22 redundant features were removed (e.g., dropping `BeneID_count` in favor of `ClaimID_count`), resulting in a final feature set of 100 robust predictors.

4 Methodology

4.1 Class Imbalance Strategy

The dataset is moderately imbalanced, with a fraud ratio of approximately 9.7:1 (9.35% fraud). To address this, multiple resampling strategies were experimentally validated:

- **Tested Methods:** SMOTE, ADASYN, Random Over-sampling, Random Under-sampling, NearMiss, and Hybrid methods (SMOTE+Tomek, SMOTE+ENN).
- **Decision:**
  - For **Tree-based models** (Random Forest, Gradient Boosting), **SMOTE** was selected to generate synthetic minority samples, preventing the model from biasing heavily toward the majority class.
  - For **Logistic Regression** and **SVM**, a **Class Weighting** strategy ('balanced') combined with **No Resampling** proved superior in preserving data integrity and maintaining precision.

4.2 Algorithm Selection

Six distinct algorithms were evaluated to determine the optimal balance between predictive power, computational efficiency, and interpretability:

1. **Decision Tree Classifier:** Evaluated as a simple, highly interpretable baseline model.
2. **Random Forest Classifier:** Chosen for its robustness to overfitting and ability to provide stable feature importance metrics through bagging.
3. **Gradient Boosting Classifier:** Selected for its ability to handle non-linear relationships and high-dimensional interactions inherent in fraud data.
4. **Logistic Regression:** Included for its transparency and to assess the linear separability of the feature space.
5. **AdaBoost Classifier:** Tested as an alternative boosting method that focuses on hard-to-classify examples.
6. **Support Vector Machine (SVM):** Evaluated for its effectiveness in high-dimensional spaces. The final model utilized the RBF kernel with `NoSampling_ClassWeight` strategy.

4.3 Experimental Design

- **Validation Scheme:** A 5-Fold Stratified Cross-Validation strategy was employed to ensure that the ratio of fraud cases remained consistent across all training folds.
- **Hyperparameter Tuning:** `GridSearchCV` was utilized to optimize parameters (e.g., tree depth, learning rate, regularization strength).
- **Optimization Metric:** Models were optimized primarily for **Recall** (Sensitivity) to maximize the detection of fraud cases, while monitoring Precision to keep false positives manageable.

5 Model Evaluation

The models were evaluated based on their ability to detect fraud (Recall) without generating excessive false alarms (Precision). The table below summarizes the performance of the six optimized models on the test set.

Table 1: Comprehensive Model Comparison (Test Set Metrics)

Model	Recall	Precision	F1-Score	ROC-AUC	PR-AUC
Gradient Boosting	0.7746	0.4583	**0.5759**	0.9198	0.5233
Decision Tree	0.7465	0.4609	0.5699	0.8821	0.4582
Random Forest	0.6620	0.4608	0.5434	0.9187	0.5349
Logistic Regression	**0.9155**	0.3846	0.5417	0.9461	**0.7113**
AdaBoost	0.4366	**0.7209**	0.5439	0.8883	0.5537
SVM (Class Weight)	0.9079	0.3966	0.5520	**0.9467**	0.6886

Key Observations:

- **SVM (NoSampling_ClassWeight):** This model emerged as a top performer for sensitivity, achieving a high Recall of 90.79% with excellent discrimination (ROC-AUC 0.9467). It utilizes the full original dataset without synthetic sampling, preserving data integrity.
- **Logistic Regression:** Showed surprising strength in fraud detection with the highest Recall (91.55%) and PR-AUC (0.7113), indicating excellent ranking capability, though with slightly lower precision than tree models.
- **AdaBoost:** Offered the highest Precision (72.09%) but missed a significant portion of fraud cases (Recall 43.66%), making it suitable only if false positives are strictly penalized.
- **Gradient Boosting:** Provided the best F1-Score (0.5759), offering a balanced approach between catching fraud and minimizing workload.

6 Conclusion and Recommendations

[This section will summarize the business impact, estimated cost savings, and final deployment recommendations.]