Examining Medical Fraud Data

Edward Liu, Bowen Ying, and Maya Narang

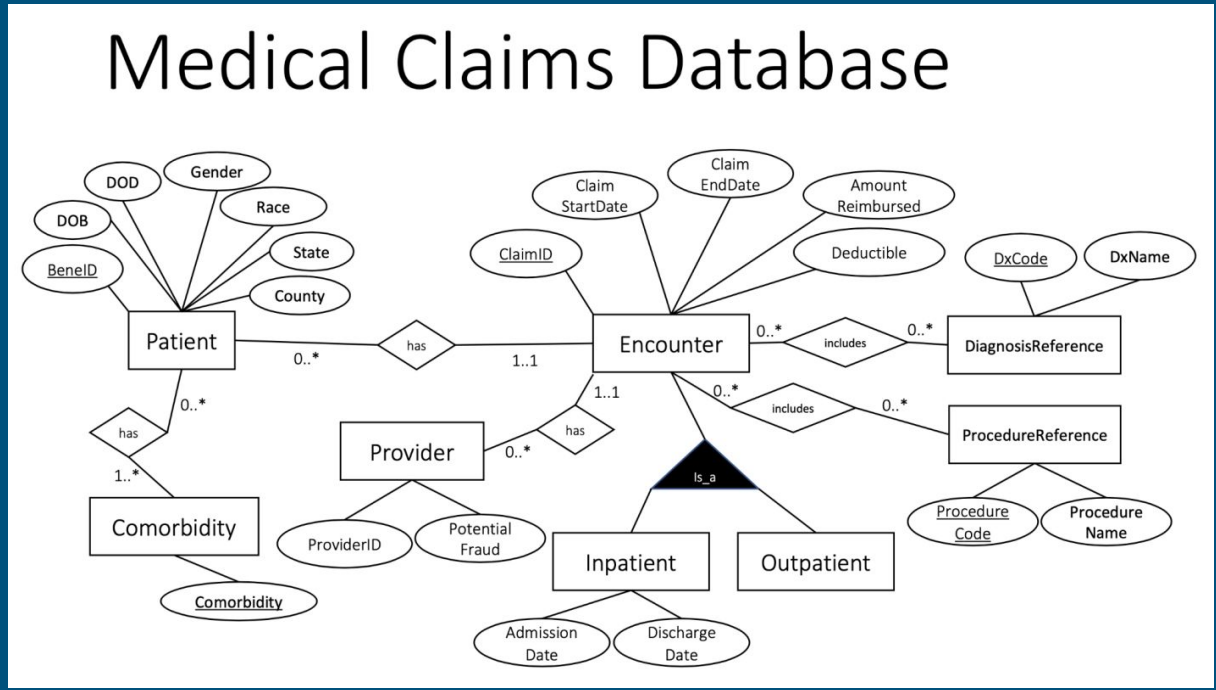
Background

- Healthcare fraud: extremely costly to healthcare system
  - Leads to increased insurance premiums, higher costs overall
- Can be committed by providers, physicians, beneficiaries, etc (focused on providers)
- Insights on when it is likely to occur could be useful for healthcare efficiency and cost reduction

Dataset & ER Diagram

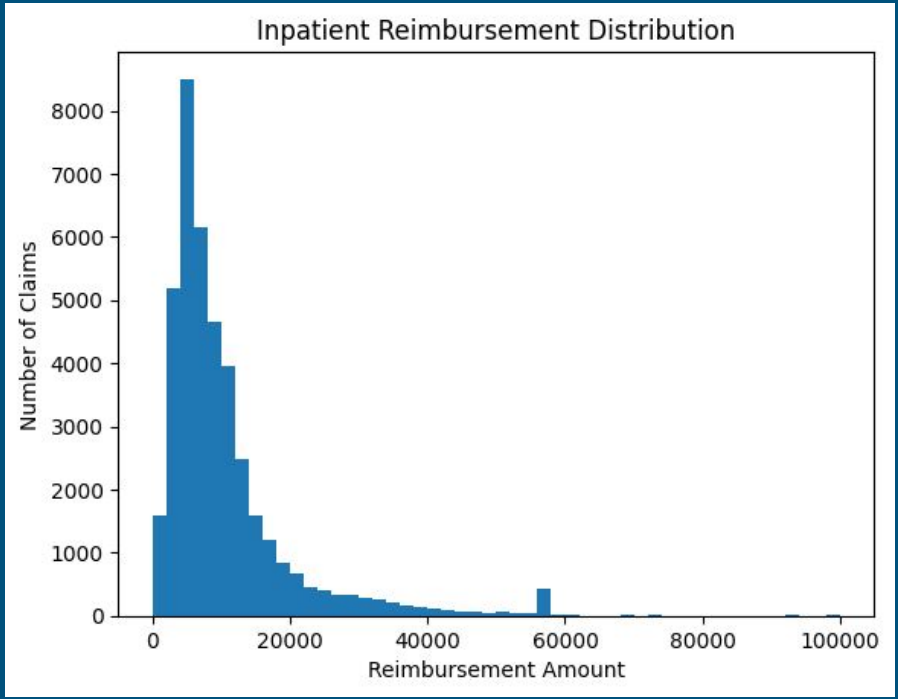
Patient claims dataset with 600,000 entries. Includes information on inpatient and outpatient claims, along with beneficiary details.

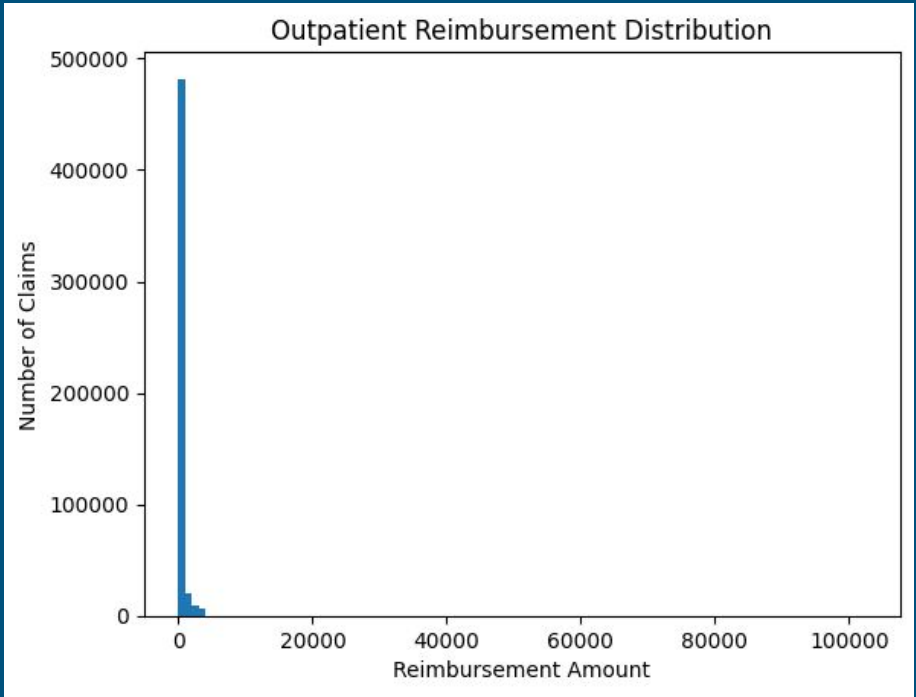
Integrated data into shown form during pre-processing.



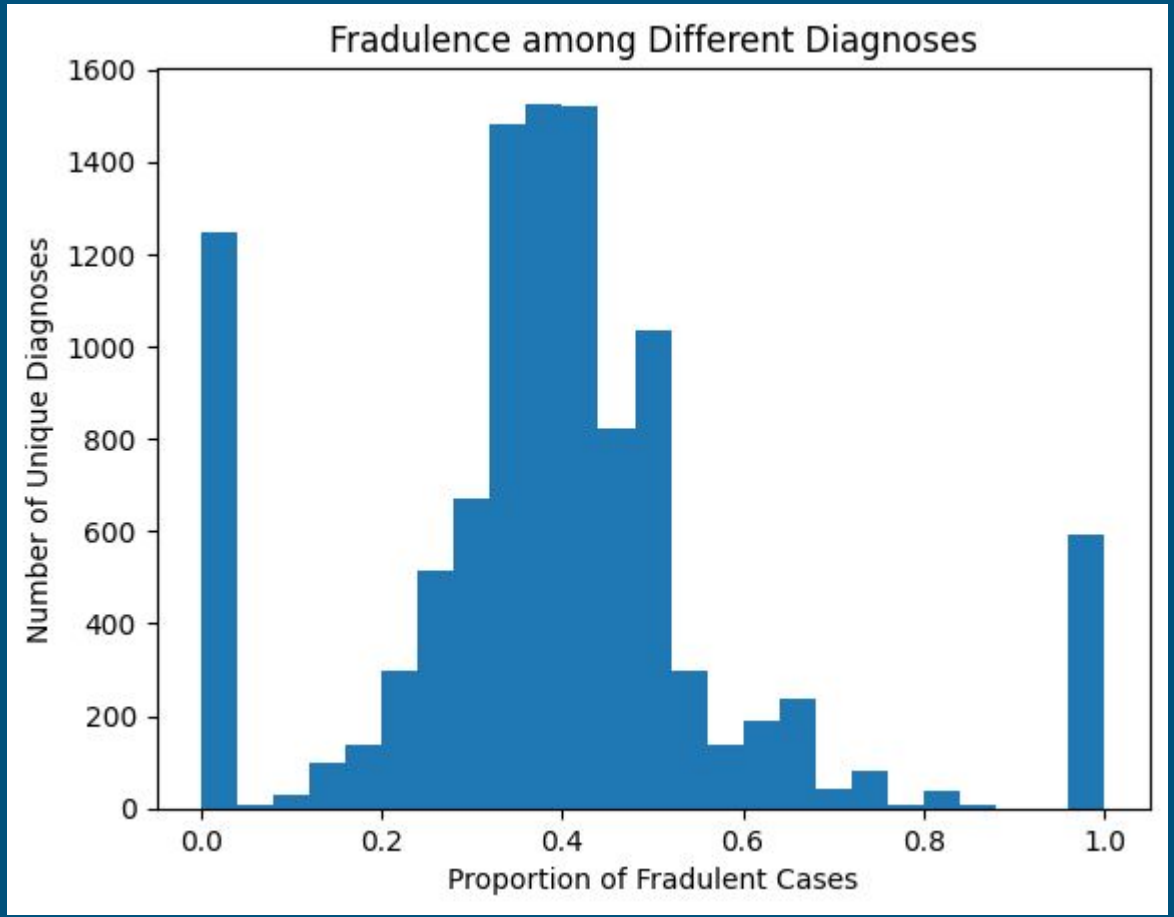
EDA

Inpatient vs. Outpatient Reimbursement



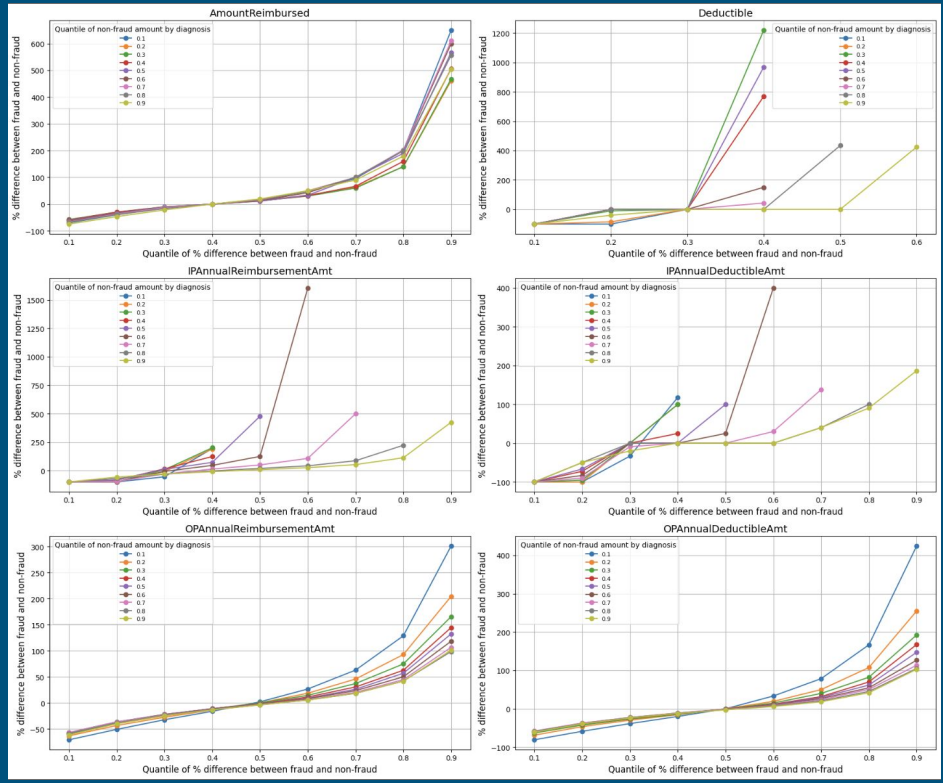


Proportion of Fraudulent Cases among Diagnoses



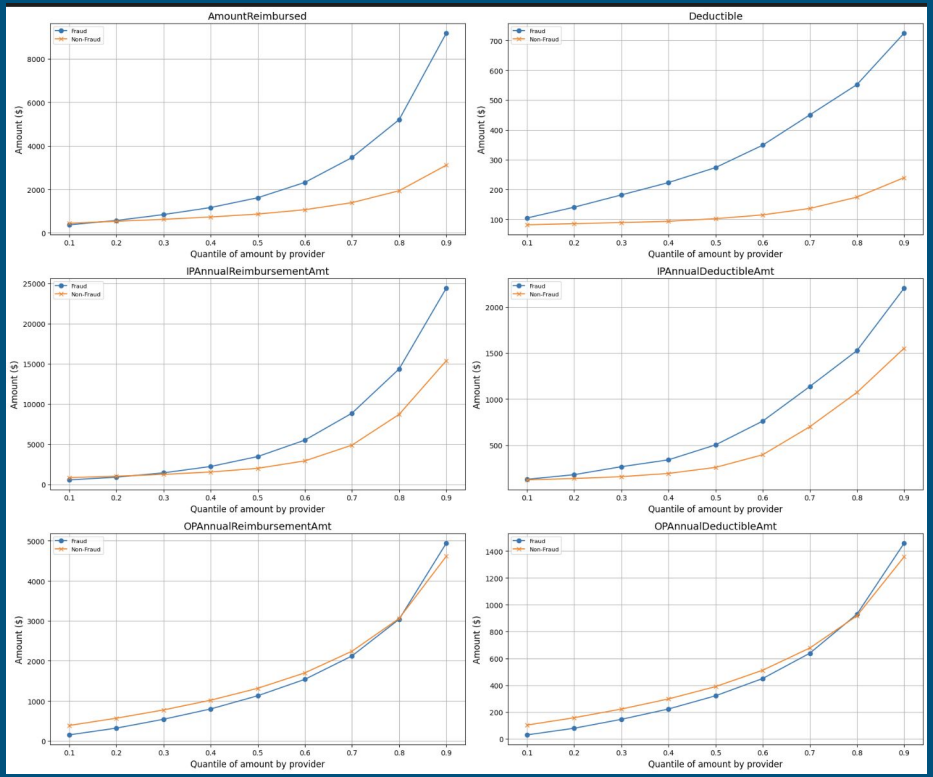
Quantile Analysis of Fraud by Diagnosis

- Non-linear relationship between fraud & reimbursement/deductible amount per diagnosis by quantile
- Most consistent / significant effects on Reimbursement Amounts (One time & Annual)



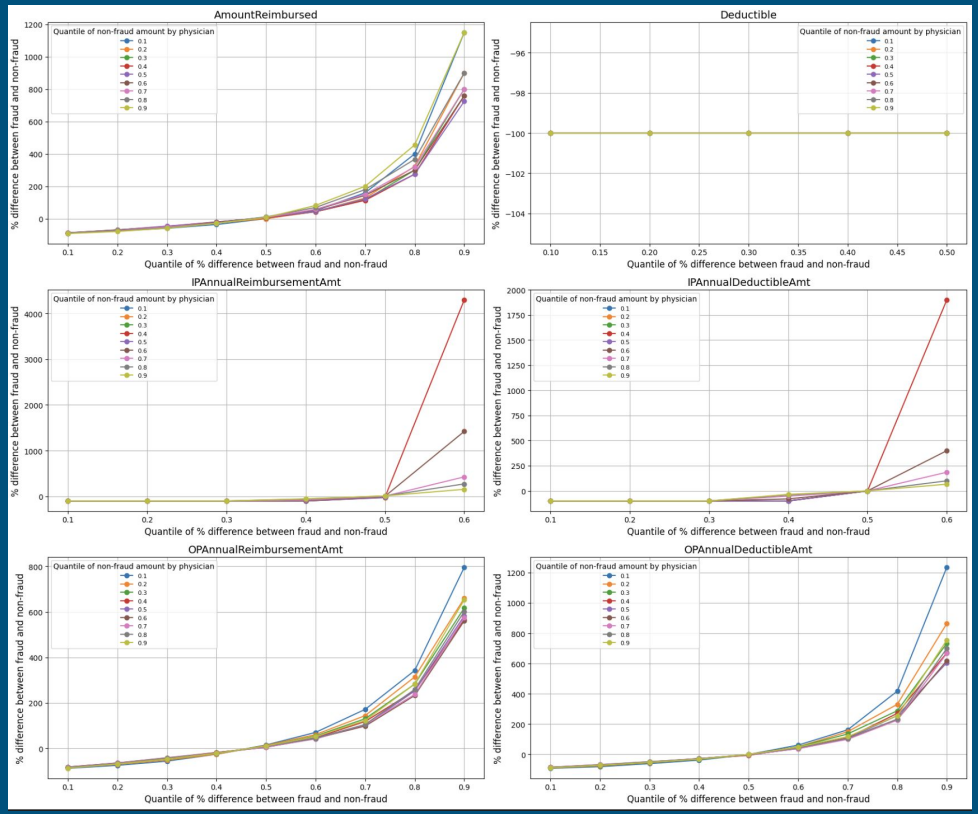
Quantile Analysis of Fraud by Provider

- Fraudulent providers generally have higher reimbursement amounts across all reimbursement quantiles

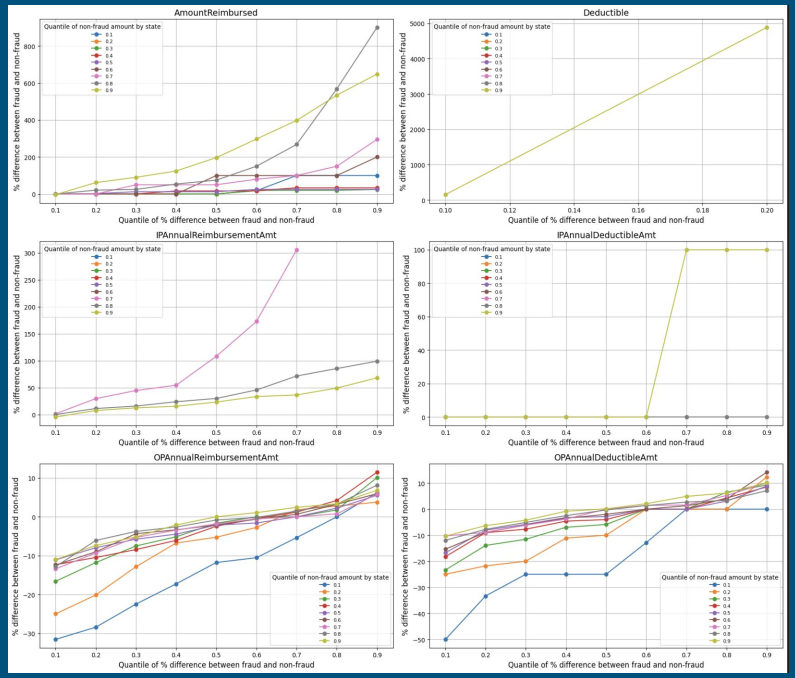


Quantile Analysis of Fraud by Physician

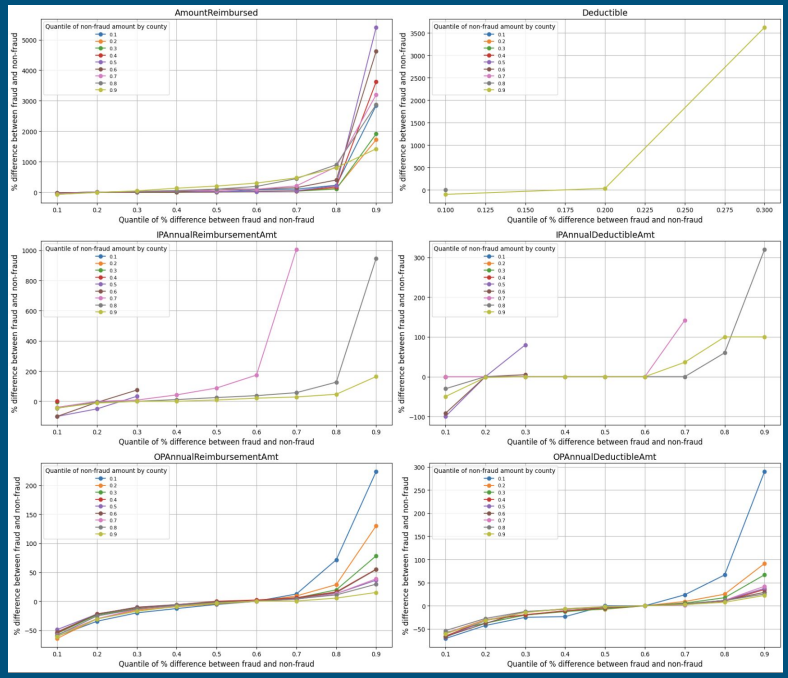
- Steep increase between fraud and non-fraud amounts at higher quantiles for reimbursement amounts (e.g. 0.8-0.9)
- Graphs for deductibles show minimal variation across quantiles, suggesting that deductibles may not be a strong indicator of fraudulent activity



Quantile Analysis of Fraud by Location (State & County)



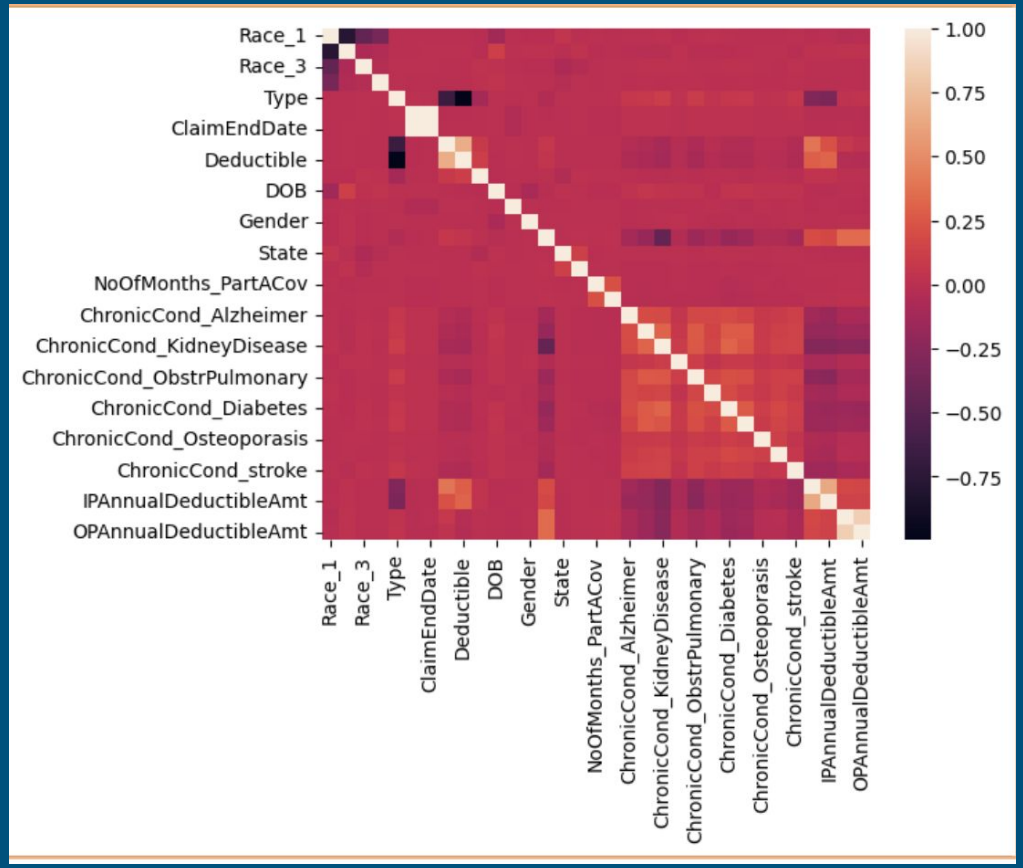
State



County

Variables & Correlation Matrix

- Not too many strong correlations
- Exceptions:
  - Claim Start Date/End Date
  - Reimbursement/Deductible amounts
- Can be handled with ensemble models like Random Forest



SQL Queries

1. Top 10 Beneficiaries with Most Frequent Claim Submission Rate – measured by claims/day

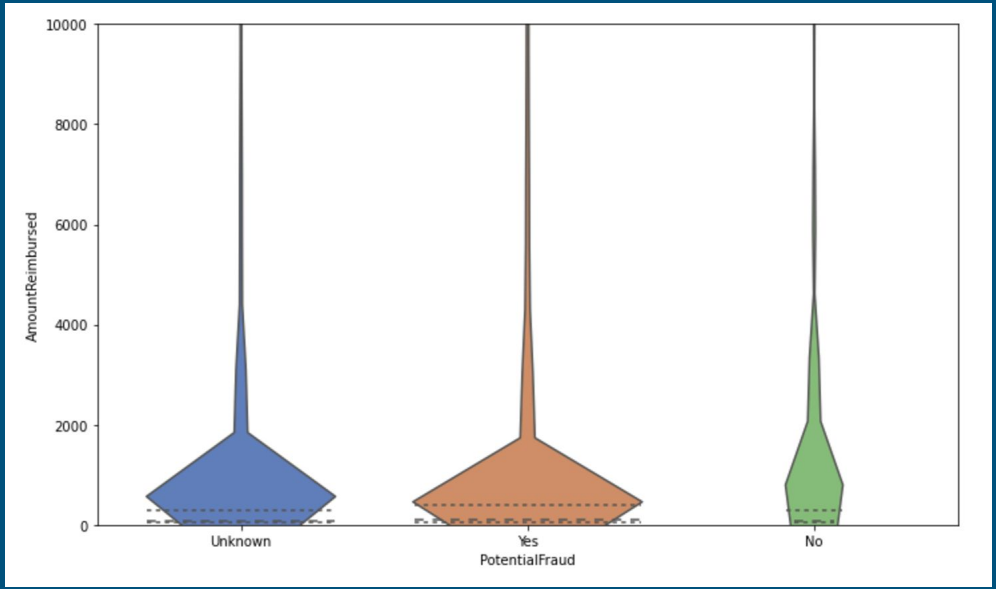
	BeneID	TotalClaims	FirstClaimDate	LastClaimDate	DaysActive	ClaimsPerDay
0	BENE56323	6	2009-08-17	2009-09-15	29.0	0.21
1	BENE57398	7	2009-01-03	2009-02-10	38.0	0.18
2	BENE61703	6	2009-01-11	2009-02-16	36.0	0.17
3	BENE42005	8	2009-02-26	2009-04-13	46.0	0.17
4	BENE41751	6	2009-01-18	2009-02-25	38.0	0.16
5	BENE59251	12	2008-12-23	2009-03-13	80.0	0.15
6	BENE22281	7	2009-01-02	2009-02-17	46.0	0.15
7	BENE128399	7	2009-10-02	2009-11-18	47.0	0.15
8	BENE40570	10	2009-01-07	2009-03-20	72.0	0.14
9	BENE75260	16	2008-12-21	2009-04-16	116.0	0.14

2. Detect Providers with Potential Fraud – those with a reimbursement amount 50% higher than the global average reimbursement

	ProviderID	TotalClaims	TotalReimbursement	AvgReimbursement
0	PRV57080	1	57000.0	57000.000000
1	PRV51814	1	57000.0	57000.000000
2	PRV57399	1	57000.0	57000.000000
3	PRV52537	1	57000.0	57000.000000
4	PRV53033	2	75000.0	37500.000000
...	...	...	...	...
1194	PRV51354	4	10440.0	2610.000000
1195	PRV53281	53	138310.0	2609.622642
1196	PRV54968	49	127820.0	2608.571429
1197	PRV52321	150	390230.0	2601.533333
1198	PRV56135	22	57230.0	2601.363636

Hypothesis Tests

```
Observed Difference in Means: 634.2917138299225
P-value: 0.0
95% Confidence Interval: [-20.58611452  20.72371041]
```



Null Hypothesis (H_0): The average amount reimbursed is the same for providers with and without fraud flags.

Alternative Hypothesis (H_1): The average amount reimbursed differs significantly between providers with and without fraud flags.

Feature Engineering

Feature Engineering

Added Features:

- **Cost/Category variables:** How extreme a reimbursement/deductibles is relative to others in some categorical variable (like: State, Physician, Diagnosis)
- **Duration/Count Variables:** Length of claim, count of diagnosis, count of chronic conditions

Encoding Categorical Variables:

- One-hot encoding of race & chronic conditions

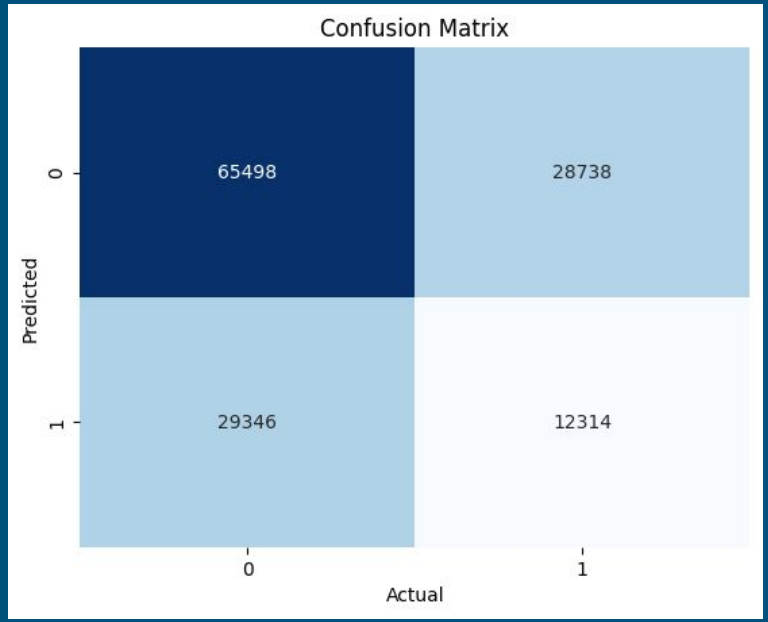
Scale Data:

- **Cost/Category** scales costs

Models

Logistic Regression (Baseline)

- **Justification**: efficient, straightforward, less prone to overfitting
- **Limitations**: assumes linear relationships between features and target
- **Results**: overall poor performance
  - Decent precision for negative (0) class
  - Very poor precision for positive (1) class



	0	1	accuracy	macro avg	weighted avg
precision	0.690587	0.299961	0.572585	0.495274	0.570837
recall	0.695042	0.295583	0.572585	0.495313	0.572585
f1-score	0.692807	0.297756	0.572585	0.495282	0.571701
support	94236.000000	41660.000000	0.572585	135896.000000	135896.000000

Gradient Boosting

```
Gradient Boosting Classifier

Gradient Boosting Accuracy: 0.8136
Gradient Boosting ROC AUC Score: 0.9029

Gradient Boosting Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.93      0.87     94236
           1       0.78      0.55      0.64     41660

    accuracy                           0.81    135896
   macro avg       0.80      0.74      0.76    135896
weighted avg       0.81      0.81      0.80    135896


Gradient Boosting Feature Importance:
AttendingPhysician                                  0.389281
AttendingPhysician_quantile_AmountReimbursed        0.121278
AttendingPhysician_quantile_OPAnnualDeductibleAmt   0.093374
ProviderID_quantile_AmountReimbursed                0.083377
AttendingPhysician_quantile_OPAnnualReimbursementAmt  0.059902
                                                      ...
ChronicCond_IschemicHeart                           0.000000
Race_2                                              0.000000
ChronicCond_rheumatoidarthritis                     0.000000
ChronicCond_stroke                                  0.000000
DxCode_quantile_OPAnnualDeductibleAmt_y             0.000000
Length: 62, dtype: float64
```

- Class 0 (Negative Class): High precision (0.82) and recall (0.93), indicating the model performs very well in predicting the negative class.
- Class 1 (Positive Class): Lower recall (0.55) compared to precision (0.78), suggesting the model struggles to identify all positive cases.

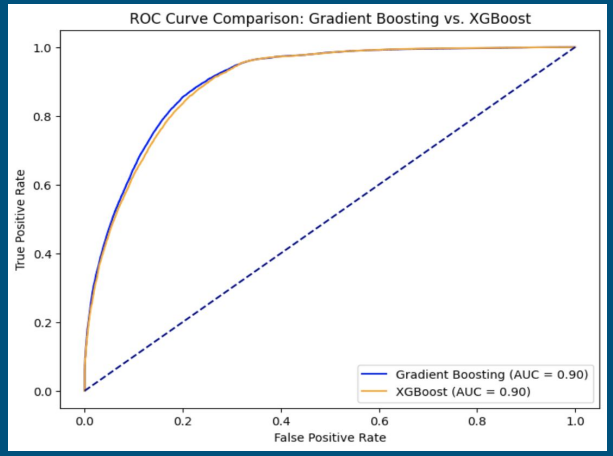
Justification: chosen for its ability to handle imbalanced datasets effectively

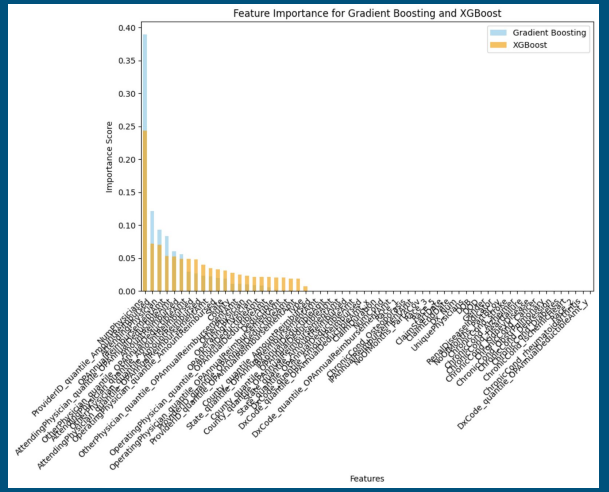
Limitations: class imbalance, performs better for negative class

Corrections: Hyperparameter tuning for optimal performance on n_estimators, learning_rate, and max_depth

Results:

- Accuracy: 81.36% overall, indicating a strong model performance but room for improvement in fraud detection.
- ROC AUC Score: 0.9029, showing the model distinguishes well between fraud and non-fraud cases.

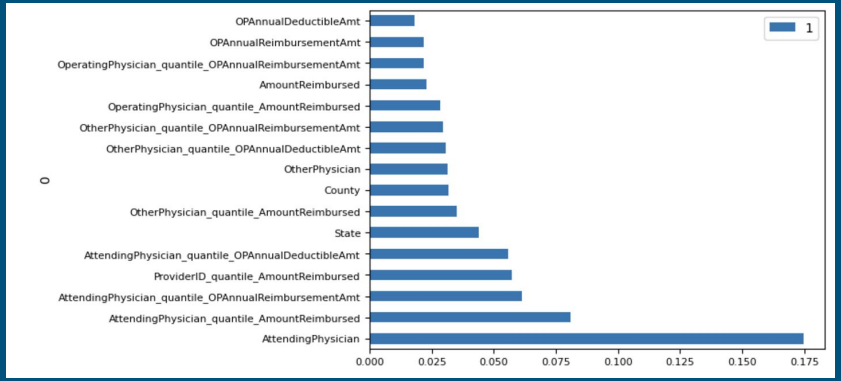




Random Forest Classifier

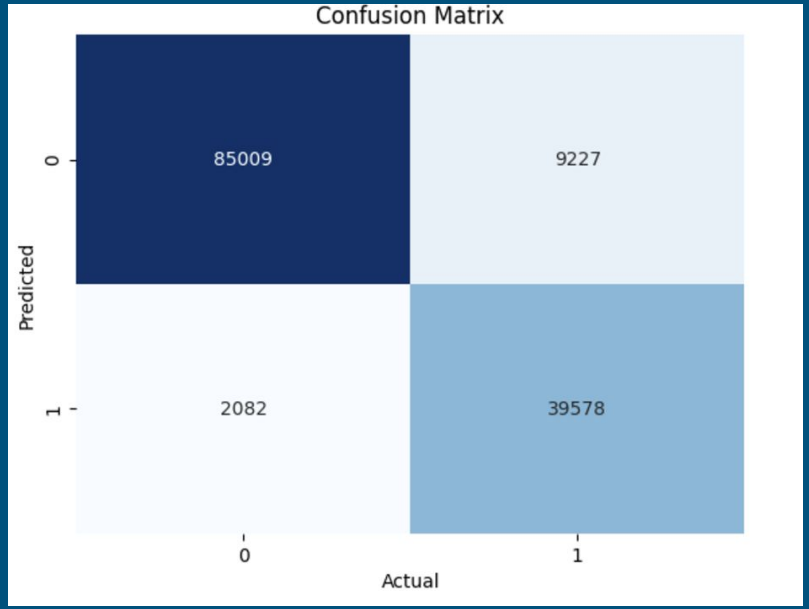
- **Justification:** Captures non-linear relationships, robust to multicollinearity
- **Limitations:** Not easily explainable, slow to run, overfitting
- **Corrections:**
  - Hyperparameter tuning:
    - n_estimators, max_depth
  - 5-Fold Cross Validation to reduce overfitting
- Results:
  - Significant variables: Physician, County, State, Cost/Category

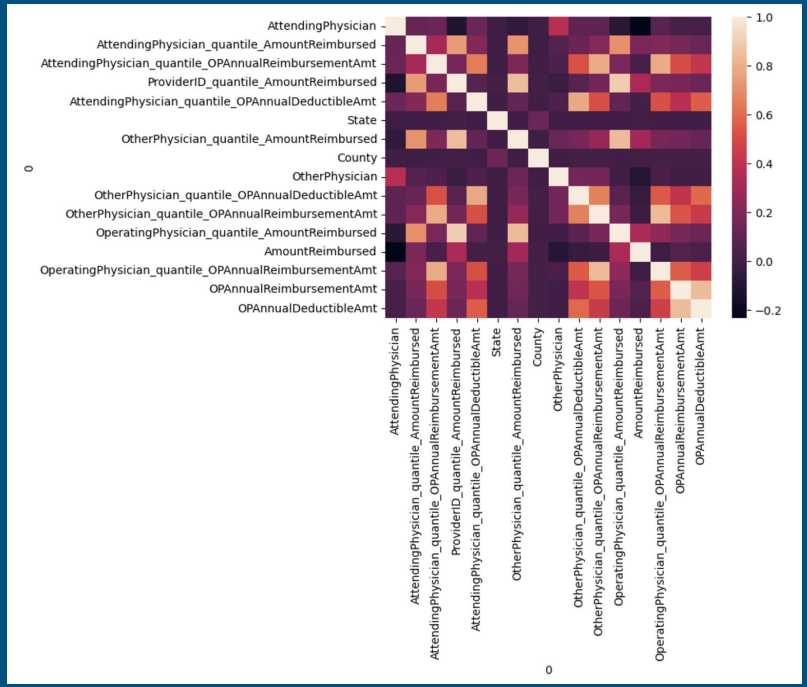
	0	1	accuracy	macro avg	weighted avg
precision	0.976094	0.810942	0.916782	0.893518	0.925465
recall	0.902086	0.950024	0.916782	0.926055	0.916782
f1-score	0.937632	0.874990	0.916782	0.906311	0.918429
support	94236.000000	41660.000000	0.916782	135896.000000	135896.000000



Random Forest Classifier

- Correlation matrix and confusion matrix





Discussion

Results

Best Model Overall: Random Forest due to its superior recall and accuracy for detecting fraudulent claims, just slow to run

Gradient Boosting Strengths: Balanced performance but struggles with recall for fraudulent claims.

Logistic Regression Weaknesses: Poor performance across all metrics; unsuitable for imbalanced datasets like fraud detection.

Top Predictors for Fraud:

- Physician, State
- Financial variables: Claim Reimbursement, Annual Reimbursement Amount
- Quantile variables: percentile of reimbursements for a given category

Next Steps

Where to go from here?

Model Optimization:

- Logistic Regression:
  - Consider regularization (L1/L2) and feature engineering to improve predictive power.
- Gradient Boosting:
  - Use techniques like Bayesian optimization to adjust learning rate, tree depth, and boosting iterations.
- Random Forest:
  - Fine-tune hyperparameters like the number of estimators, max depth, and minimum samples for splits to further improve performance.