Healthcare Provider Fraud Detection

Final Technical Report

Date: November 27, 2025

Machine Learning Project
Medicare Claims Fraud Detection

1. Introduction

Healthcare fraud is a significant problem that costs billions of dollars annually. Provider fraud, where healthcare providers submit fraudulent claims to insurance companies, represents a substantial portion of this loss. This project focuses on developing a machine learning solution to detect fraudulent healthcare providers using Medicare claims data.

The primary objective is to build a predictive model that can accurately identify potentially fraudulent providers based on their claims patterns, beneficiary characteristics, and billing behaviors. This enables proactive fraud detection and investigation, potentially saving significant resources.

2. Data Understanding

The dataset consists of Medicare claims data from multiple sources, providing a comprehensive view of provider activities:

Data Sources:
- **Beneficiary Data:** Patient demographics and chronic condition indicators
- **Inpatient Claims:** Hospital admission claims with diagnosis and procedure codes
- **Outpatient Claims:** Outpatient visit claims and associated costs
- **Provider Labels:** Binary fraud indicator for each provider

Dataset Statistics:
- Total Providers: 5,410
- Fraudulent Providers: 506 (9.4%)
- Non-Fraudulent Providers: 4,904 (90.6%)
- Inpatient Claims: ~40,000 records
- Outpatient Claims: ~500,000 records
- Beneficiaries: ~138,000 unique patients

Key Observations from Exploratory Data Analysis:
- Significant class imbalance with fraud representing only 9.4% of providers
- Fraudulent providers show higher average claim amounts and reimbursements
- Fraudulent providers tend to have more claims per beneficiary
- Chronic conditions distribution varies between fraudulent and legitimate providers
- Strong correlation between total reimbursement and claim counts

3. Feature Engineering

Provider-level features were engineered by aggregating claim and beneficiary data. This transformation is crucial as the prediction target is at the provider level, not individual claims.

Feature Categories Created:

1. Count-Based Features:
- Total number of claims (inpatient and outpatient)
- Number of unique beneficiaries served
- Number of unique physicians involved

- Claims per beneficiary ratio

2. Cost-Based Features:
- Total reimbursement amounts
- Average reimbursement per claim
- Deductible amounts paid (sum, mean, max)
- Insurance claim amounts (sum, mean, max, std)

3. Clinical Features:
- Average number of diagnoses and procedures
- Maximum number of diagnoses and procedures
- Length of stay statistics for inpatient claims
- Claim duration metrics

4. Patient Health Features:
- Average prevalence of chronic conditions (Diabetes, Heart Failure, Kidney Disease, etc.)
- Average patient age
- Overall chronic condition burden

5. Ratio Features:
- Inpatient to outpatient claims ratio
- Claims per beneficiary
- Average reimbursement per claim

Justification: These features capture billing patterns, patient complexity, and provider behavior that may distinguish fraudulent from legitimate providers. Fraudulent providers often exhibit unusual patterns such as excessive billing, treating unusually high numbers of patients, or claiming for complex procedures at abnormal rates.

4. Handling Class Imbalance

The dataset exhibits significant class imbalance with fraudulent providers representing only 9.4% of the total. This imbalance can lead models to be biased toward predicting the majority class.

Techniques Evaluated:
- **Class Weights:** Assigning higher weights to the minority class during training
- **SMOTE:** Synthetic Minority Over-sampling Technique to generate synthetic fraud examples
- **Random Oversampling:** Duplicating minority class samples
- **Random Undersampling:** Reducing majority class samples

Final Decision: Class weighting was selected as the primary technique because it:
- Maintains the original data distribution without artificial samples
- Provides stable model performance across cross-validation folds
- Works well with tree-based models (Random Forest, XGBoost)
- Avoids overfitting risks associated with synthetic data generation

5. Modeling Approach

A systematic modeling approach was employed to ensure robust and reliable fraud detection:

Train-Test Split:
- 80% training, 20% test split
- Stratified sampling to maintain fraud ratio in both sets
- Random state fixed for reproducibility

Models Trained:
- **Logistic Regression:** Baseline linear model with L2 regularization
- **Random Forest:** Ensemble of 100 decision trees with class weighting
- **XGBoost:** Gradient boosting with optimized hyperparameters

Cross-Validation Strategy:
- 5-fold Stratified K-Fold cross-validation
- Ensures each fold maintains the original fraud ratio
- Provides robust performance estimates

Performance Metrics:
- **Precision:** Proportion of predicted frauds that are actually fraudulent
- **Recall:** Proportion of actual frauds correctly identified
- **F1-Score:** Harmonic mean of precision and recall
- **ROC-AUC:** Area under the Receiver Operating Characteristic curve
- **PR-AUC:** Area under the Precision-Recall curve (critical for imbalanced data)

6. Cross-Validation Results

Cross-validation results demonstrate model performance stability across different data subsets:

Model	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
Logistic Regression	0.48 ± 0.05	0.89 ± 0.04	0.62 ± 0.04	0.96 ± 0.01	0.76 ± 0.03
Random Forest	0.61 ± 0.06	0.74 ± 0.05	0.67 ± 0.04	0.96 ± 0.01	0.75 ± 0.03
XGBoost	0.63 ± 0.05	0.71 ± 0.04	0.67 ± 0.03	0.96 ± 0.01	0.78 ± 0.02

Key Insights:

- All models achieve excellent ROC-AUC scores (>0.96), indicating strong discriminative ability
- Logistic Regression achieves highest recall (89%) but lower precision (48%)
- XGBoost provides the best balance with highest PR-AUC (0.78)
- Random Forest and XGBoost show more balanced precision-recall tradeoffs
- Low standard deviations indicate stable performance across folds

7. Test-Set Evaluation

Final model performance on the held-out test set confirms the cross-validation findings:

Model	Precision	Recall	F1-Score	ROC-AUC	PR-AUC
Logistic Regression	0.477	0.911	0.626	0.968	0.784
Random Forest	0.628	0.752	0.685	0.967	0.764
XGBoost	0.646	0.723	0.682	0.961	0.795

Performance Analysis:

- **Logistic Regression:** Highest recall (91.1%) but lowest precision (47.7%), resulting in many false positives
- **Random Forest:** Balanced performance with 62.8% precision and 75.2% recall
- **XGBoost:** Best overall performance with highest precision (64.6%) and strong PR-AUC (0.795)
- All models maintain excellent ROC-AUC scores (>0.96)

Confusion Matrix Insights:

- XGBoost correctly identifies ~72% of fraudulent providers while maintaining reasonable precision
- The model minimizes false negatives (missed frauds) while controlling false positives
- Trade-off between precision and recall is well-balanced for business requirements

8. Error Analysis

Understanding model errors provides insights into limitations and areas for improvement:

False Positives (Legitimate Providers Flagged as Fraudulent):
- Represent high-volume legitimate providers with unusual but valid billing patterns
- Often include specialists treating complex cases with high reimbursement amounts
- May include providers serving populations with high chronic disease burden
- Teaching hospitals or research institutions with atypical claim patterns

Business Impact: False positives lead to unnecessary investigations, potentially straining relationships with legitimate providers. However, they are less costly than missed fraud.

False Negatives (Fraudulent Providers Not Detected):
- Sophisticated fraudsters who mimic legitimate billing patterns
- Providers with lower fraud volumes that blend with normal activity
- New fraud schemes not represented in historical training data
- Fraudulent providers operating just below detection thresholds

Business Impact: False negatives represent the most critical error type, as they allow fraudulent activity to continue undetected, resulting in direct financial losses.

Mitigation Strategies:
- Implement a two-stage review process where high-confidence predictions trigger automatic flags
- Use model probability scores to prioritize investigations
- Combine model predictions with domain expert review
- Continuously update the model with newly identified fraud cases

9. Model Explainability

Understanding which features drive fraud predictions is crucial for model trust and actionable insights:

Top 10 Most Important Features (XGBoost):

Rank	Feature	Importance
1	Total_Reimbursement	0.2634
2	Inpatient_ClaimDuration_max	0.0507
3	Outpatient_InscClaimAmtReimbursed_sum	0.0310
4	Total_Claims	0.0239
5	Claims_Per_Beneficiary	0.0220
6	Inpatient_DeductibleAmtPaid_sum	0.0181
7	Outpatient_NumDiagnoses_max	0.0180
8	Outpatient_InscClaimAmtReimbursed_max	0.0171
9	Outpatient_ChronicCond_Cancer_mean	0.0146

10	Outpatient_ChronicCond_Diabetes_mean	0.0143

Key Feature Interpretations:

- **Total_Reimbursement (26.3%):** The dominant feature, indicating fraudulent providers typically claim significantly higher reimbursement amounts

- **Inpatient_ClaimDuration_max (5.1%):** Maximum claim duration for inpatient stays, suggesting fraudsters may extend treatment periods

- **Outpatient_InscClaimAmtReimbursed_sum (3.1%):** Total outpatient reimbursements, another strong cost-based indicator

- **Total_Claims (2.4%):** Volume of claims submitted, with fraudulent providers often showing unusually high claim counts

- **Claims_Per_Beneficiary (2.2%):** Ratio feature indicating fraudsters may bill more frequently per patient

SHAP Analysis: SHAP (SHapley Additive exPlanations) values provide instance-level explanations, showing how each feature contributes to individual predictions. High total reimbursement and claim counts consistently push predictions toward fraud, while lower values indicate legitimate providers.

10. Final Model Selection

Selected Model: XGBoost Classifier

XGBoost was selected as the final production model based on comprehensive evaluation:

Strengths:
- **Best Overall Performance:** Highest PR-AUC (0.795) on test set, critical for imbalanced classification
- **Balanced Precision-Recall:** 64.6% precision and 72.3% recall provide optimal tradeoff
- **Robust to Overfitting:** Regularization and tree pruning prevent overfitting
- **Handles Complex Patterns:** Captures non-linear relationships and feature interactions
- **Feature Importance:** Provides clear interpretability through feature importance scores
- **Scalable:** Efficient training and prediction on large datasets
- **Stable Performance:** Consistent results across cross-validation folds

Limitations:
- **False Positives:** ~35% of fraud predictions are false alarms, requiring manual review
- **False Negatives:** ~28% of fraudulent providers go undetected
- **Concept Drift:** Fraud patterns may evolve, requiring periodic model retraining
- **Data Dependency:** Performance relies on quality and completeness of claims data
- **Interpretability:** While better than deep learning, still less interpretable than logistic regression

Business Alignment:
The model aligns well with business requirements by:
- Prioritizing recall to minimize missed fraud (72.3% detection rate)
- Maintaining acceptable precision to avoid investigation overload
- Providing probability scores for risk-based prioritization
- Offering explainable predictions for compliance and auditing
- Enabling automated screening of all providers with minimal manual effort

11. Recommendations & Future Work

Deployment Recommendations:
- Implement a tiered alert system based on prediction probability scores
- Establish a feedback loop to capture investigation outcomes and retrain the model
- Set up automated monthly retraining to adapt to evolving fraud patterns
- Create a dashboard for fraud investigators showing high-risk providers and key indicators
- Integrate with existing case management systems for seamless workflow

Model Improvements:
- Incorporate temporal features to detect sudden changes in billing patterns
- Add network analysis features to identify fraud rings and collusion
- Include external data sources (provider credentials, sanctions lists, peer comparisons)
- Experiment with ensemble methods combining multiple models
- Develop separate models for different provider specialties

Additional Features to Consider:

- Geographic clustering of providers and beneficiaries
- Time-series features capturing billing pattern changes
- Peer comparison metrics (deviation from specialty norms)
- Social network features (referral patterns, shared patients)
- Text mining of diagnosis and procedure codes for unusual combinations

Real Deployment Considerations:

- **Regulatory Compliance:** Ensure model decisions are auditable and explainable
- **Fairness:** Monitor for bias across provider types, specialties, and demographics
- **Privacy:** Implement proper data governance and HIPAA compliance
- **Monitoring:** Track model performance metrics and data drift in production
- **Human-in-the-Loop:** Maintain expert review for all fraud determinations
- **Scalability:** Design infrastructure to handle growing data volumes
- **Latency:** Optimize for real-time or near-real-time fraud detection

12. Conclusion

This project successfully developed a machine learning solution for healthcare provider fraud detection, achieving strong performance with 72.3% recall and 64.6% precision on the test set. The XGBoost model demonstrates the ability to identify fraudulent providers based on their billing patterns, patient characteristics, and claim behaviors.

The model provides actionable insights through feature importance analysis, highlighting that total reimbursement amounts, claim durations, and claim volumes are the strongest fraud indicators. With proper deployment infrastructure, continuous monitoring, and expert oversight, this solution can significantly enhance fraud detection capabilities and reduce financial losses.

Future enhancements incorporating temporal patterns, network analysis, and additional data sources will further improve detection accuracy and adapt to evolving fraud schemes. The foundation established in this project provides a robust platform for ongoing fraud prevention efforts.