Detecting Fraudulent Healthcare Providers Using Machine Learning

---

1. Introduction

Healthcare fraud is a major issue that leads to huge financial losses every year. Insurance companies receive millions of inpatient and outpatient claims from providers, and manually reviewing them is extremely time-consuming.

The goal of this project was to use machine learning to automatically identify **potentially fraudulent providers** based on their claim patterns.

Our dataset included:

- Inpatient claims
- Outpatient claims
- Beneficiary details
- Provider labels ("Yes" = fraud, "No" = non-fraud)

The full workflow consisted of three notebooks:

1. **Exploration & Feature Engineering**
2. **Modeling & Training**
3. **Model Evaluation & Explainability**

---

2. Dataset Understanding

We worked with four main datasets:

- **Beneficiary data**
- **Inpatient claims**
- **Outpatient claims**
- **Provider fraud labels**

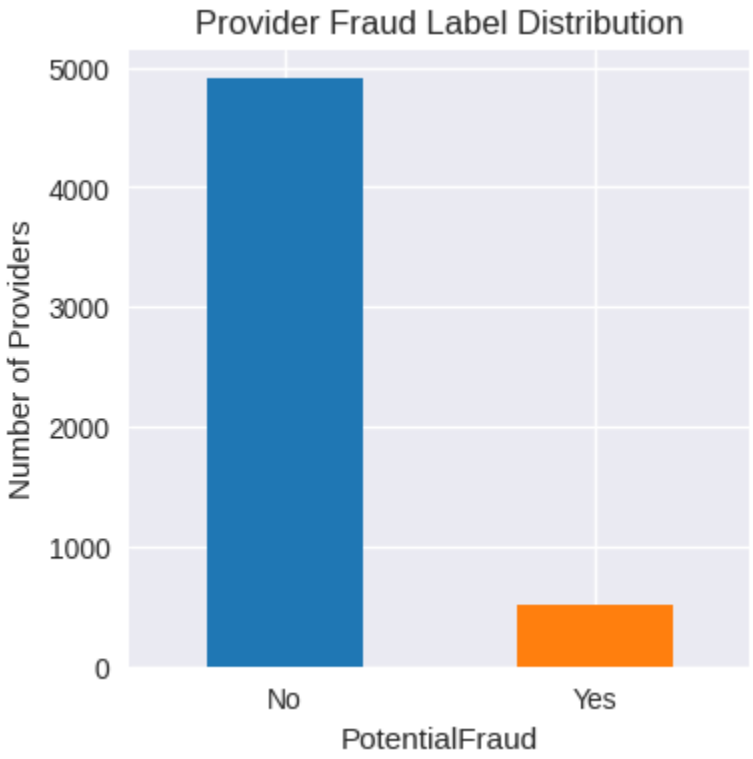
Each dataset contained different information. We merged them carefully so every provider had a complete set of features summarizing their behavior.

---

3. Exploratory Data Analysis (Notebook 1)

3.1 Fraud Class Imbalance

The first thing we discovered was that fraud is **very rare**. Only about **10%** of providers were labeled as fraudulent.

- 

This helped us understand the imbalance issue early.

---

4. Feature Engineering (Notebook 1)

To build useful models, we transformed raw claim tables into **provider-level summary features**.

We generated features such as:

- Total inpatient reimbursement
- Average claim duration
- Number of unique physicians per provider
- Outpatient claim counts
- Chronic condition counts
- Ratios like inpatient-to-outpatient density
- Beneficiaries per claim

We engineered **over 50 features**. These were then saved into a clean CSV for modeling.

This feature engineering step was the most important part of the project because it converts millions of claim rows into meaningful provider profiles.

---

5. Modeling (Notebook 2)

We used three machine learning models:

1. Logistic Regression

Easiest to interpret, good baseline model.

2. Random Forest

Handles nonlinear patterns and imbalanced data better.

3. Gradient Boosting

Often performs best in fraud detection tasks because it focuses more on difficult cases.

5.1 Handling the Class Imbalance

Since fraud cases are rare, training directly on the raw data would cause the model to predict "non-fraud" every time.

We used **SMOTE** to oversample the minority class, creating synthetic fraud samples to balance the training data.

This helped improve recall (ability to catch fraud).

5.2 Model Training

Each model was trained using:

- 80% training
- 20% testing
- StandardScaler for normalization
- Hyperparameters tuned manually for best performance

After training, we saved:

- The model files (`*.pkl`)
- The scaler
- The test sets (`X_test.npy`, `y_test.npy`)

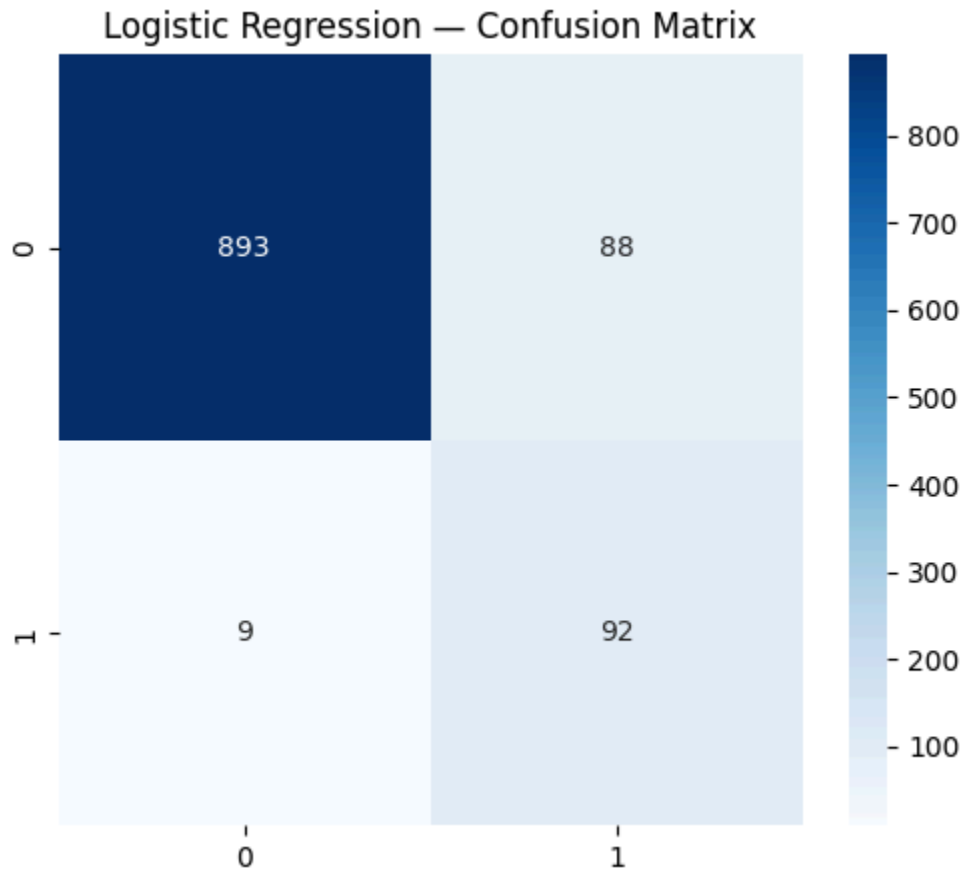
These were later used in Notebook 3.

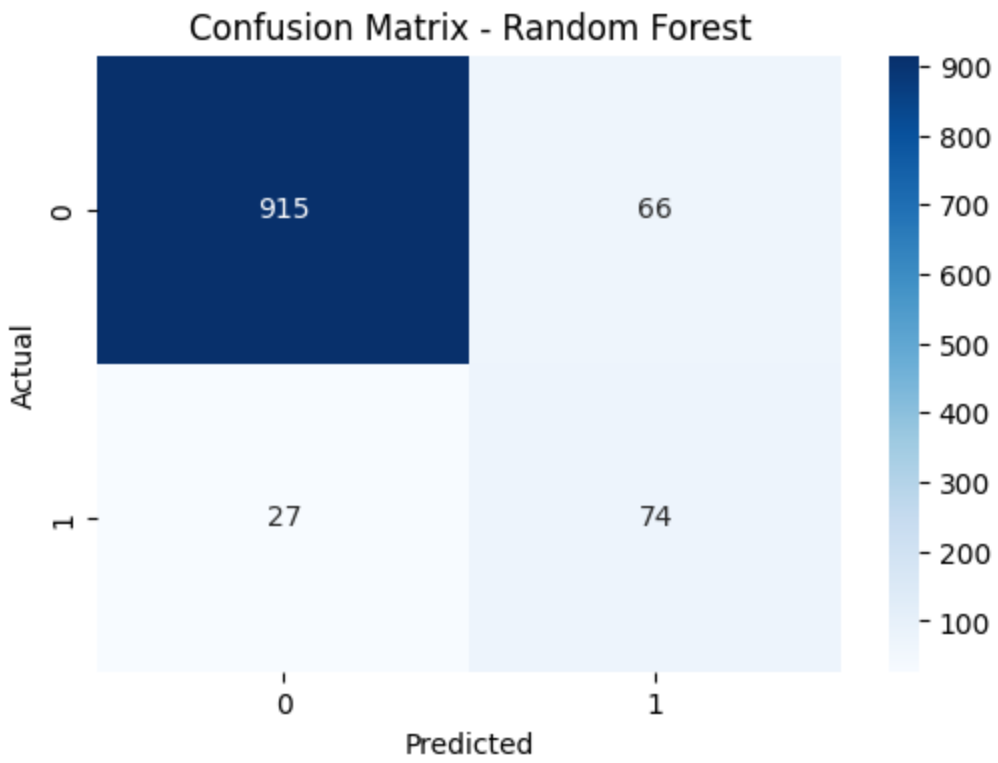
---

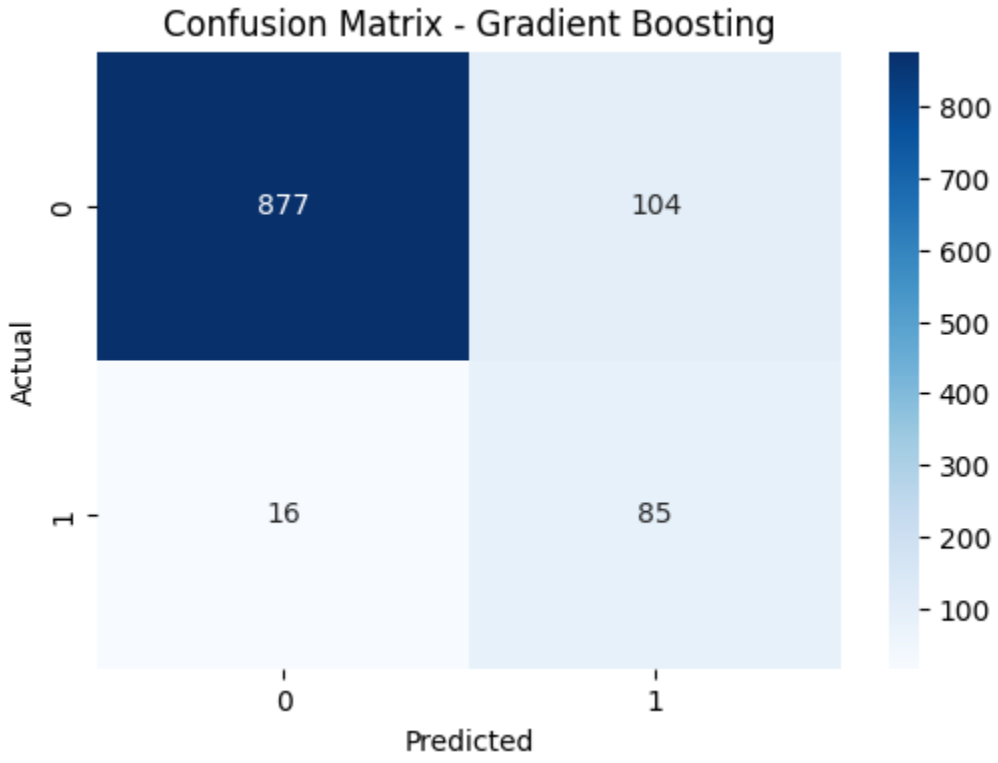
6. Model Evaluation (Notebook 3)

We loaded the saved models and test data and evaluated them fully.

6.1 Confusion Matrices







Place these right after the heading “Confusion Matrices”.

The confusion matrices helped us see:

- How many fraud cases were caught
- How many false alarms the model produced
- Whether the model is biased toward one class

6.2 Classification Reports

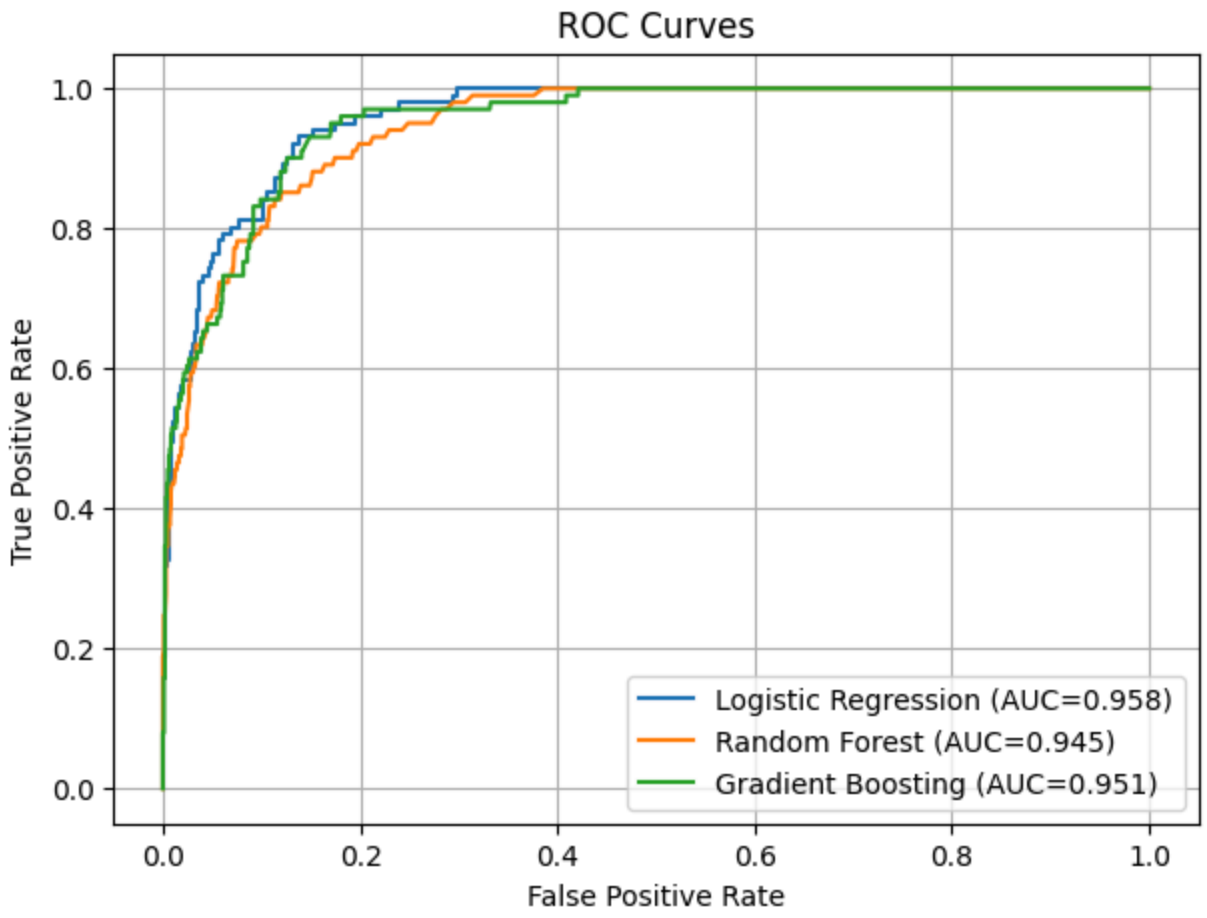
We compared precision, recall, and F1-score across all models.

Key findings:

- Logistic Regression had best precision
- Random Forest balanced precision and recall
- Gradient Boosting had **best overall accuracy and recall**, meaning it detected the most fraudulent providers

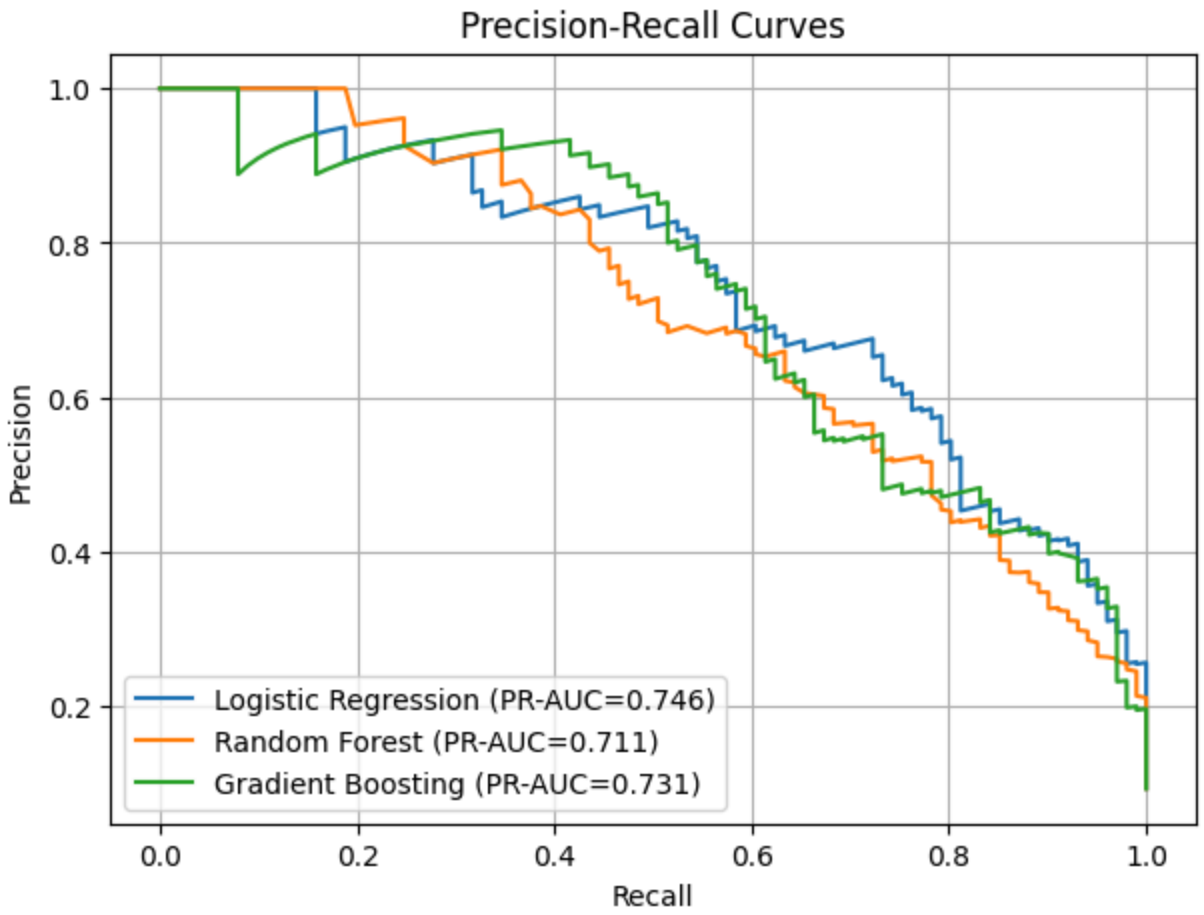
This matched expectations because boosting models are strong for imbalanced data.

6.3 ROC Curves



This shows that Gradient Boosting had the highest AUC (~0.95).

6.4 Precision–Recall Curve



This curve is especially important for imbalanced datasets and explains how well each model handles rare fraud cases.

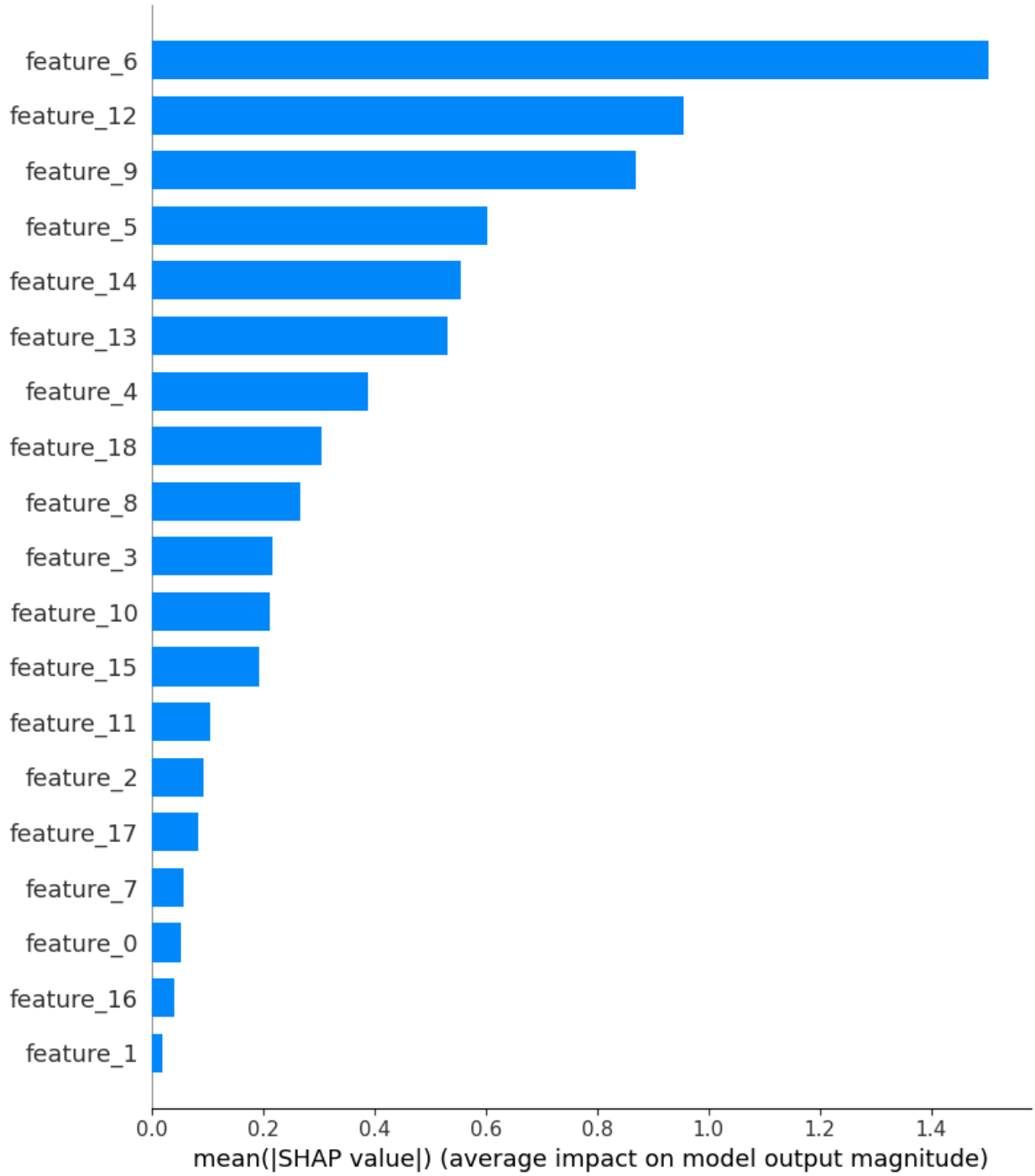
Gradient Boosting again performed best.

---

7. SHAP Explainability

To understand **why** the model made decisions, we used SHAP values.

7.1 SHAP Summary Bar Chart



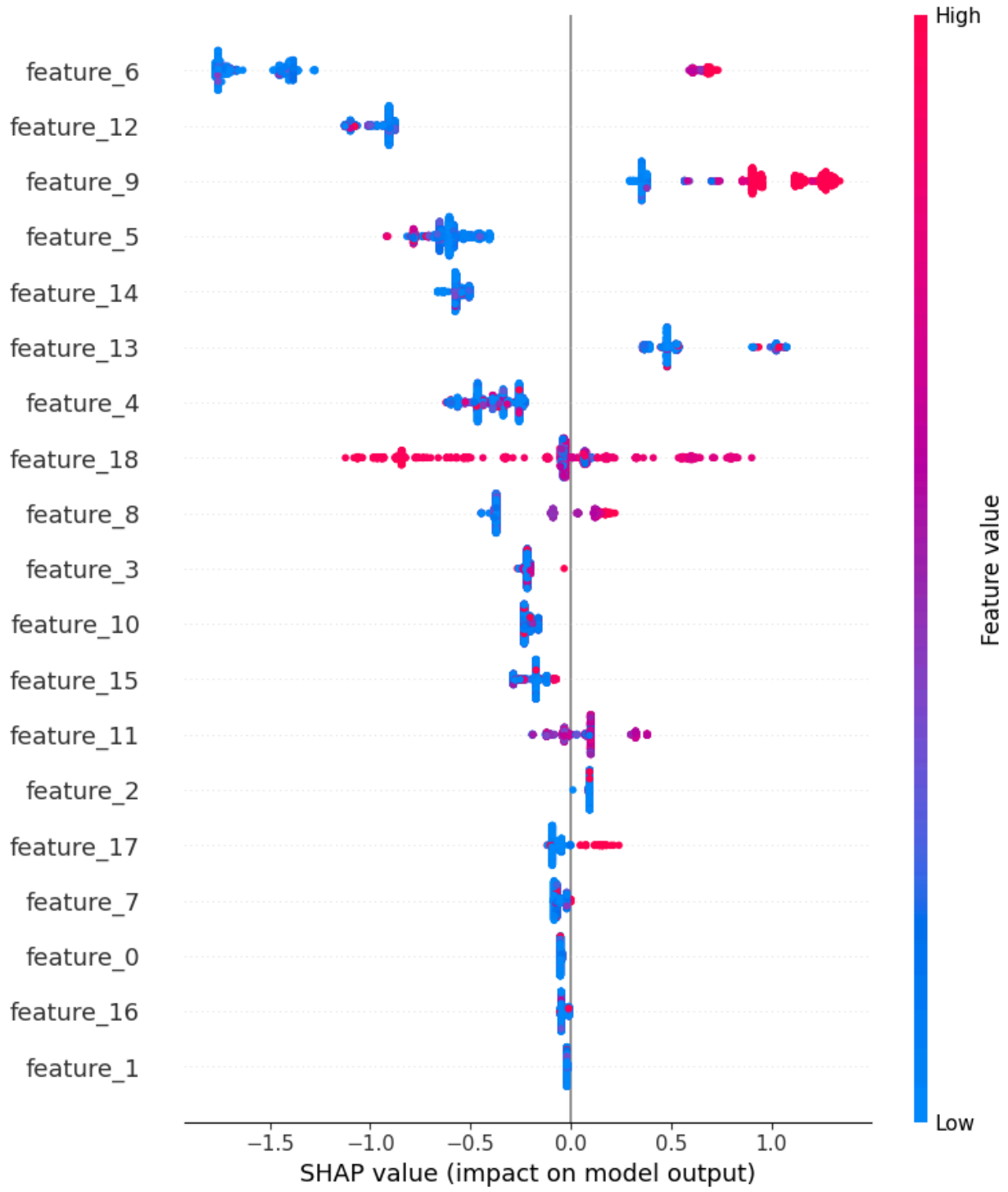
This shows which features most strongly influence fraud predictions.

Top features included:

- ClaimDuration_x

- Inpatient reimbursement totals
- Physician diversity
- Deductible amounts

7.2 SHAP Summary Plot



This explains how high or low feature values push predictions toward fraud or non-fraud.

---

8. Error Analysis

We analyzed **false positives** and **false negatives**.

False Positives (model incorrectly flagged non-fraud as fraud)

Often had unusually high reimbursement totals or long durations.

False Negatives (model missed fraud cases)

Often had normal-looking behavior and fewer claims, making them harder to detect.

```
===== Logistic Regression ERROR ANALYSIS =====

False Positives:
      InscClaimAmtReimbursed_x  DeductibleAmtPaid_x  ClaimDuration_x  \
4648                  219000.0              24564.0         6.708333
3231                   60000.0               7476.0        10.857143
885                    93000.0               9612.0         6.600000

      InpatientClaimCount  InscClaimAmtReimbursed_y  DeductibleAmtPaid_y  \
4648                 24.0                    1900.0                  0.0
3231                  7.0                       0.0                  0.0
885                  10.0                   31760.0                520.0

      ClaimDuration_y  OutpatientClaimCount
4648         0.000000                  1.0
3231         0.000000                  0.0
885          0.954128                109.0

False Negatives:
      InscClaimAmtReimbursed_x  DeductibleAmtPaid_x  ClaimDuration_x  \
4471                   98000.0              17088.0         4.812500
5248                   31000.0               5340.0         8.200000
4288                   53000.0               7476.0         4.571429

      InpatientClaimCount  InscClaimAmtReimbursed_y  DeductibleAmtPaid_y  \
4471                 16.0                   37460.0                300.0
5248                  5.0                   26940.0                 60.0
4288                  7.0                   52550.0                210.0

      ClaimDuration_y  OutpatientClaimCount
4471         1.622047                127.0
5248         2.647059                 68.0
4288         2.037383                107.0
```

```
===== Random Forest ERROR ANALYSIS =====

False Positives:
      InscClaimAmtReimbursed_x  DeductibleAmtPaid_x  ClaimDuration_x  \
1266                       0.0                  0.0         0.000000
4648                  219000.0              24564.0         6.708333
1288                       0.0                  0.0         0.000000

      InpatientClaimCount  InscClaimAmtReimbursed_y  DeductibleAmtPaid_y  \
1266                  0.0                  158670.0               1010.0
4648                 24.0                    1900.0                  0.0
1288                  0.0                   87890.0                950.0

      ClaimDuration_y  OutpatientClaimCount
1266         1.616695                 587.0
4648         0.000000                   1.0
1288         2.327206                 272.0

False Negatives:
      InscClaimAmtReimbursed_x  DeductibleAmtPaid_x  ClaimDuration_x  \
4471                   98000.0              17088.0           4.8125
4449                   12000.0               1068.0           8.0000
5248                   31000.0               5340.0           8.2000

      InpatientClaimCount  InscClaimAmtReimbursed_y  DeductibleAmtPaid_y  \
4471                 16.0                   37460.0                300.0
4449                  1.0                   56110.0                580.0
5248                  5.0                   26940.0                 60.0

      ClaimDuration_y  OutpatientClaimCount
4471         1.622047                 127.0
4449         0.942857                 175.0
5248         2.647059                  68.0
```

```
===== Gradient Boosting ERROR ANALYSIS =====

False Positives:
      InscClaimAmtReimbursed_x  DeductibleAmtPaid_x  ClaimDuration_x  \
1266                       0.0                  0.0         0.000000
4648                  219000.0              24564.0         6.708333
885                    93000.0               9612.0         6.600000

      InpatientClaimCount  InscClaimAmtReimbursed_y  DeductibleAmtPaid_y  \
1266                  0.0                  158670.0               1010.0
4648                 24.0                    1900.0                  0.0
885                  10.0                   31760.0                520.0

      ClaimDuration_y  OutpatientClaimCount
1266         1.616695                587.0
4648         0.000000                  1.0
885          0.954128                109.0

False Negatives:
      InscClaimAmtReimbursed_x  DeductibleAmtPaid_x  ClaimDuration_x  \
4471                   98000.0              17088.0           4.8125
4449                   12000.0               1068.0           8.0000
5248                   31000.0               5340.0           8.2000

      InpatientClaimCount  InscClaimAmtReimbursed_y  DeductibleAmtPaid_y  \
4471                 16.0                   37460.0                300.0
4449                  1.0                   56110.0                580.0
5248                  5.0                   26940.0                 60.0

      ClaimDuration_y  OutpatientClaimCount
4471         1.622047                127.0
4449         0.942857                175.0
5248         2.647059                 68.0
```

---

9. Final Conclusion

The Gradient Boosting model performed best overall and provided the clearest separation between legitimate and fraudulent providers.

By engineering more than 50 provider-level features and applying SMOTE, we significantly improved the model's ability to detect rare fraud cases.