Report

Data Structure and Granularity Explanation

The dataset consists of four primary tables, each representing a different level of granularity within the Medicare claims ecosystem.

1. Beneficiary Table (Train_Beneficiarydata.csv)

- Granularity: One row per patient (BeneID)
- Contains:
    - Demographics
    - Date of birth, date of death
    - Chronic conditions
- Key: BeneID
- Role: Provides patient-level health indicators that help detect patterns such as:
    - Billing dead patients
    - Unusual claims for extremely old patients

---

2. Inpatient Claims Table (Train_Inpatientdata.csv)

- Granularity: One row per inpatient claim
- Contains:
    - ClaimStartDt, ClaimEndDt
    - Admission/discharge dates
    - Reimbursement amounts
    - Diagnoses and procedures
- Key: (ClaimID, Provider, BeneID)
- Role: Captures expensive, long-duration services — often manipulated by fraudsters.

---

3. Outpatient Claims Table (Train_Outpatientdata.csv)

- Granularity: One row per outpatient claim
- Contains:
    - Visit dates
    - Reimbursement amounts
    - Procedures
- Key: (ClaimID, Provider, BeneID)
- Role: Captures high-volume, lower-cost procedures — also prone to fraud (e.g., unnecessary tests).

---

4. Provider Fraud Labels (Train_labels.csv)

- Granularity: One row per provider
- Contains:
    - Provider ID
    - PotentialFraud label (Yes/No)
- Role: Supervised learning target
- Key: Provider

Why the Provider is the Modeling Unit?

Fraud investigations are conducted at the provider level, not the patient level.

- Providers generate many claims → claim-level modeling would create leakage.
- Fraud labels are provider-level → must aggregate upwards.
- Patterns like:
    - unusually high billing
    - abnormal patient volume
    - extreme procedure diversity
      only show up AFTER aggregation.

Thus, all claim-level features must be aggregated to provider level.

Aggregation Strategy Justification

Since the fraud label exists at the *provider* level, all beneficiary and claim-level data must be aggregated to create a single feature vector per provider.

1. Why group by provider?

- Fraud is committed by providers (hospitals, clinics), not individual claims.
- Medicare flags suspicious providers, not suspicious claims.
- Aggregation ensures no data leakage from multiple claim rows per provider.

---

2. Why these statistical measures (count, mean, sum, max)?

Feature Type	Reason (Business Justification)
Counts (e.g., number of claims)	Detect overbilling, upcoding, unnecessary procedures
Means (avg payment, avg LOS)	Detect patterns compared to normal baseline
Sums (total reimbursement)	High total payouts indicate high fraud risk
Max values	Outliers are strong fraud indicators
Percentages (e.g., pct deceased)	Detect illegal billing of dead patients
Ratios	Capture abnormal provider behavior

These features are widely used in real-world healthcare fraud detection systems.

---

3. Why inpatient and outpatient must be aggregated separately

- Inpatient claims: fewer but more expensive → detect upcoding, length-of-stay fraud
- Outpatient claims: many small claims → detect test inflation, unbundling, anomalous volume

Separating them preserves meaningful fraud signals.

4. Why include beneficiary-level features

Certain fraud categories come from patient characteristics:

- Billing deceased beneficiaries
- Providers with unusual gender distributions
- Providers treating extremely old patients with many chronic diseases

These high-level behaviors are captured only when beneficiary data is merged into the claims.

Feature Engineering and Preprocessing

Categorical Encoding:

- The `common_state` (proxy for provider location) was encoded using Frequency Encoding (replacing state IDs with their count).
- We did this to to avoid one hot encoding and giving each region multiple rows

Missing Value Handling**:**

- Financial and claim related null values were handled via zero imputation.
- In claims data a missing amount often implies that nothing got paid meaning zero would be the best fit

Feature Scaling:

- All final numerical features were normalized using a StandardScaler.
- Ensures features with large ranges (like total reimbursement) do not strongly influence the model

---

Data Partitioning and Imbalance Mitigation

- Data Split:
  - The providers were split using a stratified approach into: Training (60%), Validation (20%), and Test (20%) sets.
  - Stratification ensures that all sets have the same fraud percentage as we don’t want to train a model on a training set with no fraud or test one with a test set that has no fraud

- SMOTE Application:
  - The class imbalance was solved using SMOTE (Synthetic Minority Over-sampling Technique), applied only to the Training Set.

- This Balances the fraud class for effective training as it creates more “fraud” cases based on the ones already present as to not have the model just guess not fraud everytime

- Evaluation Metrics:
  - The model was optimized for PR-AUC (Precision-Recall Area Under Curve), ROC-AUC (Receiver Operating Characteristic - Area Under the Curve), Recall and F1 score, as these metrics are more reliable than Accuracy in imbalanced fraud detection problems.
  - We used these metrics to make sure that if the model just guessed not fraud everytime the we would be able to tell as accuracy here wouldve done the opposite

---

Model Selection and Parameter Refinement

- Candidate Algorithms:
  - Three models were tested:
    - Logistic Regression,
    - Random Forest
    - Gradient Boosting.

- Out of those three we chose Random Forest to tune as it had the highest F1 score with a good PR-AUC and percision while Logistic Regression had a very low percision and gradient boosting had a low F1 score and percision

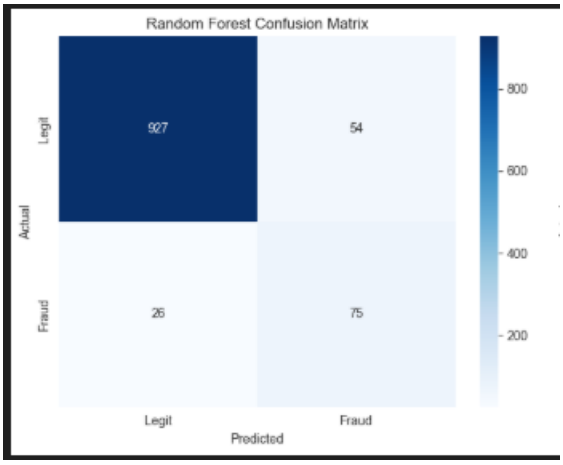
```
...
    --- Model Performance Comparison (Test Set) ---
```

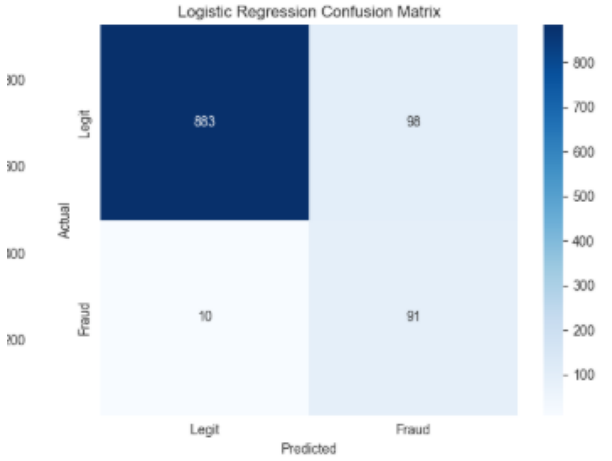
	Model	Precision (Fraud)	Recall (Fraud)	F1-Score (Fraud)	ROC-AUC	PR-AUC
0	Random Forest	0.581395	0.742574	0.652174	0.941987	0.694782
1	Logistic Regression	0.481481	0.900990	0.627586	0.956409	0.763430
2	Gradient Boosting	0.509317	0.811881	0.625954	0.947159	0.717192
3	Best Model (from tuning)	0.555556	0.742574	0.635593	0.943132	0.710088

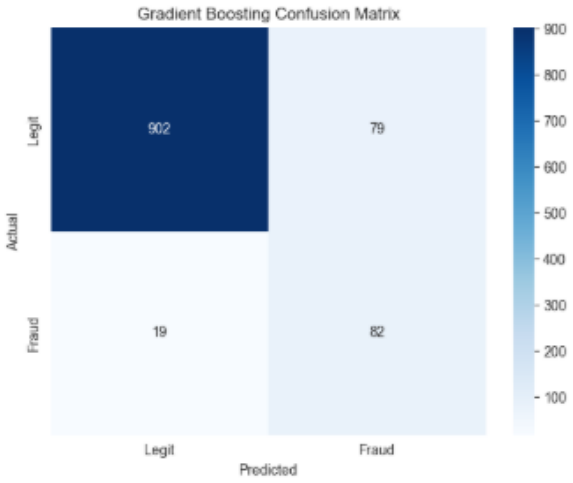
- Tuning (Random Forest):
  - The final, optimal parameters were:
    - `n_estimators=100`
    - `max_depth=None`
    - `min_samples_split=2`.

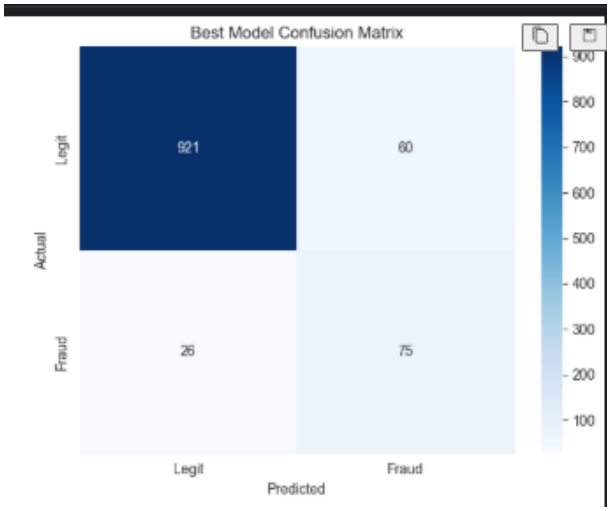
EVALUATION AND ERROR ANALYSIS

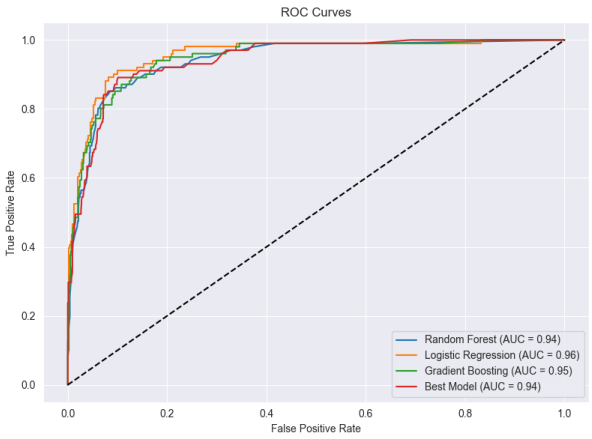
- Found 54 False Positives and 26 False Negatives (Random Forest).
- Total Value of Fraud Missed by Model (False Negatives): $2,474,200.00

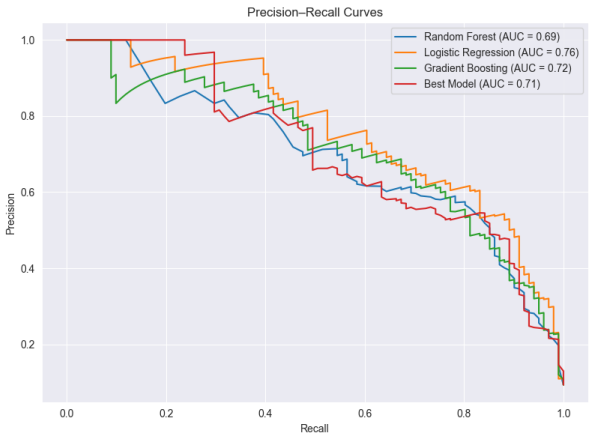












Metric	FN Case 0 (Fraud Missed)	FN Case 1 (Fraud Missed)	FN Case 2 (Fraud Missed)
Num Unique Patients	12	2	33
Num Inpatient Claims	14	6	42
Total Inpatient Payment	$137K	$38K	$413K
Avg Length of Stay	6.4 days	9.5 days	5.1 days
Num Outpatient Claims	0	14	68
State Frequency (State Risk)	182 (High)	12 (Low)	52 (Medium)

Low Volume:

- FN Cases 0 and 1 have very low unique patient counts (12 and 2). The model likely weighted the low volume too heavily, causing it to dismiss the risk. In particular, Case 1 has only 6 claims, which is insufficient to trigger the fraud threshold based on counts alone.

Specific Scenario Risk:

- FN Case 0 shows zero outpatient claims (`num_outpatient_claims=0`). This suggests the provider may be focusing on inpatient-only fraud, making their profile appear less "busy" or complex than high-volume fraudsters who manipulate both claim types.

Metric	FP Case 0 (False Alarm)	FP Case 1 (False Alarm)	FP Case 2 (False Alarm)
Num Unique Patients	45	73	13
Num Inpatient Claims	51	88	14
Total Inpatient Payment	$597K	$919K	$119K
Total Outpatient Payment	$55K	$197K	$38K
State Frequency (State Risk)	62	99 (High)	30

Legitimate High Volume:

- FP Case 1 stands out with the highest number of unique patients (73) and claims (88 inpatient, 704 outpatient), resulting in the highest total payment ($919K). The model is trained to flag high volume as risk, so a legitimately large medical center is triggering the alarm.

High-Risk Location:

- FP Case 1 also operates in a state with a high fraud frequency (`state_freq=99`). The model likely interprets the combination of high volume and high-risk location as fraud, even though the provider's per-claim metrics may be normal.