DataOrbit — Healthcare Provider Fraud Detection Project

Members:

Youstina Raouf - 13001755

Chantal Sherif - 13007034

Ali Wahsh - 13006943

Omar Al-Dahabi 13007802

---

1. Project Overview

Healthcare fraud costs the U.S. system over $68 billion annually. The Centers for Medicare & Medicaid Services (CMS) can investigate only a fraction of suspicious cases. This project aims to detect high-risk healthcare providers using data-driven methods, balancing interpretability, performance, and business impact.

Objectives:

1. Detect fraudulent providers from multi-table claims data.
2. Handle class imbalance (~10% of providers are fraudulent).
3. Provide explainable predictions for investigators.
4. Demonstrate business value by prioritizing high-risk providers.

---

2. Data Understanding & Feature Engineering (Notebook 1)

Datasets Used:

- `Train_Beneficiarydata.csv` — Patient demographics, coverage, chronic conditions.
- `Train_Inpatientdata.csv` — Hospital admission claims with financial, procedural, physician info.
- `Train_Outpatientdata.csv` — Outpatient claim data.
- `Train_labels.csv` — Provider-level fraud labels (`Yes/No`).

Key Identifiers:

- `BeneID` links patients to claims.
- `Provider` links claims to fraud labels.

2.1 Data Exploration

- **Checked data quality:** missing values, duplicates, inconsistent data.
- Explored distributions of claims, claim amounts, provider sizes.
- Compared **fraudulent vs legitimate providers**:
    - Fraudulent providers often had higher claim volumes and unusual billing patterns.
    - Some upcoding and unbundling patterns were detectable.

2.2 Data Cleaning and Merging Strategy

To prepare the Medicare claims datasets for analysis and modeling, we performed a structured series of cleaning and merging operations. Since the fraud labels are assigned at the **provider level**, but the raw data is distributed across **beneficiary**, **inpatient**, and **outpatient** claim tables, the datasets must first be standardized and integrated into a unified structure.

```
Combined claims shape: (117849, 31)
After merging beneficiaries: (117849, 55)
Final merged dataset shape: (117849, 56)
```

	BeneID	ClaimID	ClaimStartDt	ClaimEndDt	Provider	InscClaimAmtReimbursed	AttendingPhysician	OperatingPhysician
0	BENE11001	CLM46614	2009-04-12	2009-04-18	PRV55912	26000	PHY390922	NaN
1	BENE11001	CLM66048	2009-08-31	2009-09-02	PRV55907	5000	PHY318495	PHY318495
2	BENE11001	CLM68358	2009-09-17	2009-09-20	PRV56046	5000	PHY372395	NaN
3	BENE11011	CLM38412	2009-02-14	2009-02-22	PRV52405	5000	PHY369659	PHY392961
4	BENE11014	CLM63689	2009-08-13	2009-08-30	PRV56614	10000	PHY379376	PHY398258

5 rows × 56 columns

2.3 Missing Data Handling

Healthcare claims data often contains missing values due to incomplete reporting, optional fields, or inconsistencies across providers and claim types. Properly handling missing data is essential to ensure modeling stability, prevent bias, and preserve the interpretability of downstream features.
The following strategy was applied based on the semantic meaning of each feature category.

2.4 Provider-Level Feature Aggregation

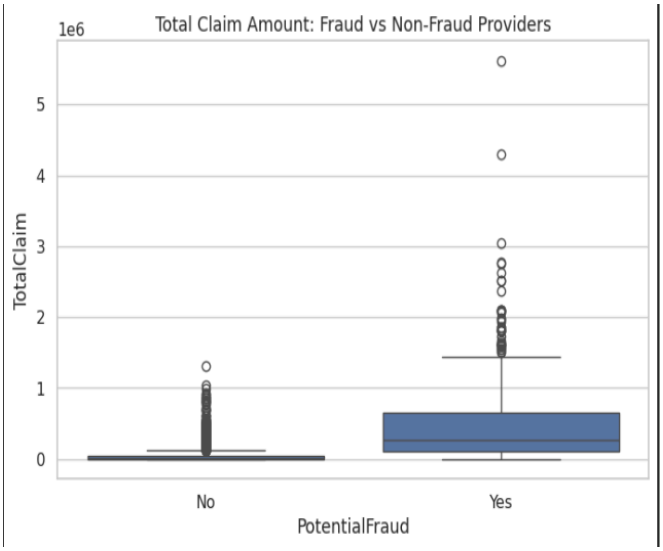
A critical step in transforming the raw Medicare claims data into an effective modeling dataset involved aggregating claim-level information into **provider-level features**, since fraud labels are assigned at the provider level. This aggregation captures behavioral patterns, billing tendencies, and service characteristics that can distinguish fraudulent providers from legitimate ones.

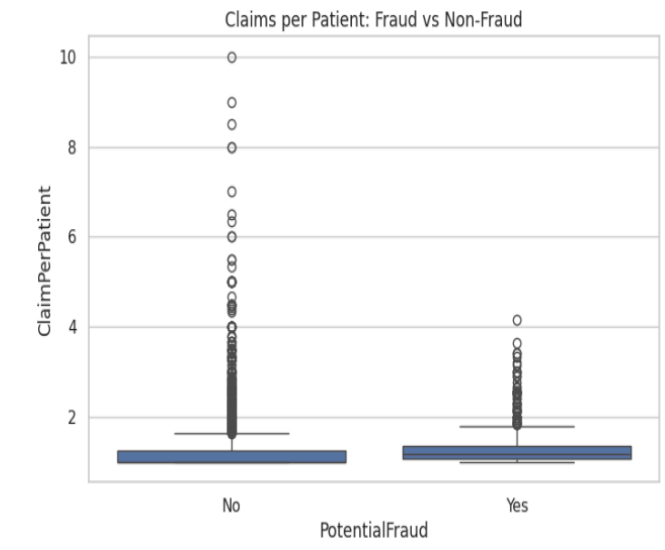
Provider	TotalClaim	AvgClaim	MaxClaim	StdClaim	NumClaims	UniquePatients	InpatientClaims	OutpatientClaims
PRV51001	97020.0	16170.000000	42000.0	18221.893425	6	6	5	1
PRV51003	578940.0	8513.823529	57000.0	8463.176255	68	59	62	6
PRV51004	960.0	320.000000	500.0	230.651252	3	3	0	3
PRV51005	11040.0	190.344828	2600.0	452.478421	58	17	0	58
PRV51007	19440.0	2777.142857	10000.0	3895.988169	7	5	3	4

Resulting dataset: `provider_features.csv` (used for modeling).

2.5 Fraud vs Non-Fraud Providers

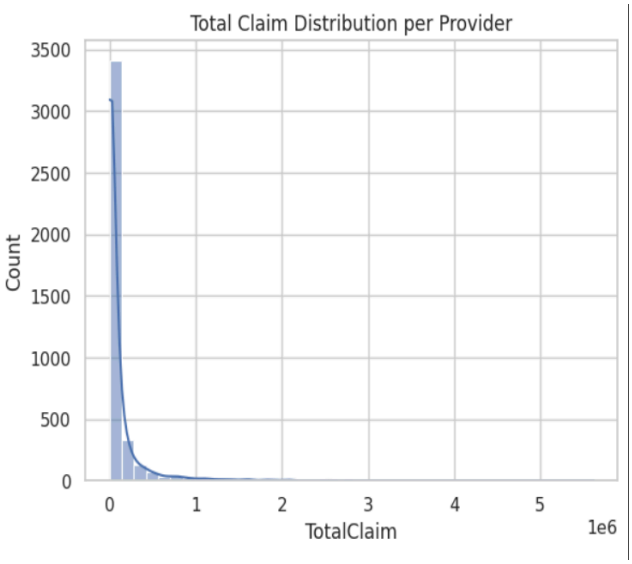
Boxplots were generated to compare `TotalClaim` and `ClaimPerPatient` between fraud and non-fraud providers. This allows visual assessment of how claim amounts differ between the two groups and helps identify features potentially indicative of fraud.

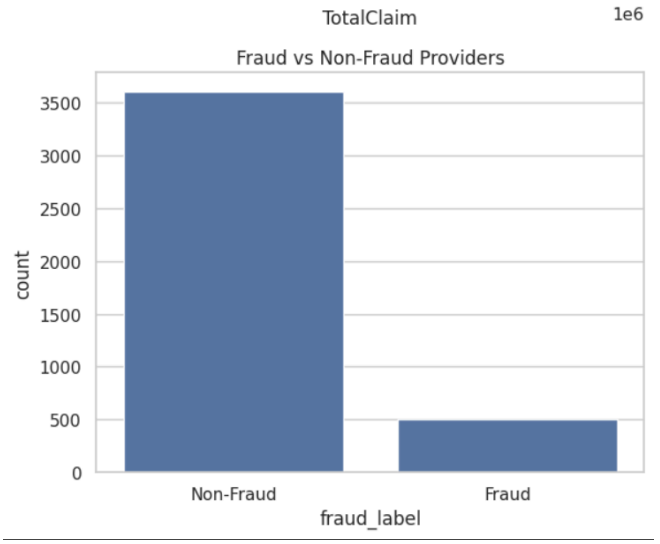




2.6 Distribution of Total Claim Amounts

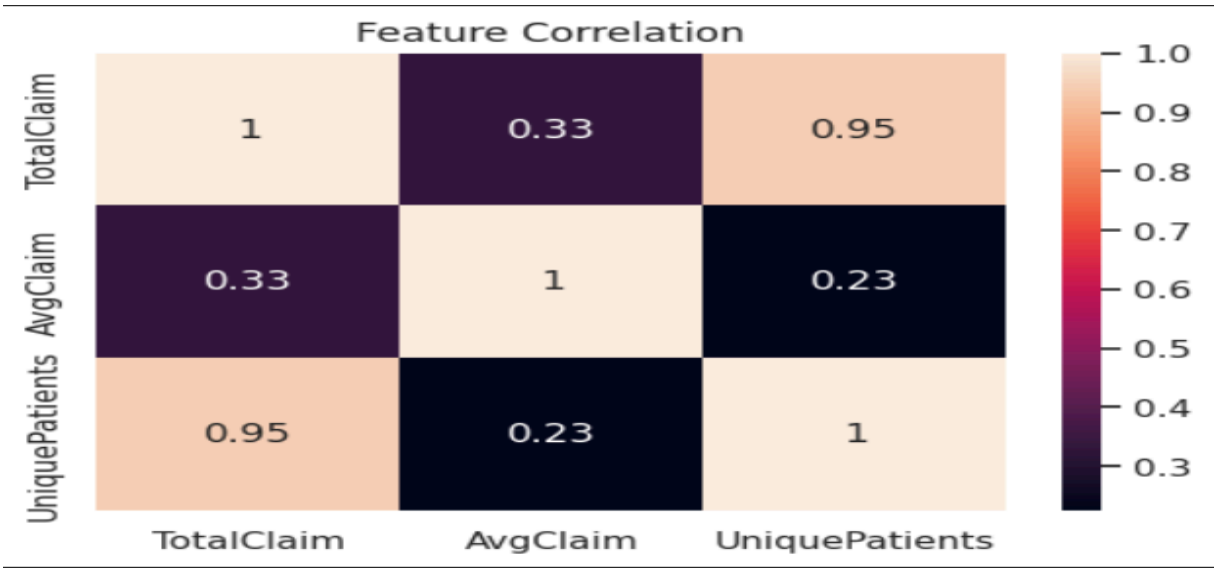
To understand the financial behavior of providers, we examined the distribution of the total claim amounts submitted by each provider. The following plot was generated using a histogram with kernel density estimation (KDE):





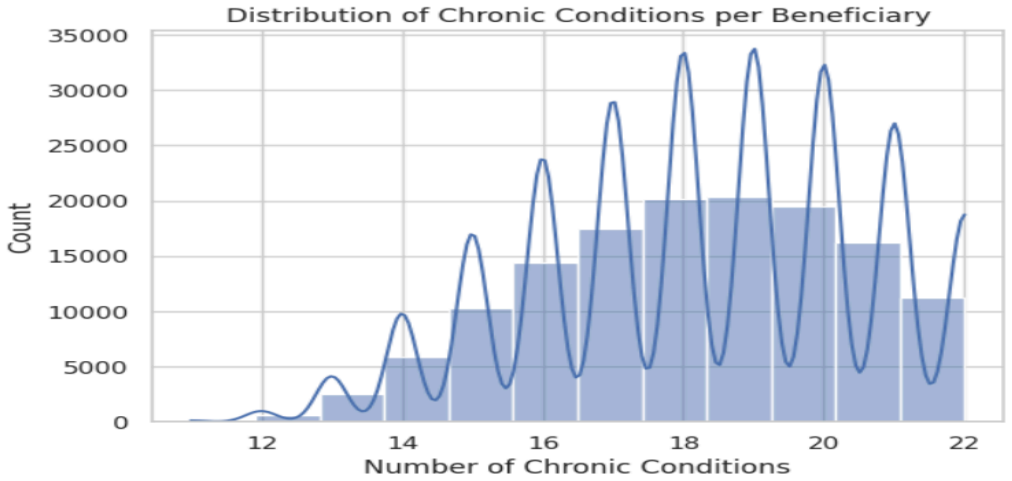
2.7 Feature Correlation

A correlation heatmap was generated for `TotalClaim` , `AvgClaim`, and `UniquePatients` to examine their relationships. This helps identify how strongly these numeric features are related, which can guide feature selection for modeling.



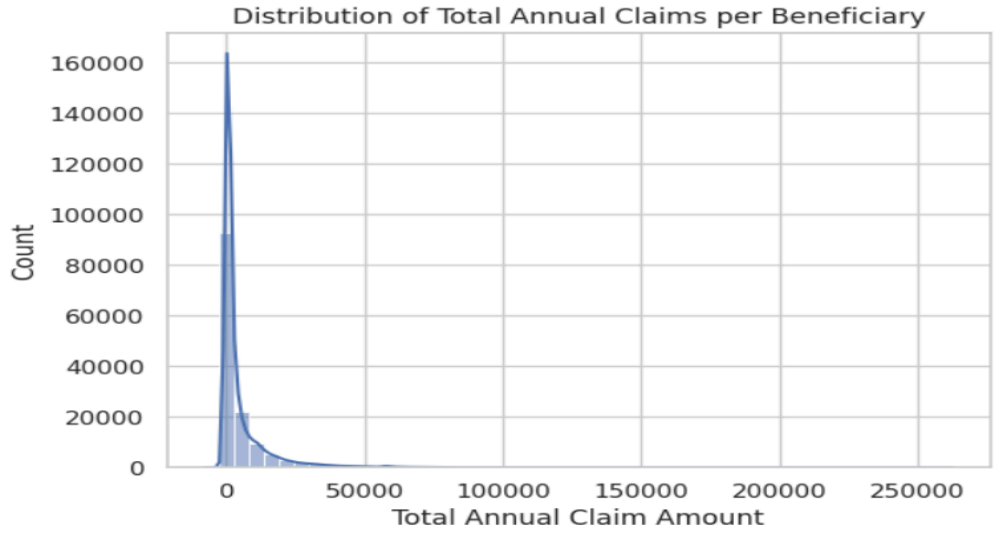
2.8 Chronic Conditions per Beneficiary

The number of chronic conditions for each beneficiary was calculated by summing across 11 chronic condition indicators. A histogram was generated to visualize the distribution of `NumChronic` across all beneficiaries. This analysis provides insight into the overall health burden of the population and can inform risk assessment and modeling strategies.



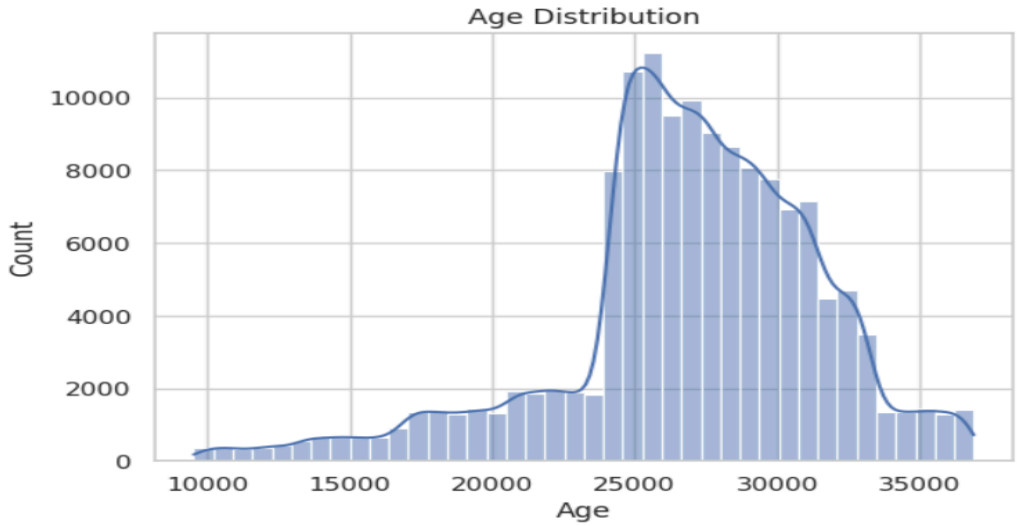
2.9 Total Annual Claims per Beneficiary

The total annual claims for each beneficiary were calculated by summing inpatient (`IPAnnualReimbursementAmt`) and outpatient (`OPAnnualReimbursementAmt`) reimbursement amounts. A histogram of `TotalAnnualClaims` was generated to visualize the distribution across all beneficiaries, providing insight into claim patterns and overall healthcare utilization.



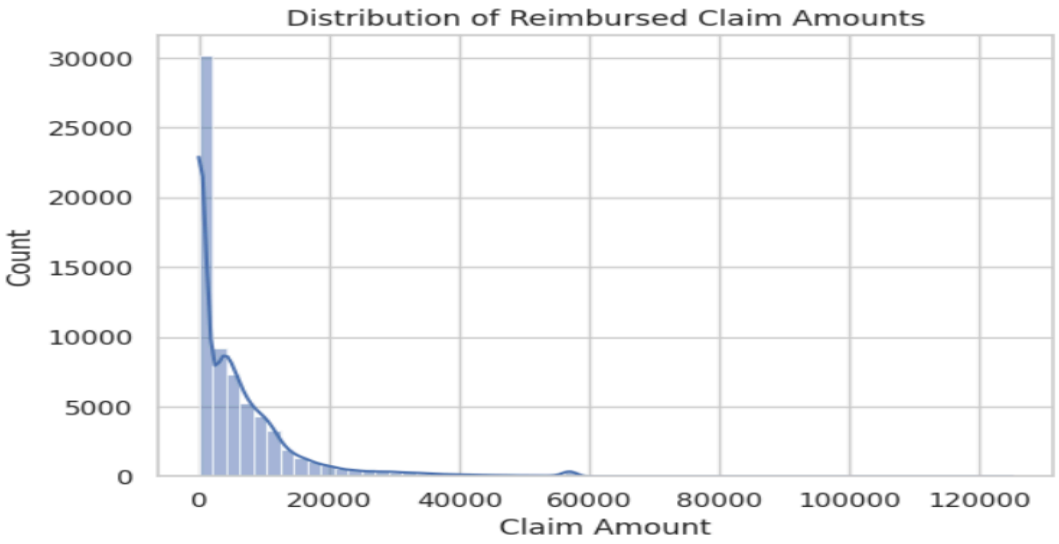
2.10 Age Distribution of Beneficiaries

The age of each beneficiary was calculated based on their date of birth (`DOB`) relative to December 31, 2009. A histogram was generated to visualize the age distribution, providing an overview of the population demographics and highlighting the spread of ages among beneficiaries.



2.11 Distribution of Reimbursed Claim Amounts

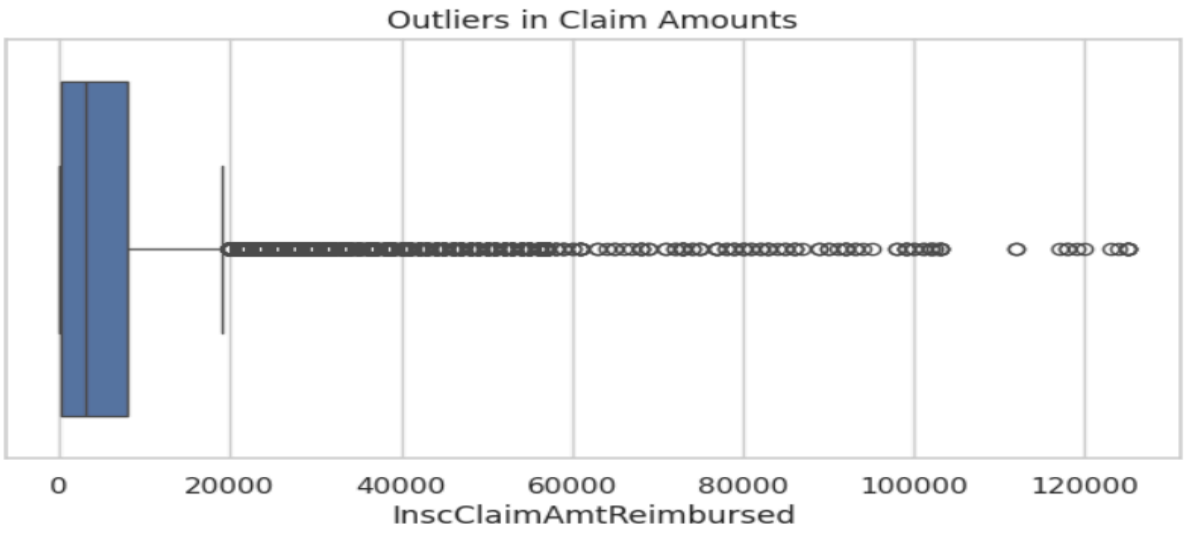
A histogram was generated to visualize the distribution of reimbursed claim amounts (`InscClaimAmtReimbursed`) across all claims. This provides insight into the typical claim sizes, highlights the range of claim values, and can help identify patterns or outliers in reimbursement data.



2.12 Outliers in Claim Amounts

A boxplot was created for the reimbursed claim amounts (`InscClaimAmtReimbursed`) to identify potential outliers. This visualization

highlights unusually high or low claim values, which may indicate data anomalies or cases requiring further investigation for fraud detection or data quality assessment.

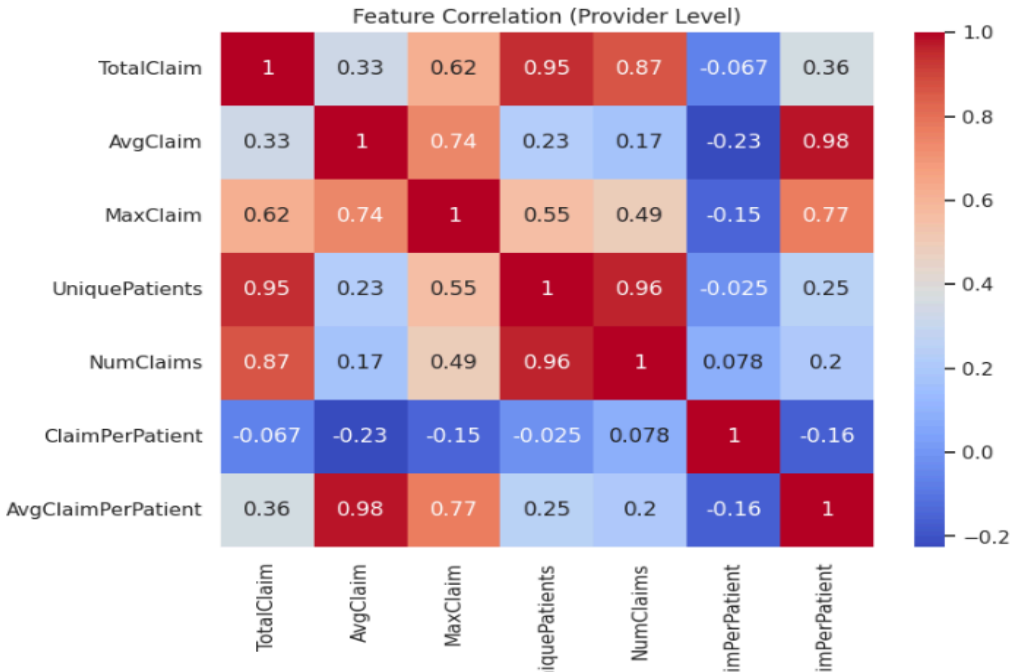


2.13 Encoding Fraud Labels

The `PotentialFraud` column was encoded into a numeric `fraud_label` for modeling purposes, with `Yes` mapped to `1` and `No` mapped to `0`. Missing values were filled with `0`, assuming non-fraud in the absence of a label. The resulting label distribution provides an overview of the balance between fraudulent and non-fraudulent providers in the dataset.

2.14 Provider-Level Feature Correlation

A correlation heatmap was generated for key numeric provider-level features, including `TotalClaim`, `AvgClaim`, `MaxClaim`, `UniquePatients`, `NumClaims`, `ClaimPerPatient`, and `AvgClaimPerPatient`. This visualization highlights the strength and direction of linear relationships between features, which can inform feature selection and modeling strategies for fraud detection.



---

3. Modeling (Notebook 2)

3.1 Train-Test Split

- Stratified split: 80% train, 20% test, preserving class imbalance.

3.2 Preprocessing

- ColumnTransformer:
    - Numeric features → StandardScaler
    - Categorical features → OneHotEncoder

3.3 Models Trained

- Logistic Regression (baseline)
- Random Forest (tuned hyperparameters)
- Gradient Boosting (Sklearn GB)
- XGBoost (tree-based, high interpretability)

3.4 Metrics

- Metrics computed: **Precision, Recall, F1-score, ROC-AUC, PR-AUC**
- Class imbalance considered (PR-AUC prioritized).

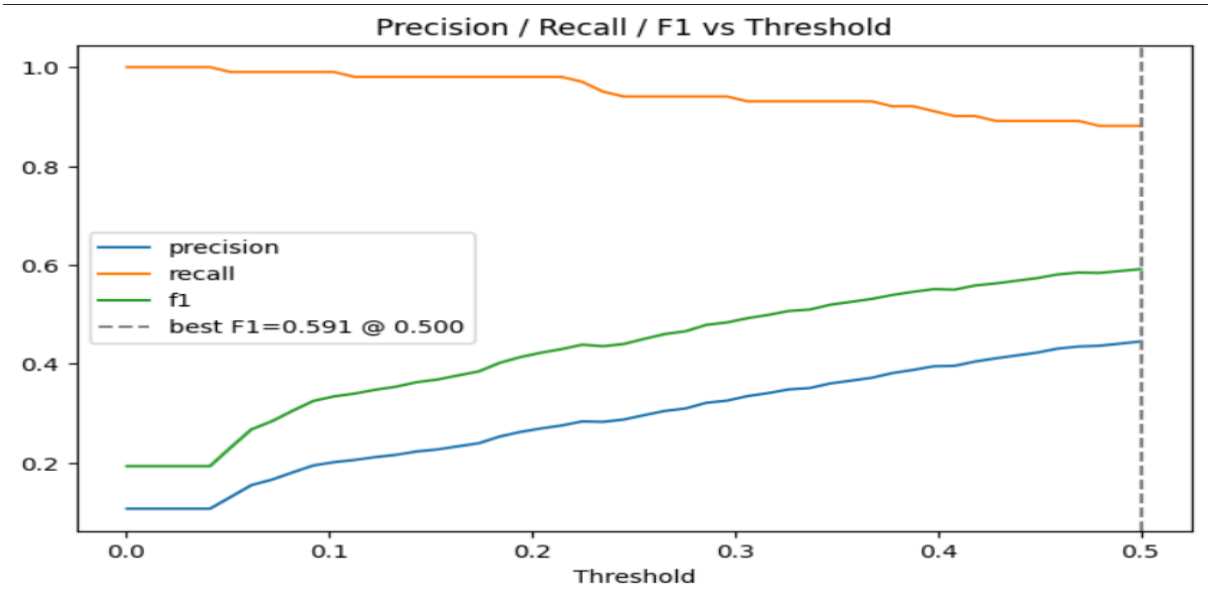
3.5 Model Comparison Table

	Model	Precision	Recall	F1-score	ROC-AUC	PR-AUC	Notes
0	Logistic Regression	0.4450	0.8812	0.5914	0.9345	0.7077	Selected model (PR-AUC)
1	XGBoost	0.4607	0.8119	0.5878	0.9322	0.6983	Comparison / interpretability
2	Random Forest	0.6615	0.4257	0.5181	0.9240	0.6670	Tuned RF
3	Gradient Boosting	0.6944	0.4950	0.5780	0.9309	0.6555	Sklearn GB

Observation: Logistic Regression achieves highest PR-AUC, making it the primary model for evaluation.

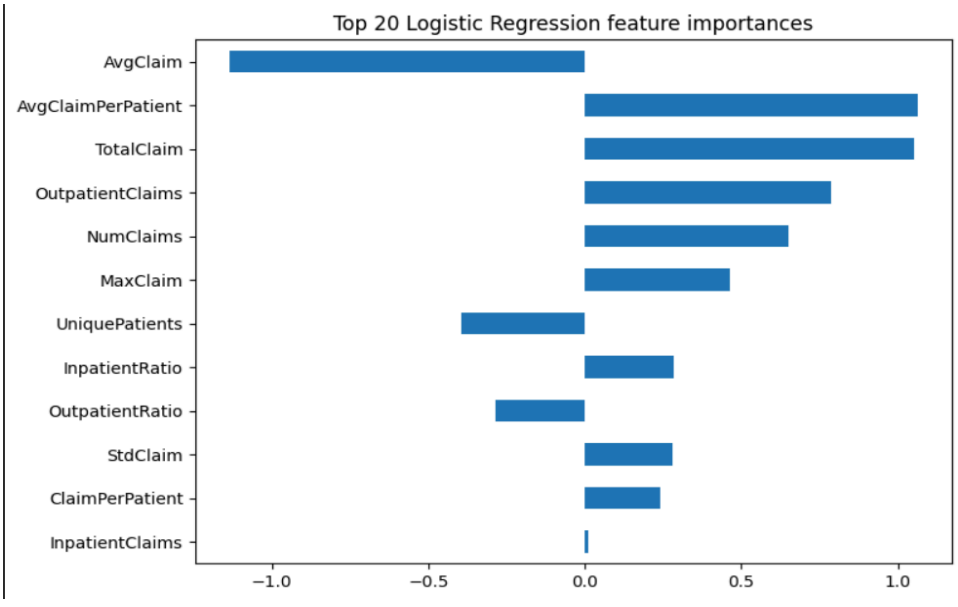
3.6 Threshold Analysis

In fraud detection, using the default threshold of 0.5 is rarely optimal—especially for imbalanced datasets where the positive class (fraudulent providers) represents a small minority. To determine the most effective decision boundary for classifying providers as fraudulent, we conducted a threshold sweep analysis by computing precision, recall, and F1-score across all thresholds between 0 and 1.



3.7 Logistic Regression Feature Importance Analysis

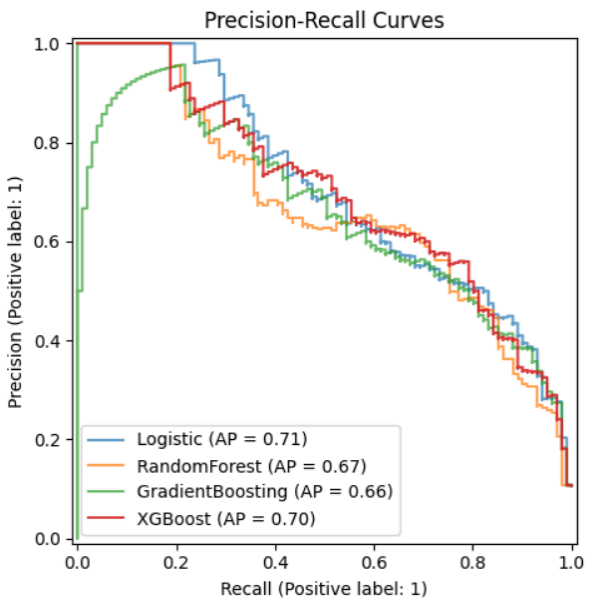
To ensure interpretability—an essential requirement when working with regulatory agencies such as CMS—we analyzed the feature importance of the Logistic Regression model. Logistic Regression offers a direct way to inspect model behavior through its learned coefficients, allowing investigators to understand which provider-level characteristics most strongly contribute to fraud predictions.

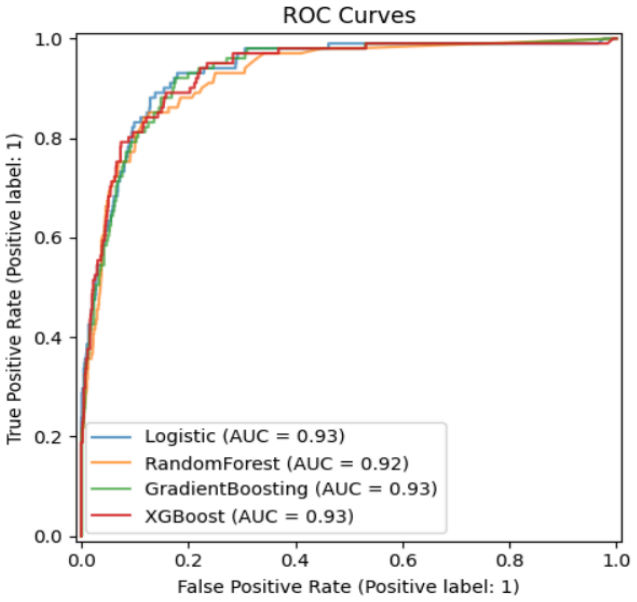


AvgClaim	-1.135438
AvgClaimPerPatient	1.062116
TotalClaim	1.053049
OutpatientClaims	0.788205
NumClaims	0.648829
MaxClaim	0.461223
UniquePatients	-0.396261
InpatientRatio	0.284030
OutpatientRatio	-0.283993
StdClaim	0.280734
ClaimPerPatient	0.242658
InpatientClaims	0.008920

3.8 Model Comparison Using Precision–Recall and ROC Curves

To thoroughly compare the performance of all trained models—Logistic Regression, Random Forest, Gradient Boosting, and XGBoost—we plotted both **Precision–Recall (PR) curves** and **Receiver Operating Characteristic (ROC) curves** on the same figure. This side-by-side visualization provides a comprehensive assessment of each model's ability to distinguish fraudulent from legitimate providers across all thresholds.





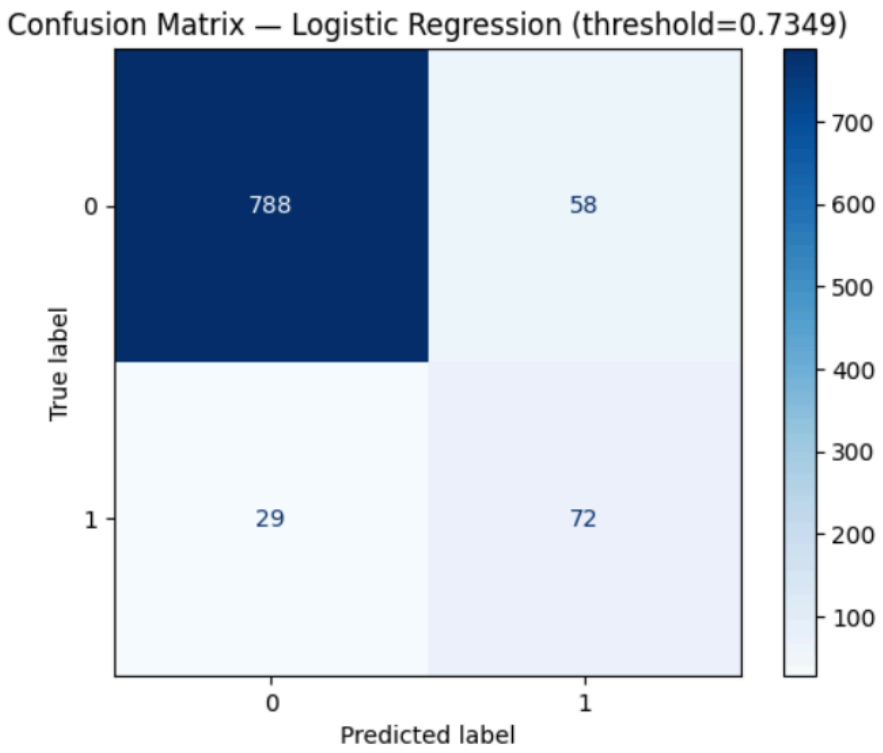
4. Evaluation (Notebook 3)

4.1 Logistic Regression (Primary Model)

Metrics:

- Precision: 0.445
- Recall: 0.881
- F1-score: 0.591
- ROC-AUC: 0.933
- PR-AUC: 0.7077

Confusion Matrix:



4.2 Error Analysis

We analyzed the misclassifications of the selected model (Logistic Regression based on PR-AUC) to understand potential sources of errors.

=== False Positives (top 3) ===

	TotalClaim	AvgClaim	MaxClaim	StdClaim	NumClaims	UniquePatients	InpatientClaims	OutpatientClaims
3491	905560	5880.259740	57000	8950.800158	154	138	92	62
120	867010	9126.421053	57000	9678.732434	95	92	90	5
2792	814770	9585.529412	57000	10903.959472	85	80	81	4

=== False Negatives (top 3) ===

	TotalClaim	AvgClaim	MaxClaim	StdClaim	NumClaims	UniquePatients	InpatientClaims	OutpatientClaims	ClaimPerPatient
713	136000	10461.538462	22000	6715.844258	13	12	13	0	1.083333
4401	160000	8421.052632	19000	5378.101580	19	19	19	0	1.000000
1718	79240	3773.333333	22000	6408.863654	21	14	7	14	1.500000

Observations:

- **False Positives**: These are legitimate providers with unusually high claim volumes or atypical billing patterns. For example, some providers had a very

high number of claims or abnormal ratios of inpatient to outpatient claims.

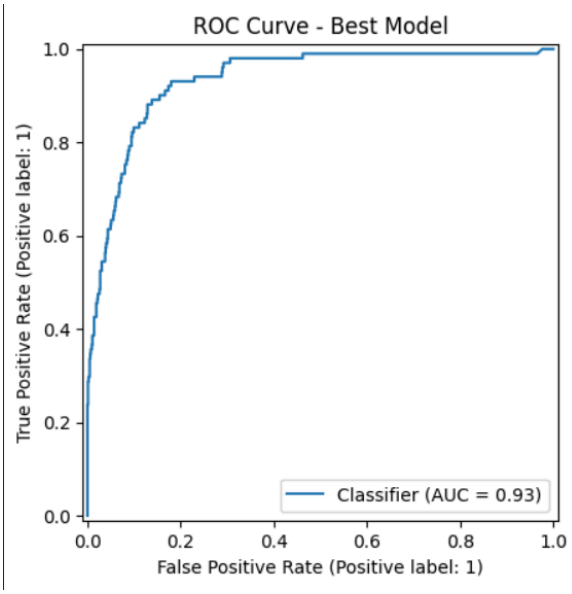
- **False Negatives**: These are fraudulent providers whose patterns are subtle and not fully captured by the current features. Many had low claim counts or unusual claim distributions, making detection more challenging.

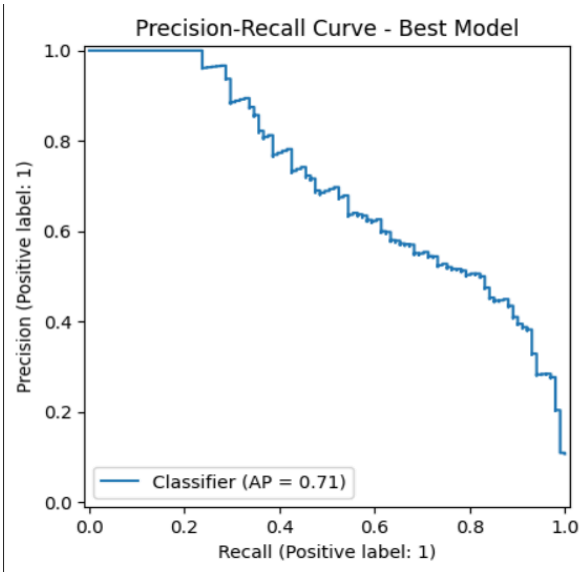
Future Improvements:

- Incorporate temporal trends (e.g., claim patterns over time).
- Include patient-level patterns to detect subtle fraud.
- Explore network or relational features among providers and patients to improve discrimination

4.3 ROC and Precision–Recall Curve Analysis

To further evaluate the performance of the selected model, we plotted both the Receiver Operating Characteristic (ROC) curve and the Precision–Recall (PR) curve using the predicted fraud probabilities. These visualizations help assess model behavior under class imbalance and provide deeper insight beyond point metrics such as F1-score or accuracy.





4.4 Discussion

This project shows that combining multiple healthcare claims tables into a single provider-level dataset can effectively support Medicare fraud detection. Key findings include:

Model Behavior & Trade-offs

- Logistic Regression achieved the highest PR-AUC (0.708), making it well-suited for imbalanced fraud detection.
- It had high recall (0.881), capturing most fraudulent providers, but moderate precision (0.445), meaning some legitimate providers were flagged.
- Tree-based models (XGBoost, Random Forest) were competitive but showed slightly lower PR-AUC and occasional overfitting.
- This suggests that fraud patterns may be roughly linearly separable, and that transparent models like Logistic Regression help interpret results.

4.5 Data-driven Patterns

Exploratory analysis revealed that fraudulent providers tend to:

- Submit higher inpatient and outpatient claims.
- Have higher average claims and more variability in billing.
- Show unusual ratios, such as high inpatient-to-outpatient billing or high-cost procedures.
- Serve more patients with chronic conditions, which may indicate upcoding.

These patterns align with known fraud behaviors like inflated services and unnecessary procedures.

4.6 Business Impact

Reduced Financial Losses

- The model helps identify high-risk providers, potentially reducing the $68+ billion annual fraud losses.
- Investigations can focus on the highest-risk cases.

Enhanced Investigation Efficiency

- Provider risk scores allow CMS to prioritize audits.

- Logistic Regression's interpretability helps investigators understand why a provider was flagged.

4.7 Minimizing False Positives

- False positives create unnecessary investigations, provider dissatisfaction, and administrative burden.

- The model aims to balance detection while limiting disruption to legitimate providers.

4.8 Limitations

- No temporal features: Cannot detect sudden changes or claim trends over time.

- Limited clinical context: Missing diagnosis codes, procedure codes, provider specialties, and geographic data.

- Class imbalance: Only ~10% of providers are fraudulent, affecting model stability.

- Basic feature engineering: More advanced features could improve performance.

4.9 Future Work

- Temporal modeling: Add time-based features like trends and rolling averages.

- Graph-based detection: Build networks of providers, patients, hospitals, and claims to detect collusion.

- Anomaly detection: Use methods like Isolation Forest, LOF, or Autoencoders to find unseen fraud.

- Explainability: Use SHAP values to show why providers are flagged.

- Cost-sensitive optimization: Penalize missing fraud more heavily than flagging innocent providers.

5.0 Conclusion

- Built an end-to-end fraud detection system using provider-level features.
- Addressed class imbalance and tested multiple models.
- Logistic Regression was chosen for its high PR-AUC and explainability.
- Key insights: fraudulent providers submit more claims, have higher average billing, and serve more chronic-condition patients.
- Model performance:
    - Precision: 0.445
    - Recall: 0.881
    - F1-score: 0.591
    - ROC-AUC: 0.933
    - PR-AUC: 0.708

This solution is effective, interpretable, and operationally viable for CMS to prioritize high-risk providers and reduce fraud losses.