## **Executive Summary**

This report documents the end-to-end modeling, validation, and selection process for a healthcare fraud detection system at the provider level. The goal is to flag providers with abnormal patterns indicative of potential fraud, using a rigorous, auditable, and business-aligned workflow. The entire pipeline—from feature engineering through model training, validation, threshold setting, and business impact analysis—was designed to be transparent, reproducible, and ready for production deployment.

## **Data Preparation & Feature Engineering**

* **Datasets:** Inpatient, outpatient, and beneficiary claims for Medicare providers; labels for known fraud.
* **Initial EDA:** Automated profiling and custom analysis confirmed severe class imbalance (fraud = 9.3%), strong skew in claim volumes, and high feature relevance for claim, demographic, and billing pattern metrics.
* **Feature Engineering:** 30+ provider-level features were engineered, including:  
  + Total/average/median/max reimbursement
  + Claim counts (total, inpatient, outpatient)
  + Inpatient/outpatient ratios
  + Unique beneficiary count, average age, % deceased
  + Claim diversity metrics, high-risk claim indicators
* **Feature Quality:** Nulls were handled (imputed with 0), distributions and correlations validated, and all features documented for business transparency.

## **Train/Test Splitting & Class Imbalance Handling**

* **Stratified 80/20 train/test split:** Maintained 9.3% fraud in both sets (train: 4,328, test: 1,082 providers).
* **Balanced training set:** Majority class downsampled for 50/50 fraud/non-fraud (405 each) to enable minority pattern learning.
* **Evaluation:** All models were trained on balanced data and evaluated on imbalanced test data to reflect real-world business conditions.

## **Model Training & Tuning**

Four model types were trained, tuned, and compared:

| **Model** | **AUC** | **F1** | **Recall** | **Precision** |
| --- | --- | --- | --- | --- |
| Logistic Regression | 0.955 | 0.55 | 0.88 | 0.40 |
| Random Forest | 0.954 | 0.53 | 0.88 | 0.37 |
| XGBoost | 0.951 | 0.51 | 0.87 | 0.36 |
| Decision Tree | 0.896 | 0.50 | 0.84 | 0.35 |

**Model selection criteria:**

* High recall (detecting most frauds)
* Manageable precision (controlling investigation load)
* Business interpretability
* Reproducibility and stability (cross-validated AUC)

### **Hyperparameter Tuning**

* **Random Forest:** Tuned n\_estimators ([100, 300, 500]) and max\_depth ([10, 20, None]); best: 300 trees, max\_depth=10.
* **XGBoost:** Tuned learning\_rate ([0.01, 0.1, 0.2]), n\_estimators, max\_depth; best: 0.01, 500 trees, depth=5.
* **Decision Tree:** Trained at max\_depth=6 for interpretability.

## **Key Insights From Feature Importances**

* **Top predictors** across all models:  
  + total\_reimb, total\_deductible, max\_reimb, total\_claims, and avg\_deductible
* **Fraudulent providers** have significantly higher claim volume and reimbursement.
* **Patient demographic features** (e.g., % deceased, diabetes rate) contribute but are less dominant than financial metrics.

## **Visualizations**

* **ROC and Precision-Recall Curves:** All models show high area under the curve, indicating strong separation of fraud vs non-fraud.
* **Threshold tuning:** Business-ready plot produced for precision and recall at different thresholds.

## **Threshold Selection & Business Constraints**

* **Investigation capacity:** Stakeholders can review ~100 flagged providers/month.
* **Threshold selection:** At threshold = 0.8945 (XGBoost), ~100 providers are flagged per batch.
* **Business impact:** Maximizes fraud capture (recall ~0.87), while controlling review volume.

## **Interpretability & Business Communication**

* **Decision Tree:** Shallow tree model trained and visualized for compliance and business rules extraction.
* **Feature dictionary, pipeline, and visualizations** delivered for end-to-end transparency.

## **Deployment Recommendations**

* **Model to deploy:** XGBoost with (learning\_rate=0.01, n\_estimators=500, max\_depth=5).
* **Threshold for investigation:** 0.8945.
* **Artifacts delivered:**
  + /models/xgb\_fraud\_model.pkl
  + /models/xgb\_threshold.txt
  + /reports/model\_comparison.csv
  + ROC/PR/feature importance plots for all models
  + Final decision tree visualization
* **Feature pipeline:** Documented in /reports/feature\_engineering\_summary.txt

## **Monitoring & Maintenance**

* **Monitor:** Monthly flagged count, fraud confirmation rate, feedback from investigators.
* **Trigger retraining:** If flagged count, fraud capture rate, or input distributions change significantly.
* **Audit:** All splits, parameters, and data transformations are version-controlled and reproducible.

## **Next Steps**

* **Optional explainability:** Deploy SHAP or TreeExplainer for local and global explanation to investigators.
* **Error analysis:** Analyze false positives/negatives for future improvements.
* **Automation:** Deploy batch scoring pipeline for new provider claims monthly/quarterly.

## **Appendices**

* **All code, plots, and artifacts are stored in Google Drive under /models/ and /reports/.**
* **Modeling decisions and deviations from plan are tracked in MLflow runs and included in markdown reports.**

## **Summary**

This project delivers a robust, explainable, and business-aligned fraud detection model—ready for production and stakeholder review. All steps, from EDA to deployment recommendation, are documented, reproducible, and designed for real-world insurance fraud challenges.