# DataOrbit HealthCare Provider Fraud Detection

## Technical Report & Project Documentation

Date: December 3, 2025

Repository: fraud\_detection\_project

## 1. Executive Summary

DataOrbit was contracted to develop a data-driven solution for detecting fraudulent healthcare providers. The objective was to replace legacy rule-based systems with a machine learning pipeline capable of identifying sophisticated fraud patterns while minimizing false positives.

Key Outcome:

Our team successfully developed a Logistic Regression model utilizing class\_weight='balanced' and L2 regularization.

* **Performance:** The final model achieved a **Recall of 1.0** (capturing 100% of known fraud in the test set) and a **Precision of ~0.95**.
* **Business Impact:** This model serves as a highly effective first-pass filter, ensuring no potential fraud is missed while maintaining a manageable audit workload.

## 2. Data Exploration & Feature Engineering

### 2.1 The Challenge: Granularity Mismatch

The primary data engineering challenge was the structural mismatch between the input data (Claim-Level) and the prediction target (Provider-Level).

* **Raw Data:** 500,000+ transactional claims (Inpatient/Outpatient).
* **Target:** ~5,400 distinct Providers flagged as Yes/No for fraud.

### 2.2 Aggregation Strategy

To resolve this, we implemented a robust **Aggregation Pipeline** consolidating transactional behaviors into provider profiles.

Rationale: Fraud is rarely a single event but a pattern of behavior over time.

Implementation Steps:

1. **Beneficiary Enrichment:** Merged patient demographics (Train\_Beneficiarydata.csv) onto claims *before* aggregation to preserve patient context.
2. **Statistical Summarization:** Grouped data by Provider to calculate specific risk indicators.
3. **Feature Categories Created:**
   * **Financial Velocity:** TotalReimbursement, AvgReimbursement (Captures profit-seeking behavior).
   * **Patient Demographics:** AvgAge, ChronicCond\_KidneyDisease\_Prevalence (Detects if a provider targets vulnerable populations).
   * **Operational Patterns:** InpatientRatio, AvgClaimDuration (Detects anomalies in hospital stays).

### 2.3 Exploratory Data Analysis (EDA) Insights

* **Insight 1 (Financials):** Fraudulent providers exhibited significantly higher mean reimbursements and claim counts compared to legitimate ones.
* **Insight 2 (Patient Health):** A higher prevalence of ischemic heart disease and Renal Failure was observed in the patient base of fraudulent providers, suggesting possible upcoding (billing for more severe conditions than treated).

## 3. Modelling Methodology

### 3.1 Class Imbalance Strategy

Problem: The dataset was highly imbalanced (~9% Fraud, ~91% Non-Fraud).

Decision: We rejected Undersampling (loss of data) and Oversampling/SMOTE (risk of creating synthetic noise).

Solution: We utilized Cost-Sensitive Learning (class\_weight='balanced').

* **Rationale:** This method penalizes the model significantly more for missing a fraud case (False Negative) than for flagging a legitimate doctor (False Positive), directly aligning the algorithm with the business goal of maximizing Recall.

### 3.2 Algorithm Selection

We tested three distinct model architectures to evaluate the trade-off between complexity and interpretability.

|  |  |  |  |
| --- | --- | --- | --- |
| **Algorithm** | **Pros** | **Cons** | **Outcome** |
| **Logistic Regression** | Highly interpretable; explicit feature weights; robust to noise. | Assumes linear relationships. | **Selected (Best Performer)** |
| **Random Forest** | Handles non-linearities; generally robust. | "Black box"; struggled with Recall in this specific dataset. | Discarded |
| **Gradient Boosting** | High potential accuracy. | Prone to overfitting on smaller datasets. | Discarded |

### 3.3 Experimental Log & Trials

|  |  |  |  |
| --- | --- | --- | --- |
| **Trial ID** | **Configuration** | **Metric Focus** | **Result & Insight** |
| **Exp-01** | **Baseline:** Logistic Regression (No class weights). | Accuracy | **Failed.** Accuracy was 90%+, but Recall was near 0. The model simply guessed "No Fraud" for everyone. |
| **Exp-02** | **Tree Models:** Random Forest (Default params). | PR-AUC | **Underperformed.** The model struggled to isolate the minority class, achieving a Recall of only ~0.53. |
| **Exp-03** | **Ensemble:** Gradient Boosting. | F1-Score | **Mixed.** Better Recall (0.86) but poor Precision (0.50). Too many false alarms. |
| **Exp-04** | **Final Config:** Logistic Regression (class\_weight='balanced', C=0.01). | Recall/PR-AUC | **Success.** Achieved perfect Recall (1.0) and high Precision (0.95). The linear decision boundary effectively separated the high-cost fraudsters. |

## 4. Evaluation & Error Analysis

### 4.1 Quantitative Performance (Test Set)

The final Logistic Regression model produced the following results:

* **Recall:** 1.00 (Captured 101/101 Fraudulent Providers).
* **Precision:** 0.95 (Minimal False Positives).
* **F1-Score:** 0.97.

### 4.2 Error Analysis: The Cost of Mistakes

**False Negatives (Type II Error):**

* *Count:* 0
* *Implication:* The model did not miss a single fraudulent provider in the test set. This is the ideal outcome for a screening tool, preventing financial loss.

**False Positives (Type I Error):**

* *Count:* Very Low (< 5 cases in validation)
* *Implication:* A small number of legitimate providers were flagged for audit.
* *Root Cause Analysis:* These providers were typically large research hospitals or specialized clinics dealing with terminally ill patients (high DeceasedRatio, high Reimbursement), mimicking the cost patterns of fraud.
* *Mitigation:* These cases can be easily cleared by a human auditor reviewing the specific nature of the facility.

## 5. Conclusion:

The project successfully delivered a robust fraud detection model. By prioritizing **Recall** through cost-sensitive learning and utilizing a transparent **Logistic Regression** framework, we satisfied the dual requirements of high detection rates and model explainability.