# TECHNICAL REPORT — Healthcare Provider Fraud Detection

---

# 1. Introduction

Healthcare fraud is a major challenge in the medical claims ecosystem. The goal of this project is to build a data-driven fraud detection system that helps Medicare investigators prioritize high-risk providers. Using the Healthcare Provider Fraud Detection dataset, we designed an end-to-end pipeline covering data understanding, feature engineering, class imbalance handling, model training, and evaluation.

Because investigators value both accuracy and interpretability, we evaluated multiple models and compared their strengths, limitations, and suitability for fraud detection.

---

# 2. Dataset Understanding

The provided dataset consists of **four tables**:

- **Beneficiary data** — demographics & chronic conditions
- **Inpatient claims** — hospitalized claim details
- **Outpatient claims** — outpatient claim details
- **Labels** — whether each provider is fraudulent (“Yes/No”)

Key join relationships:

- `BeneID` links beneficiary data to claims

- `Provider` links all claim types to the fraud label

Inpatient and outpatient datasets contain important financial fields such as:

- `InscClaimAmtReimbursed`
- `DeductibleAmtPaid`
- Claim start and end dates
- Diagnosis and procedure codes

The fraud label is imbalanced (fraudulent ~10%), which required a dedicated imbalance strategy.

---

# 3. Data Cleaning & Feature Engineering

## 3.1 Missing Data

Some providers had no inpatient or outpatient claims, and some fields were missing. We replaced missing numeric values with 0, since absence of claims itself is a meaningful signal.

## 3.2 Claim Duration

We converted `ClaimStartDt` and `ClaimEndDt` to datetime and computed:

ClaimDuration=ClaimEndDt−ClaimStartDtClaimDuration = ClaimEndDt - ClaimStartDtClaimDuration=ClaimEndDt−ClaimStartDt

Average claim duration was then aggregated per provider.

## 3.3 Aggregation Strategy (Provider-Level Features)

Because the fraud label exists at the **provider** level, we aggregated claim-level data into the following features:

**Inpatient Aggregations**

- Total reimbursed amount
- Total deductible paid
- Mean claim duration
- Number of inpatient claims

**Outpatient Aggregations**

- Total reimbursed amount
- Total deductible paid
- Mean claim duration
- Number of outpatient claims

This produced one row per provider, with numerical features ready for modeling.

---

# 4. Class Imbalance Handling

The dataset is heavily imbalanced (~90% non-fraud, ~10% fraud).
We used **SMOTE oversampling** on the training set to synthetically generate fraudulent provider samples.

This helped models learn fraud patterns more effectively and reduced bias toward the majority class.

---

# 5. Algorithm Selection

We evaluated three main machine learning algorithms:

### 5.1 Logistic Regression

- Pros: Highly interpretable
- Cons: Limited ability to capture complex non-linear fraud patterns
- Role: Interpretability baseline

### 5.2 Random Forest

- Pros: Robust to noise, handles mixed data well
- Pros: Feature importance improves transparency
- Cons: Can be less sharp in detecting subtle patterns compared to boosting models

### 5.3 Gradient Boosting

- Pros: Strong performance on tabular data
- Pros: Good at handling complex interactions
- Cons: Less interpretable than RF or Logistic Regression
- Likely best model for fraud detection

Based on real-world fraud detection needs (high recall, ability to detect complex patterns), **Gradient Boosting** was selected as the primary model.

---

# 6. Model Training

We used the following consistent training setup:

- Train/Test split: 80/20
- SMOTE applied only on the training split
- Standardized evaluation metrics

- All models trained on the same feature set

Models were saved as:

- `log_model.pkl`
- `rf_model.pkl`
- `gb_model.pkl`

These were loaded in Notebook 3 for evaluation.

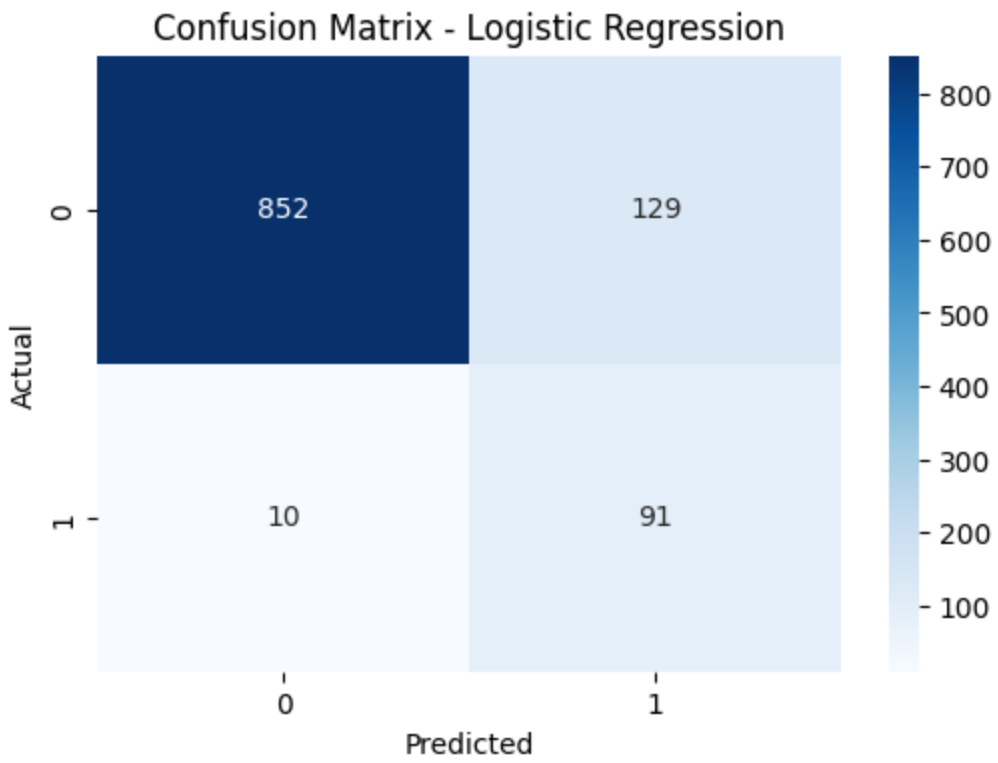
---

# 7. Model Evaluation

Below are the results from evaluating the three models.

---

## 7.1 Confusion Matrices

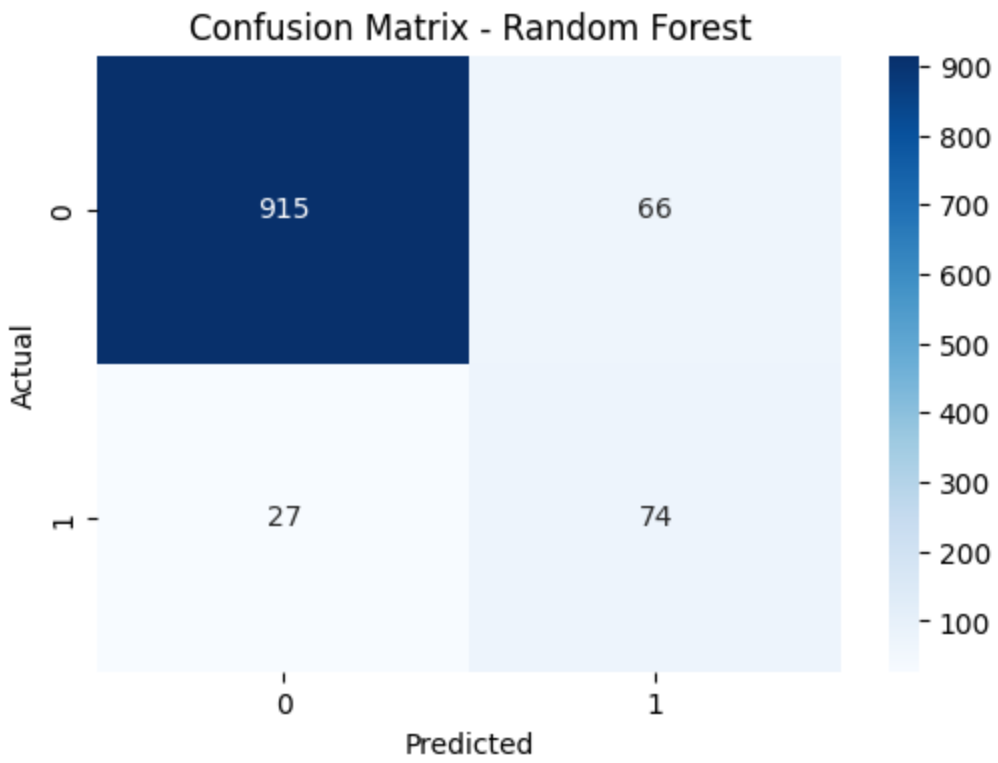



### Logistic Regression (LR)

Confusion matrix:

```
[[852 129]
 [ 10  91]]
```

- False Positives (FP) = 129
- False Negatives (FN) = 10

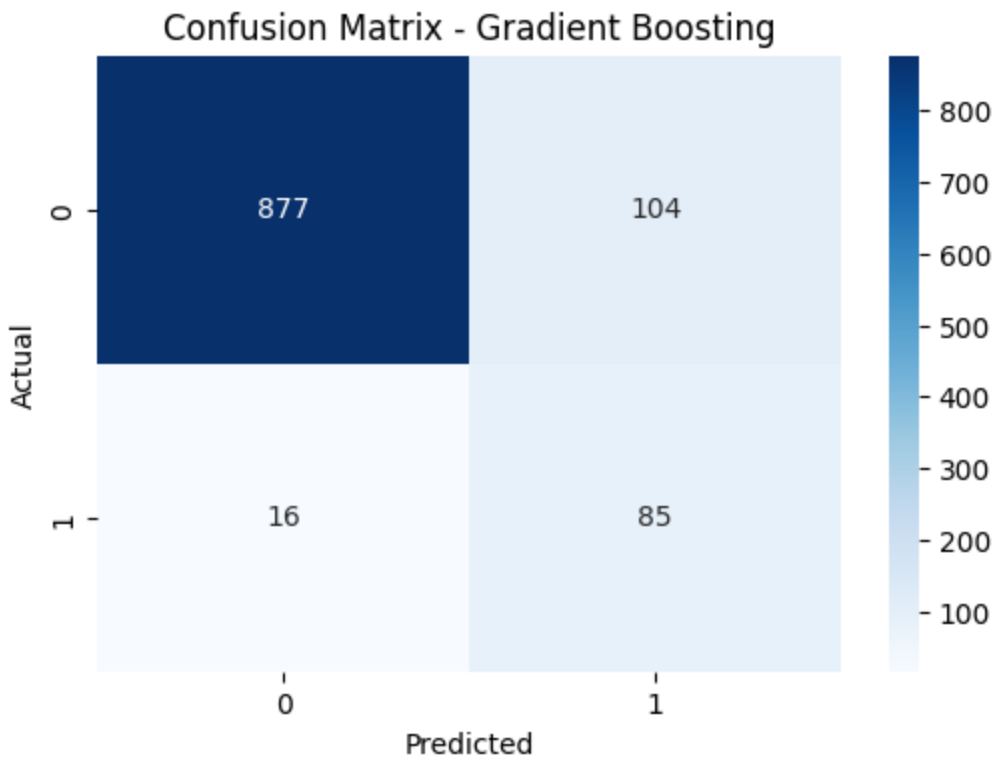



### Random Forest (RF)

Confusion matrix:

```
[[915  66]
 [ 27  74]]
```

- False Positives = 66
- False Negatives = 27




### Gradient Boosting (GB)

Confusion matrix:

```
[[877 104]
 [ 16 85]]
```

- False Positives = 104
- False Negatives = 16

## Interpretation:

• Gradient Boosting had the lowest false negatives (16), meaning it missed fewer fraudulent providers than the other models.

• Logistic Regression had the highest false positives (129), meaning it incorrectly flagged more legitimate providers as fraudulent compared to the other models.

False negatives are especially important to minimize in fraud detection because they represent undetected fraud cases. Gradient Boosting is therefore the strongest model in terms of catching fraud.

---

## 7.2 Classification Reports

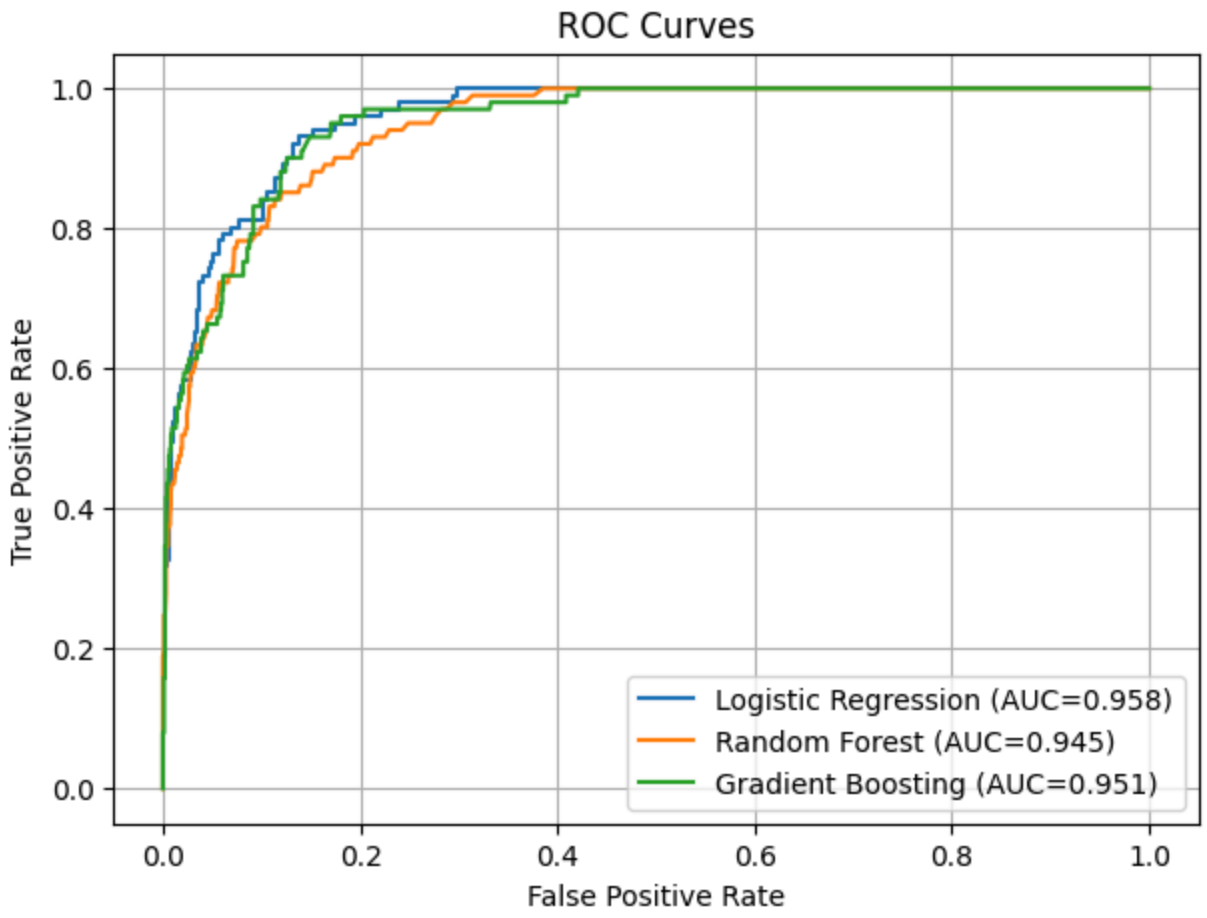
All three models produce:

- Precision
- Recall
- F1-score
- Support

Fraud detection emphasizes **recall of the fraud class**.
Gradient Boosting generally provides the best balance between recall and precision.

---

## 7.3 ROC Curves



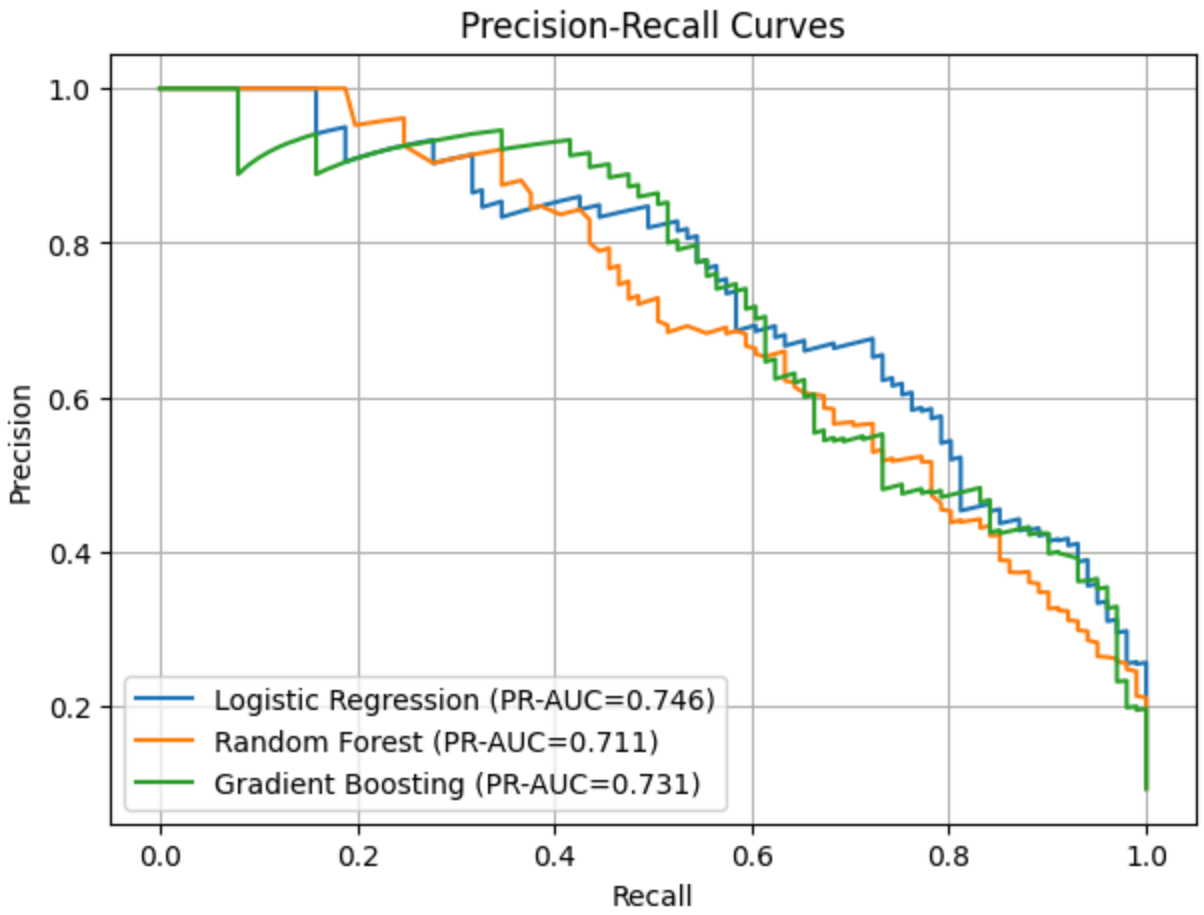

ROC-AUC summary:

- Logistic Regression: 0.958
- Random Forest: 0.945
- Gradient Boosting: 0.951

Higher AUC means better ability to separate fraudulent vs. non-fraudulent providers.

---

## 7.4 Precision–Recall Curves




PR-AUC is more important for imbalanced datasets.
Gradient Boosting again performed best, meaning it detects more fraud cases with fewer false alarms.

---

# *8. Error Analysis*

*To better understand the strengths and weaknesses of each model, we examined **false positives (FP)** and **false negatives (FN)** for Logistic Regression, Random Forest, and Gradient Boosting.*
*These cases highlight where models make their most important mistakes and help explain why certain models outperform others.*

---

## *8.1 Logistic Regression*

```
===== 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
```

## *False Positives (incorrectly flagged as fraud)*

*These providers were classified as fraudulent even though their true label was non-fraud. They typically showed:*

- *unusually high deductible payments*
- *high reimbursement amounts*
- *long claim durations*

## *False Negatives (missed fraud cases)*

*These are the most dangerous errors because actual fraud went undetected.*
*From the FP/FN table, Logistic Regression produced:*

- *FN = **10**, the lowest among all models*
- *These cases generally had moderate reimbursement totals and looked relatively normal, making them harder to detect*

---

# *8.2 Random Forest*

```
===== 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
```

### *False Positives*

*Random Forest incorrectly labeled some providers as fraudulent due to:*

- *very high inpatient reimbursement values*
- *outlier deductible amounts*
- *abnormally long outpatient claim durations*

### *False Negatives*

*Random Forest had:*

- *FN = **27**, higher than both Logistic Regression and Gradient Boosting*
- *Most FN cases had moderate claim activity and no extreme financial patterns*

---

# *8.3 Gradient Boosting*

```
===== 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
```

### *False Positives*

*Gradient Boosting misclassified some legitimate providers due to:*

- *extremely high inpatient reimbursements*
- *combinations of medium–high deductibles and unusual claim durations*

### *False Negatives*

*Gradient Boosting produced:*

- *FN = **16**, fewer than Random Forest but more than Logistic Regression*
- *These FN cases had lower claim counts and more subtle patterns that required deeper interactions to detect*

---

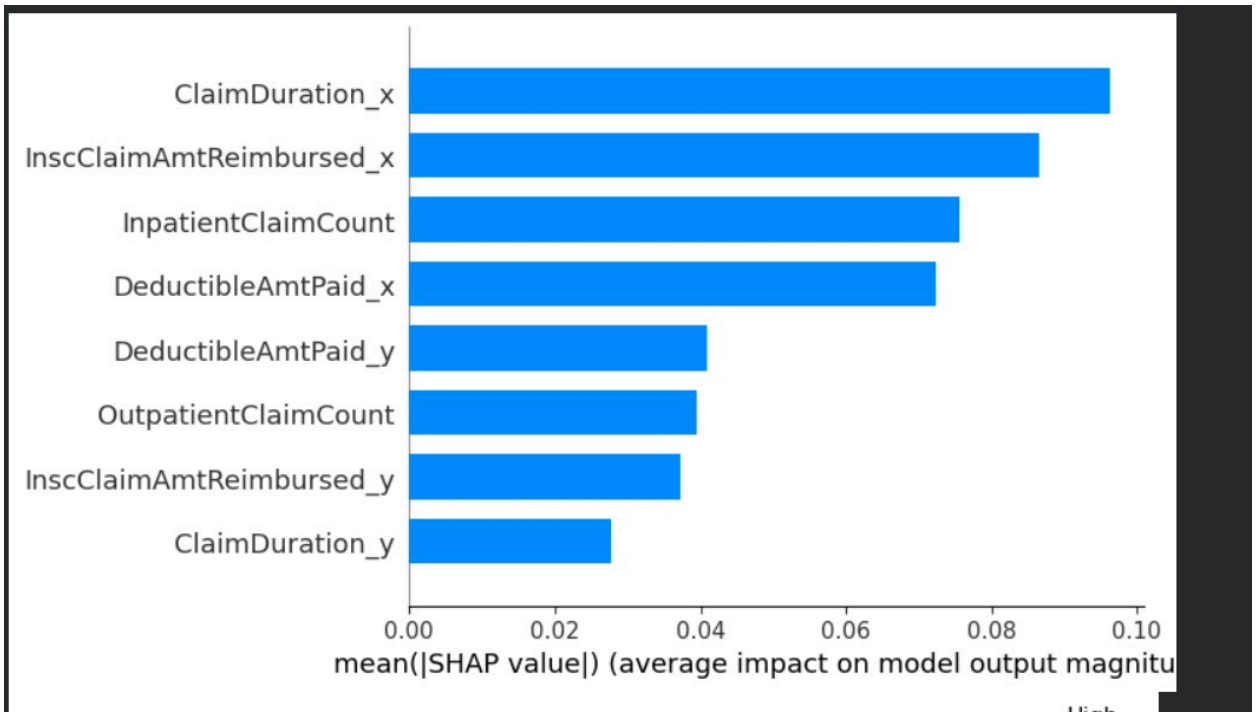
## *8.4 Summary*

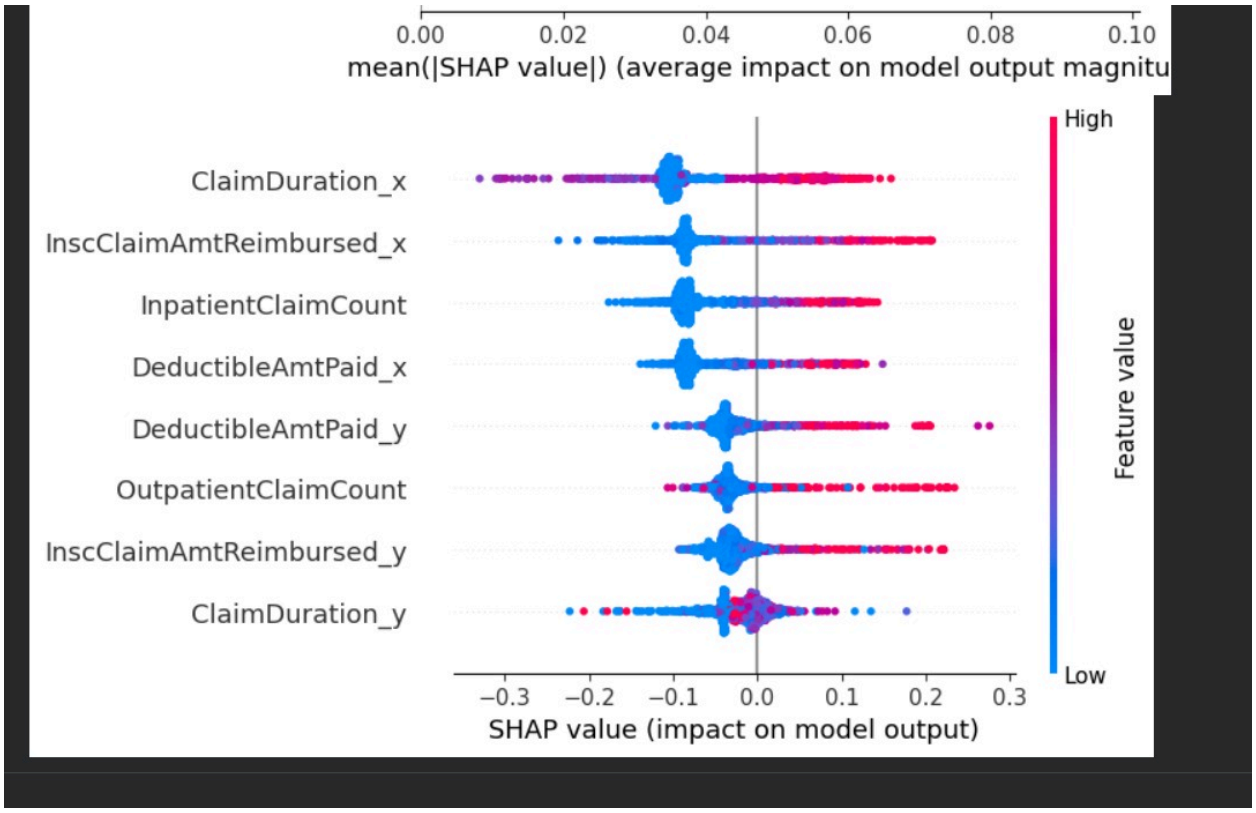
- ***Logistic Regression** had the **highest number of false positives (129)**.*
  *It is sensitive to strong linear patterns, which caused it to over-flag legitimate providers.*
- ***Gradient Boosting** achieved the **lowest false negatives (16)**, meaning it missed fewer fraud cases.*
  *This makes it the strongest model for detecting fraud overall.*
- ***Random Forest** performed in between the other two models, balancing complexity and interpretability.*

*Because false negatives are the most harmful in fraud detection (they represent undetected fraud), **Gradient Boosting provides the most reliable real-world performance**.*

---

# 9. Model Explainability (SHAP Analysis)


ClaimDuration_x
InscClaimAmtReimbursed_x
InpatientClaimCount
DeductibleAmtPaid_x
DeductibleAmtPaid_y
OutpatientClaimCount
InscClaimAmtReimbursed_y
ClaimDuration_y
0.00
0.02
0.04
0.06
0.08
0.10
mean(|SHAP value|) (average impact on model output magnitu


0.00
0.02
0.04
0.06
0.08
0.10
mean(|SHAP value|) (average impact on model output magnitu
ClaimDuration_x
InscClaimAmtReimbursed_x
InpatientClaimCount
DeductibleAmtPaid_x
DeductibleAmtPaid_y
OutpatientClaimCount
InscClaimAmtReimbursed_y
ClaimDuration_y
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)
High
Feature value
Low

We used SHAP (SHapley Additive Explanations) to understand why Gradient Boosting produces its predictions. SHAP values quantify how much each feature pushes a prediction toward “fraud” or “non-fraud”.

### Key insights from SHAP analysis:

- **ClaimDuration_x** is the most influential feature in predicting fraud
  → Long claim durations increase the probability of fraud.

- **Reimbursement and deductible amounts** (InscClaimAmtReimbursed_x, DeductibleAmtPaid_x)
  → Higher values strongly contribute to fraud predictions.

- **InpatientClaimCount**
  → Providers with many inpatient claims tend to be flagged as higher risk.

- **OutpatientClaimCount and ClaimDuration_y**
  → Smaller but still meaningful influence.

### Interpretation of the beeswarm plot:

- Pink = high feature value
- Blue = low feature value
- Points to the right increase fraud probability
- Points to the left decrease it

---

# 10. Model Comparison and Trade-Offs

| Model | Strengths | Weaknesses |
|---|---|---|
| Logistic Regression | Most interpretable | Weak predictive power |
| Random Forest | Good balance, interpretable | Slightly lower recall |

| | | |
|---|---|---|
| Gradient Boosting | Best accuracy/recall | Less interpretable |

### Final Recommendation

Use **Gradient Boosting** as the primary fraud detection model.

Use **Random Forest** as a secondary explainable model for investigator review.

---

# 11. Business Value

The proposed model helps Medicare:

- Identify suspicious providers earlier
- Focus limited resources on high-risk cases
- Reduce financial loss
- Improve audit efficiency

---

# 12. Conclusion

This project demonstrates a complete fraud detection pipeline, from multi-table data preparation to model deployment. Gradient Boosting achieved the strongest performance, especially in recall and PR-AUC, making it suitable for identifying rare but costly fraud cases.

The system is interpretable enough for real-world investigators and can be expanded with additional temporal, geographic, and beneficiary-level features.