Detecting Inconsistencies in Healthcare Provider Data

Randy Pantinople
Anjali Pathak
Jay Kim

What is Healthcare Fraud?

- A type of white-collar crime
- Involves filing of dishonest health care claims to turn a profit
- Impacts the healthcare system both financially and in the way how the integrity and value of the country's health care system is being perceived

Significance of the study

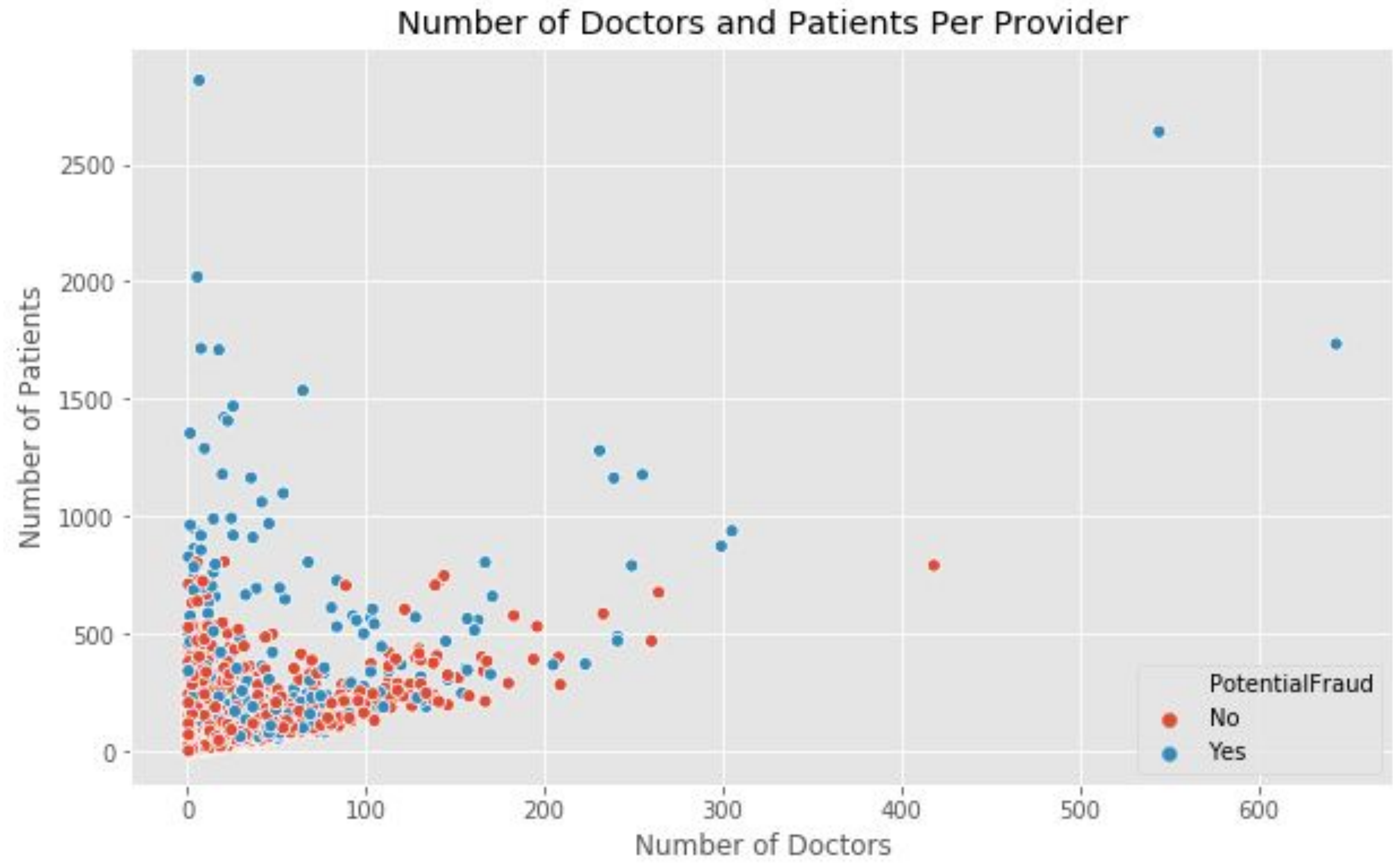
- NHE (National Health Expenditure) grew 4.6% to $3.6 trillion in 2018 for billions of claims ($11,172 per person)
- The National Health Care Anti-Fraud Association (NHCAA) estimates that the financial losses due to health care fraud are in the tens of billions of dollars each year.
- Through our project, we will be uncovering the types of ways in which providers commit healthcare fraud

Introduction to the Dataset

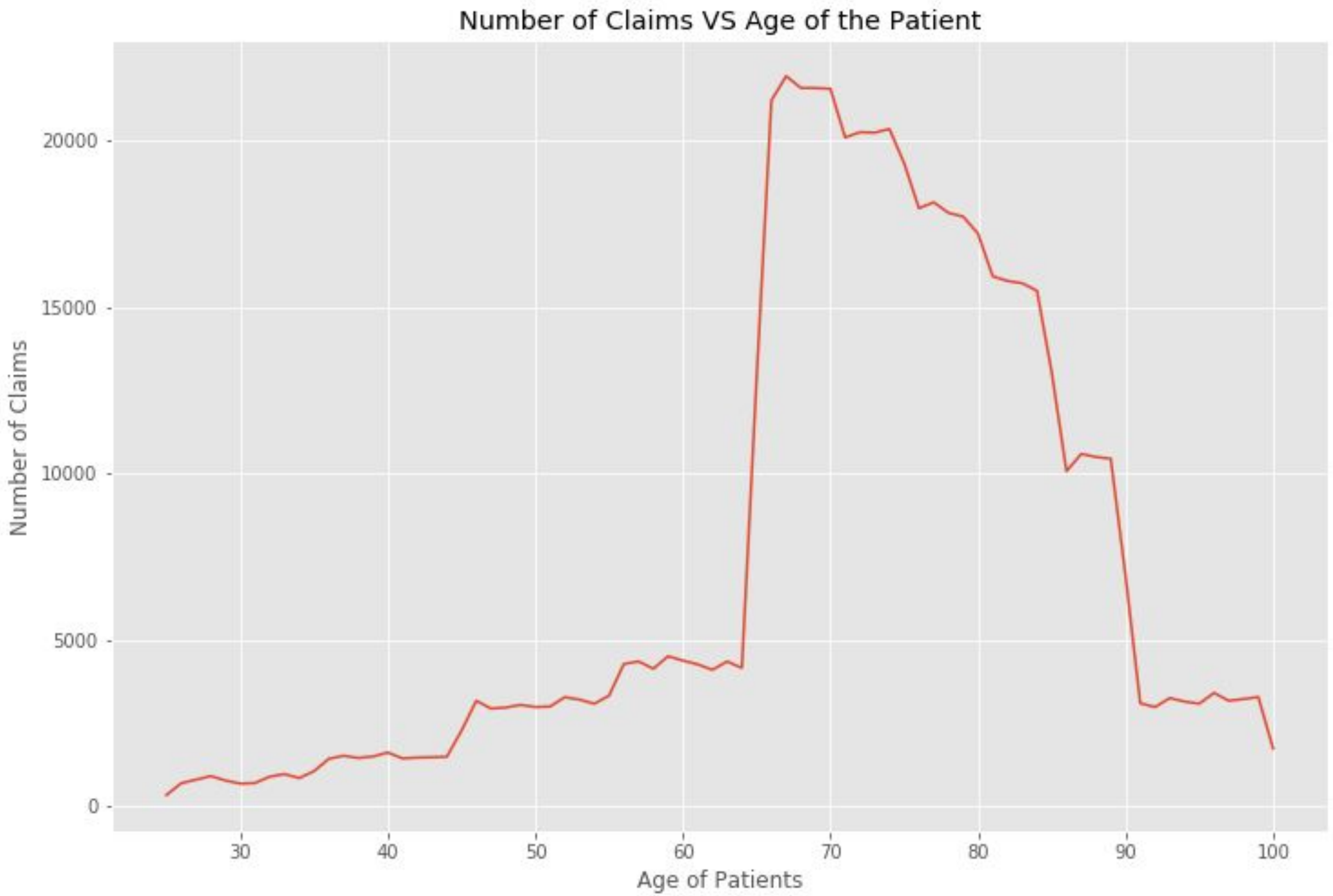
- Total number of claims : 558,211
- Total number of providers : 5,410
- Different types of providers : Inpatients only, outpatients only, and both
- Inpatient - Patients who had been formally admitted to hospitals
- Outpatient - Patients who had not been formally admitted to hospitals

Exploratory Data Analysis

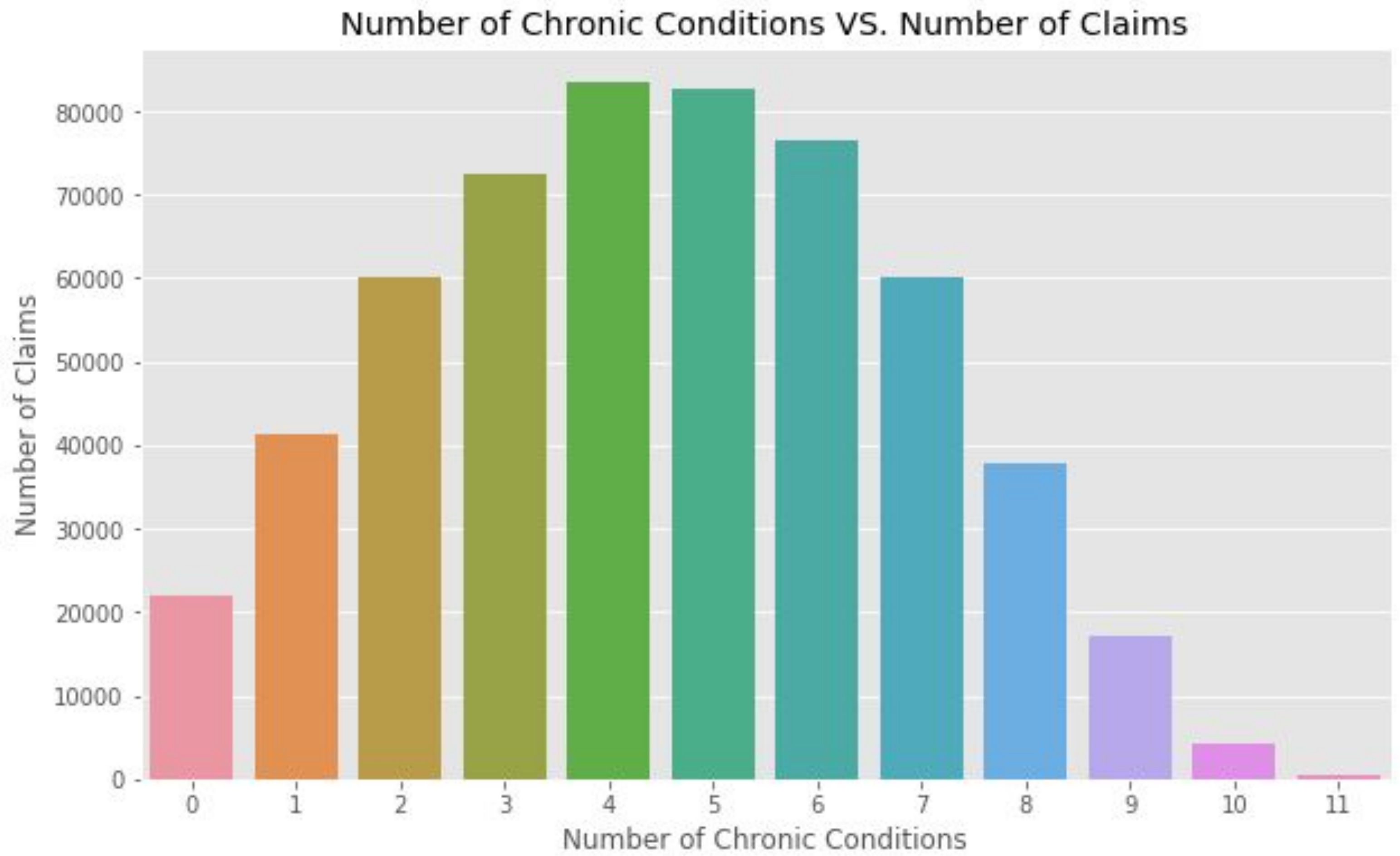
How does the number of doctors and patients affect the probability of encountering potentially fraudulent providers?


Number of Doctors and Patients Per Provider
Number of Patients
Number of Doctors
0
500
1000
1500
2000
2500
0
100
200
300
400
500
600
PotentialFraud
No
Yes

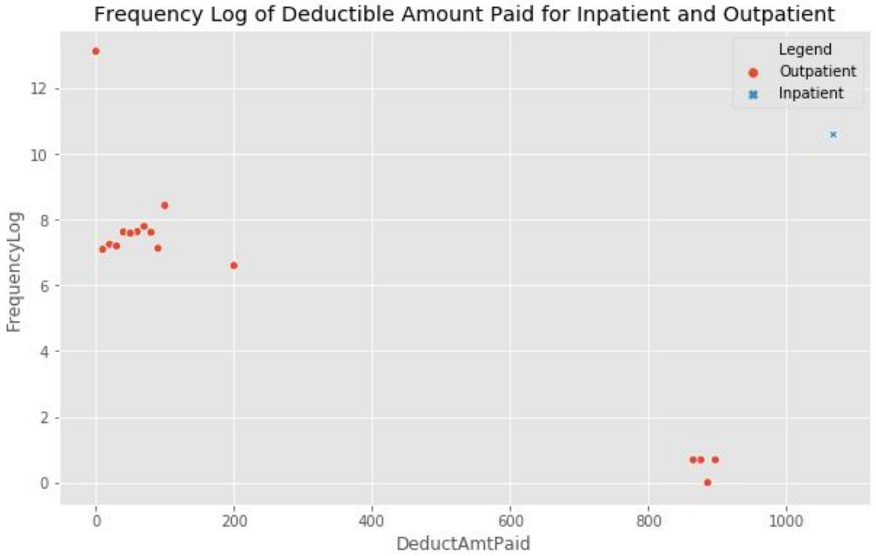
How does the number of claims differ for different ages of patients?

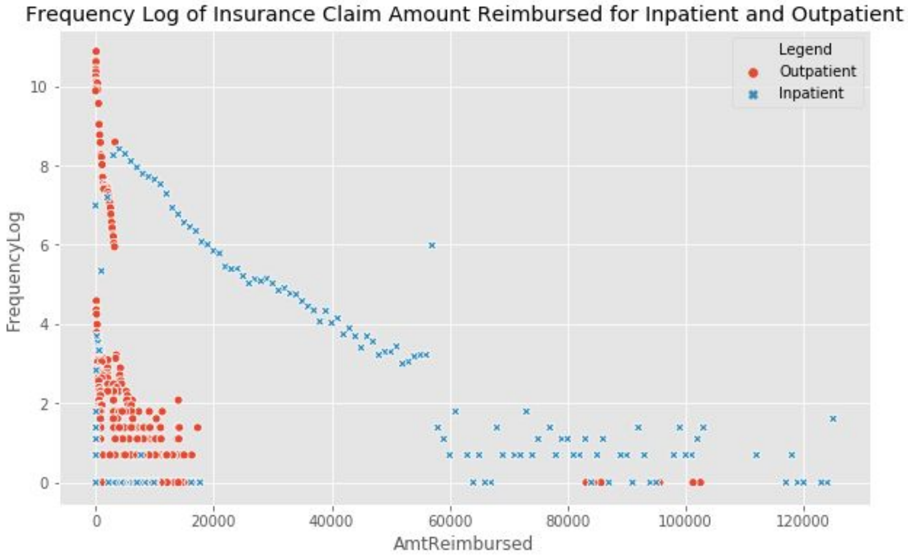

Number of Claims VS Age of the Patient
Number of Claims
0
5000
10000
15000
20000
30
40
50
60
70
80
90
100
Age of Patients

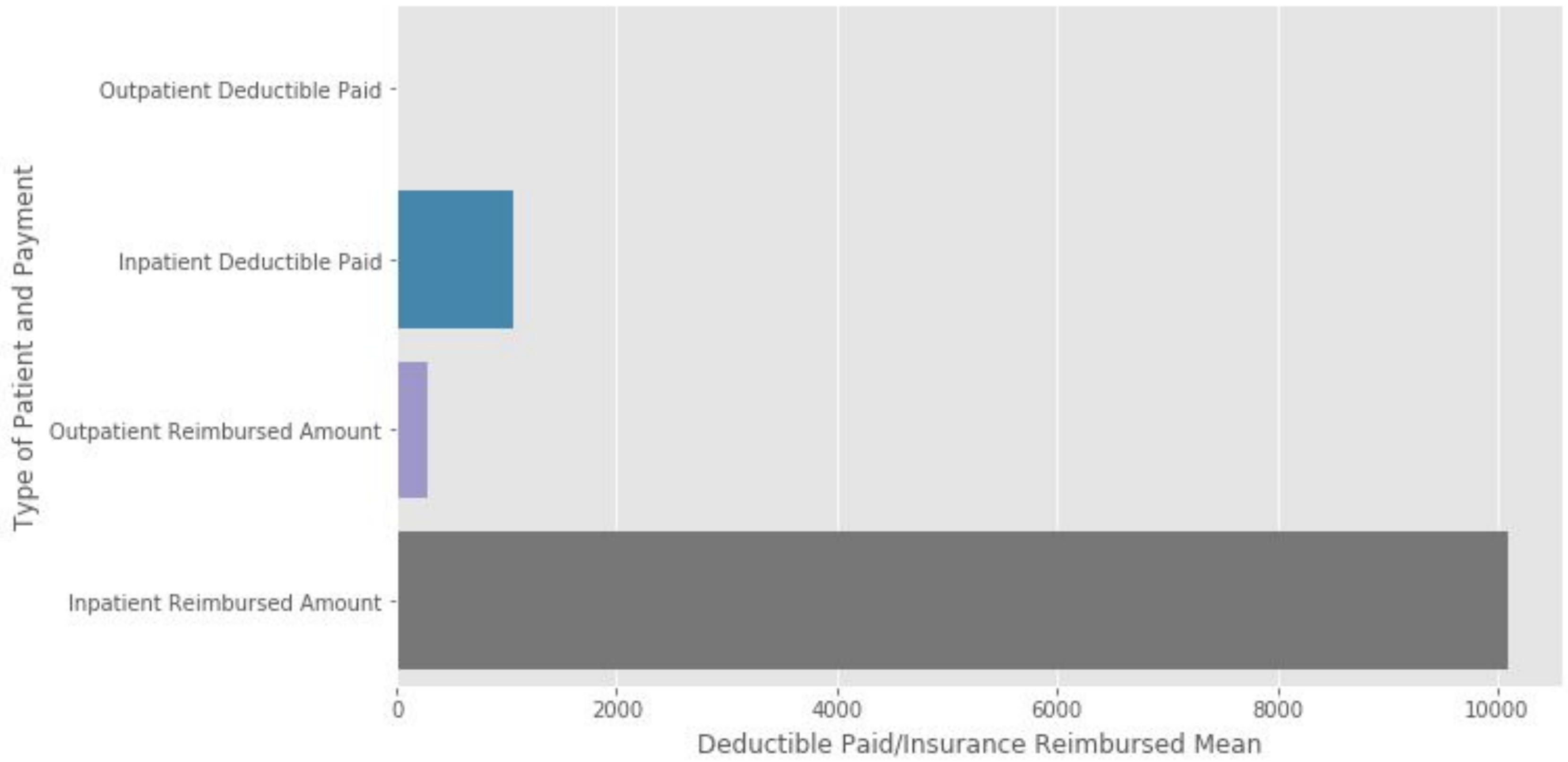
Do patients with more chronic conditions have more claims than those with less chronic conditions?


Number of Chronic Conditions VS. Number of Claims
Number of Claims
80000
70000
60000
50000
40000
30000
20000
10000
0
0
1
2
3
4
5
6
7
8
9
10
11
Number of Chronic Conditions

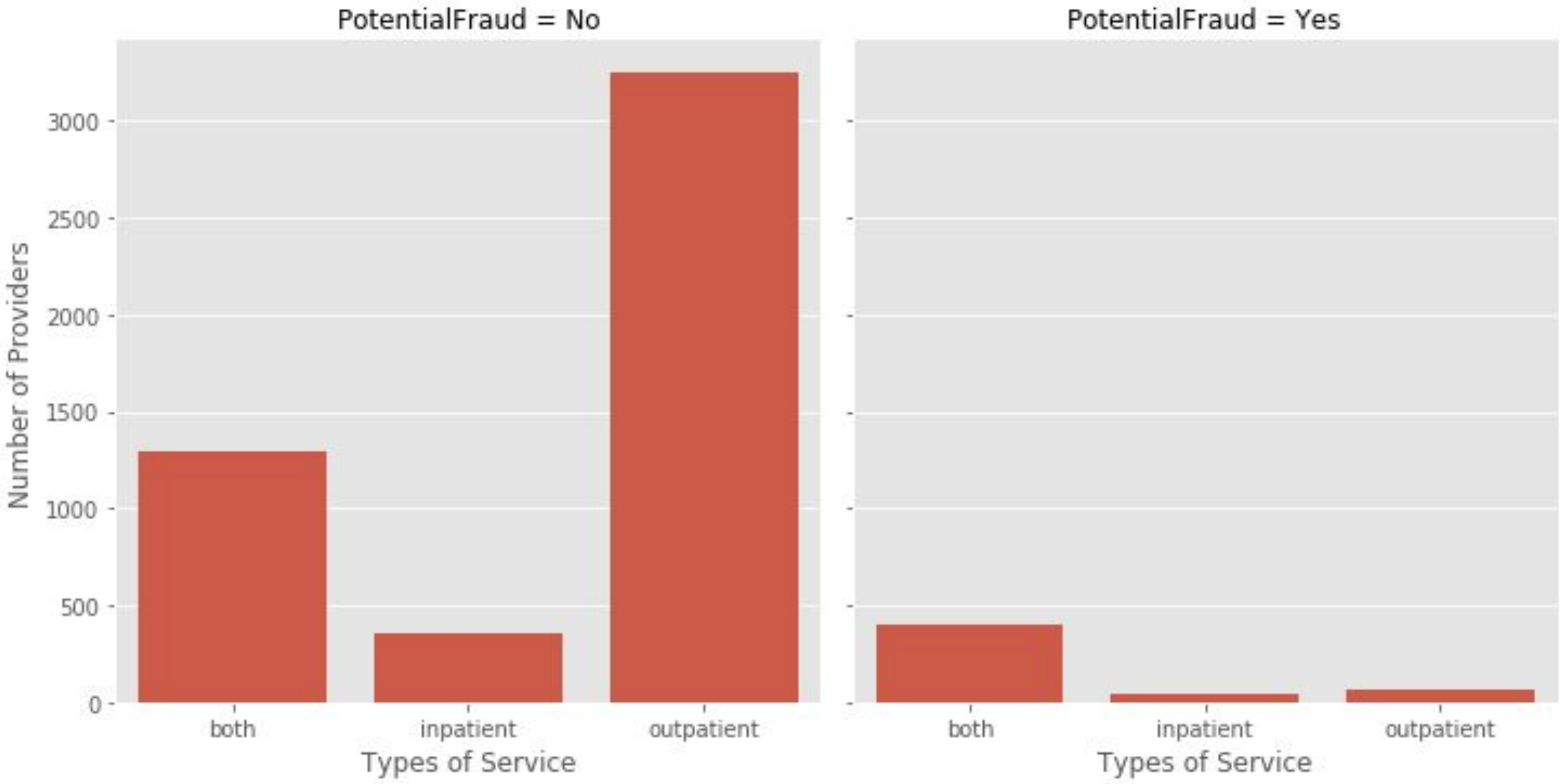
How are deductible amounts and insurance reimbursed amounts distributed for inpatient and outpatient?


Frequency Log of Deductible Amount Paid for Inpatient and Outpatient
Legend
Outpatient
Inpatient
FrequencyLog
DeductAmtPaid
0
2
4
6
8
10
12
0
200
400
600
800
1000

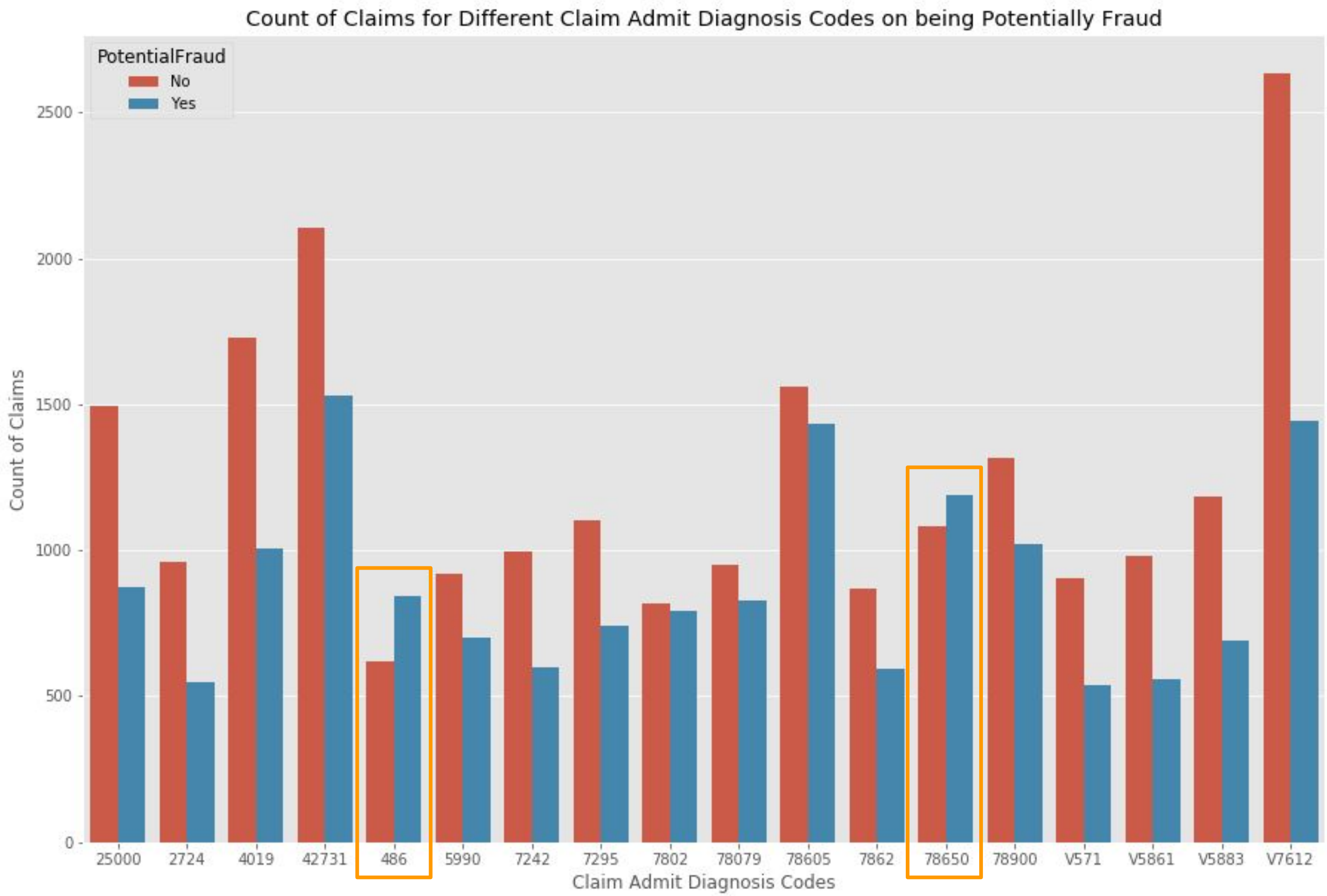

Frequency Log of Insurance Claim Amount Reimbursed for Inpatient and Outpatient
Legend
Outpatient
Inpatient
FrequencyLog
AmtReimbursed
0
2
4
6
8
10
0
20000
40000
60000
80000
100000
120000


Outpatient Deductible Paid
Inpatient Deductible Paid
Outpatient Reimbursed Amount
Inpatient Reimbursed Amount
Type of Patient and Payment
0
2000
4000
6000
8000
10000
Deductible Paid/Insurance Reimbursed Mean

Analysis of Types of Services

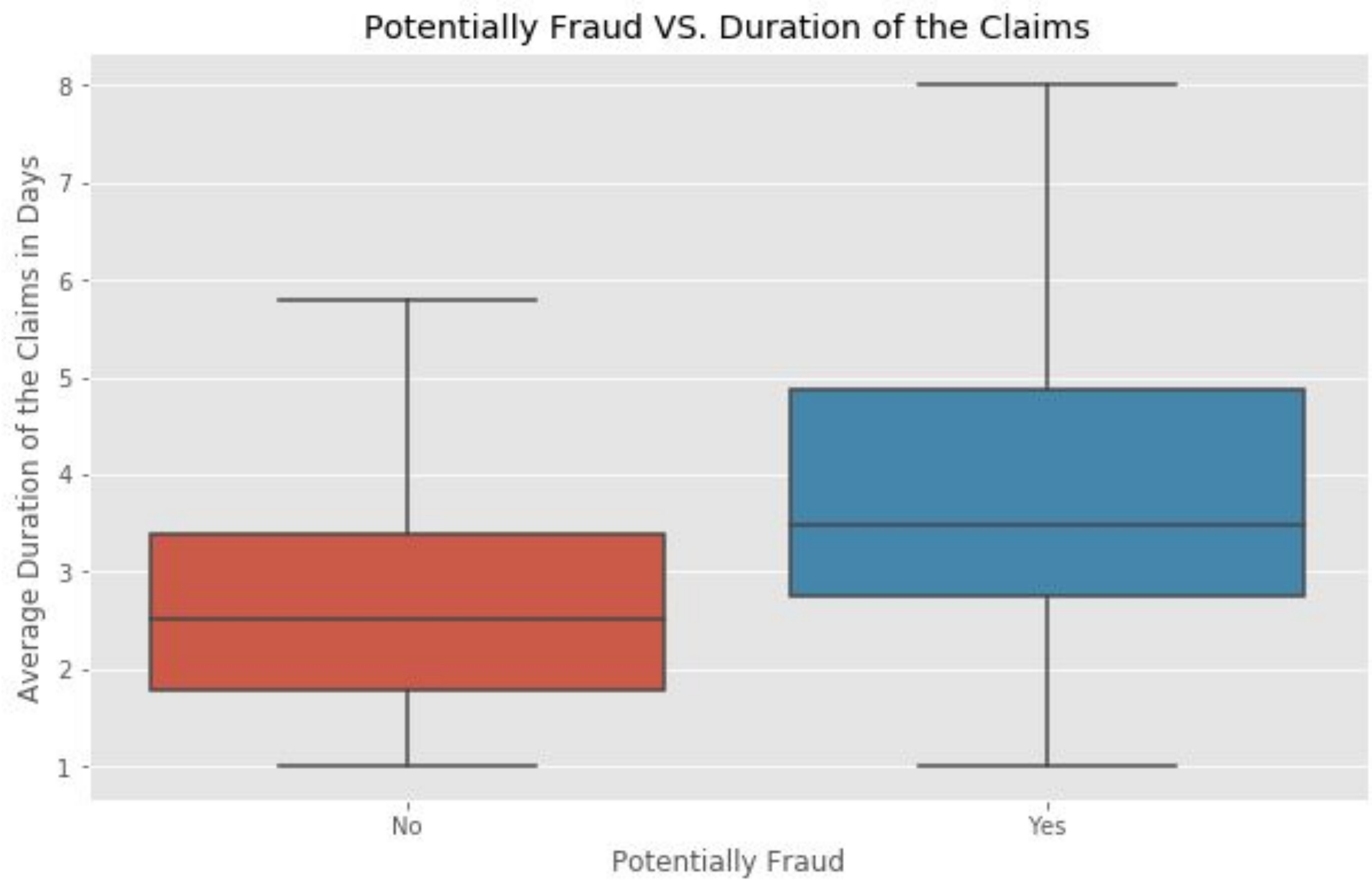

PotentialFraud = No
PotentialFraud = Yes
Number of Providers
3000
2500
2000
1500
1000
500
0
both
inpatient
outpatient
Types of Service
both
inpatient
outpatient
Types of Service

Evaluating the Relationship between Number of Claims and Claim Admit Diagnosis Codes



Count of Claims for Different Claim Admit Diagnosis Codes on being Potentially Fraud
PotentialFraud
No
Yes
Count of Claims
2500
2000
1500
1000
500
0
25000
2724
4019
42731
486
5990
7242
7295
7802
78079
78605
7862
78650
78900
V571
V5861
V5883
V7612
Claim Admit Diagnosis Codes

Distribution of Claims’ Average Duration


Potentially Fraud VS. Duration of the Claims
Average Duration of the Claims in Days
8
7
6
5
4
3
2
1
No
Yes
Potentially Fraud

Feature Engineering

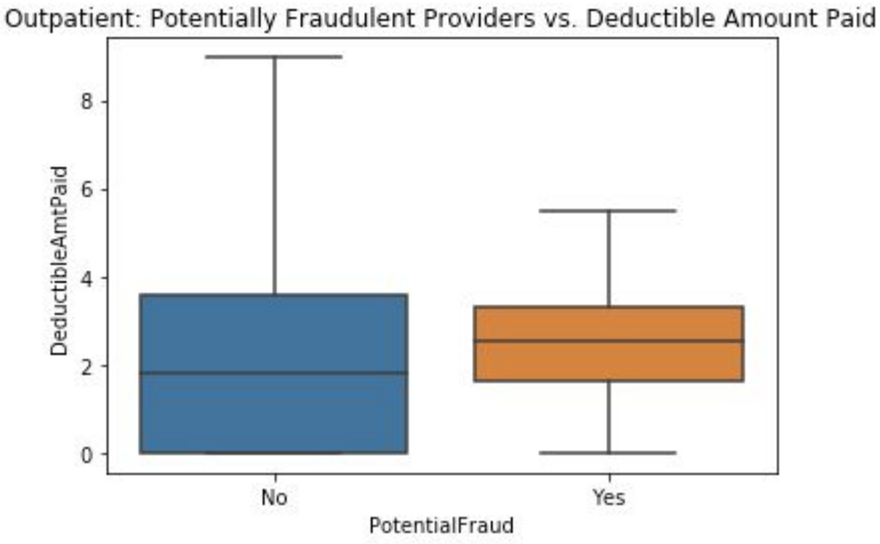
Feature Engineering

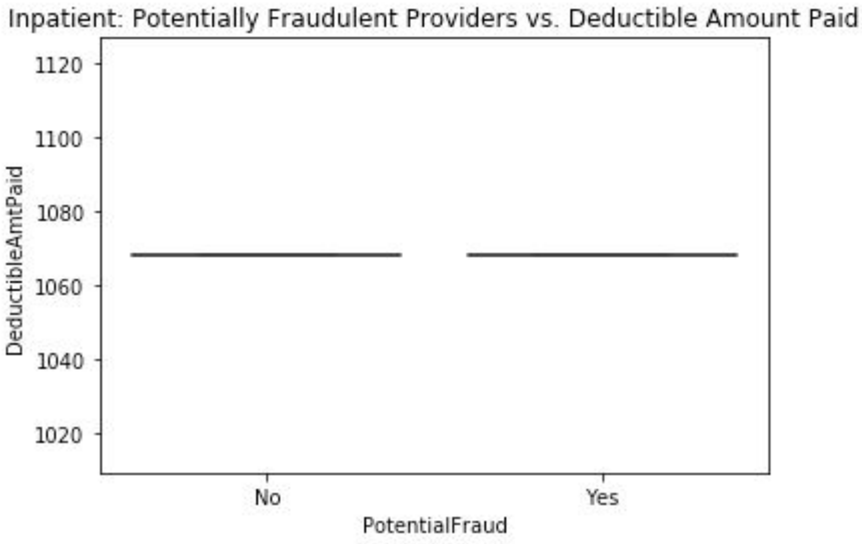
- Datasets provided based on patients and claims
- Aggregated inpatient, outpatient, beneficiary, and fraud datasets
- Created new dataset based on providers
- Flow:

EDA ⇒ Engineer ⇒ Feature Selection ⇒ EDA ⇒ Engineer ⇒ Model ⇒ Analysis

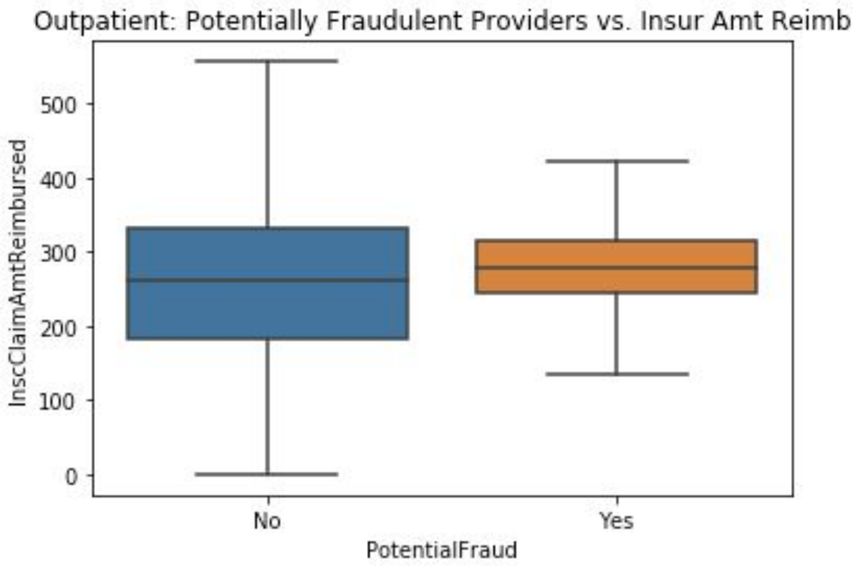
Features’ Categories		
Days Admitted	Financial	Age
Race	Type of Service	Claims
States	Counties	Chronic Conditions
Diagnosis Codes	Procedure Codes	Gender
Number of Patients	Number of Doctors	Attending / Operating Physicians

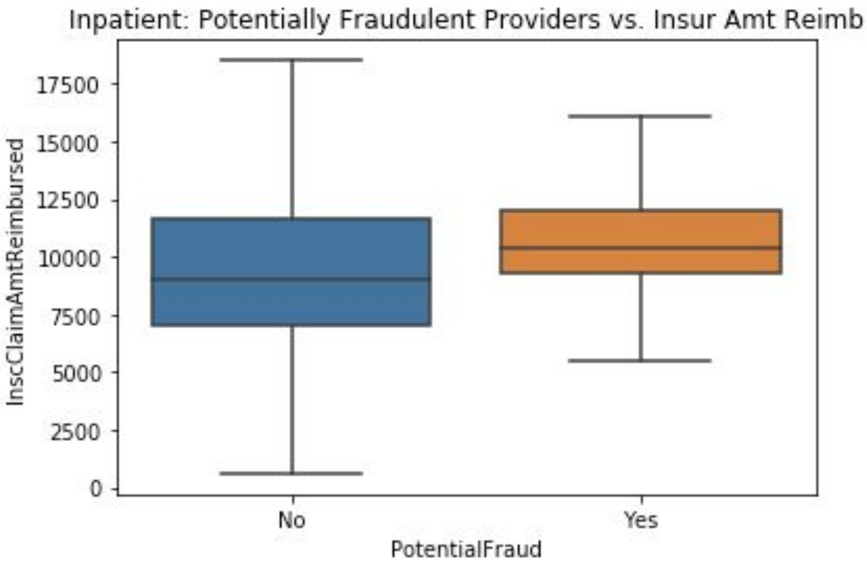
Assessing Fraudulent Providers Based on Deductible Amount Paid



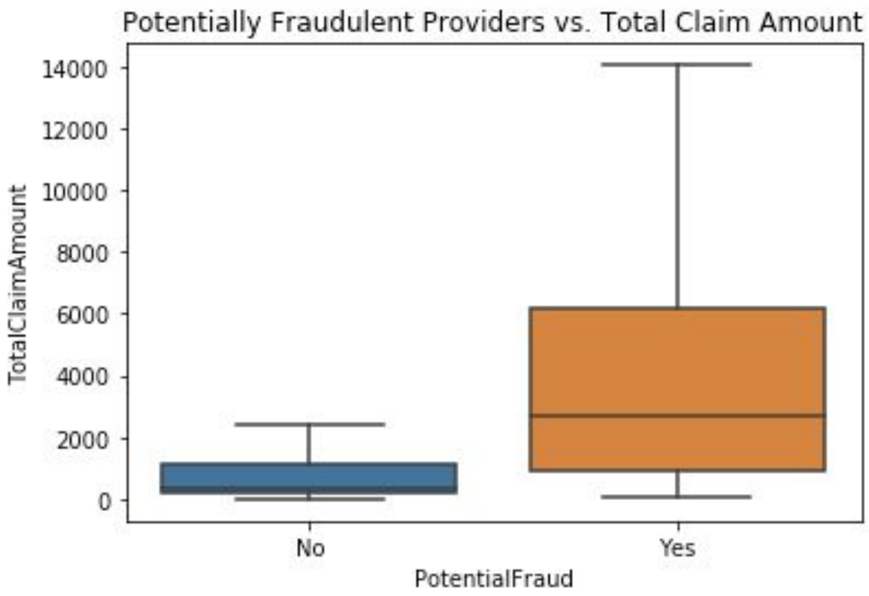


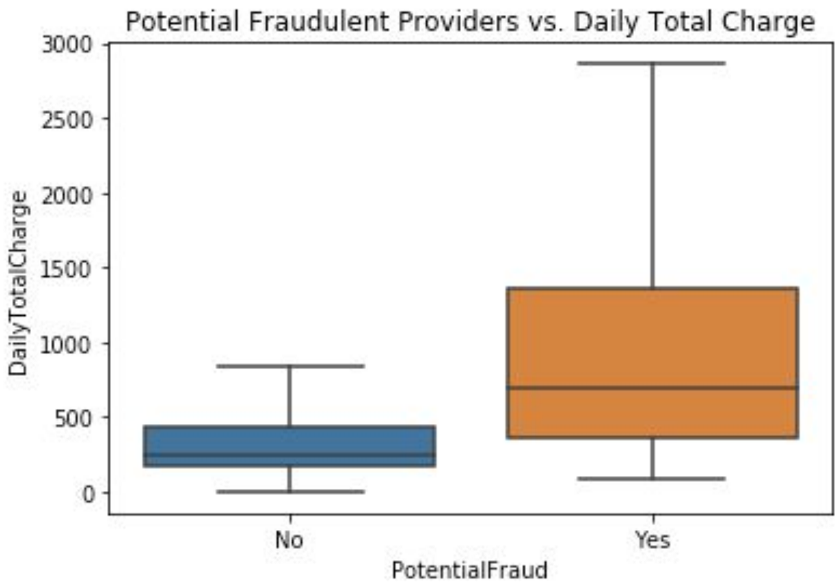
Assessing Fraudulent Providers Based on Insurance Claim Amount Reimbursed



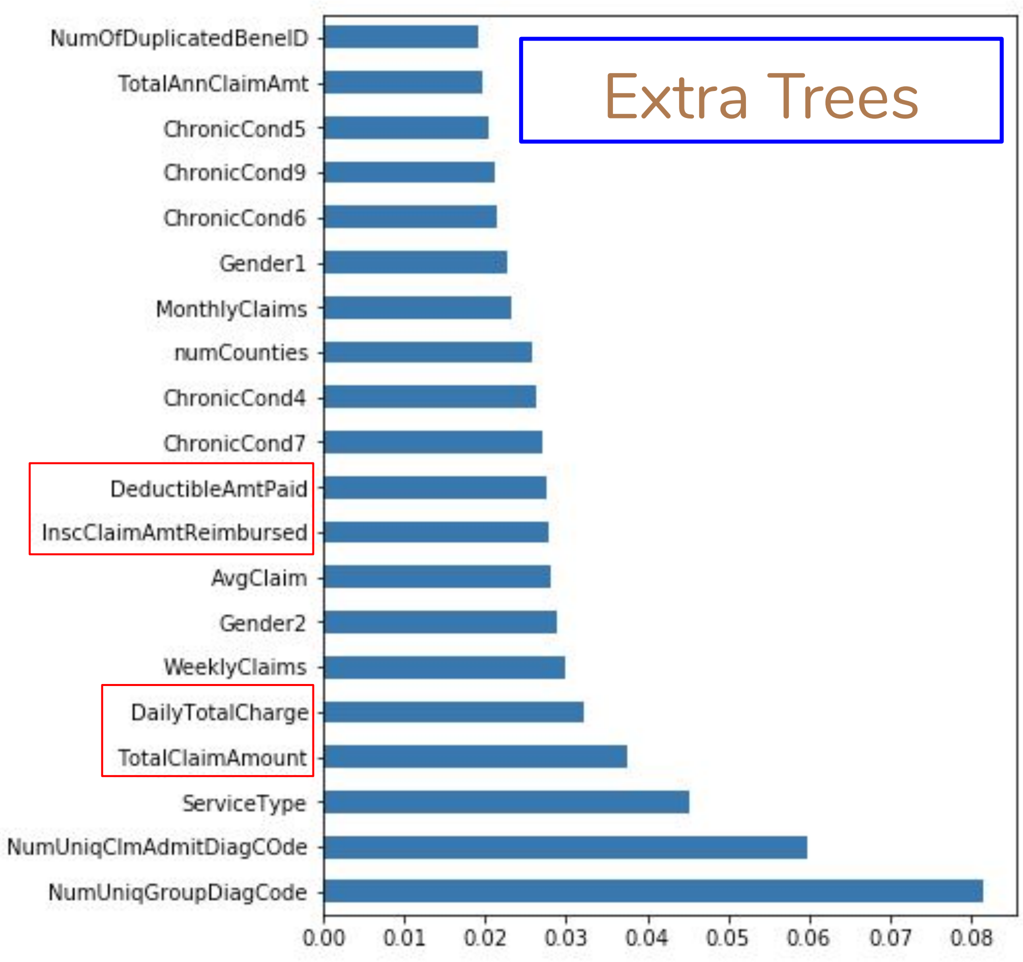


Combining Features: Assessing Potentially Fraudulent Providers Based on Total Claim Amounts





Feature Importance/Selection Example

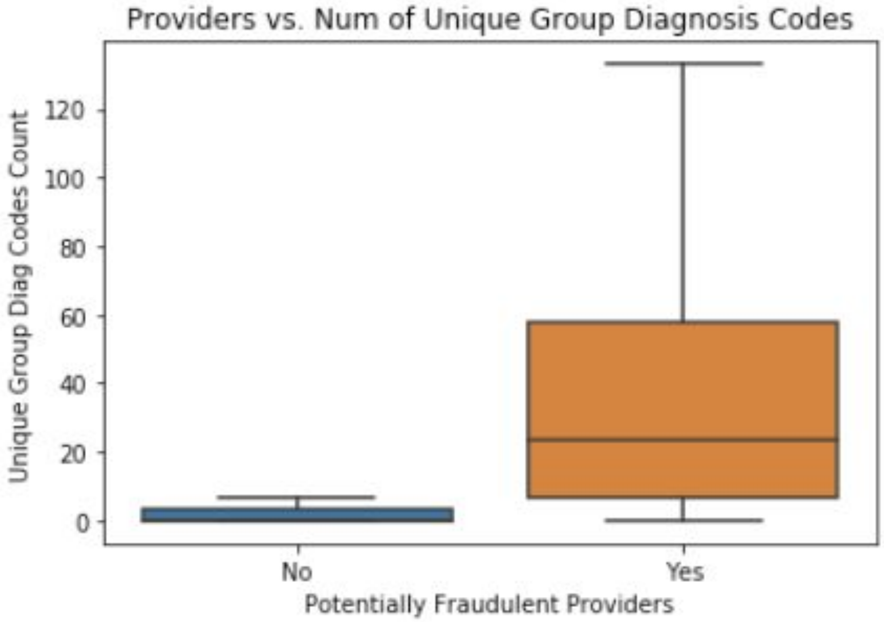


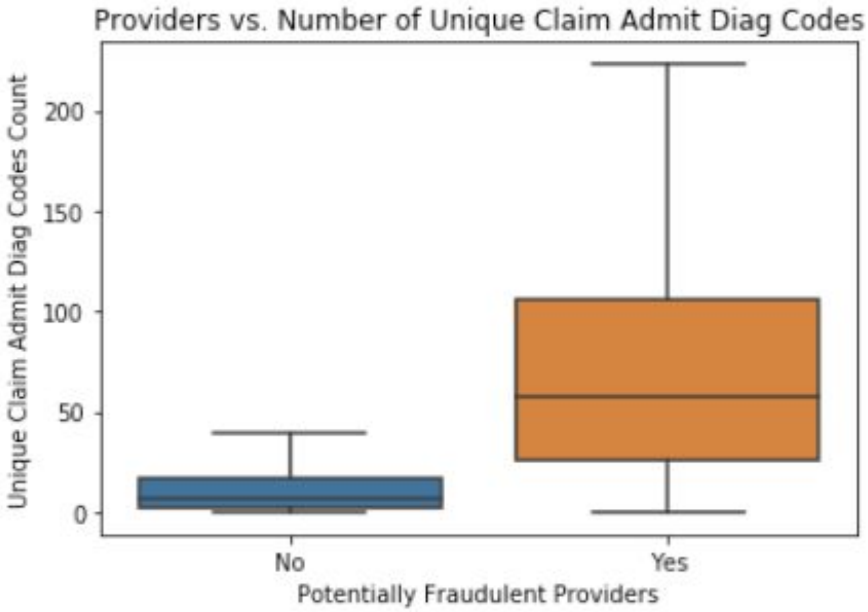
- Extra Trees Classifier
- Lasso Regression for Feature Importance
- Determined which features to retain and which to drop

Lasso

- total features: 55
- selected features: 42
- features with coefficients shrank to zero: 13

Analyzing Group Diagnosis Codes and Claim Admit Diagnosis Codes as Means of Detecting Fraudulent Providers

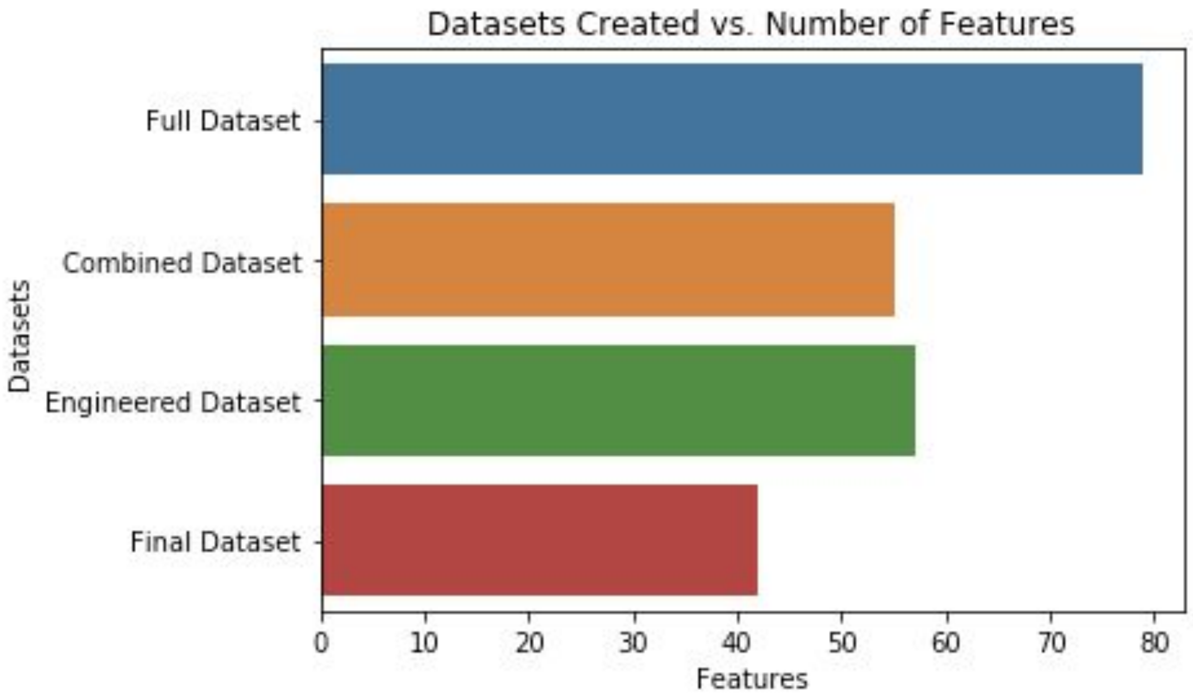




Assessing Validity of Features Using Logistic Regression

Features	Train Accuracy Score	Test Accuracy Score
• Number of Duplicated Beneficiary IDs • Patients with 12 Chronic Conditions	0.65	0.63
• Number of Duplicated Beneficiary IDs • Patients with 12 Chronic Conditions • Total Claim Amount	0.76	0.76
• Number of Duplicated Beneficiary IDs • Patients with 12 Chronic Conditions • Total Claim Amount • NumUniqGroupDiagCode	0.85	0.85

Final Dataset Going into Machine Learning Models



- Inpatient, Outpatient, Beneficiary datasets = 79 features
- Combined above datasets = 55 features
- Our engineered dataset with most features = 57 features
- Final engineered dataset = 42 features

Machine Learning Models

Stochastic Gradient Descent Classifier

Best parameters:
Alpha = 0.01 , penalty : l2

Cross validation score : 0.869

Performance score : 0.863

Support Vector Classifier

Best parameters:
C= 4300, degree= 3, kernel = poly

Cross validation score : 0.958

Performance score : 0.964

Random Forest Classifier

Best parameters:
Criterion = entrophy max_depth= 30
Min_samples_leaf = 4, min_samples_split =6
N_estimators = 70

Cross validation score = 0.974

Performance score = 0.978

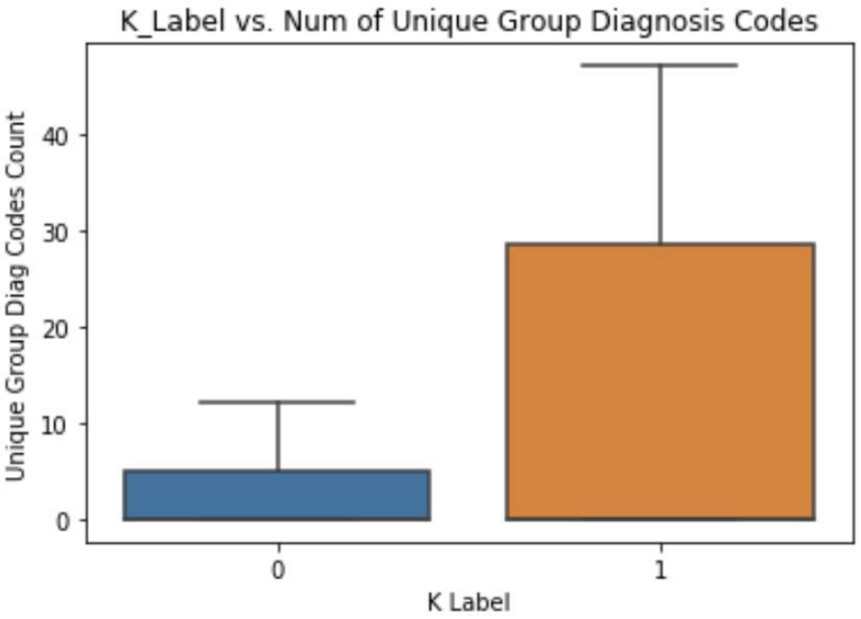
Gradient Boosting Classifier

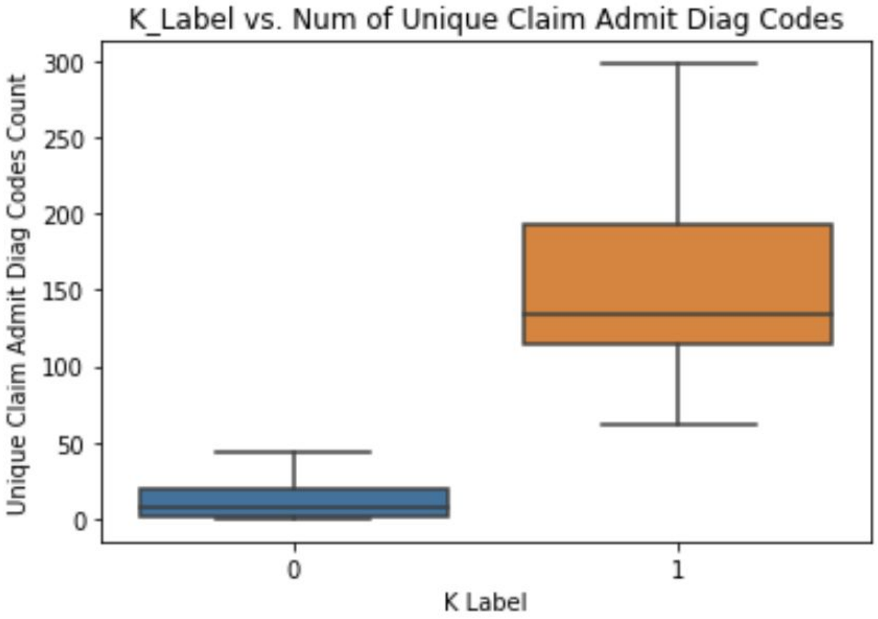
Best parameters:
Min_samples_split = 8 min_samples_leaf= 6
Learning rate = 0.56 n_estimators = 1500
Max_features = 5 max_depth = 25

Cross validation score = 0.979

Performance score = 0.982

Clustering Using K-Means: Unlabelled Dataset



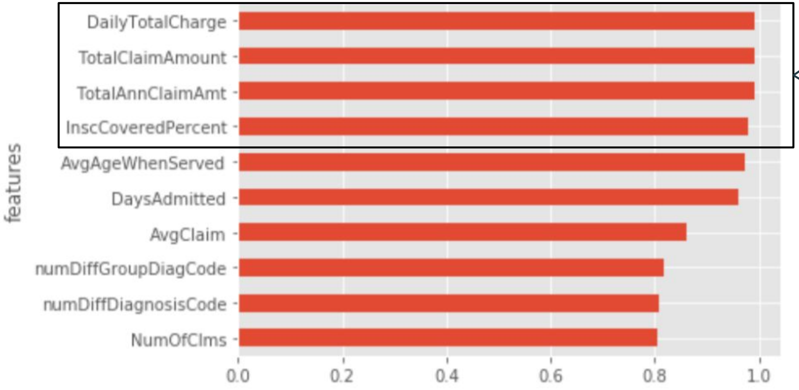


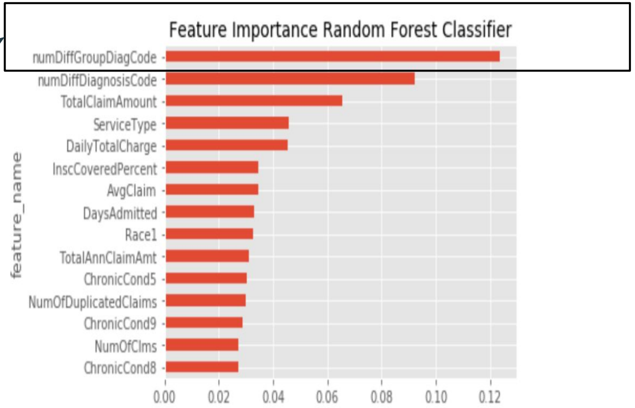
★ Label 1 is the minority class in the test dataset

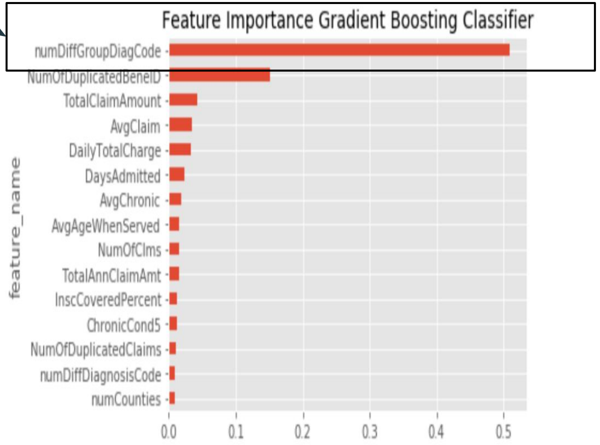
Final Analysis

Unary classification vs Random Forest and Gradient Boosting:

Most Important Features







Why do the Gradient Boosting and Random Forest models choose ‘Number of Group Diagnosis Codes’ as the most important feature?

Diagnosis Related Group Code (DRG)

- Diagnosis Related Group Code (DRG): means of classifying patients under a particular group
  - Same group: patients likely to need similar level of hospital resources
- Each DRG has a payment weight assigned to it
  - Allows hospital to determine how much it can charge for its services

Where could a possible anomaly come from?

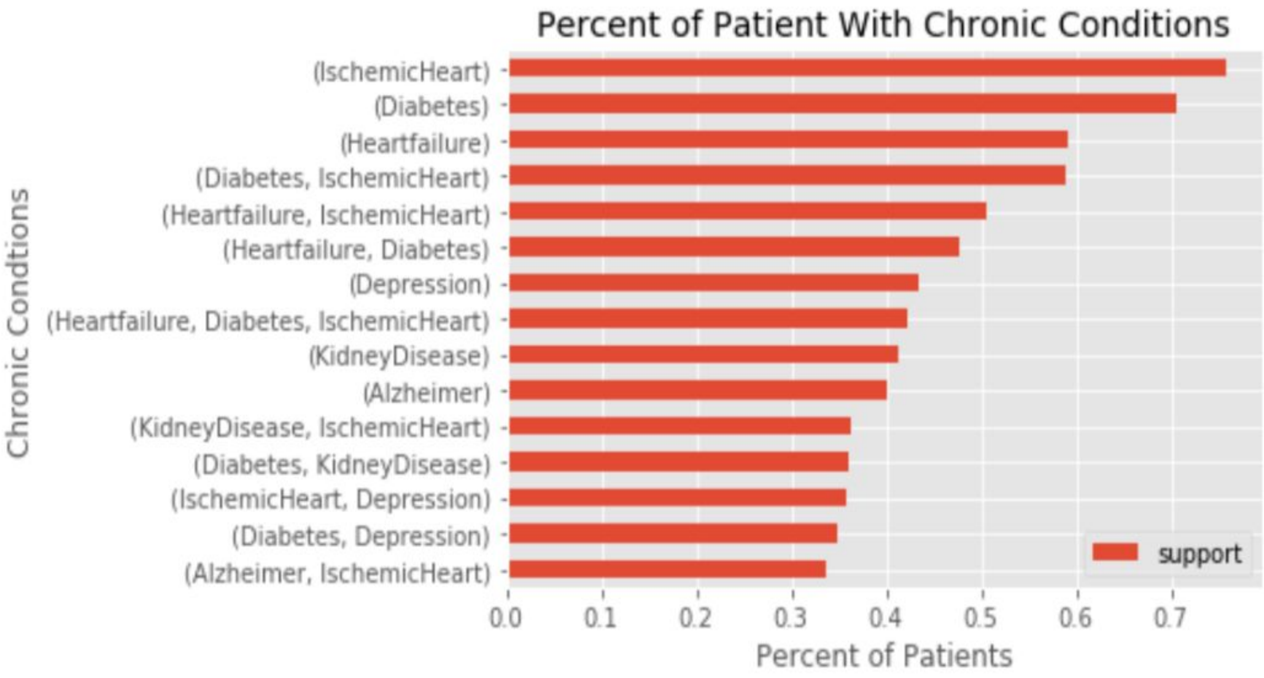
- Why do we care about the total number of unique group diagnosis code?

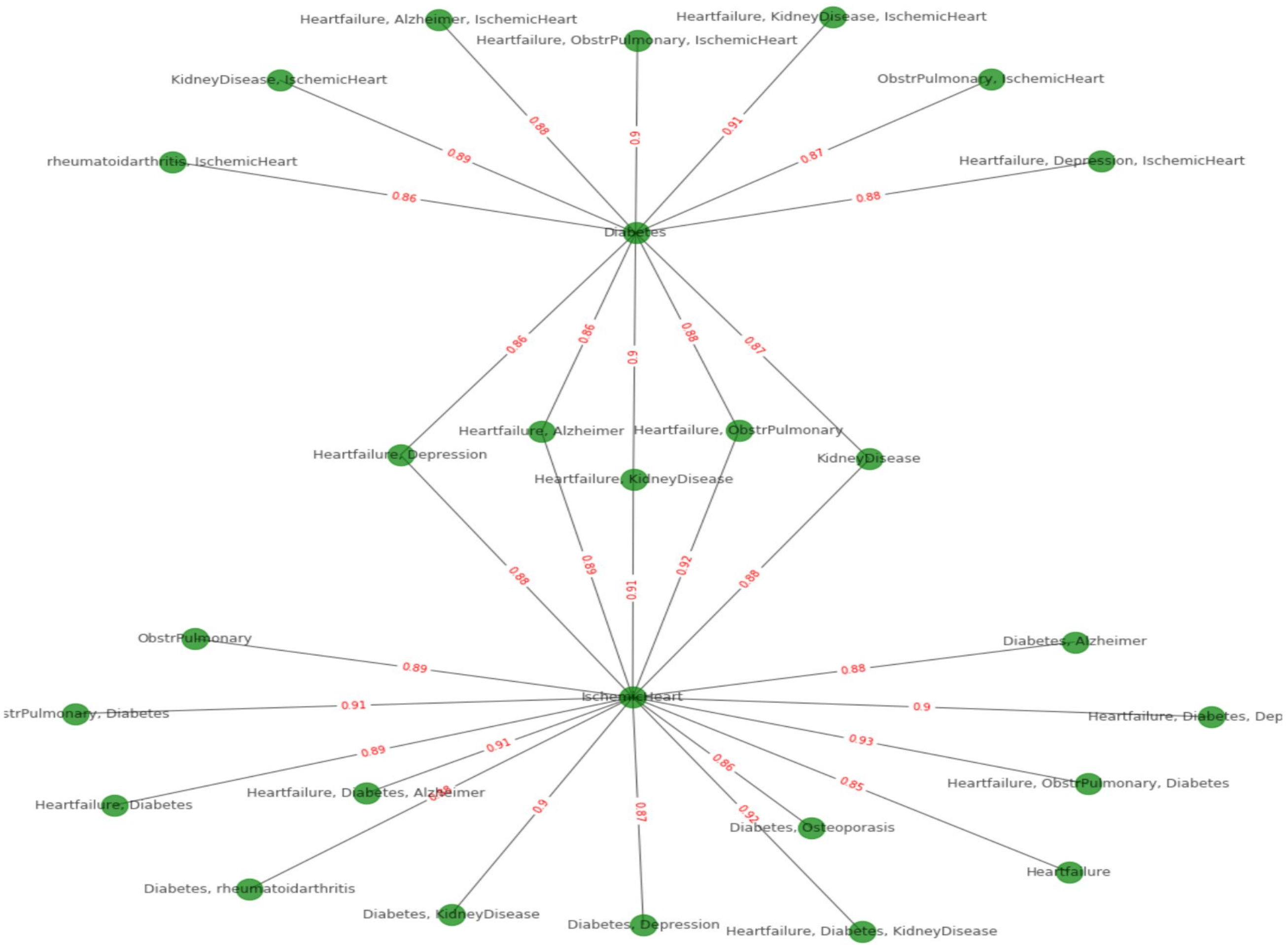
Upcoding

Unbundling

Recommendations

Market Basket Analysis





Heartfailure, Alzheimer, IschemicHeart
Heartfailure, ObstrPulmonary, IschemicHeart
Heartfailure, KidneyDisease, IschemicHeart
KidneyDisease, IschemicHeart
ObstrPulmonary, IschemicHeart
rheumatoidarthritis, IschemicHeart
Heartfailure, Depression, IschemicHeart
Diabetes
Heartfailure, Alzheimer
Heartfailure, ObstrPulmonary
Heartfailure, Depression
KidneyDisease
Heartfailure, KidneyDisease
ObstrPulmonary
Diabetes, Alzheimer
IschemicHeart
strPulmonary, Diabetes
Heartfailure, Diabetes, Dep
Heartfailure, ObstrPulmonary, Diabetes
Heartfailure, Diabetes
Heartfailure, Diabetes, Alzheimer
Diabetes, Osteoporasis
Heartfailure
Diabetes, rheumatoidarthritis
Diabetes, KidneyDisease
Diabetes, Depression
Heartfailure, Diabetes, KidneyDisease

Conclusion

- Most Important Features for Detecting Fraudulent Providers:
    - Unique Group Diagnosis Codes
    - Unique Claim Admit Diagnosis Codes
    - Total Claim Amount
    - Service Type
- Future Work:
    - Hypertuning K-Means Model to Affirm Whether our Label Assumptions are Correct
        - Also use K-Means to identify new features
    - Further analyze fraudulent providers using Market Basket Analysis, and use correlations to create new features

Questions?