DataOrbit: Healthcare Provider Fraud Detection

1. Project Context

1.1 Executive Summary

Healthcare fraud imposes a massive financial burden on the U.S. system, costing over $68 billion annually. CMS currently relies on reactive, rule-based systems that struggle to detect sophisticated schemes. Data Orbit was tasked with developing a proactive machine learning solution. Our objective was to flag high-risk providers based on behavioral patterns in claims data while minimizing false positives to protect legitimate reputations.

2. Data Analysis & Engineering

2.1 Data Architecture

The source data provided by CMS was highly relational, consisting of four distinct files: Labels (Provider fraud status), Beneficiary (Patient demographics), Inpatient Claims, and Outpatient Claims.

A critical challenge was the granularity mismatch: Fraud labels exist at the 'Provider' level, but the behavioral data exists at the 'Claim' level. To bridge this, we engineered a data pipeline that aggregated transactional claim lines into provider-level behavioral fingerprints.

2.2 Data Integrity & Imbalance

We identified significant missing data in fields such as 'Date of Death' (DOD) and procedure codes. These were handled logically (e.g., missing DOD implies the patient is alive). A major modeling challenge was Class Imbalance: out of 5,410 providers, only ~9.3% were flagged as fraudulent. This required specific resampling strategies during training.

2.3 Feature Engineering

We moved beyond raw data by creating composite features that describe provider behavior:

- Volume Metrics: Total claim counts and distinct beneficiary counts.
- Financial Metrics: Average reimbursement and deductible amounts.
- Network Metrics: 'Claims per Beneficiary' (churning) and 'Inpatient-to-Outpatient Ratios'.

These features allow the model to detect providers who bill excessively compared to their peer group.

3. Methodology and Modeling

3.1 Preprocessing Strategy

To ensure a robust evaluation, we employed a stratified 75/25 train/test split. This ensured that the ratio of fraud cases in our testing environment perfectly matched the real-world distribution.

3.2 Addressing Imbalance (SMOTE)

Because the minority class (Fraud) was so small, standard models would bias towards the majority class. We applied SMOTE (Synthetic Minority Over-sampling Technique) to the training data only. This generated synthetic examples of fraud providers, allowing the algorithm to learn the decision boundary more effectively without contaminating the test set.

3.3 Algorithm Selection

We experimented with two distinct modeling approaches:

- Baseline (Logistic Regression): Chosen for its high interpretability. It performed well on Recall (catching fraud) but suffered from very low Precision (too many false alarms).
- Primary (Random Forest): A non-linear tree-based ensemble. This model outperformed the baseline because fraud patterns are rarely linear; they often involve complex interactions between volume, cost, and patient demographics. The Random Forest successfully captured these nuances.

4. Model Evaluation Matrix

Metric	Logistic Regression	Random Forest	Interpretation
ROC-AUC	0.92	0.92	Both models rank fraud risks effectively. The Area Under the Curve is identical, but ranking capability is only one part of the story.
Precision	0.40	0.51	Random Forest is significantly better here. It is correct 51% of the time it flags fraud, compared to only 40% for LR. This reduces auditor workload.
Recall	0.85	0.71	Logistic Regression catches more fraud (85%) but generates too many false alarms. RF misses some cases (71%) but is much more trustworthy.
F1-Score	0.54	0.60	The F1-Score (harmonic mean of Precision and Recall) favors Random Forest, confirming it is the more balanced and deployment-ready model.

5. Error Analysis & Conclusion

5. Error Analysis
Our analysis of the model's errors revealed specific patterns:
- False Positives (Type I Error): The model occasionally flagged legitimate providers who had very high claim volumes (e.g., large urban hospitals). These entities statistically resemble fraud rings due to sheer volume.
- False Negatives (Type II Error): The model missed "quiet" fraudsters-providers who bill fraud frequently but in very small amounts, staying below the statistical radar.

6. Conclusion: Why Random Forest Prevailed
Upon reviewing the results, the Random Forest (RF) model is the clear recommendation for deployment. Its superiority over the Logistic Regression baseline is driven by three key technical factors:

1. Handling Non-Linearity & Interactions:
Fraud is rarely defined by a single variable. For example, high claim volume alone is not fraud (it might just be a busy hospital). However, high claim volume combined with low distinct beneficiaries indicates "churning." Linear models struggle to map these complex interactions without manual feature engineering. Random Forest naturally creates decision branches that isolate these specific combinations of behaviors.

2. Robustness to Noise:
Healthcare data is inherently noisy. Logistic Regression attempts to fit a smooth curve to the data, which allows outliers (like extremely expensive but legitimate surgeries) to skew the results, generating False Positives. The Random Forest ensemble method averages out these anomalies across hundreds of trees, resulting in a more stable decision boundary.

3. Operational Efficiency (Precision):
In fraud detection, the cost of a False Positive is high-it wastes the time of expert investigators. While Logistic

Regression caught more fraud, it also accused 60% of the providers it flagged falsely. Random Forest achieved a Precision of 51%, meaning every second alert it generates is actionable fraud. This efficiency makes it a far more practical tool for the CMS audit team.