## **Overview**

This report documents the exploratory data analysis (EDA) conducted on the cleaned, provider-level feature set for the Healthcare Fraud Detection project. It is designed for both technical teams and business stakeholders, combining automated profiling and targeted visual analysis.

## **What Was Done**

* Loaded all cleaned data from /data/processed/.
* Aggregated claim and beneficiary features at the provider level (e.g., total/mean reimbursement, claim counts, distinct beneficiaries).
* Generated an automated EDA profiling report (provider\_profile\_report.html) to analyze distributions, correlations, outliers, and target relationships.
* All steps follow the Healthcare Fraud Detection EDA standards from the knowledge base.

## **Key Visuals and Insights**

**1. Target Distribution (Class Imbalance):**

* Fraudulent: **506 providers (9.35%)**
* Legitimate: **4,904 providers (90.65%)**
* Class imbalance confirmed—precision-recall focus needed in modeling.

**2. Feature Distributions:**

* **Total/Mean Claim Amounts:** Heavily right-skewed; a small number of providers account for outsized reimbursements.
* **Claim Counts:** Most providers have few claims; a minority submit very high volumes.

**3. Outlier Detection:**

* Multiple numeric features display significant outliers (especially reimbursement and deductible sums).
* Outliers may represent either large clinics or potential fraud—will be flagged for business review.

**4. Correlation Analysis:**

* High positive correlation between sum of reimbursements, claim count, and sum of deductibles.
* Mean values show less correlation—some providers process high-value but fewer claims.

**5. Feature-Target Relationships:**

* Fraudulent providers have:  
  + Avg unique beneficiaries: **215.9** (fraud) vs **46.9** (legitimate)
  + Higher claim-to-beneficiary ratios
  + Higher average claims per provider: **420.5 (fraud)** vs **70.4 (legitimate)**
  + Higher total reimbursements: **$584,350 (fraud)** vs **$53,194 (legitimate)**
* These separations match published benchmarks and support their use in modeling.

## **Risk Stratification (from Knowledge Base)**

* **Low Risk:** 4,298 providers (1.9% fraud rate)
* **Medium Risk:** 548 providers (17.2% fraud rate)
* **High Risk:** 255 providers (43.9% fraud rate)
* **Very High Risk:** 309 providers (70.6% fraud rate)

## **Business Context for Outliers**

* High-activity providers may be legitimate (e.g., regional hospitals or clinics) or fraud signals.
* Each flagged outlier will be reviewed by the business team for operational understanding.
* Decision: Outliers are **not dropped** at this stage; they are flagged and described in the feature table.

## **Key Takeaways**

* **Top features that separate fraud from legitimate:**
  + Claim count per provider
  + Total and mean reimbursement amounts
  + Unique beneficiary count
  + Claim-to-beneficiary ratio
  + Deductible totals
* **Recommended Modeling Approach:**
  + Use provider-level aggregated features with tree-based models (XGBoost, Random Forest)
  + Focus on precision-recall metrics, with threshold tuning
  + Apply stratified sampling or class weights due to 9.35% fraud rate
* **Data Quality/Limitation Notes:**
  + No missing in engineered features
  + Outlier-rich, long-tailed numeric distributions—careful model interpretation needed
  + Some categorical enrichments (e.g., provider specialty) could further improve modeling if available