Healthcare Provider Fraud Detection: Technical Report

1. Introduction Healthcare fraud is a critical issue that leads to substantial financial losses for insurance providers and government programs. The objective of this project is to build a predictive machine learning model capable of identifying potentially fraudulent healthcare providers. By analyzing patterns in claims data—such as claim duration, reimbursement amounts, and beneficiary details—we aim to flag suspicious providers for further investigation. This system is intended to act as a "triage" tool, prioritizing high-risk cases for human auditors to maximize resource efficiency.

2. Data Exploration and Feature Engineering The analysis is based on a dataset comprising beneficiary demographics, inpatient/outpatient claims, and provider labels. A key challenge was the difference in data granularity: raw data existed at the individual claim level, while fraud labels existed at the provider level.

To address this, we aggregated all transactional data to the **Provider Level**. The following features were engineered to capture suspicious behavior:

- **Volume Metrics:** TotalClaimCount, TotalInscClaimAmtReimbursed.
- **Financial Ratios:** Average reimbursement per beneficiary and average deductible paid.
- **Interaction Metrics:** AvgLengthOfStay (duration of inpatient admissions) and UniqueBeneficiaryCount.
- **Patient Health:** NumChronicConditions_Mean (the average health status of a provider's patients).

Exploratory analysis revealed a strong positive correlation between claim volume and fraudulent activity. However, the dataset was highly imbalanced, with fraudulent providers representing only roughly 9.35% of the total population, necessitating careful handling during modeling.

3. Methodology and Modeling We treated this as a binary classification problem (Fraud vs. Legitimate). Three supervised learning algorithms were trained and evaluated:

1. **Logistic Regression:** Used as a baseline. While interpretable, it struggled to capture the complex, non-linear relationships inherent in fraud patterns.
2. **Random Forest:** A robust ensemble method that handled outliers and non-linearities better than the baseline.
3. **Gradient Boosting Classifier (GBM):** This model builds decision trees sequentially, with each tree correcting the errors of the previous ones.

The models were trained using 5-fold Stratified Cross-Validation to ensure stability. The **Gradient Boosting Classifier** was selected as the final model due to its superior performance in distinguishing between classes.

4. Evaluation and Results The final model was evaluated on a held-out test set to simulate real-world performance.

- **ROC-AUC Score: 0.9536** This score indicates excellent discriminative ability. The model is highly effective at ranking fraudulent providers higher than legitimate ones.
- **Precision (Fraud Class): 74%** When the model flags a provider as fraudulent, it is correct 74% of the time. This is a strong result that minimizes the administrative burden of investigating false alarms.
- **Recall (Fraud Class): 54%** At the default decision threshold (0.5), the model successfully identifies 54% of all actual fraud cases. While decent, this suggests that nearly half of the fraud cases—likely those with more subtle patterns—are being missed.

5. Error Analysis We analyzed the model's errors to understand its limitations:

- **False Positives (Legitimate providers flagged as fraud):** The model occasionally misclassifies legitimate, high-volume providers (e.g., large hospitals) as fraudulent. This is because they share characteristics with fraudsters, such as high total reimbursements and high beneficiary counts.
- **False Negatives (Fraudulent providers missed):** The model struggles to detect "low-intensity" fraud. These are providers who commit fraud but keep their claim volumes and lengths of stay low to avoid detection. They mimic the statistical profile of a small, legitimate practice.

6. Conclusion and Recommendations The Gradient Boosting model serves as a robust foundation for a fraud detection system, capturing the majority of high-impact fraud cases. To improve operational utility, we recommend the following:

- **Adjust the Threshold:** The default threshold of 0.5 is conservative. Lowering the threshold to roughly 0.30 would significantly increase the Recall (catching more fraud), which is financially justifiable even if it results in slightly more investigations.
- **Normalize Features:** To reduce False Positives among large facilities, future iterations should normalize volume features (e.g., calculating "Claims per Bed" or "Reimbursement per Region") rather than using raw totals.
- **Incorporate Temporal Data:** Adding features that detect spikes in activity (e.g., weekend claims or rapid-fire billing) could help identify the "low-intensity" fraud that is currently being missed.