

Fighting Fraud, Protecting Policyholders

Using machine learning to detect Medicare fraud and strengthen targeted prevention strategies

Group 4 | March 11th, 2025



Medicare

Agenda

1

Problem Definition & Objective

Overview of the business problem and primary goal guiding the analysis

2

Approach & Methodology

Discussion of the data processing, feature engineering, and model selection

3

Modeling Results

Summary of the key findings and performance of the tested models

4

Analysis & Limitations

Analysis of the business implications and potential areas of enhancement

Agenda

1

Problem Definition & Objective

Overview of the business problem and primary goal guiding the analysis

2

Approach & Methodology

Discussion of the data processing, feature engineering, and model selection

3

Modeling Results

Summary of the key findings and performance of the tested models

4

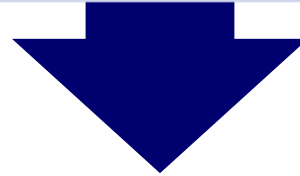
Analysis & Limitations

Analysis of the business implications and potential areas of enhancement

Problem Definition and Objective: Using machine learning to identify fraudulent claims

Problem:

Medicare fraud, including false claims, overbilling, and service misrepresentation, led to \$2 billion in taxpayer losses in 2024 and increased policyholder premiums by \$25–\$50. These fraudulent activities strain healthcare resources, inflate costs and undermine the integrity of the system.



Objective (What Machine Learning Can Do):

We will develop a machine learning model to identify fraudulent claims given high-dimensional insurance data. Specifically, by leveraging anomaly detection and classification models, we can enable Medicare and the Department of Health and Human Services to engage in targeted prevention efforts against fraud.

Agenda

1

Problem Definition & Objective

Overview of the business problem and primary goal guiding the analysis

2

Approach & Methodology

Discussion of the data processing, feature engineering, and model selection

3

Modeling Results

Summary of the key findings and performance of the tested models

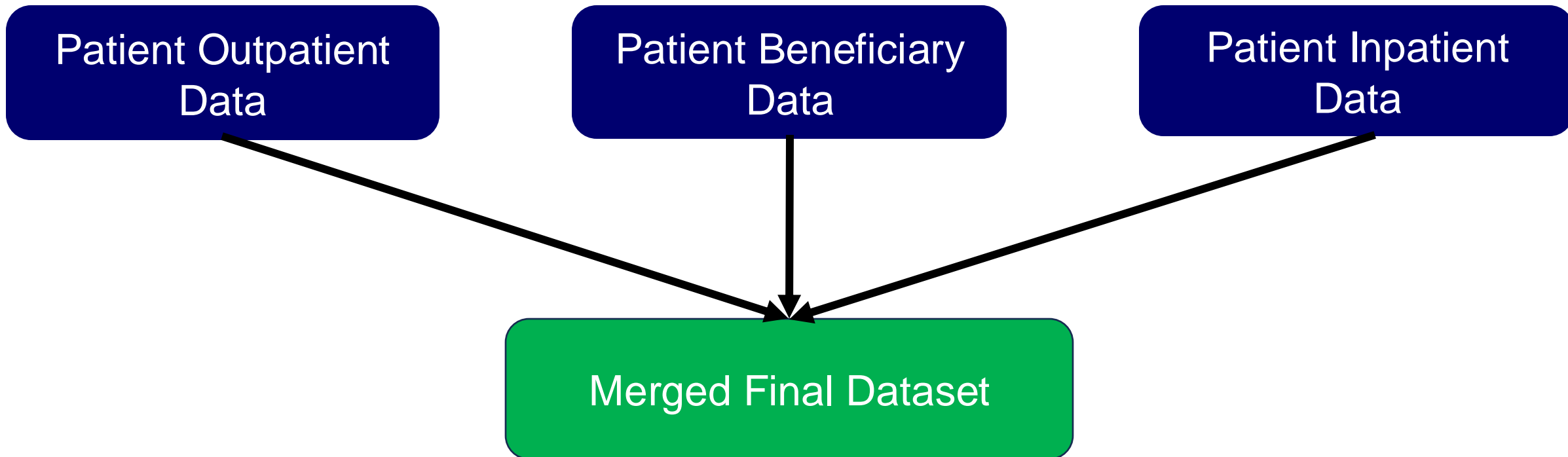
4

Analysis & Limitations

Analysis of the business implications and potential areas of enhancement

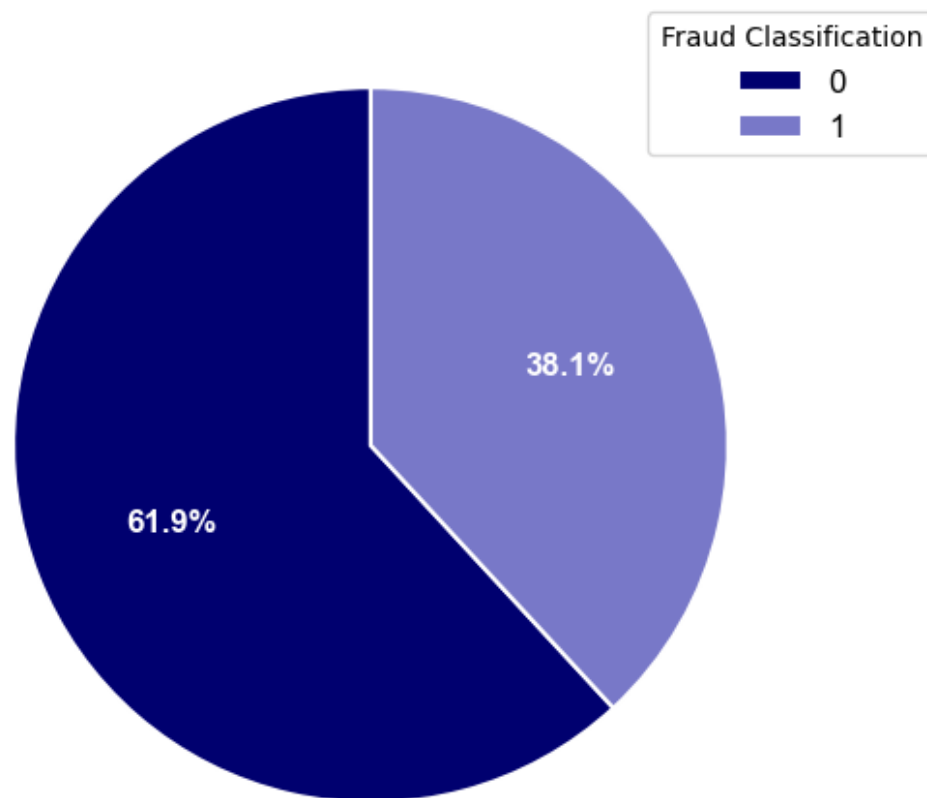
From Kaggle, we collected inpatient, outpatient, and beneficiary data to understand Medicare fraud

To detect fraudulent healthcare providers, we collect and analyze Medicare claims data from inpatient (hospital stays), outpatient (non-hospital stays), and beneficiary records. This data includes hospital stays, medical procedures, billing details, and patient demographics.



The balanced fraud distribution does not match real-world fraud frequency, raising generalization concerns

Percentage of Potentially Fraudulent Observations



But, **only about 5-7%** of Medicare claims were found to be **actually fraudulent** in 2024

Data Preparation and Feature Engineering: High-Level Overview & Approach

Data Preparation

- + Handle Missing Values (simple imputation)
- + Handle outliers (use scaling/transformation)
- + Encode categorical columns (e.g. claim codes)

Feature Engineering

- + Extract date features to remove datetime types
- + Discretize age into bins
- + Log transform skewed distributions



Apply Feature Scaling for Use in Modeling

We used a 4-pronged approach, shortlisting a comprehensive range of modeling options



Linear-Based Models

- Logistic Regression



Tree-Based Models

- Random Forest
- XG Boost



Probabilistic Models

- Gaussian Naïve Bayes
- Bernoulli Naïve Bayes



Ensemble Methods

- Voting Model (LR, RF, GNB)
- Stacked Model (LR, SVM)

Agenda

1

Problem Definition & Objective

Overview of the business problem and primary goal guiding the analysis

2

Approach & Methodology

Discussion of the data processing, feature engineering, and model selection

3

Modeling Results

Summary of the key findings and performance of the tested models

4

Analysis & Limitations

Analysis of the business implications and potential areas of enhancement

Ultimately, XGBoost performed the best among the tested models in 3/5 metric categories




Medicare

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
Logistic Regression	81.07%	78.35%	69.63%	73.73%	86.30%
Random Forest	93.17%	99.02%	82.92%	90.25%	98.42%
Gaussian Naïve Bayes	80.11%	81.98%	61.39%	70.21%	80.43%
Bernoulli Naïve Bayes	62.83%	56.48%	11.30%	18.83%	53.17%
Voting Model	94.76%	99.73%	86.51%	92.65%	-
Logistic Regression Stacked	97.64%	99.82%	93.58%	96.81%	97.31%
SVM Stacked	97.65%	99.79%	94.03%	96.82%	-
XGBoost	97.54%	99.86%	93.68%	96.67%	99.82%

While the XGBoost model performed best overall, precision and recall were not equally balanced

		True Label	
		Not Fraud	Fraud
Predicted Label	Not Fraud	69,184	2,701
	Fraud	57	40,034



Modeling Performance Metrics	
	97.5% Overall Accuracy
	99.9% Weighted Average Precision
	93.7% Weighted Average Recall

Agenda

1

Problem Definition & Objective

Overview of the business problem and primary goal guiding the analysis

2

Approach & Methodology

Discussion of the data processing, feature engineering, and model selection

3

Modeling Results

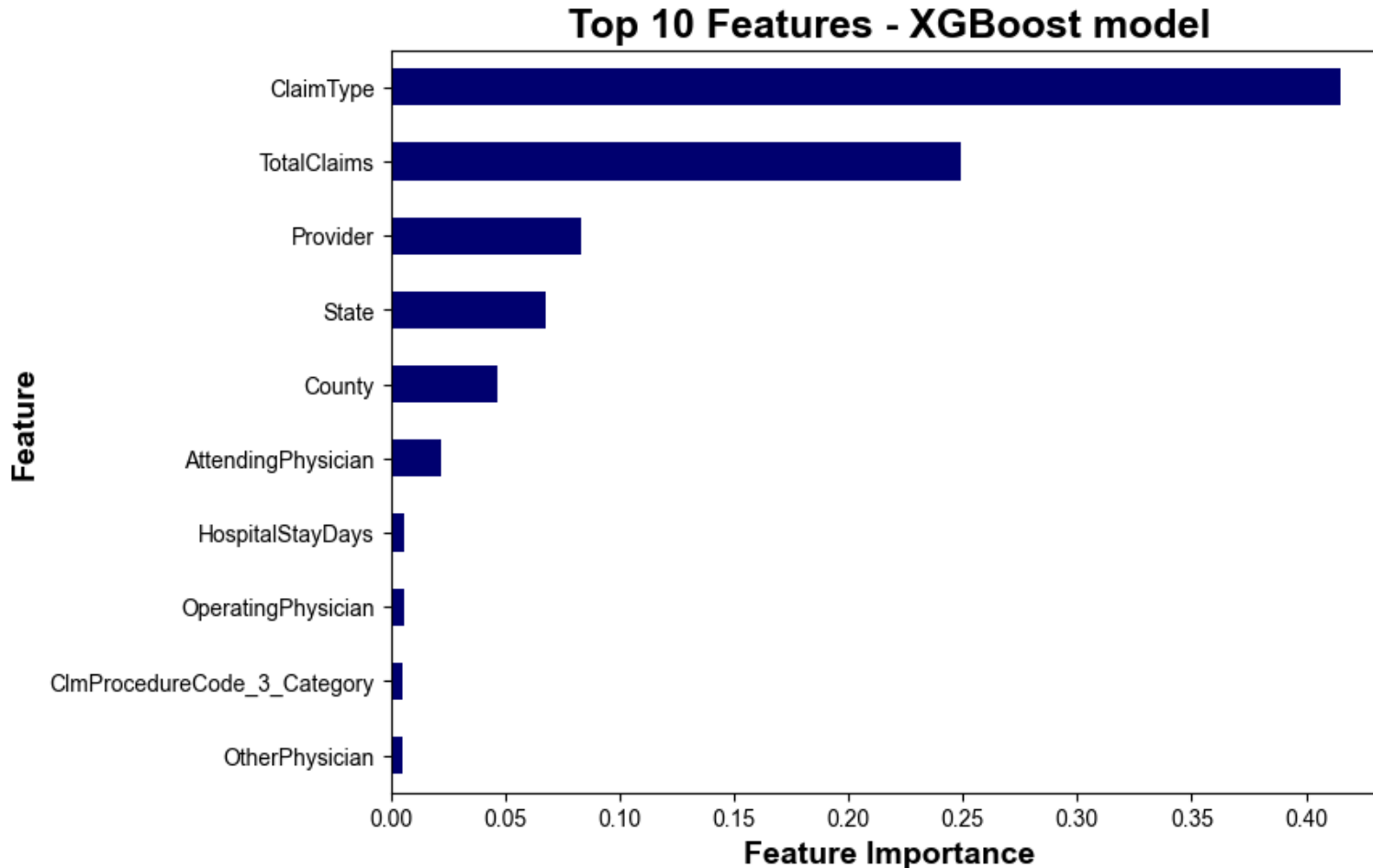
Summary of the key findings and performance of the tested models

4

Analysis & Limitations

Analysis of the business implications and potential areas of enhancement

Claim type, total claims, and the provider should be prioritized most in future fraud investigations



Claim type is the #1 predictor of fraud

The top 3 predictors describe 75% of fraud

Optimizing Fraud Detection: Examining the trade-off between precision and recall



Medicare

Fine tuning was a balance between the cost of investigating fraud and the benefits of identifying it



PRECISION

Of the cases predicted to be fraudulent,
99% were actually fraudulent



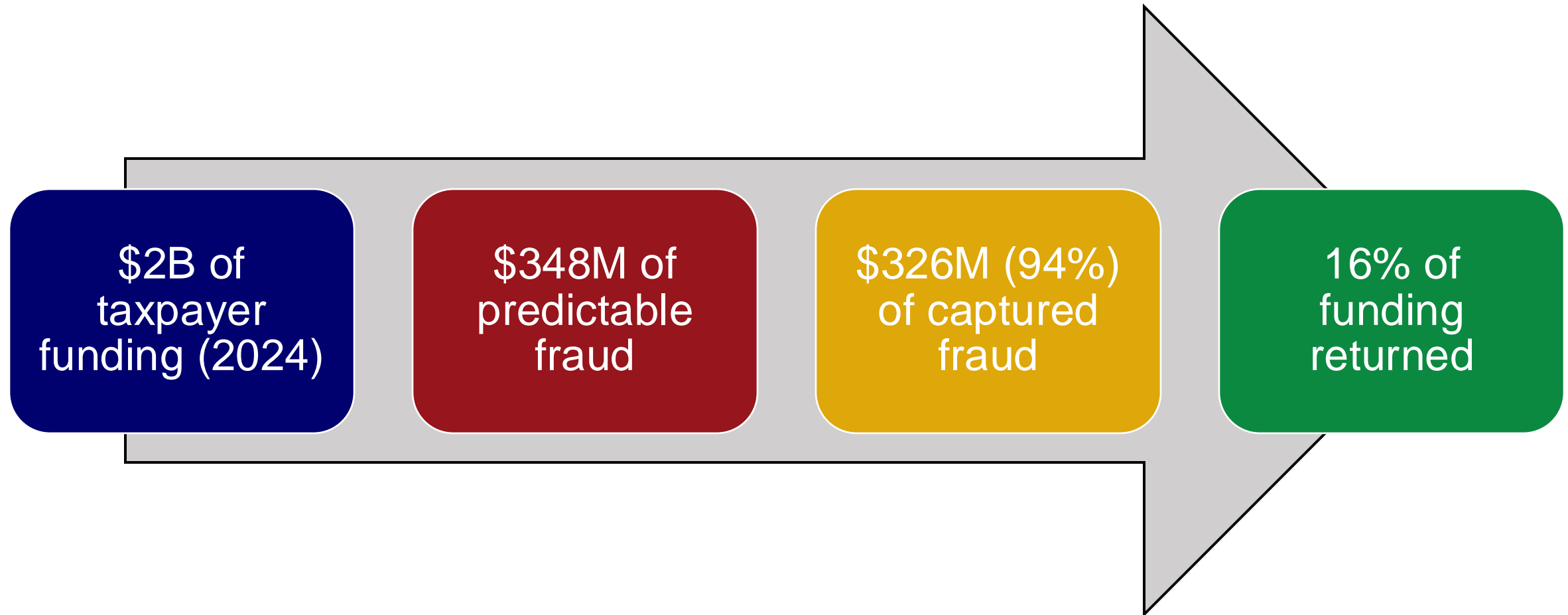
RECALL

Our model accurately
predicts 94% of fraud cases

Model deployment could lead to the recovery of approximately 16+% of taxpayer funding



Medicare



More information about the data and business priorities will further enhance the model's value

1

Data Validation & Clarification

How is the data sourced and validated?

What is the exact definition of '*PotentialFraud*'?

2

Business Objective Refinement

What is the cost of fraud investigations?

Should the model prioritize recall or precision to maximize impact?



Medicare

Questions?

Appendix

Key Takeaways from our Exploratory Data Analysis of the datasets:

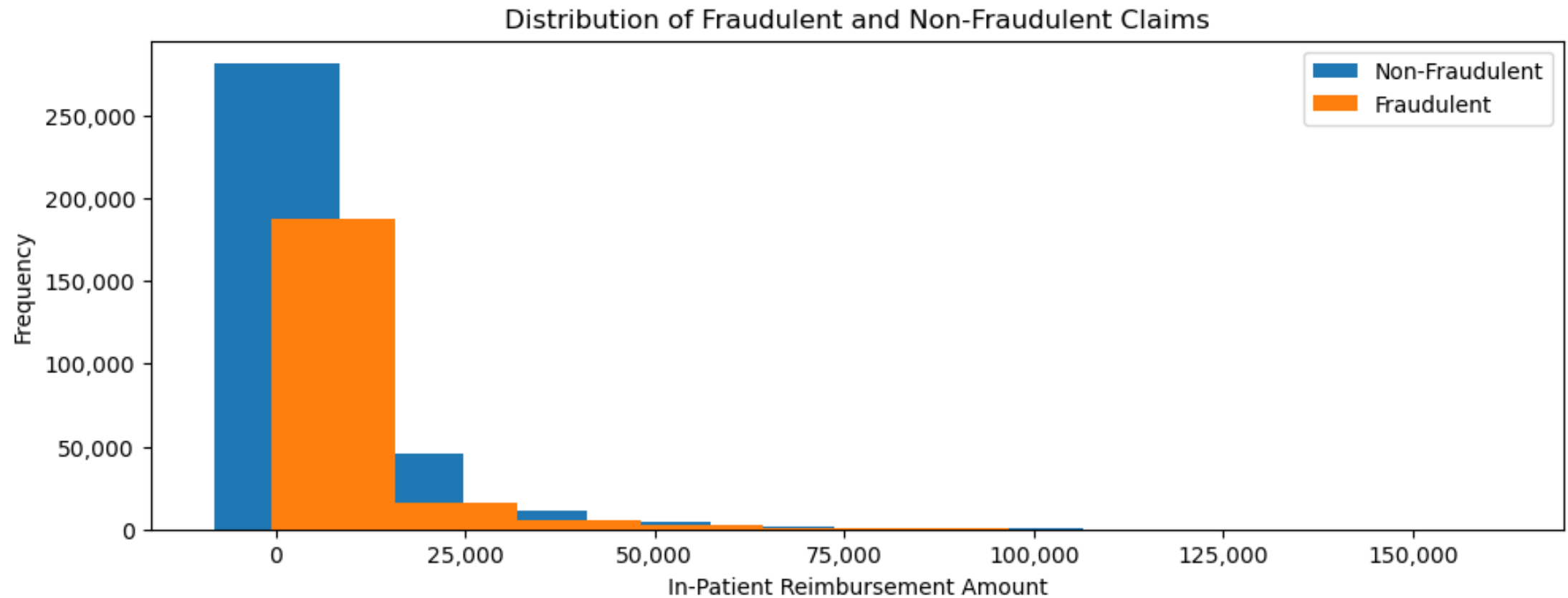
1. Fraudulent claims have a higher reimbursement amount
2. Several columns have missing values, but many can be attributed to a sparse structure
3. There are some outliers in all datasets, but this is not unexpected given the variation in the cost of medical services and needs
4. Many categorical features require pre-processing to be used meaningfully in modeling – particularly the diagnosis and claims codes
5. The fraud distribution is balanced, raising generalizability concerns

Specific Data Preparation and Feature Engineering Processing Steps:

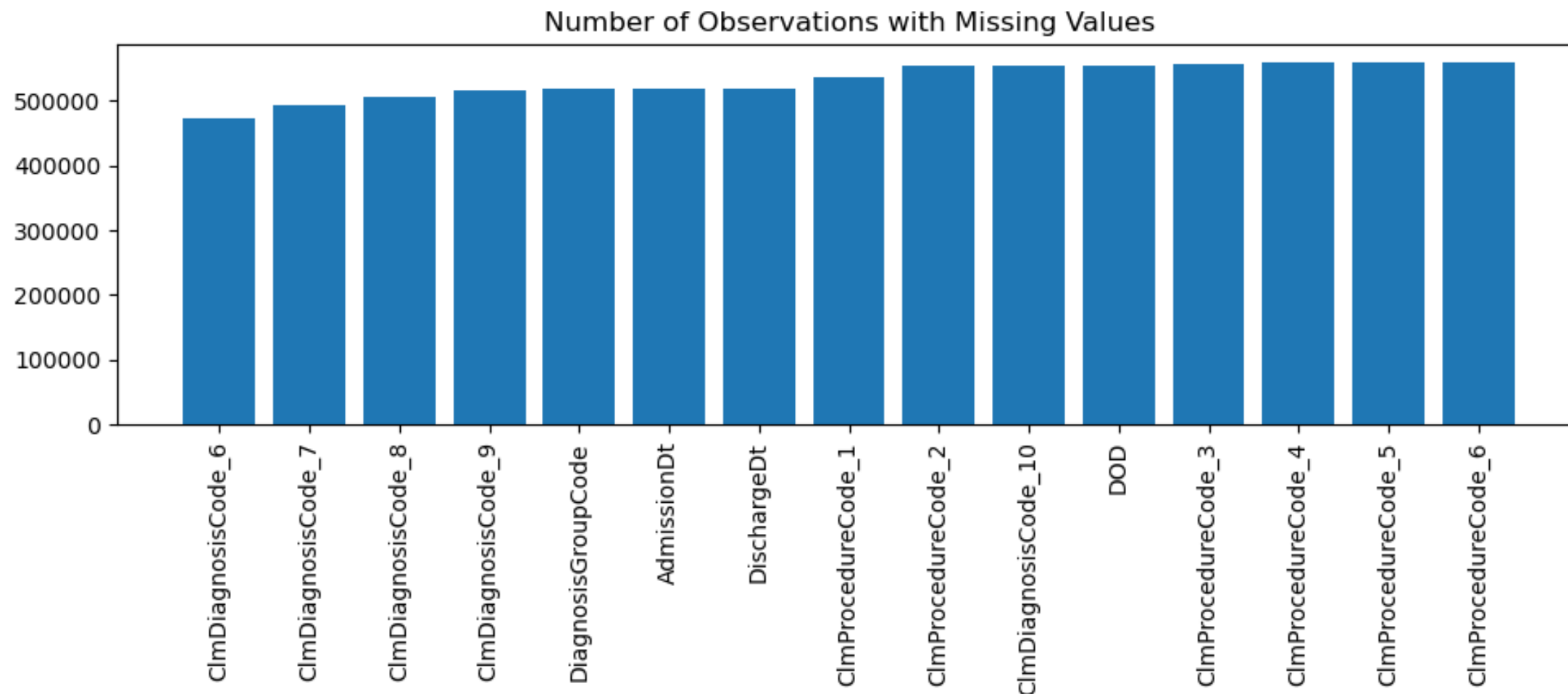
Stage	Step	Detail	Original Data Example	Processed Data Example
Data Cleaning	Dataset Merging	Joined the inpatient, outpatient, beneficiary, and provider datasets	-	-
	Missing Value Handling	Kept all rows to avoid data loss but replaced NaN with 0 values	ClmProcedureCode_2: NaN	ClmProcedureCode_2: 0
	Outlier Handling / Transformation	Retained all outliers and instead applied log-transformation to claim and deductible columns	InscClaimAmtReimburse: 11000 InscClaimAmtReimburse: 3000	InscClaimAmtReimburse: 4.04 InscClaimAmtReimburse: 3.48
	Categorical Encoding	Categorically encoded extracted diagnosis form claim and procedure codes	ClmProcedureCode_2: 41071	Acute Myocardial Infarction (AMI) [Diagnosis Family] Subendocardial Infarction [Diagnosis Subtype Column]
Feature Engineering	Discretization	Binned age and claim features to for use in non-learn models	DOB: 1943-01-01	Age: 82 Age_Bin: 80-85
	Feature Creation	Created new features and made feature interactions	-	Avg_Claim_Provider Age_Diagnosis_Family

Fraudulent claims have a higher reimbursement amounts on average, but both are highly skewed distributions

Medicare



Several columns have missing values, but this can be attributed to collection structure, not quality issues



There are outliers, but this is not unexpected given the variation in the cost of medical services and needs

Distribution of Select Columns

