Healthcare Provider Fraud Detection - Report

1. Introduction

Objective

This project aims to develop an end-to-end data-driven fraud detection pipeline for identifying potentially fraudulent healthcare providers in the U.S. Medicare system. Fraudulent activities, such as billing for non-rendered services, upcoding, and submitting claims for deceased patients, cost the U.S. healthcare system billions annually. This model seeks to assist investigators by automating the process of detecting fraudulent providers, focusing on minimizing false positives to avoid unnecessary investigations.

Scope

We employ machine learning techniques to classify providers as either fraudulent or non-fraudulent based on historical Medicare claims data. Our solution addresses the challenge of class imbalance, where only a small fraction of providers are fraudulent, and aims to provide explainable predictions for transparency in the decision-making process.

---

2. Data Exploration and Preprocessing

Dataset Overview

The dataset consists of the following files:

- Train_Beneficiarydata.csv: Contains demographic information, chronic conditions, and coverage details for beneficiaries.
- Train_Inpatientdata.csv: Contains hospital admission claims, including financial and procedural details.
- Train_Outpatientdata.csv: Contains outpatient claim data, such as visits, tests, and procedures.
- Train_Labels.csv: Contains fraud labels indicating whether a provider is fraudulent.

Exploratory Data Analysis (EDA)

- Data Inspection: We performed an initial check of the datasets, looking at their shapes and the first few rows to ensure that the data was loaded correctly.
- Missing Value Analysis: Missing data was analyzed and imputed where necessary, particularly in numeric columns where missing values were replaced with the mean.
- Feature Distribution: Key statistics were explored, such as the distribution of fraudulent vs. non-fraudulent providers, average reimbursement amounts, and the number of claims per provider.
- Class Imbalance: The dataset has a severe class imbalance, with only about 10% of providers labeled as fraudulent.

Feature Engineering

- Inpatient and Outpatient Data: For both inpatient and outpatient data, date columns were converted to datetime format, and new features such as claim durations and the number of diagnosis codes were derived.
- Beneficiary Data: Beneficiary data was enriched with features such as age (calculated at the end of 2009) and deceased status.
- Provider-level Aggregation: We aggregated claim-level data (from both inpatient and outpatient datasets) into provider-level records. Features like total reimbursement amount, claim count, and the number of beneficiaries per provider were computed.

---

3. Model Selection and Training

Algorithm Choice

- Logistic Regression: Chosen as a baseline model due to its simplicity and interpretability, helping us understand feature importance and the relationship between inputs and predictions.
- Random Forest: Selected for its robustness to overfitting and ability to handle high-dimensional datasets. It performs well on non-linear relationships in the data.
- Gradient Boosting: Chosen for its high performance, especially in handling class imbalances by focusing on misclassified samples in each iteration.
- Support Vector Machine (SVM): Although computationally expensive, SVM was tested to understand its potential in separating classes with non-linear boundaries.

Handling Class Imbalance

- SMOTE (Synthetic Minority Over-sampling Technique) was applied to oversample the minority class (fraudulent providers) in the training data.
- Class Weighting: All models, including Logistic Regression, Random Forest, and SVM, were trained with class_weight="balanced" to give more importance to the minority class.

Hyperparameter Tuning

- Randomized Search was used to optimize hyperparameters such as the number of estimators in Random Forest and Gradient Boosting, as well as the depth of trees in the Decision Tree classifier.

Model Training

- The models were trained on the data, and their performance was validated using a separate validation set. All models were evaluated on metrics suitable for imbalanced datasets, such as Precision, Recall, F1-score, ROC-AUC, and PR-AUC.

---

4. Model Evaluation

Evaluation Metrics

The models were evaluated on the test set using the following metrics:

- Precision: Measures the proportion of fraud predictions that were correct.
- Recall: Measures the ability of the model to detect fraudulent providers.
- F1-Score: A balance between Precision and Recall.
- ROC-AUC: Measures the trade-off between true positive rate and false positive rate.
- PR-AUC (Average Precision): Focuses on the positive (fraudulent) class, which is more relevant for imbalanced datasets.

Results

Model	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
Logistic Regression	0.72	0.65	0.68	0.75	0.70
Random Forest	0.78	0.76	0.77	0.80	0.76
Gradient Boosting	0.80	0.80	0.80	0.85	0.81

Model Comparison

- Best Model: Gradient Boosting outperformed the other models with the highest PR-AUC and ROC-AUC. It also provided the best balance between precision and recall.
- Logistic Regression: While interpretable, it had the lowest performance on the test set, especially in terms of recall.

Model Interpretation

- Feature Importances: For the models that supported feature importance (e.g., Random Forest and Gradient Boosting), we visualized the top features. Features like total reimbursement amount, the number of claims, and the number of beneficiaries were the most important in predicting fraud.

---

5. Error Analysis

False Positives (FP) and False Negatives (FN)

- False Positives: These are legitimate providers flagged as fraudulent. The top false positives were examined, and common characteristics were found, such as a high number of claims with low reimbursement amounts.
- False Negatives: These are fraudulent providers that were missed by the model. The top false negatives had features like high reimbursement sums and a large number of beneficiaries, but the model failed to identify them.

Insights

- False Positives: Unnecessary investigations could be triggered, leading to resource wastage and reputational damage.

- False Negatives: Missed fraud could result in significant financial losses. Enhancing the model with additional features or refining the class imbalance strategy could reduce false negatives.

---

6. Trials and Experiments

Log of Trials

- Trial 1: Implemented different preprocessing strategies such as SMOTE and class weighting.
    - Insight: SMOTE was effective in improving the recall of the models, particularly in the Random Forest and Gradient Boosting models.
- Trial 2: Tested different models, including Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting.
    - Insight: Gradient Boosting consistently outperformed the other models, providing the best trade-off between precision and recall.
- Trial 3: Tuned hyperparameters for Random Forest and Gradient Boosting.
    - Insight: Hyperparameter tuning significantly improved model performance, especially for Gradient Boosting.

---

7. Conclusion

Summary of Findings

The project successfully developed a fraud detection system that identifies high-risk fraudulent healthcare providers. Gradient Boosting was selected as the best model due to its high performance on imbalanced datasets and ability to handle complex relationships between features.

Future Work

- Further improvements can be made by including more features, such as temporal patterns (e.g., claim trends over time) and geographic data.