Health Insurance Fraud Detection

By Tania Ghosh and Emily Wang

Background

- The National Health Care Anti-Fraud Association (NHCAA) estimates that the financial losses due to health care fraud are in the **tens of billions of dollars** each year
- Healthcare fraud translates to higher premiums, out-of-pocket expenses, and reduced benefits or coverage for consumers, as well as higher costs for employers providing benefits to employees

Objectives

1. Identify health insurance providers with potentially fraudulent claims
2. Extract the most important features in predicting fraud
3. Translate our findings into business cost savings

Our Machine Learning Approach

EDA & Pre-Processing → Feature Engineering → Modeling → Extract Insights & Analysis

EDA & Pre-Processing

- ➔ Gather initial insights
- ➔ Consolidate multiple datasets
- ➔ Impute missing values
- ➔ Dummify/encode categorical attributes

Feature Engineering

- ➔ Add and modify features as part of an iterative process

Modeling

- ➔ Evaluate and compare various binary classification models
- ➔ Tune hyperparameters to further refine the best models

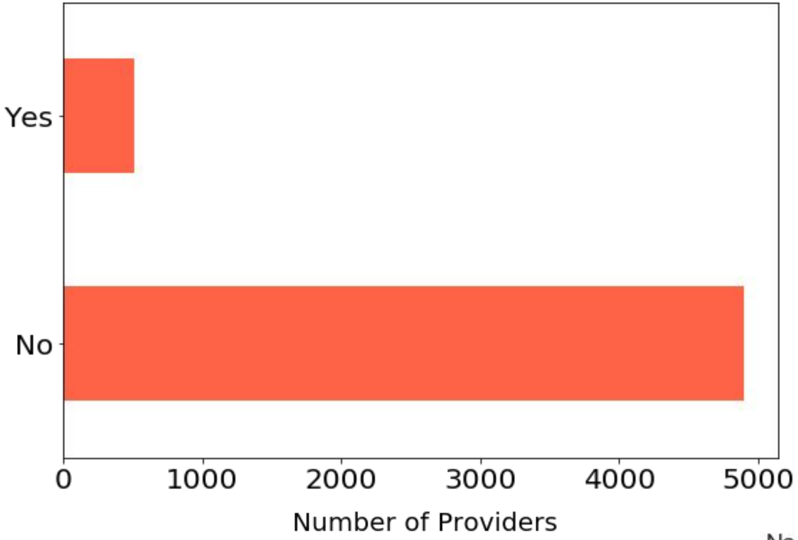
Extract Insights & Analysis

- ➔ Evaluate feature importance to extract key insights
- ➔ Convert our findings into business recommendations for cost savings

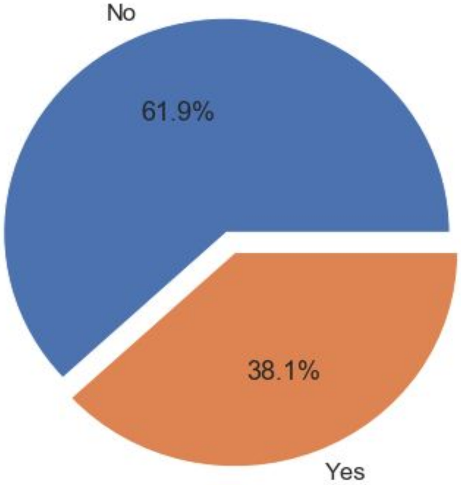
Data Overview

- ➢ Beneficiaries (138,556)
- ➢ Inpatients (40,474)
- ➢ Outpatients (517,737)
- ➢ Providers (5,410)

The records span one full year of claims(2009).



Proportion of Fraudulent Providers

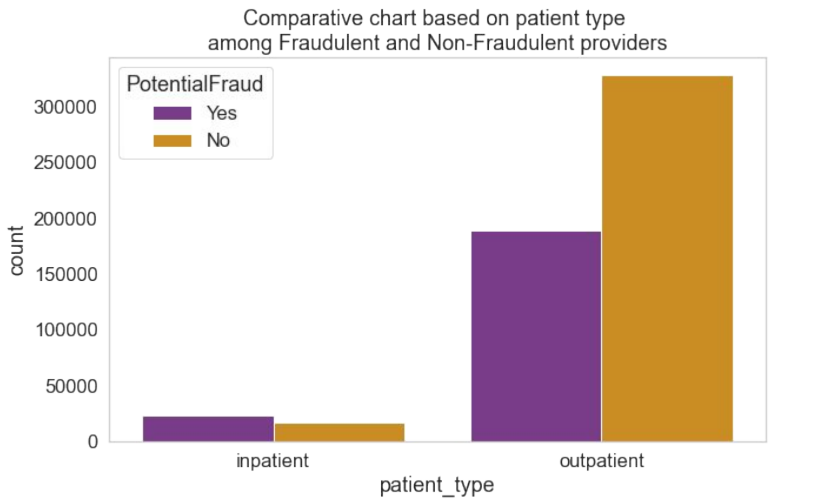


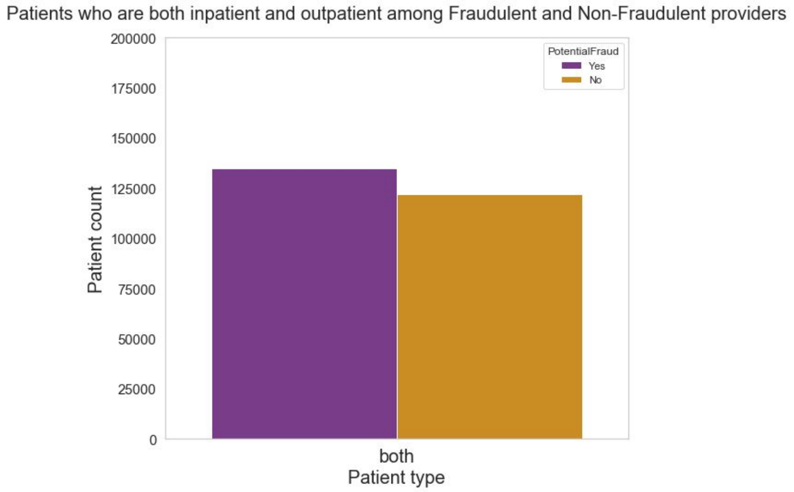
Proportion of Fraudulent Claims

EDA

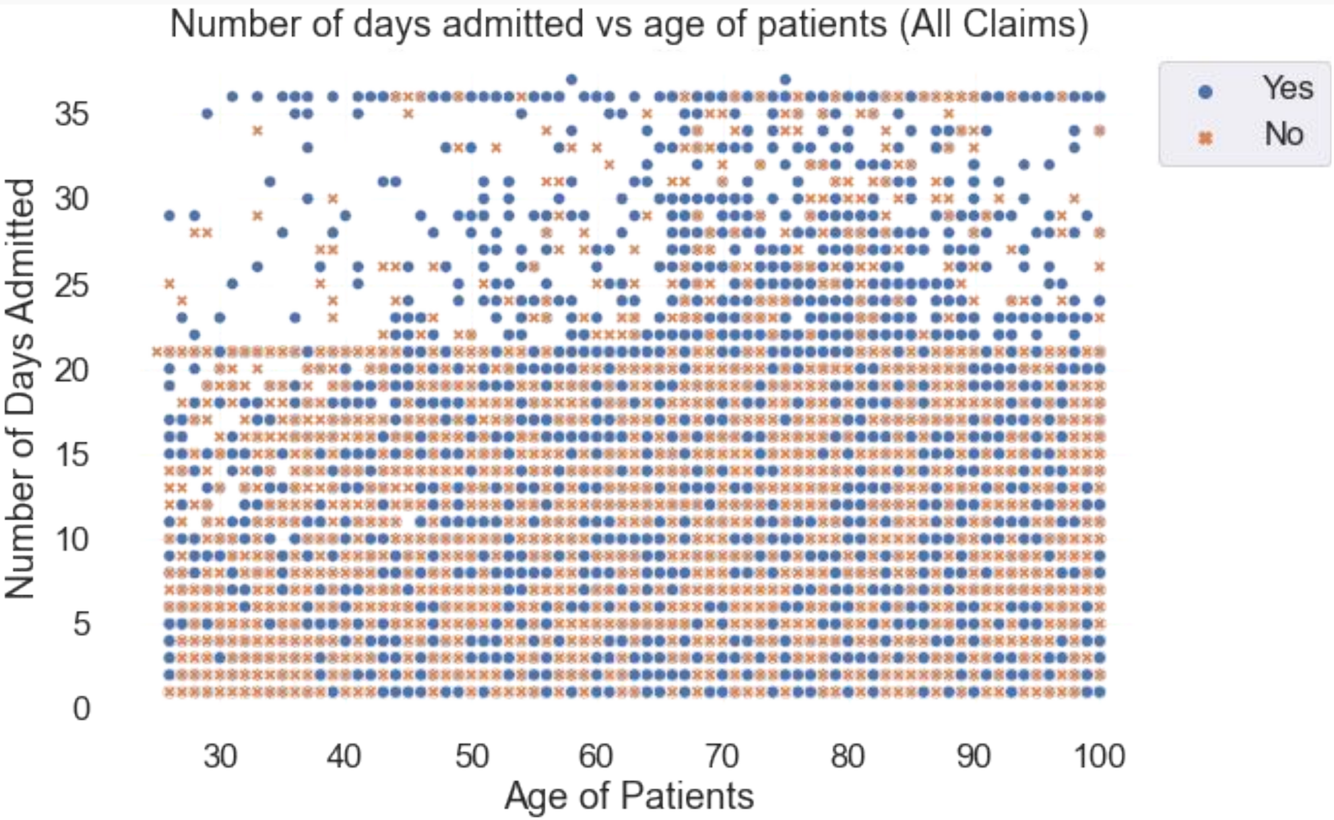
Inpatient vs Outpatient breakdown

	Inpatient	Outpatient	Both
Fraud	23402	189394	134682
No Fraud	17072	328343	121782



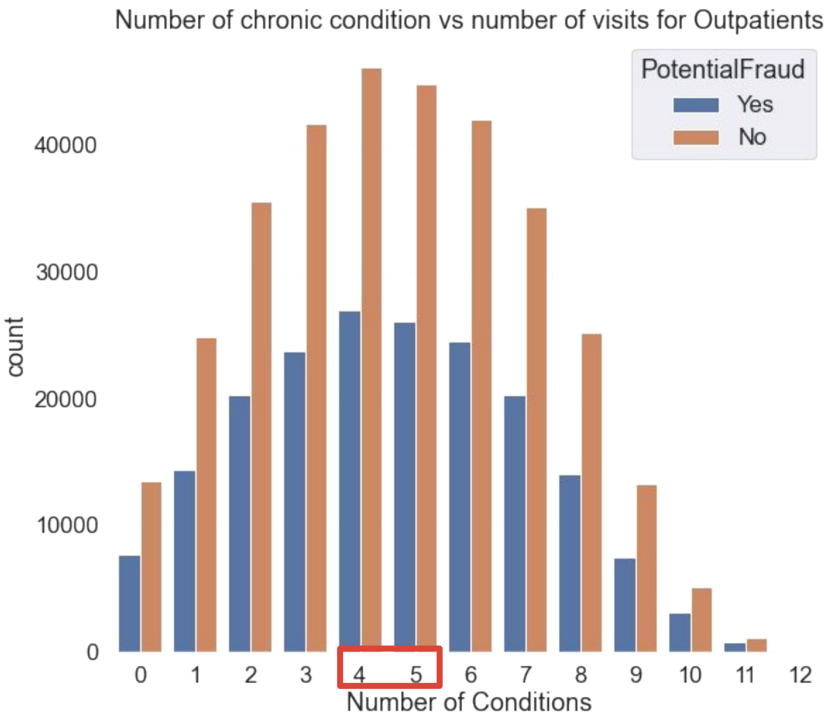


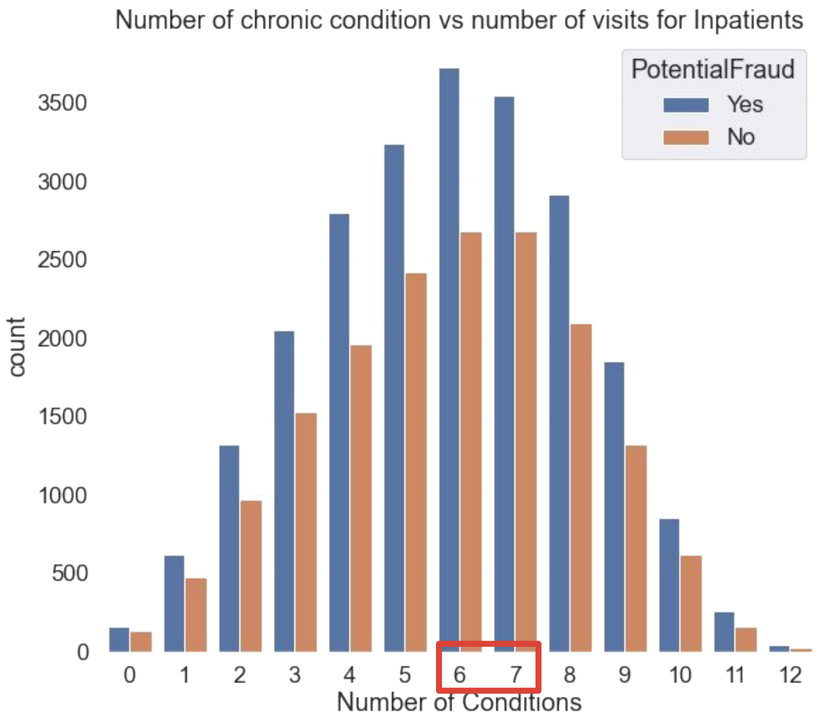
EDA: Number of Admitted Days



The fraudulent claims are mostly for patients admitted for longer hospital stays

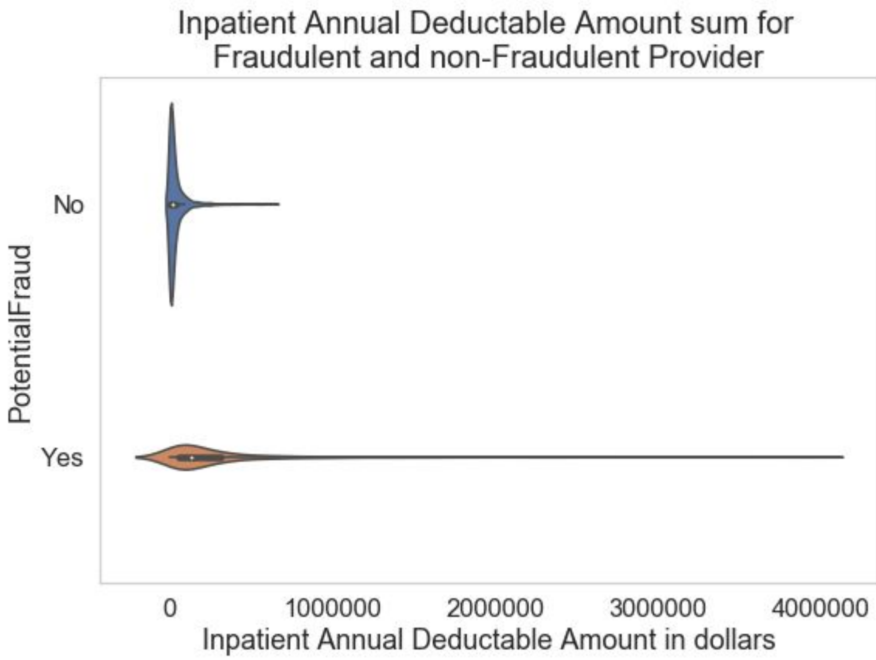
EDA: Factors that Influence Number of Visits

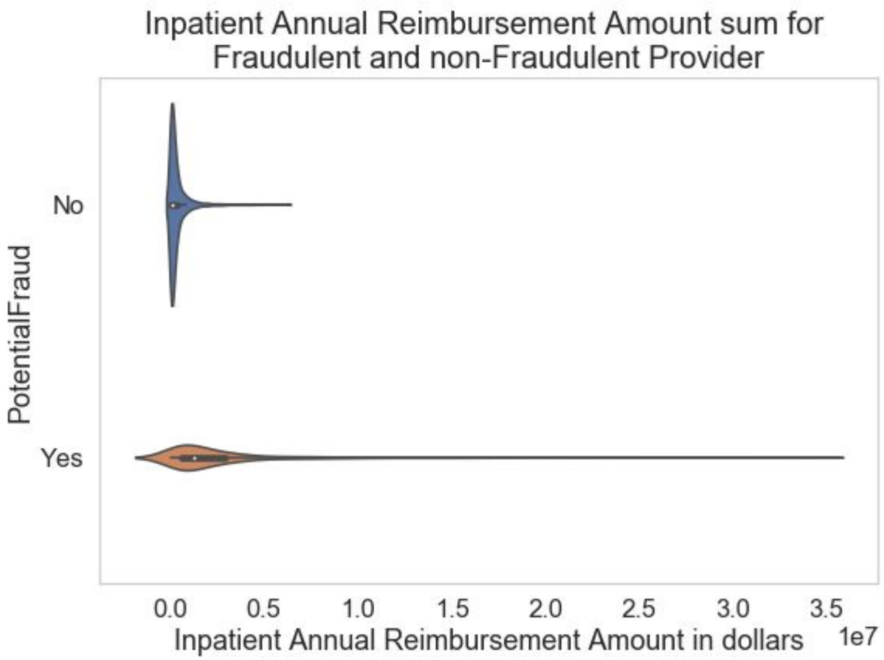




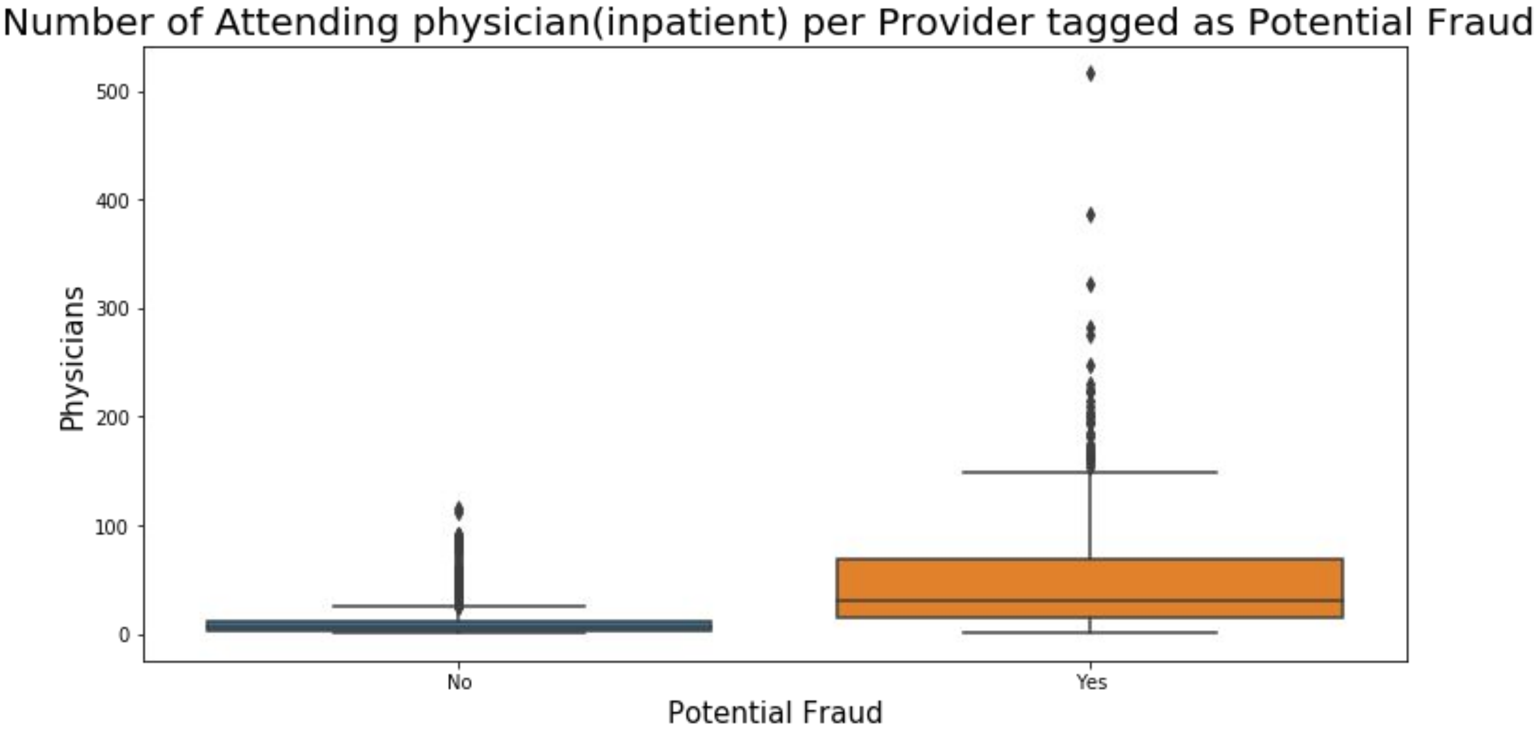
Chronic conditions: heart disease, stroke, ischemic heart disease, cancer, diabetes, depression, renal & kidney disease, Alzheimer's, obstructive pulmonary, osteoporosis, rheumatoid arthritis

EDA: Inpatient Reimbursements & Deductibles





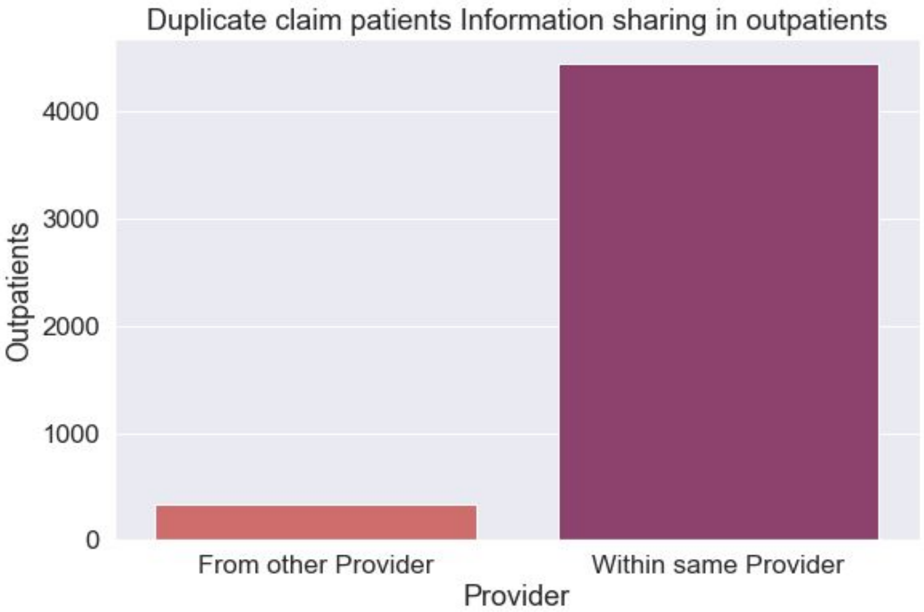
EDA: How big is the providers network?

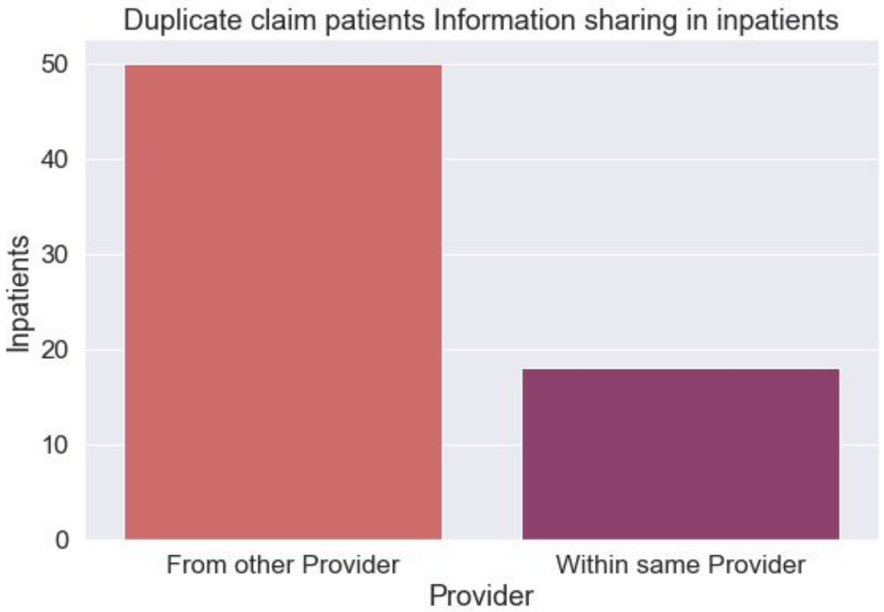


Providers with the widest network of doctors are mostly fraudulent

EDA: Duplicate Claims

Interestingly, inpatient providers take information from other providers more often than duplicating their own claims. It is the opposite for outpatients.





Feature Engineering

Patient Count - per provider

Mean Age - of patients per provider

State count - number of states are connected with each provider

Patient type - inpatient, outpatient, both

Phy_count - Physician count per provider

No_phy - count of cases with no physician for each provider

Chronic mean - mean of chronic condition are taken care by each provider

Days_admitted - number of days admitted

claim _count = number of claims per provider

Duplicate claims -count per provider

Patient duplicate claims - number of patients involved in duplicate claims per provider

Mean Revenue per day

Mean Coverage

Mean Annual Amount - Inpatient and outpatient Reimbursement and Deductible amount

Mean Total Amount charged

Balancing the Data

We experimented with 5 different approaches:

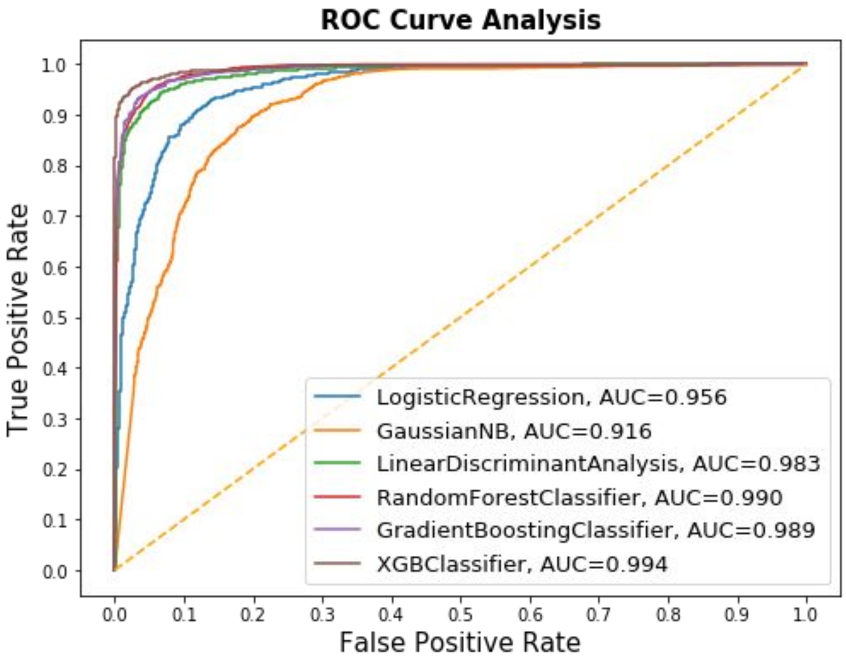
- **"Balanced"** parameter weight while applying machine learning models
- **Undersampling** - Edited Nearest Neighbor
- **Undersampling** - Random Undersampling
- **Oversampling** - Synthetic Minority Oversampling Technique (SMOTE)
- **Oversampling** - Random Oversampling

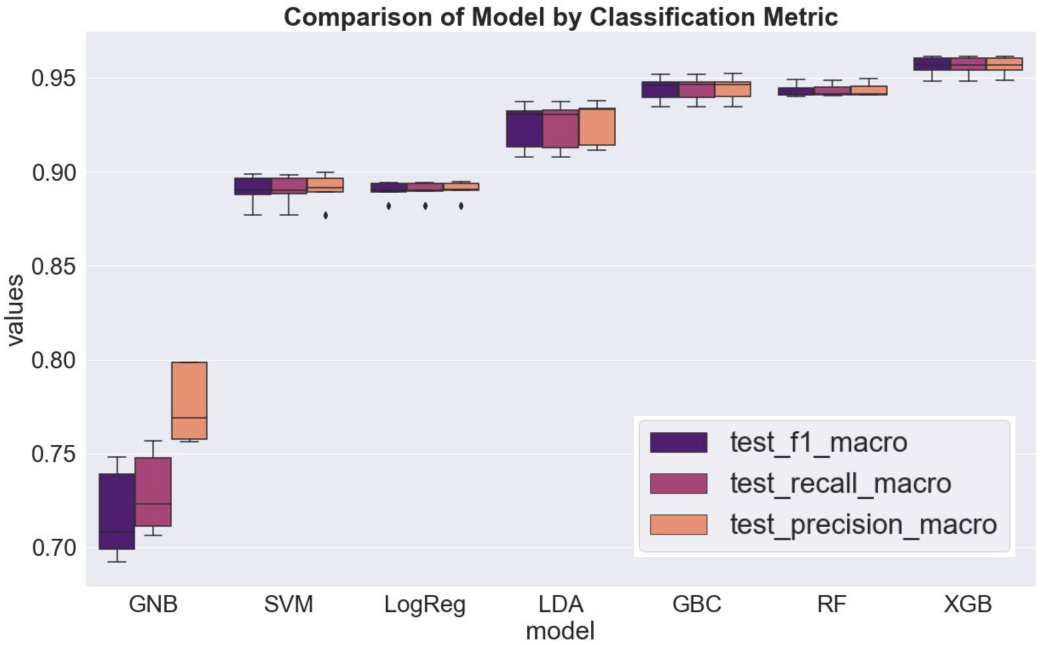
Machine Learning Models

Binary Classification

- Logistic Regression
- Linear Discriminant Analysis
- Gaussian Naive Bayes
- Support Vector Machine
- Random Forest
- Gradient Boosting
- XGBoost

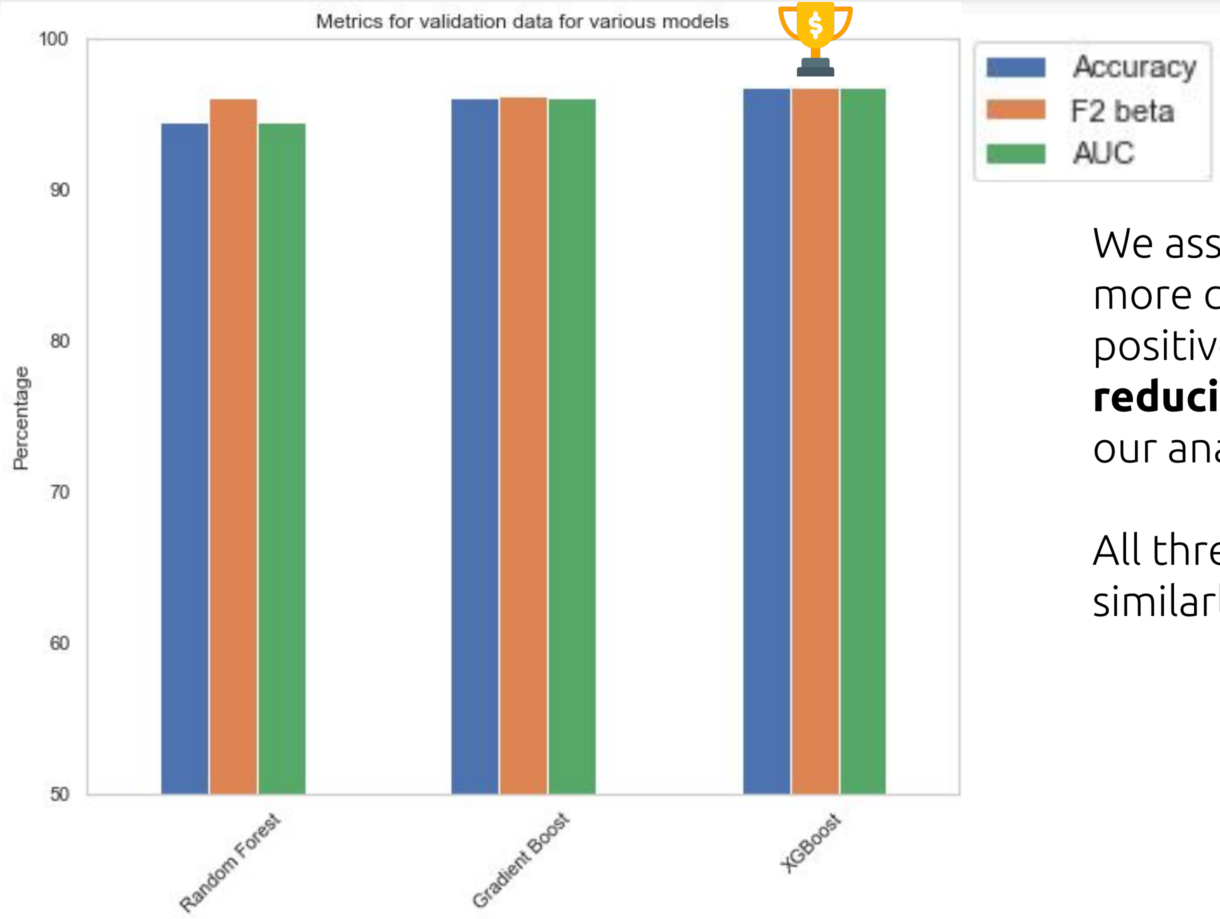
Model Performance





Random Forest, **Gradient Boost** and **XGBoost classifier** were selected for further hyperparameter tuning

Top 3 Models after Hyperparameter Tuning



We assume false negatives are more costly than false positives and thus prioritized **reducing false negatives** in our analysis.

All three models performed similarly well.

Best Model - XGBoost

Accuracy score - 96.7%

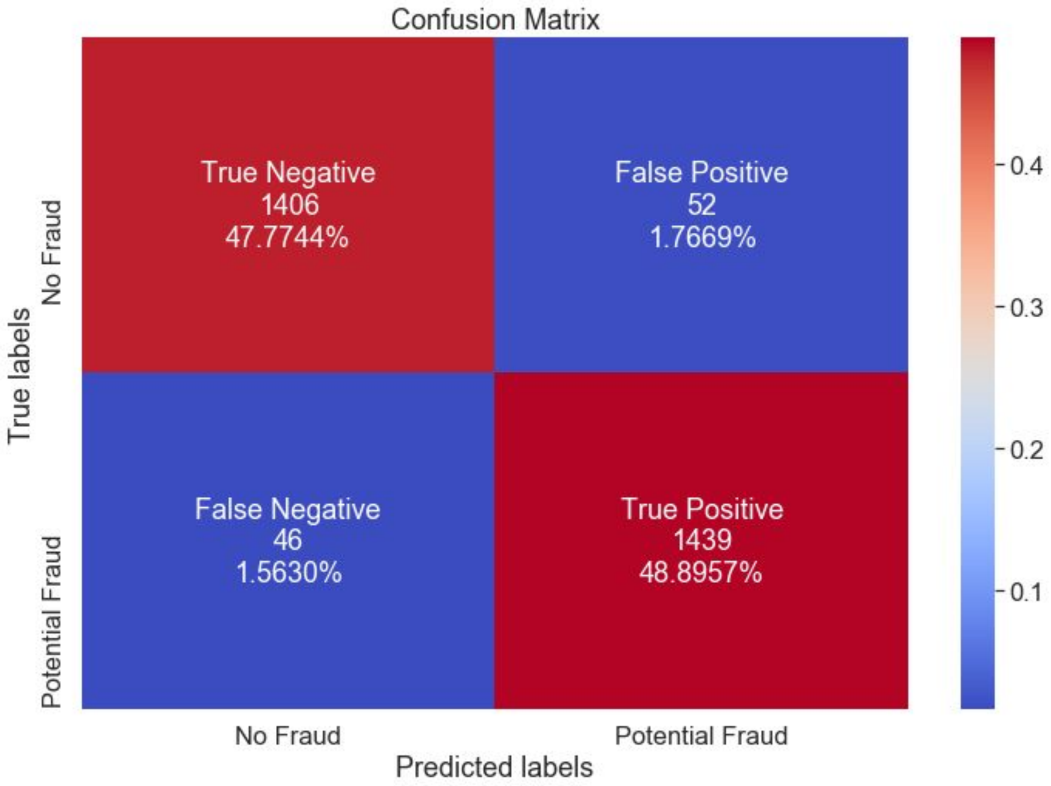
Precision - 96.5%

Recall - 96.9%

F1 - 96.7%

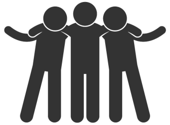
F2 beta - 96.8%

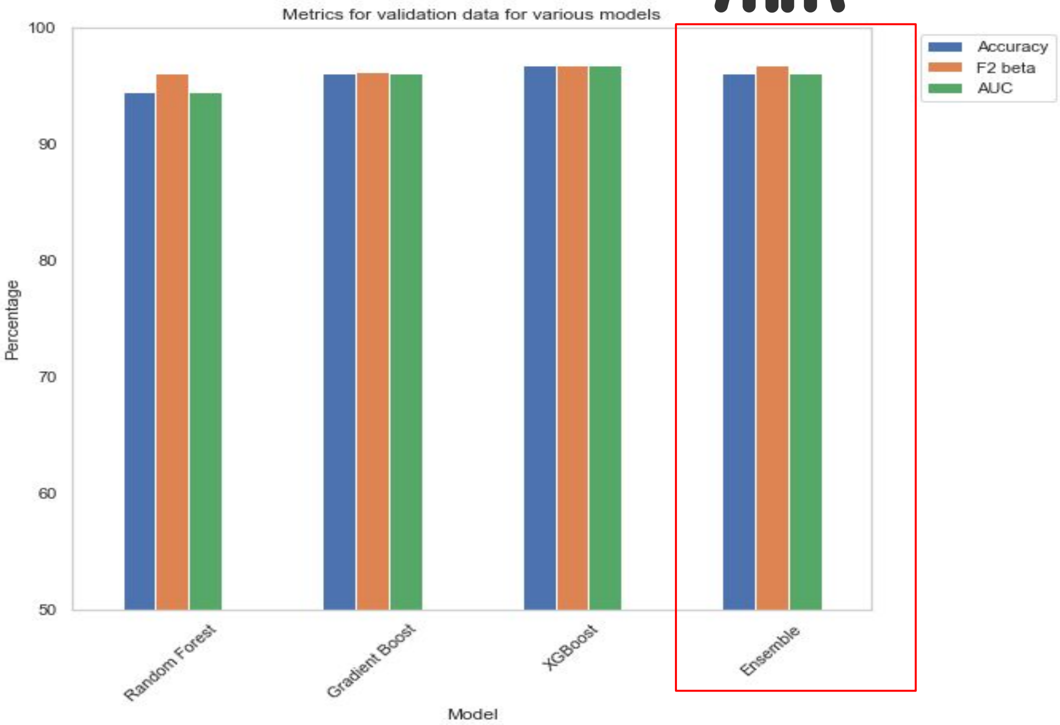
AUC - 96.7%



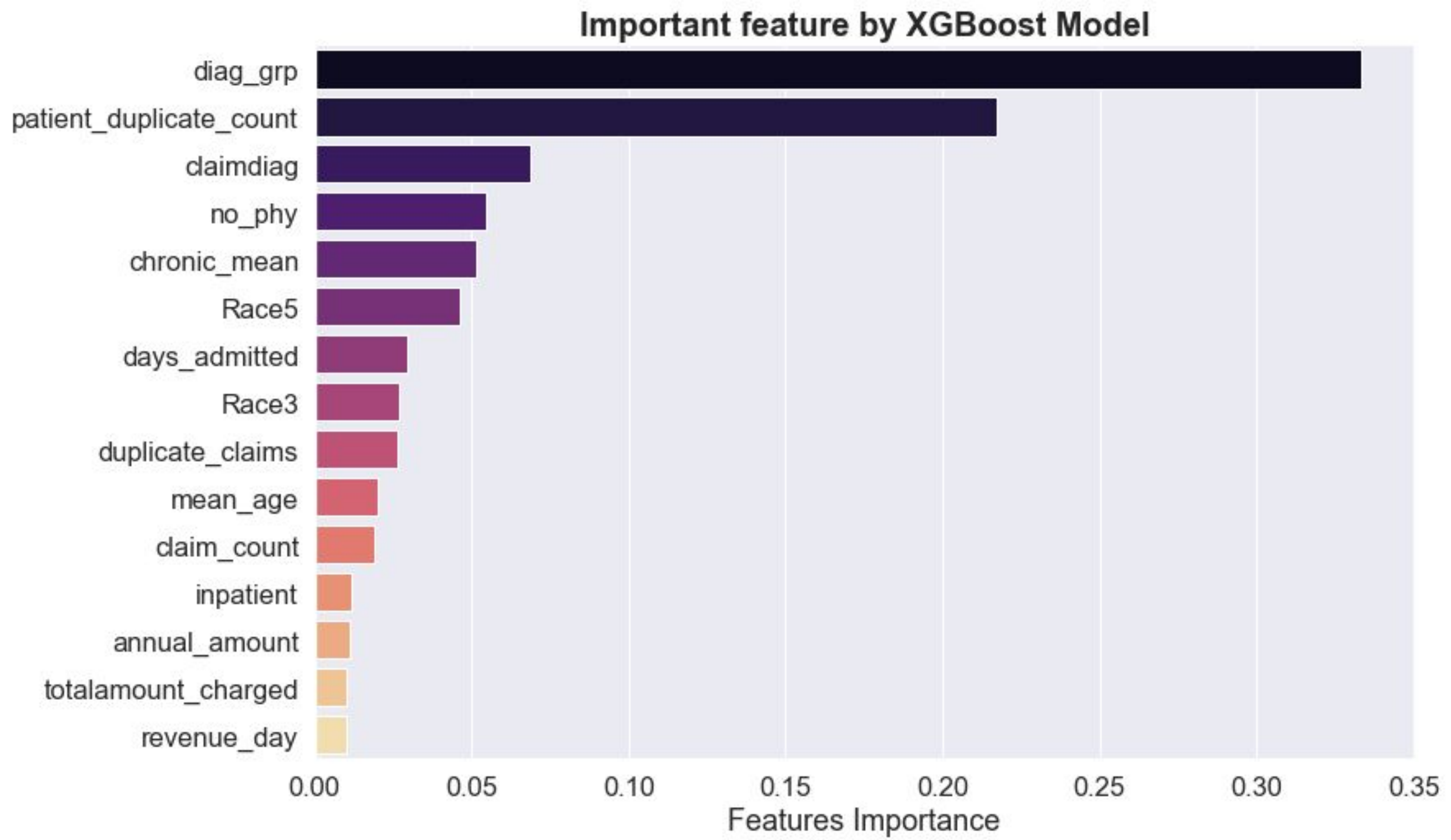
Ensembling

Using a simple Voting Classifier, we combined our top 3 models in an effort to further improve performance

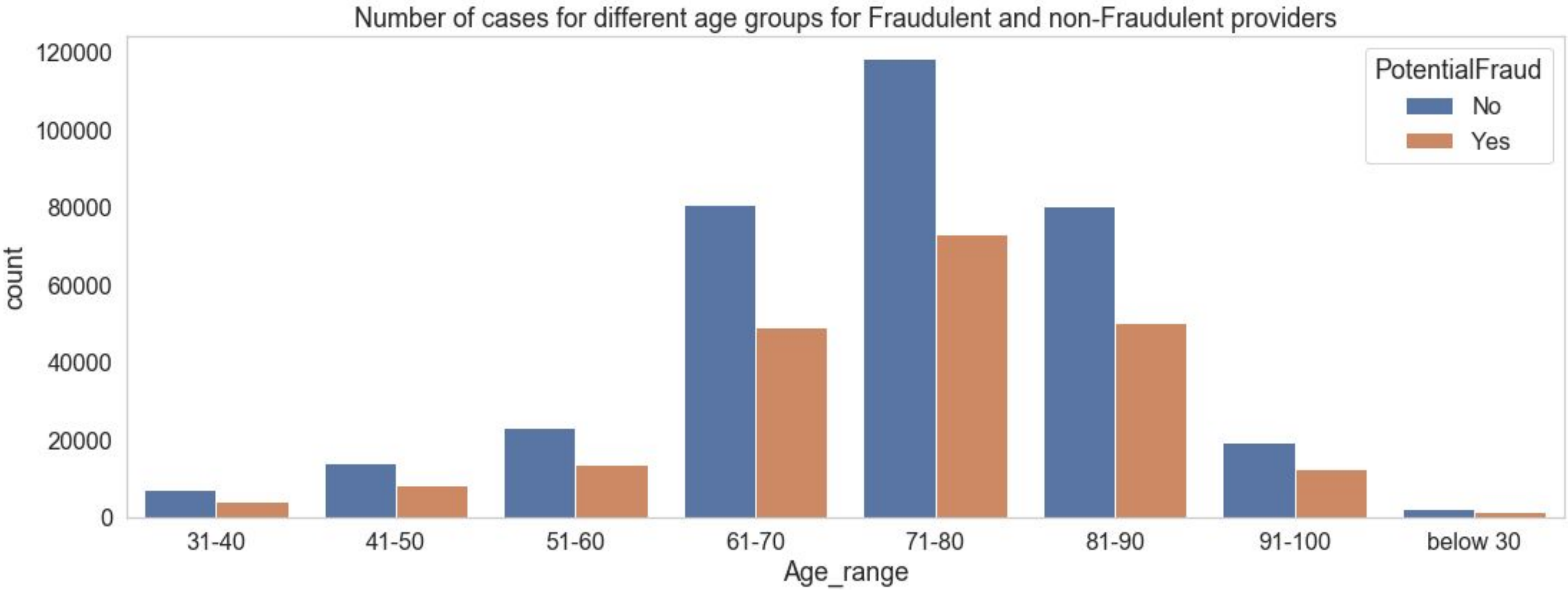




Feature Selection



Patient Age Profile



Top 10 Diagnostic Codes

4019 - Unspecified Hypertension

25000 - Diabetes Mellitus

2724 - Hyperlipidemia

V5869 -long term use of other medication

42731- Atrial Fibrillation(rapid heart rate)

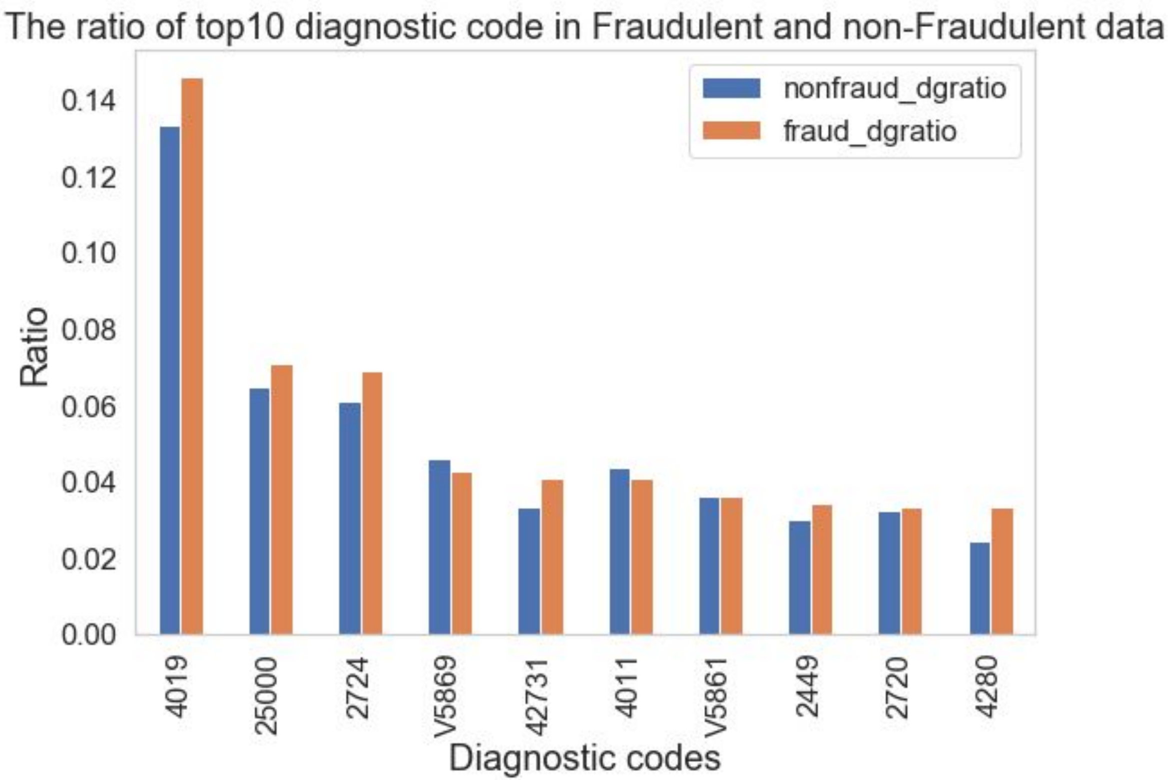
4011 - Benign essential hypertension

V5861 - Long term use anticoagulants

2449 - Unspecified acquired hypothyroidism

2720 - Hypercholesterolemia

4280 - Congestive heart failure



Top 10 Procedure Codes

4019 -diagnostic procedure on lymphatic structures

2724 -biopsy of mouth

9904 - Transfusion of packed cells

8154 - Total knee replacement

66 - removal of fallopian tube

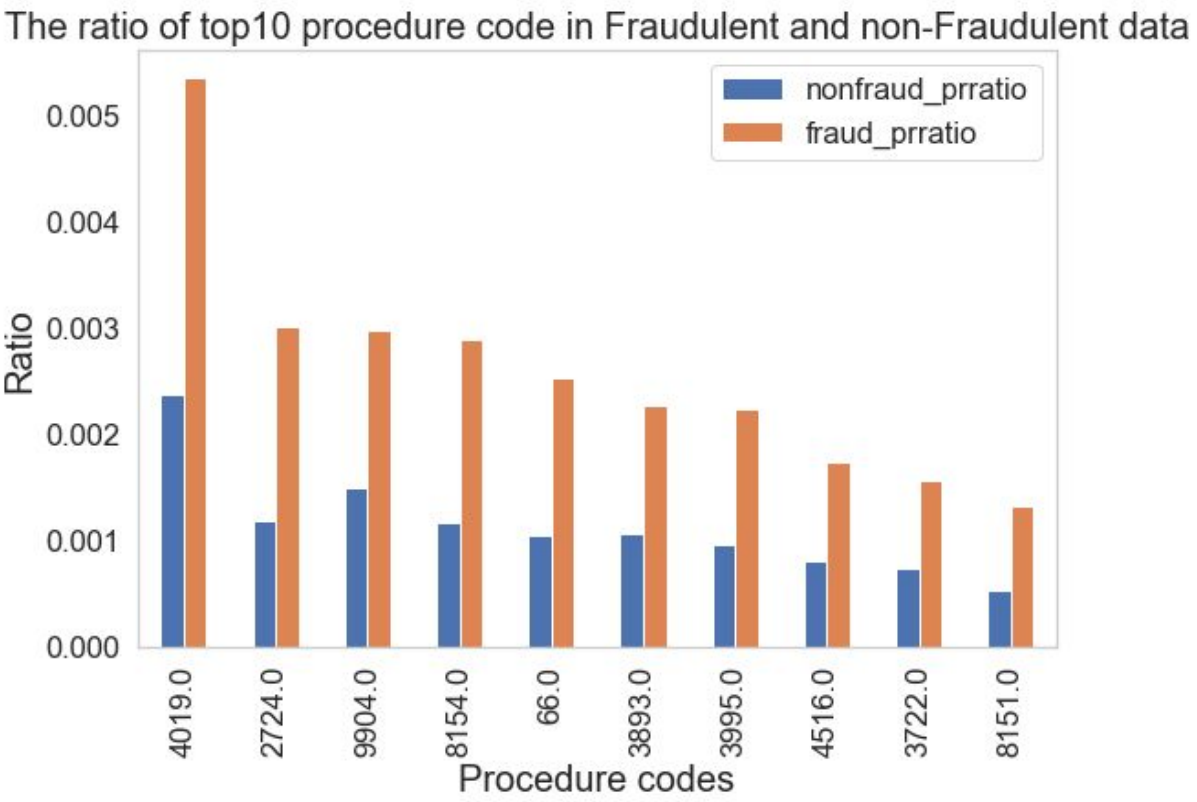
3893 - Venous catheterization

3995 - Hemodialysis

4516- Esophagogastroduodenoscopy

3722 - Left heart cardiac catheterization

8151 - Total hip replacement

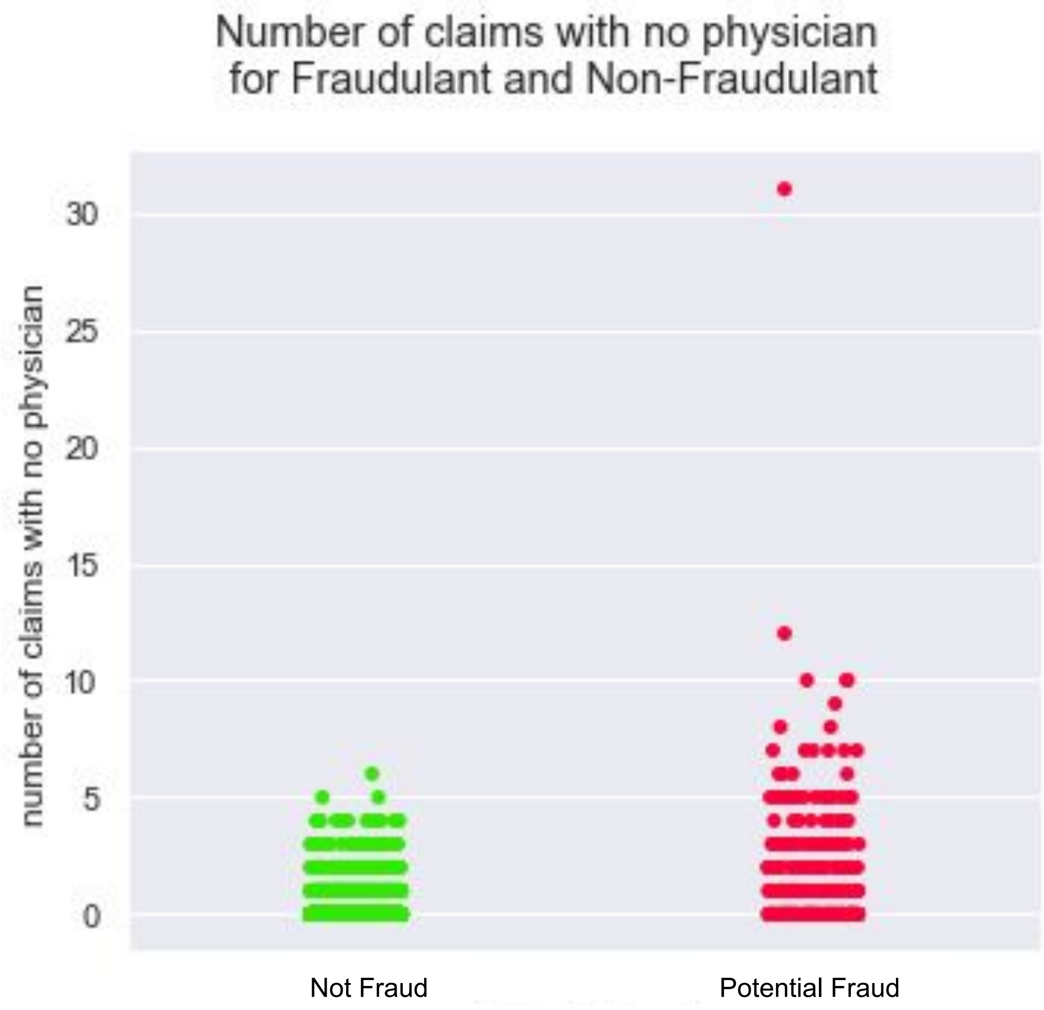


No Physician

Interestingly, the fraudulent claims have more instances of patients recorded as not seen by any physician

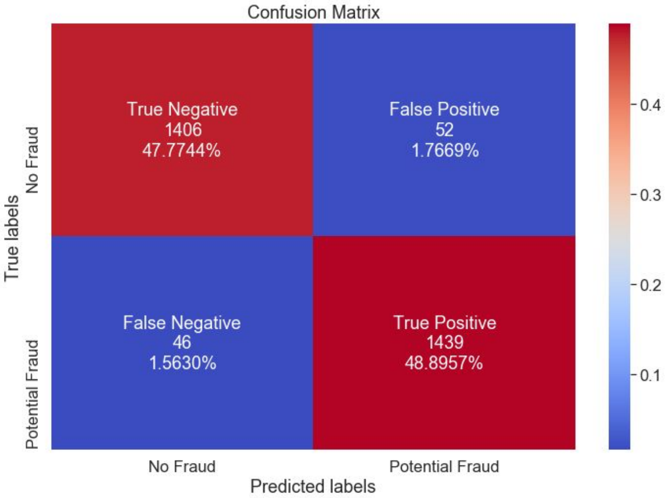
The maximum is 31 claims put by an provider(PRV51459) with no physician.

On the other hand only 5-6 claims with no physician was found with non-fraudulent providers



How much money can be saved?

Quantifying our XGBoost Model



\$998 = Avg cost per claim
\$58 = 2 hrs x 29 = Assumed cost to investigate a claim
103 = Avg number of claims per provider

(TP x 998) - (FP + TP) x 58 - (58 * FN)
Scaled up to the dataset sample size

≈ **\$5.2M** USD per year

Takeaways: Recommendations for Health Insurance Companies

- Consider establishing a extra checkpoint for when the most common diagnostic and procedure codes come up
- Closely monitor any duplicate claims, as well as claims submitted with no physician
- Fraudulent inpatient claims are significantly more prevalent than outpatient. Focus the majority of investigatory resources on inpatients.

Further Analysis

- We can further combine some of our models using advanced stacking or ensembling techniques, and consider incorporating other combinations
- Healthcare fraud can be classified into categories such as duplicate claims, “upcoding”, and billing for services never rendered. Further analysis that sorts the predicted fraud into these categories could provide more robust insights
- Covid-19 has brought about a new and unique set of fraud challenges, and it might be valuable to re-run analysis on more recent data to understand these developments