# **📑 SHAP Model Explainability Notebook – Documentation**

## **1. Overview**

This notebook delivers a full **model interpretability analysis** for our healthcare fraud detection project, using **SHAP (SHapley Additive exPlanations)** to explain and audit the predictions of our final, production-ready XGBoost model.  
 Interpretability is critical for trust, regulatory compliance, and actionable business insight in fraud detection.

## **2. Workflow & Methodology**

### **2.1. Model and Data Loading**

* **Loaded the trained XGBoost model** (xgboost\_fraud\_model.json) previously saved from the modeling notebook.
* **Validation set** (features and labels) was loaded from parquet files to ensure consistency and reproducibility with the final evaluation metrics.

### **2.2. SHAP Explainer Setup**

* Used shap.TreeExplainer (optimized for XGBoost) to compute SHAP values for each provider in the validation set.
* SHAP values quantify, for each provider and feature, **how much that feature “pushed” the model towards or away from predicting fraud.**

### **2.3. Global Feature Importance**

* **SHAP summary bar plot** and **beeswarm plot** were generated to show which features are most important overall in fraud detection.  
  + The bar plot ranks features by mean absolute SHAP value (average impact).
  + The beeswarm plot shows the direction of impact (whether high or low feature values increase fraud risk).

### **2.4. Local (Per-Provider) Explanations**

* For a sample provider flagged as fraud, generated **force and waterfall plots** to visualize exactly which features contributed most to the prediction.
* Printed all feature values for transparency and auditability.

## **3. Key Findings**

### **3.1. Global Insights**

* **Top predictors of fraud:**
  + total\_reimb (total reimbursement): By far the most powerful indicator; providers with unusually high reimbursements are at greatest risk.
  + kidney\_rate, heartfail\_rate: High prevalence of certain chronic conditions among patients signals possible fraud.
  + inpatient\_claims, claims\_per\_bene, median\_reimb: Patterns of high volume and value of claims are strong fraud signals.
* **SHAP beeswarm plot** confirmed that **high values** of financial and claim-count features drive the model toward fraud predictions.

### **3.2. Local (Per-Provider) Explanation**

* For a specific provider flagged as fraud:  
  + **Exceptionally high total reimbursement and high inpatient claims** were the dominant drivers of the fraud prediction.
  + Other features (e.g., claims per beneficiary, diabetes rate) had smaller mitigating or amplifying effects.
  + The model’s decision was fully transparent and could be audited step-by-step.

## **4. Business & Compliance Value**

* **SHAP provided actionable insights**: Audit teams can prioritize investigation of providers with similar feature patterns.
* **Transparency & trust:** SHAP enables clear justification for every model prediction, supporting regulatory compliance and fair decision-making.
* **Model auditability:** All plots, feature values, and explanations are saved and can be reviewed for each provider.

## **5. Best Practices Applied**

* **Data, model, and splits were versioned and reproducible.**
* **Interpretability tools (SHAP) were applied both globally and locally.**
* **Documentation is business-friendly, technically sound, and ready for audit or stakeholder review.**

## **6. Next Steps / What’s Missing**

* Integrate SHAP visualizations into reporting dashboards or case review tools.
* Apply batch SHAP explanations to all flagged providers for systematic auditing.
* Optionally, use SHAP feature importances to further refine or simplify the model if desired.

## **7. Summary Statement**

**This notebook closes the loop on model transparency for healthcare fraud detection.  
 All predictions can now be explained and defended, aligning with the highest standards for compliance, audit, and business trust.**