Identifying and Predicting Healthcare Fraud

by **Lucas Kim**, **Ryan Park**, and **Sita Thomas**
NYCDSA Capstone Project, October 2020

What Is Healthcare Fraud?

- Fake or altered medical claims
- It is difficult to identify
- It increases medical costs
- Reducing fraud saves money

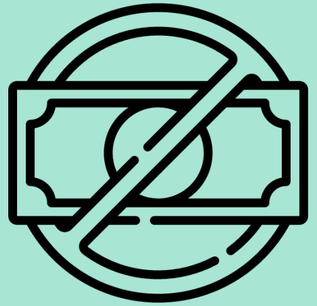
Audience and Goals

WHO IS THIS PROJECT FOR?

A healthcare insurance company

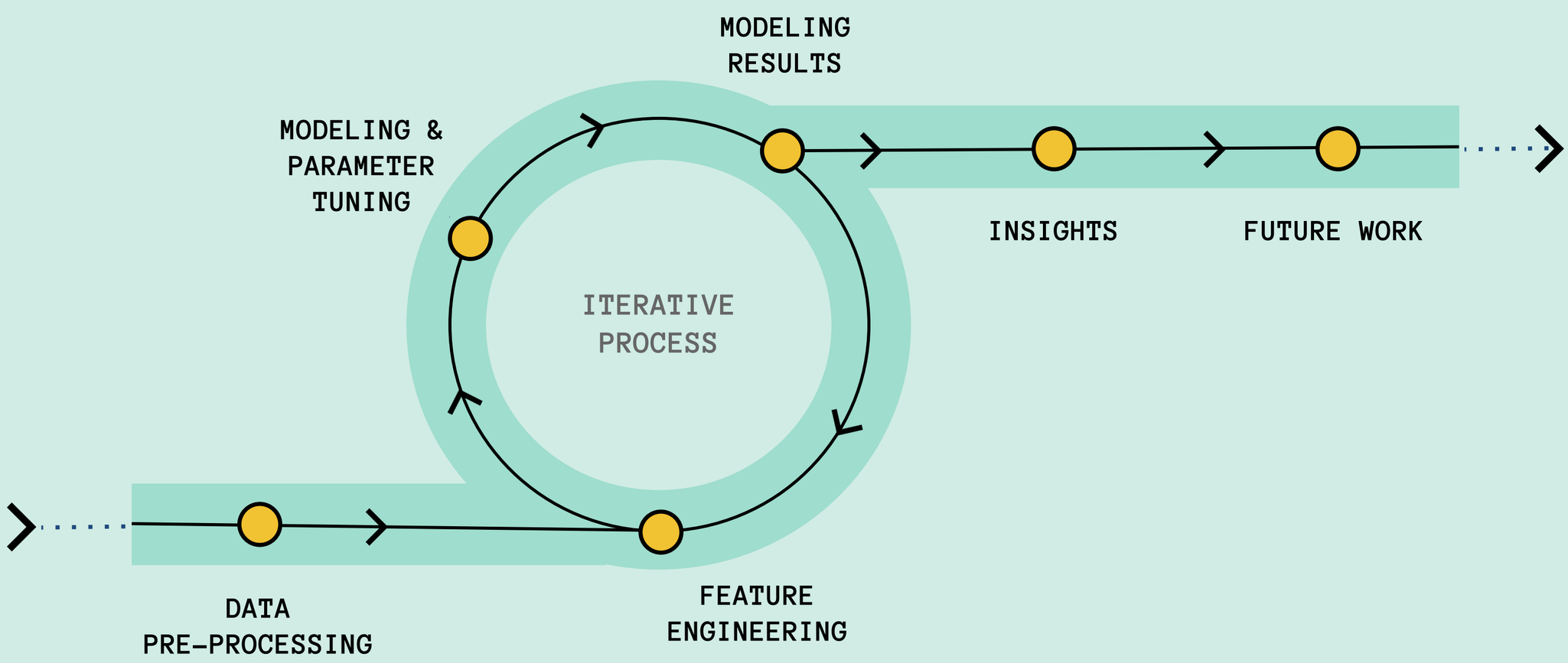
WHAT ARE THE PROJECT GOALS?

Flag potentially fraudulent claims



Minimize financial loss

Workflow


MODELING
RESULTS
MODELING &
PARAMETER
TUNING
INSIGHTS
FUTURE WORK
ITERATIVE
PROCESS
DATA
PRE-PROCESSING
FEATURE
ENGINEERING

Pre-processing: Claims Dataframe

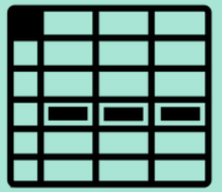
Inpatient Claim Data

Outpatient Claim Data

Patient Data

Potential Fraud Flag

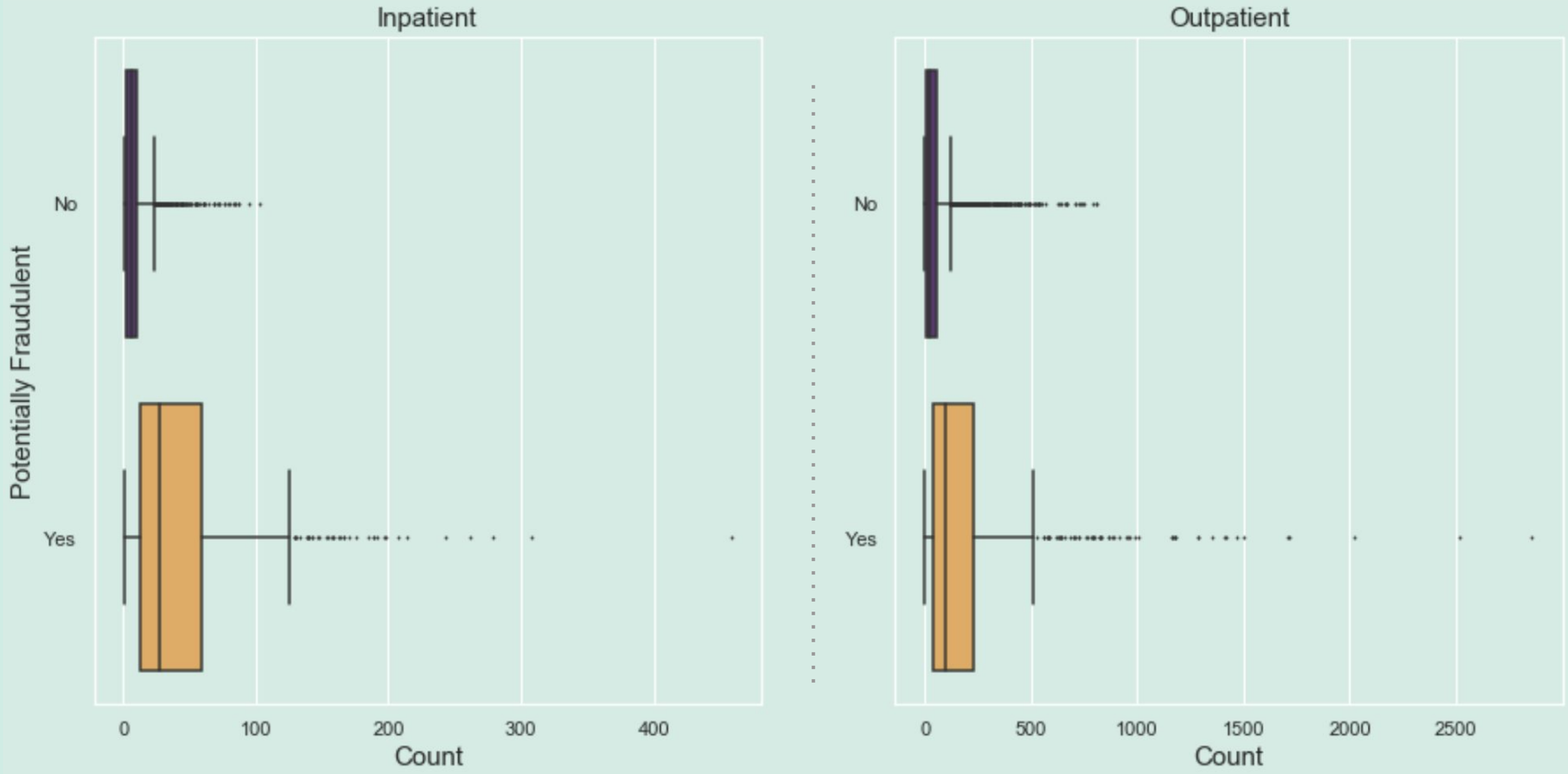
4 .CSV FILES



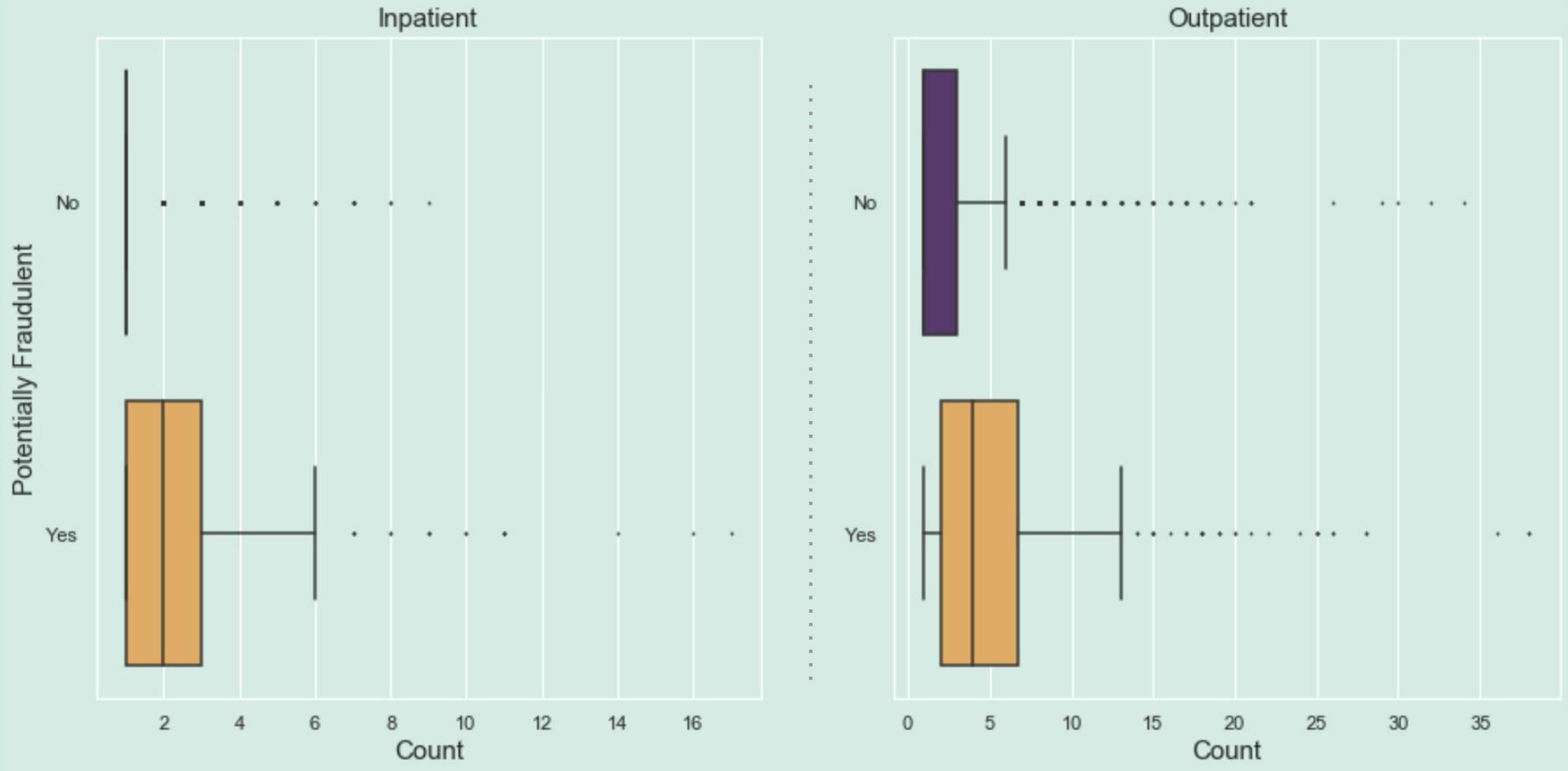
550,000+ CLAIMS

56 FEATURES

EDA: Number of Unique Patients Per Provider



EDA: Number of Unique States Per Provider



Assumptions

01

1C0
1010
01

ENCODED FEATURES
FROM 1 AND 2 TO 1 AND 0

02

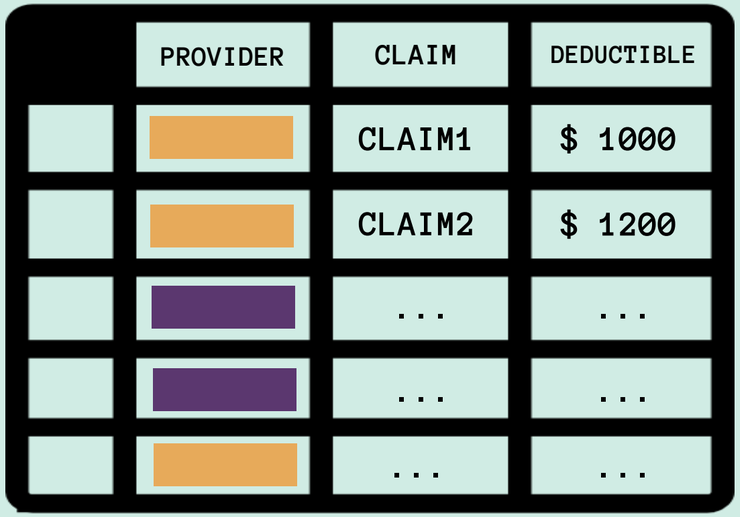
DUPLICATED CLAIMS HAVE
THE SAME SET OF PROCEDURE
AND DIAGNOSIS CODES

03

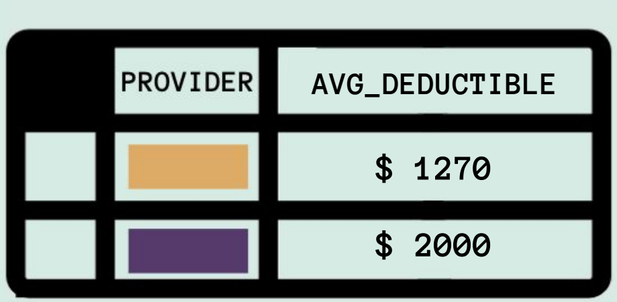
N/A

MISSING VALUES
ARE ALL MNAR

Pre-processing: Providers Dataframe



	PROVIDER	CLAIM	DEDUCTIBLE
		CLAIM1	$ 1000
		CLAIM2	$ 1200
		...	...
		...	...
		...	...



	PROVIDER	AVG_DEDUCTIBLE
		$ 1270
		$ 2000

550,000+ claims
56 features

5,410 providers
87 features

Model Scoring Metrics

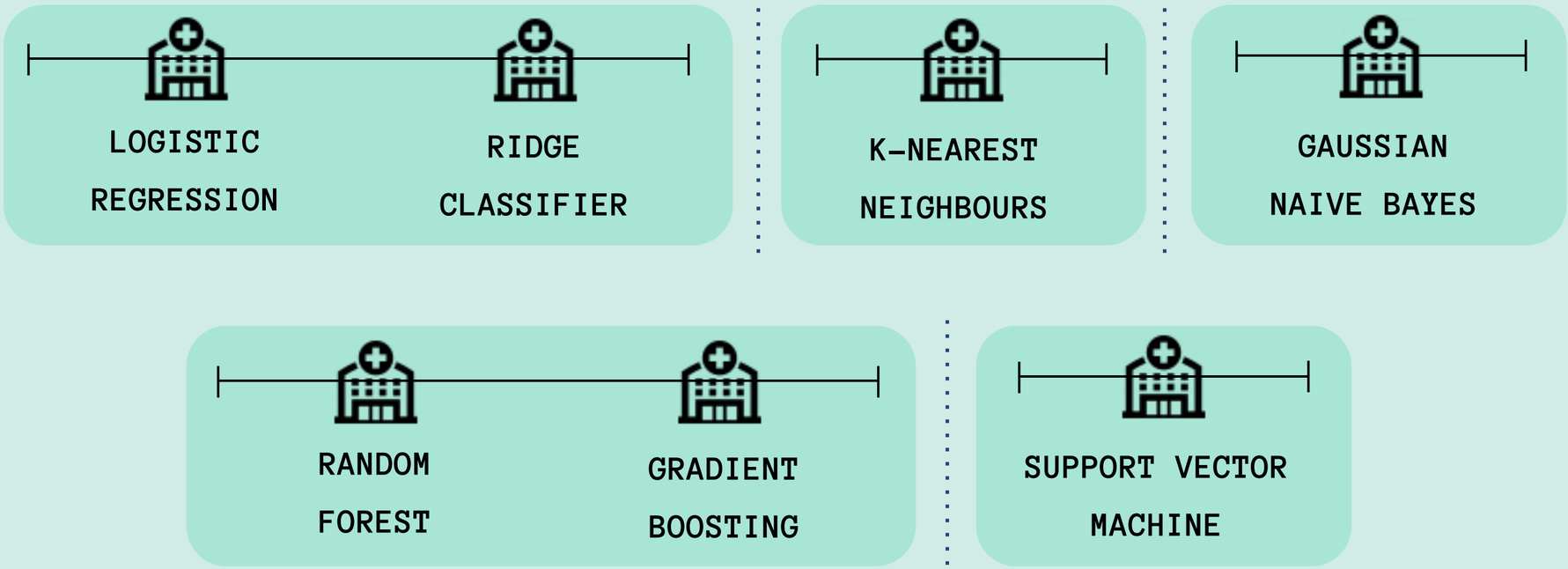
RECALL SCORE

Minimize false negatives

MONEY SAVED

Balance false negatives
and false positives
to lower review costs

Classification Models



Linear Models

- Logistic Regression
  - L1-norm penalty (Lasso)
  - Reduced features from 87 to 13
  - Recall score: Train 0.9181

    Test 0.9145

- Ridge Classifier
  - Improved positive features
  - Recall score: Train 0.9294

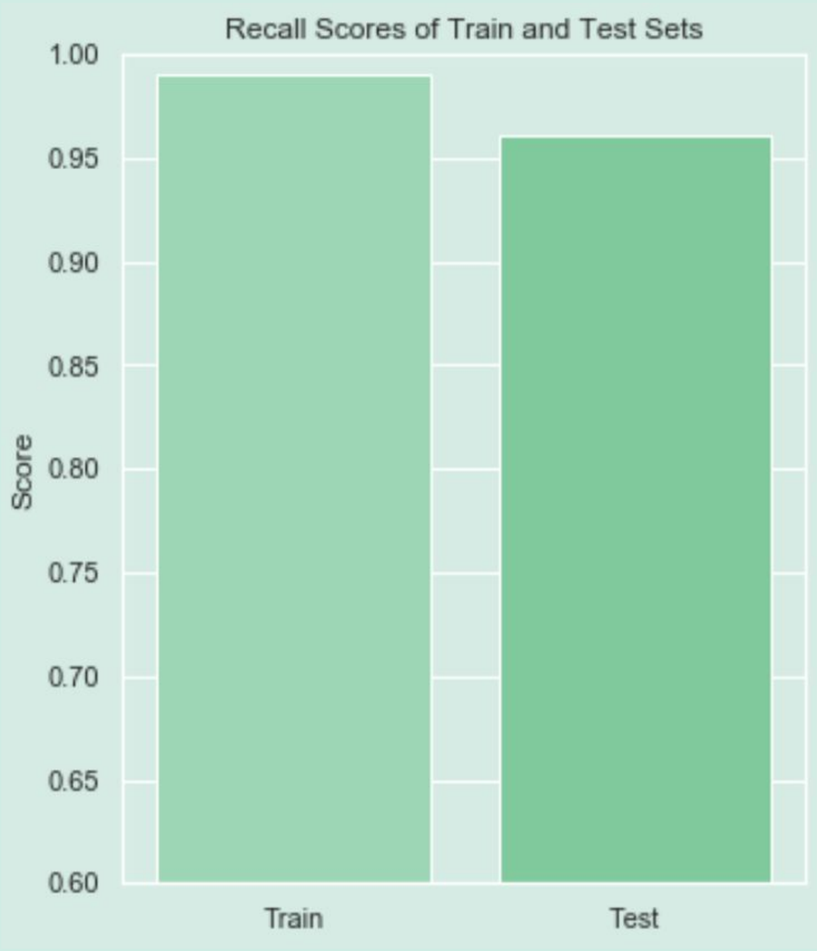
    Test 0.9145

K-Nearest Neighbors

- Highly interpretable model
- Fewest false negatives

  but many false positives
- Recall score: Train 0.9898

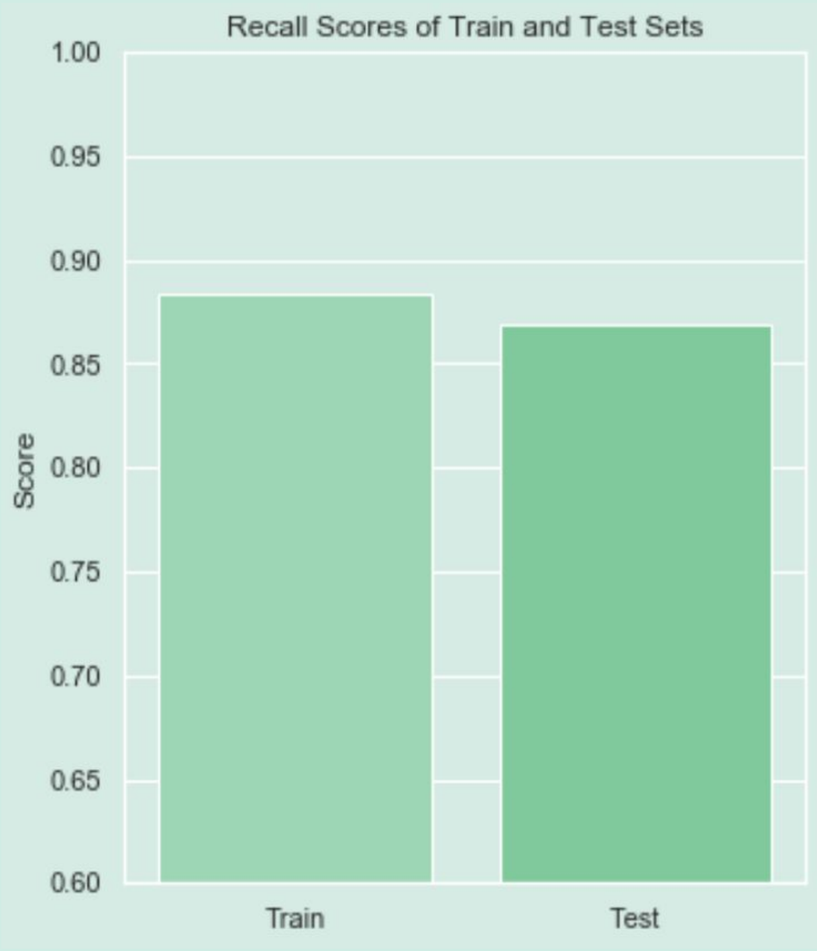
  Test 0.9605



Gaussian Naive Bayes

- Assumes features are independent
- Minimal parameter tuning required
- Recall score: Train 0.8842

  Test 0.8684



Tree-Based Models

- Capture non-linear relationships
- Show feature importances

- Random Forest
  - Recall score: Train 0.9096

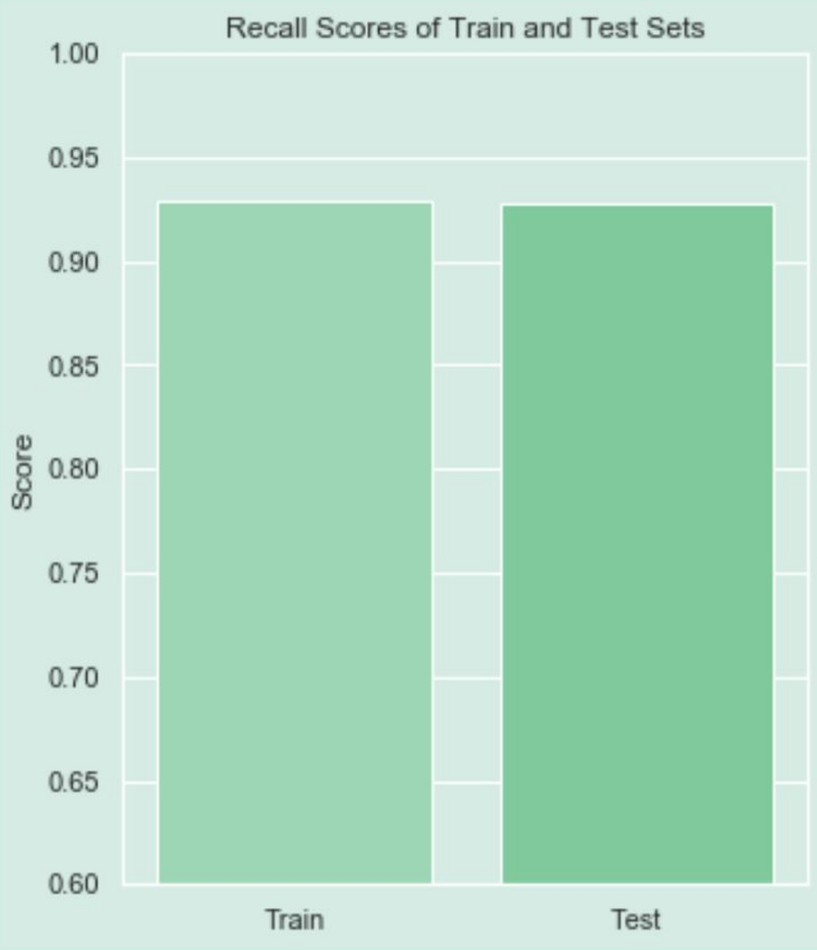
    Test 0.9013
- Gradient Boosting Decision Trees
  - Complex but computationally expensive
  - Recall score: Train 0.9265

    Test 0.9211

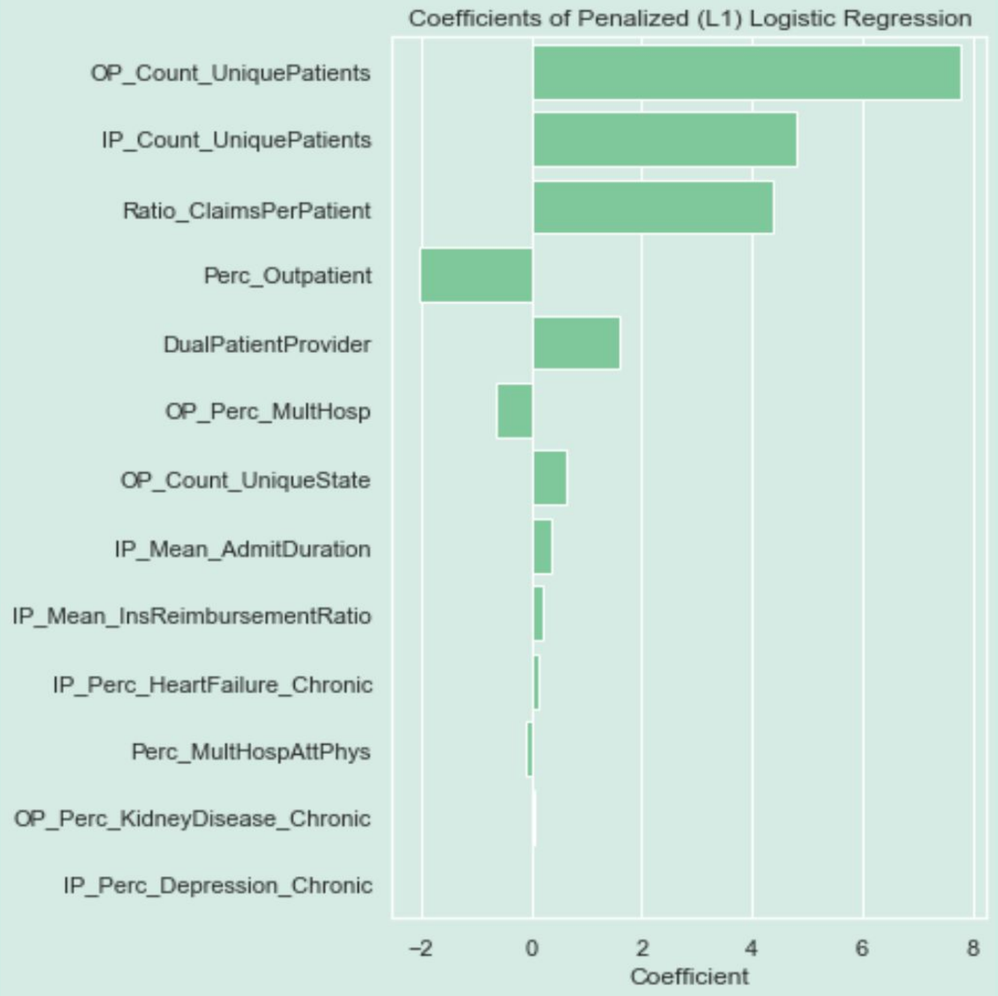
Support Vector Machine

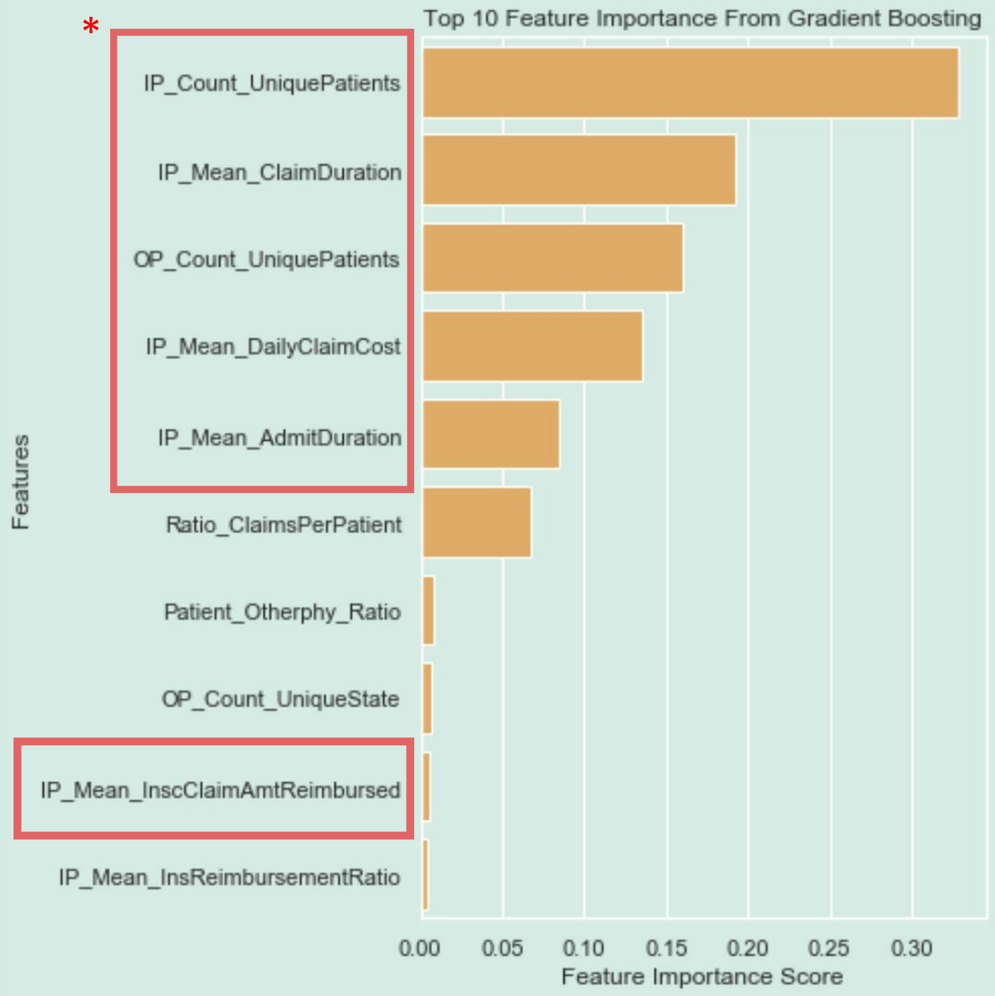
- Computationally less expensive
- Expands feature space
- Avoids bias/variance tradeoff
- Recall score: Train 0.9294

  Test 0.9276



Modeling Iterations 2 and 3





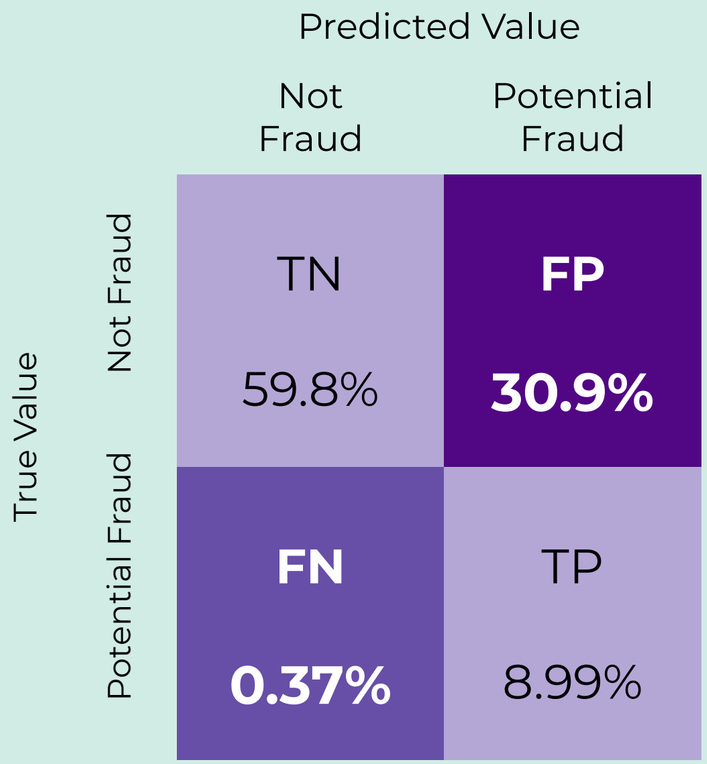
* *Boxed features also appear in the top 10 feature importances for Random Forest*

Production Model Selection

Saves the most money

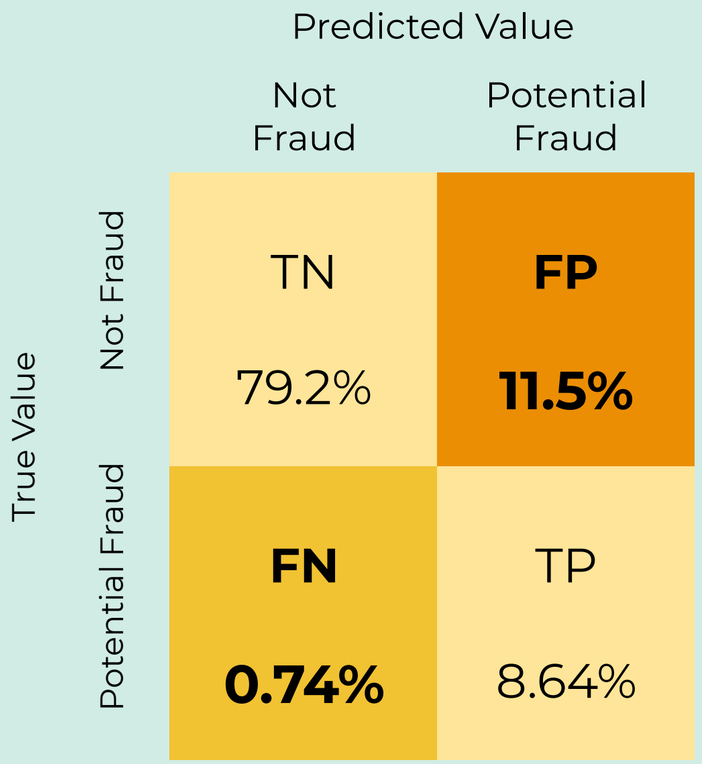
Highest recall score

Confusion Matrices



K-Nearest Neighbors

Saves $86,000*

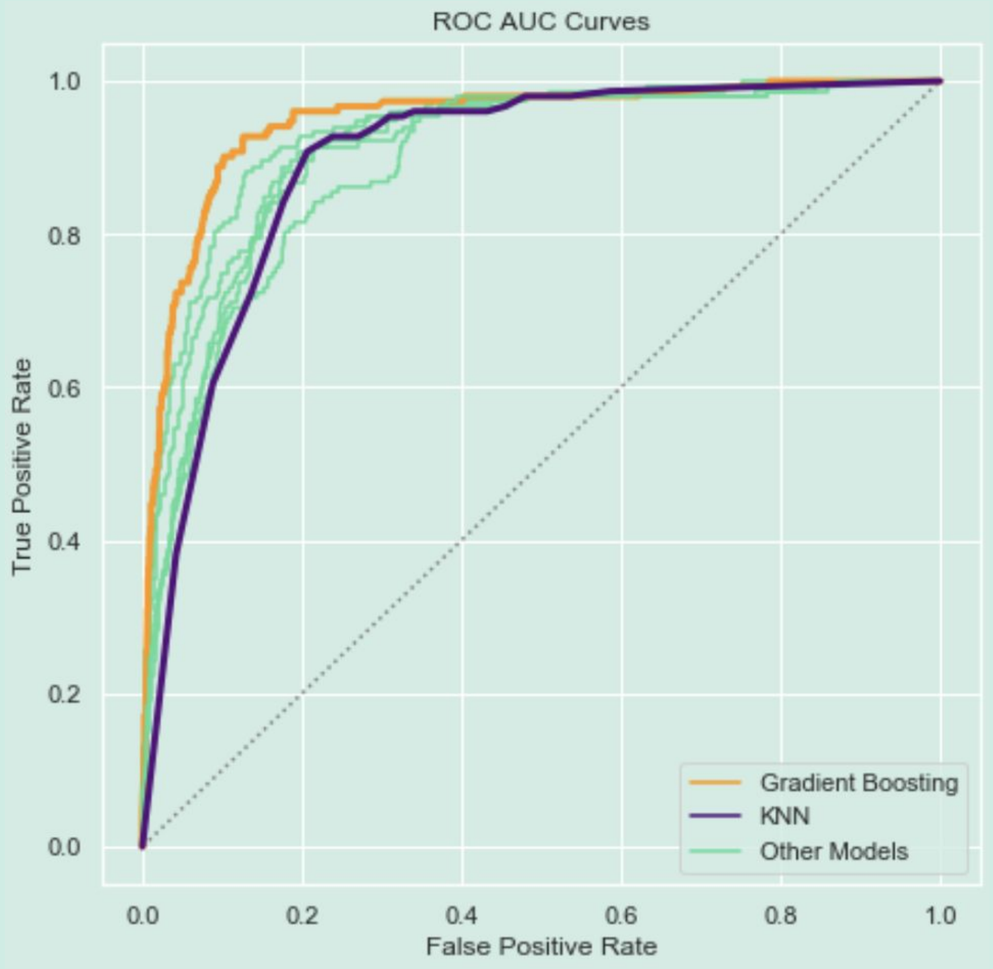


Gradient Boosting

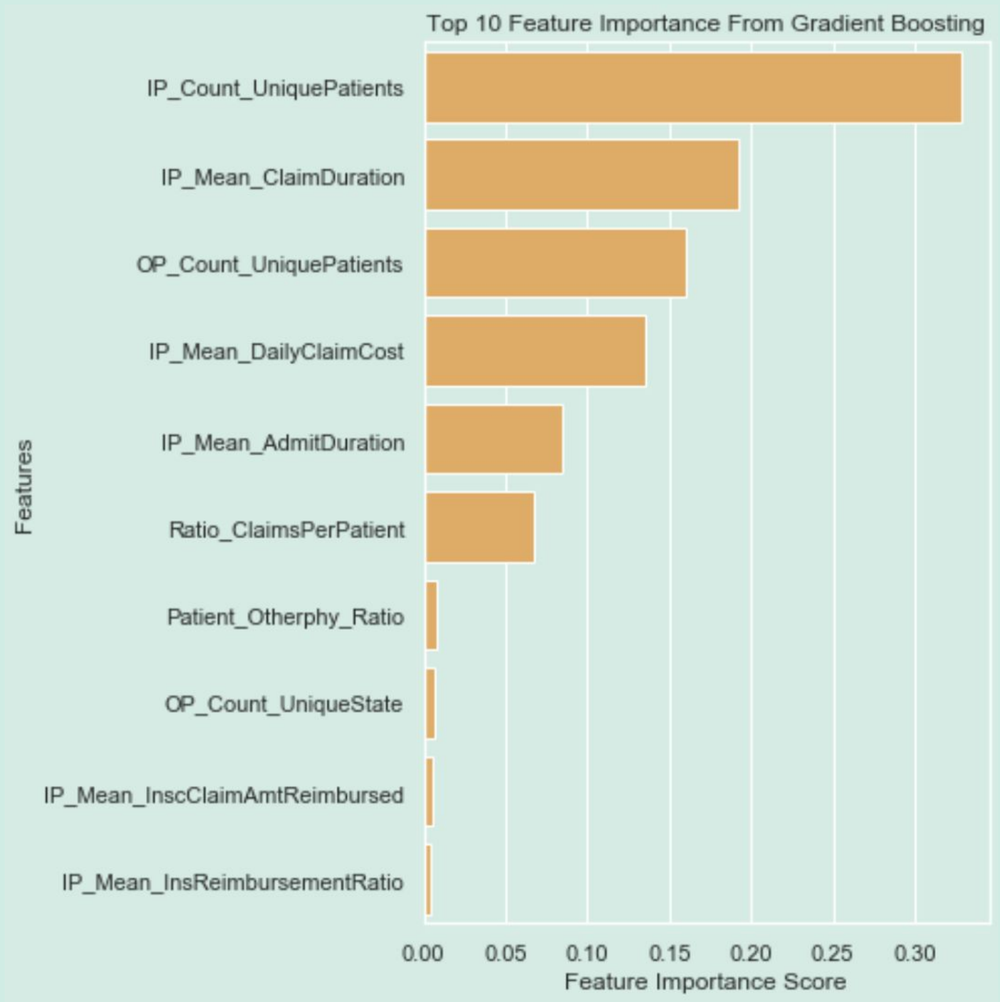
Saves $100,000*

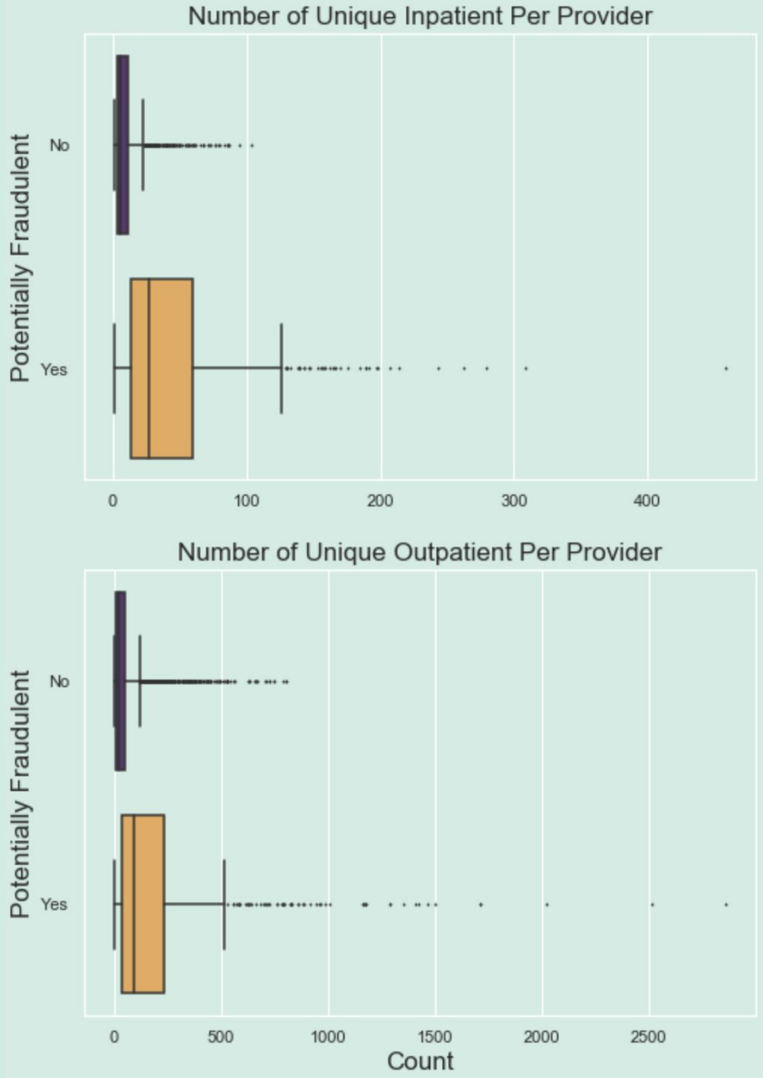
** Savings are calculated with confusion matrix value counts, using the formula:*
(TP) x ($1,000 Average Claim Cost) - (FP + TP) x ($28 Hourly Wage) x (3 Hours) - (FN) x ($1000 Average Claim Cost)

ROC/AUC Curve

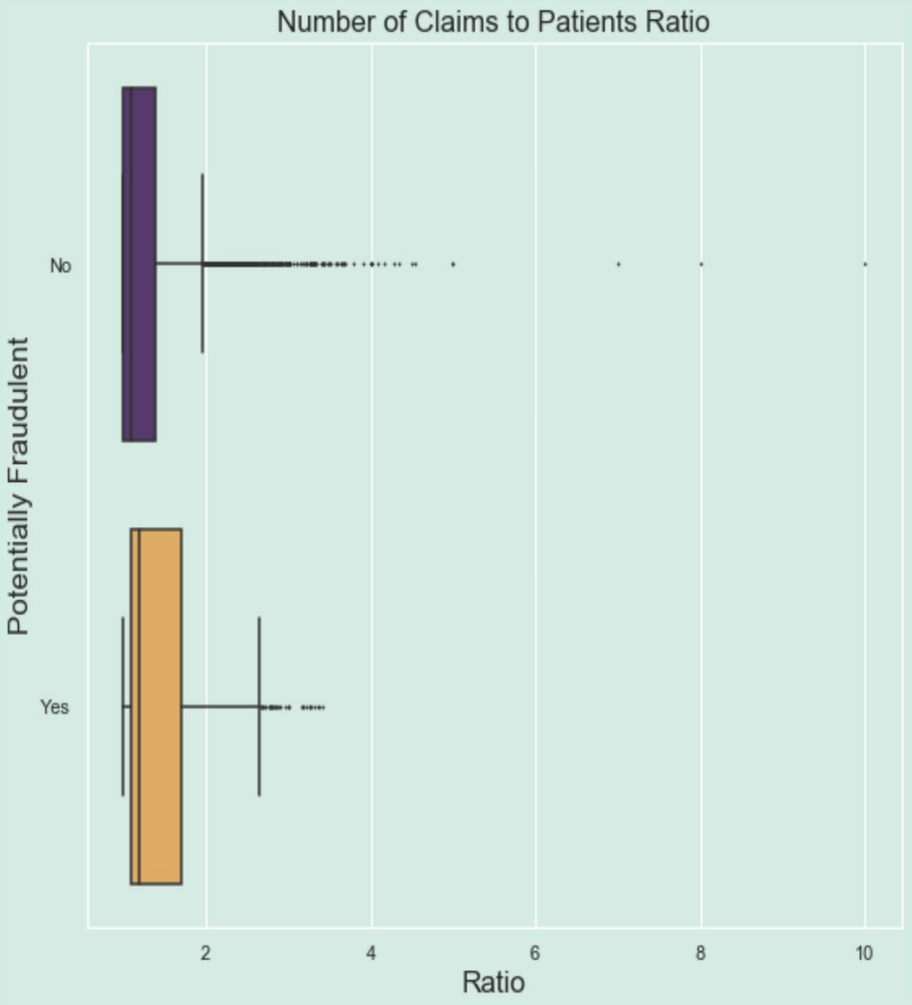


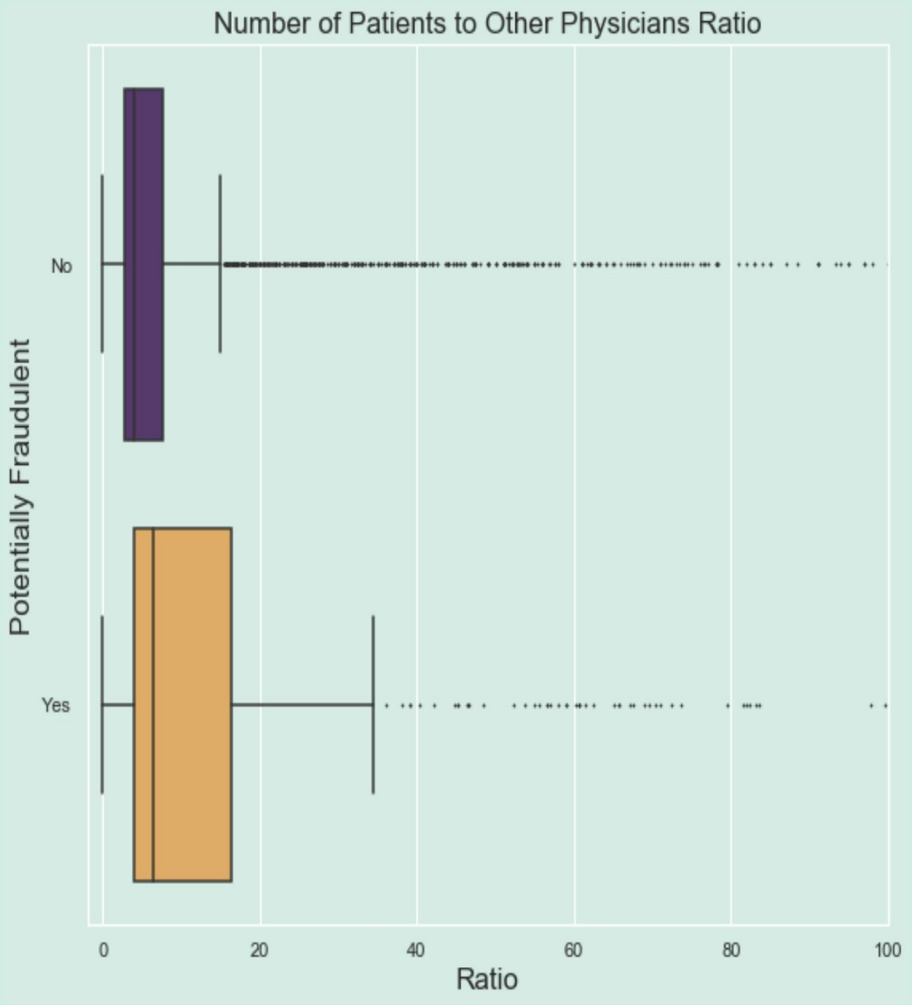
Insights: Feature Importances



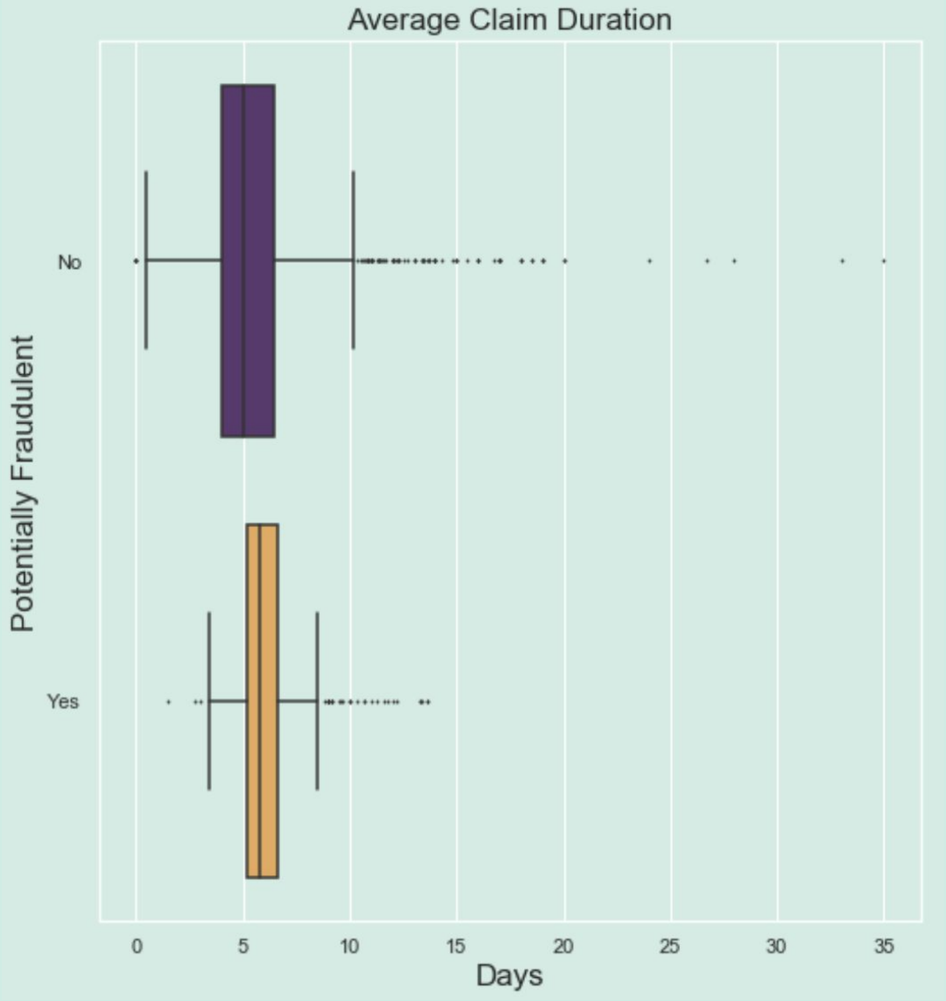


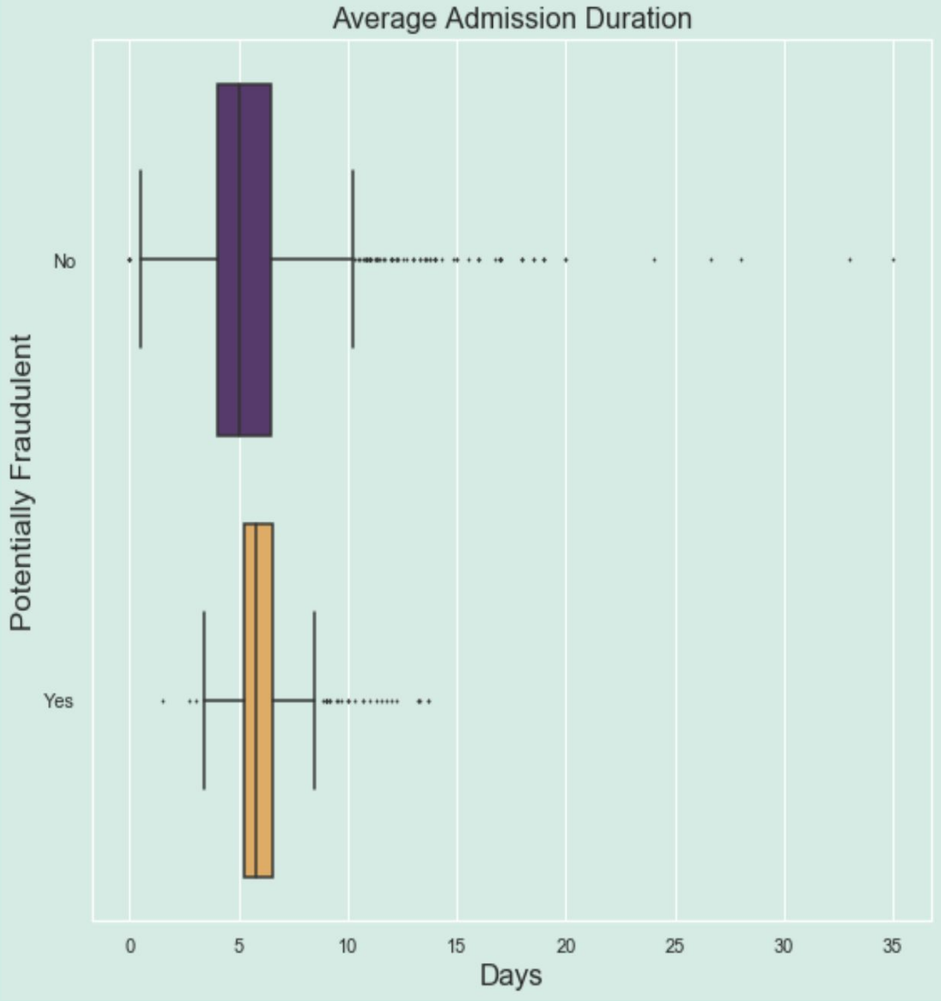
Insights: Patients



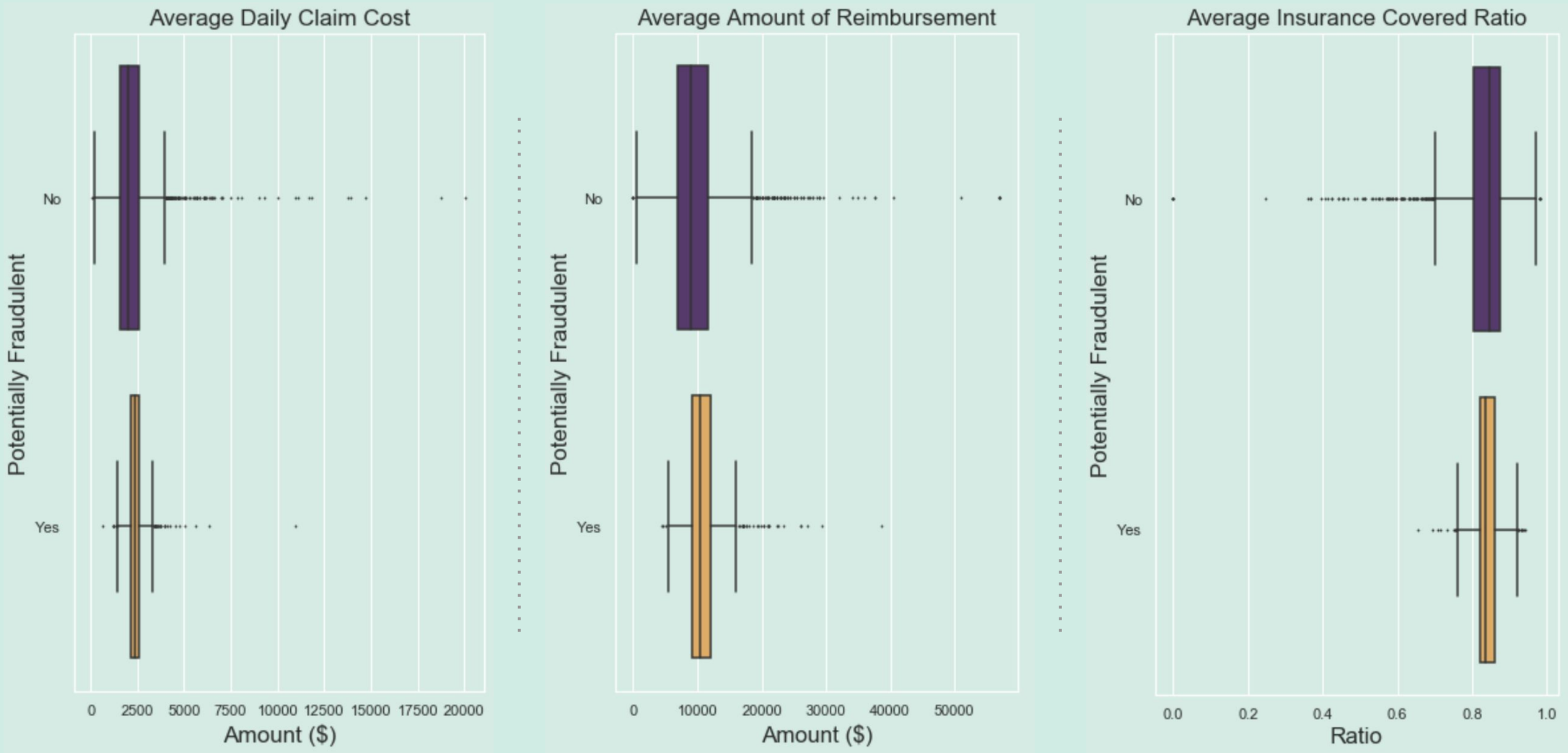


Insights: Duration



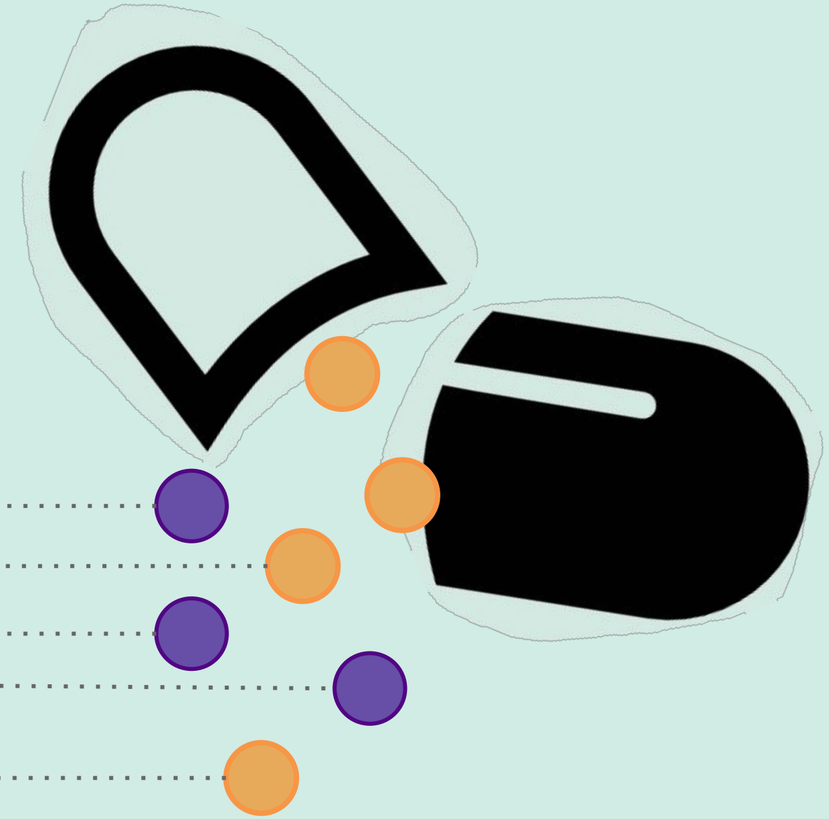


Insights: Cost

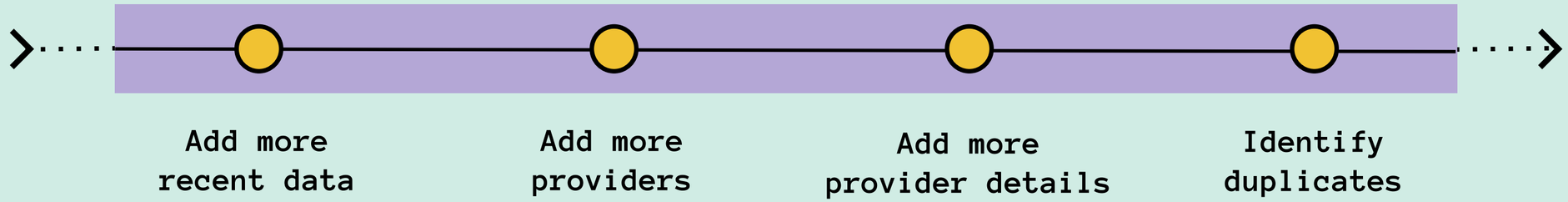


Further Insights

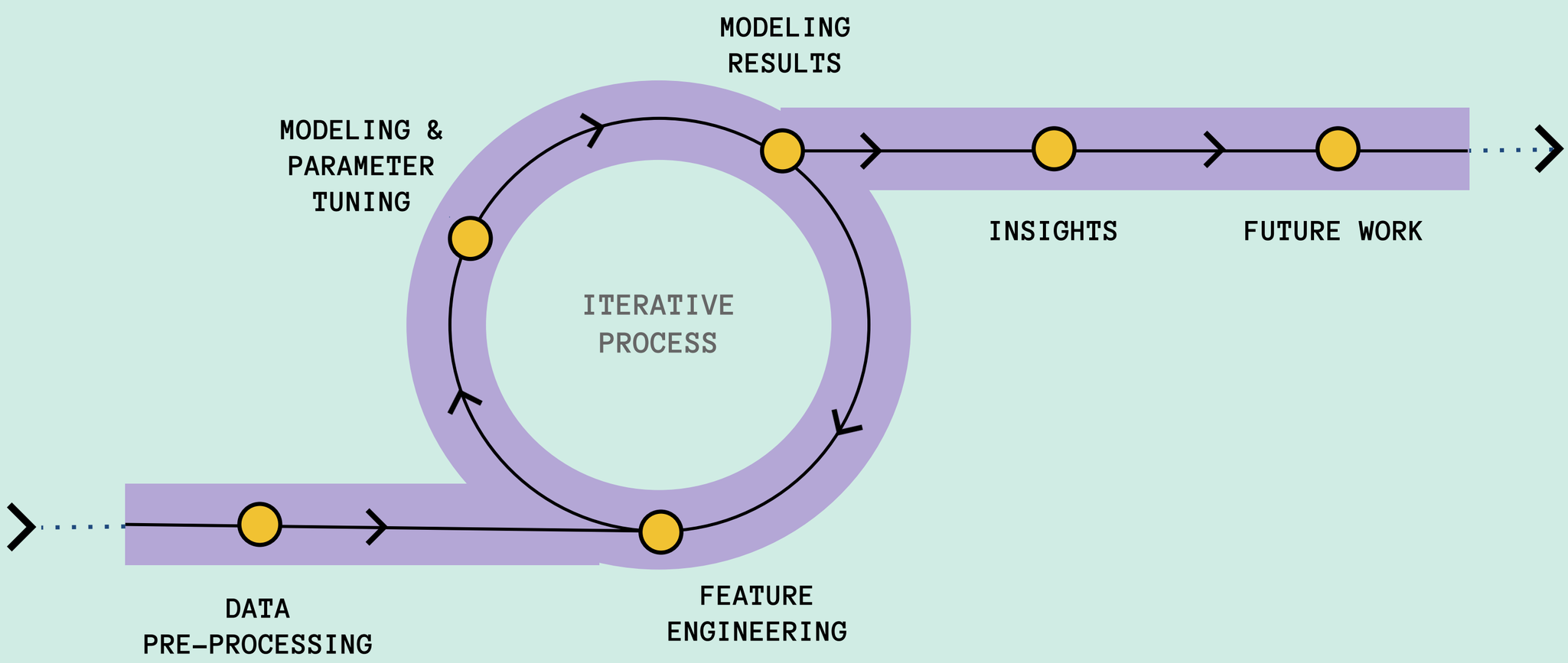
- Limit number of contracts with providers offering inpatient services
- Cancel contracts with providers who demonstrate fraudulent behaviour
- Regularly audit providers
- Recruit doctors to review claims
- Incentivize patients to review claims



Future Work


Add more recent data
Add more providers
Add more provider details
Identify duplicates

Summary


MODELING
RESULTS
MODELING &
PARAMETER
TUNING
INSIGHTS
FUTURE WORK
ITERATIVE
PROCESS
DATA
PRE-PROCESSING
FEATURE
ENGINEERING

Thank you!

LUCAS KIM

/kiml1

RYAN PARK

/ryanhpark

SITA THOMAS

/sitathomas