DataOrbit: Technical Report

1. Project Context

1.1 Overview

Healthcare fraud costs the U.S. system over $68 billion annually. CMS currently relies on reactive, rule-based systems. DataOrbit has developed a proactive machine learning solution to flag high-risk providers based on behavioral patterns, minimizing false positives to protect legitimate reputations.

2. Data Analysis & Engineering

2.1 Data Sources and Structure

The data provided by CMS was relational:

- Labels: Provider fraud status.
- Beneficiary: Patient demographics.
- Inpatient/Outpatient: Claims data.

Key Insight: Fraud exists at the Provider level, but data is at the Claim level. Aggregation was performed to create a provider-centric view.

2.2 Data Quality

Significant missing data in 'Date of Death' (DOD) was handled logically. Class imbalance was high (only ~9.3% fraud), necessitating resampling strategies like SMOTE to ensure robust training.

2.3 Feature Engineering

We transformed transactional data into provider behavioral fingerprints:

- Volume: Total claim counts and distinct beneficiaries.
- Financial: Average reimbursement amounts and deductibles.
- Network: Claims per beneficiary and physician connections.
- Service Mix: Ratio of inpatient to outpatient services.

3. Methodology and Modeling

3.1 Preprocessing

We used a stratified 75/25 train/test split to maintain fraud ratios in evaluation.

3.2 Handling Imbalance

We applied SMOTE (Synthetic Minority Over-sampling) to the training data only. This allowed the model to learn fraud patterns without overfitting, while keeping the test data pure and realistic.

3.3 Model Selection

- Experiment A (Baseline): Logistic Regression. Resulted in high recall but low precision (many false alarms).
- Experiment B (Primary): Random Forest. Achieved superior balance, capturing non-linear fraud patterns better than linear models.

4. Evaluation

4.1 Performance Metrics

We evaluated models on the 25% held-out test set. The table below summarizes the results:

Metric	Logistic Reg.	Random Forest	Interpretation
ROC-AUC	0.92	0.92	Both rank risk effectively.
Precision	0.40	**0.51**	RF reduces false alarms.
Recall	0.85	0.71	LR catches more but flags innocents.
F1-Score	0.54	**0.60**	RF is the balanced choice.

4.2 Error Analysis (Confusion Matrix)

Detailed breakdown of the Random Forest predictions on the test set:

- True Positives (90): Fraudsters correctly identified.
- False Positives (85): Legitimate providers flagged. Likely due to high volumes in large hospital systems.
- False Negatives (37): Fraudsters marked safe. Likely 'quiet' fraudsters with low volume.
- True Negatives (1141): Legitimate providers correctly ignored.

5. Conclusion

The Random Forest model successfully identifies 71% of fraud cases with a Precision of 51%, significantly outperforming the baseline. This ensures the audit team focuses on high-probability targets, optimizing operational efficiency.