2020 Master of Management Analytics Online Data Analysis Competition
Case Challenge: Medical Insurance Fraud Investigation

Team J: Nhat Pham, Robin Deng

Preliminary Write-up

October 2020

A. Problem Introduction

1. Size and Scope of Medicare Fraud

Medical provider fraud is one of the biggest problems facing the US public healthcare system. Medicare, one of the largest public insurance programs primarily for people aged 65 or older, accounts for a majority of public healthcare spending and has been the target of many fraud schemes. Extensive records are collected by Medicare on its utilization and costs; however, it is still vulnerable to fraud with fewer than 5% of Medicare claims being audited.

Analysis of Medicare data has shown that many cases of fraud involved physicians and associated providers. They adopt ways in which an ambiguous diagnosis code is used to charge Medicare for unnecessary procedures and drugs. This results in increased costs, raising insurance premiums for all Americans.

Fraud has been a common socialism topic for years, and it is wide in every region. The Federal Bureau of Investigation (FBI) estimates that **3-10% of all healthcare expenditure** is from fraud [1]. To make it clearer, the Canadian Life and Health Insurance Association (CLHIA) announced that fraud in the health care industry costs Ontario taxpayers **over $100 million** yearly [2]. Those great losses in medicare overburdens the healthcare system and drives accelerated insurance inflation.

2. Problem Statement

Analysis of Medicare data has shown that many cases of fraud involved physicians and associated providers. They adopt ways in which an ambiguous diagnosis code is used to charge Medicare for unnecessary procedures and drugs. This results in increased costs, raising insurance premiums for all Americans.

We would like to investigate fraud among Medicare providers and provide data-driven recommendations on how to combat Medicare fraud.

Make a recommendation on the criteria that should be used to select which medical providers are audited, given the limited audit resources available.

Some of the most common types of fraud by providers are:
- Billing for services that were not provided
- Duplicate submission of a claim for the same service
- Misrepresenting the service provided
- Charging for a more complex or expensive service than was actually provided
- Billing for a covered service when the service actually provided was not covered

B. Exploratory Data Analysis

1. Dataset Overview

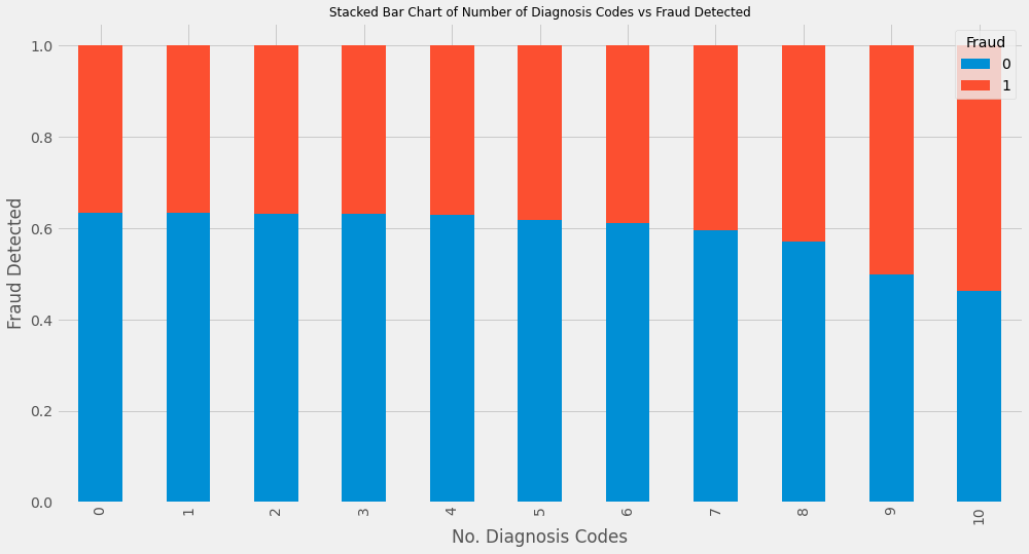
The provided dataset contains a sample of anonymized data for beneficiaries, as well as their inpatient and outpatient insurance claims between Dec 2008 and Dec 2009. A separate table of health care provider flags show those who appear to have engaged in fraud.

The dataset has 212,796 claims with fraud providers, 345,415 claims with non-fraud providers (38.1% fraud, 61.9% no fraud). The proportion of claims with fraud vs non-fraud providers is skewed but not imbalanced.

2. Hypothesis Development

We examined each variable, then based on description of common types of fraud stated above, we developed hypotheses that might explain relationship between several variables and likelihood of fraud and visualized these potential correlations:

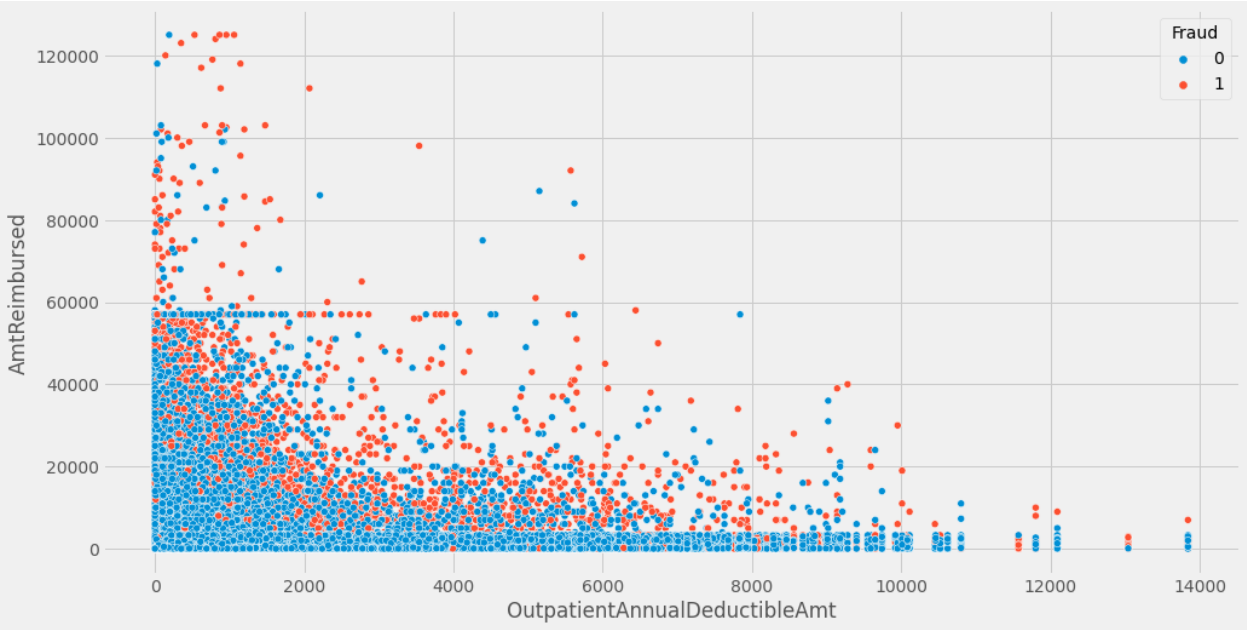
a. Hypothesis 1: Claims with more diagnosis codes tend to charge more and consequently have higher reimbursement amount. In other words, providers might list out unnecessary conditions to get more money.



From the stacked bar chart above, it seems that claims with more diagnosis codes are more likely to be from fraudulent providers.

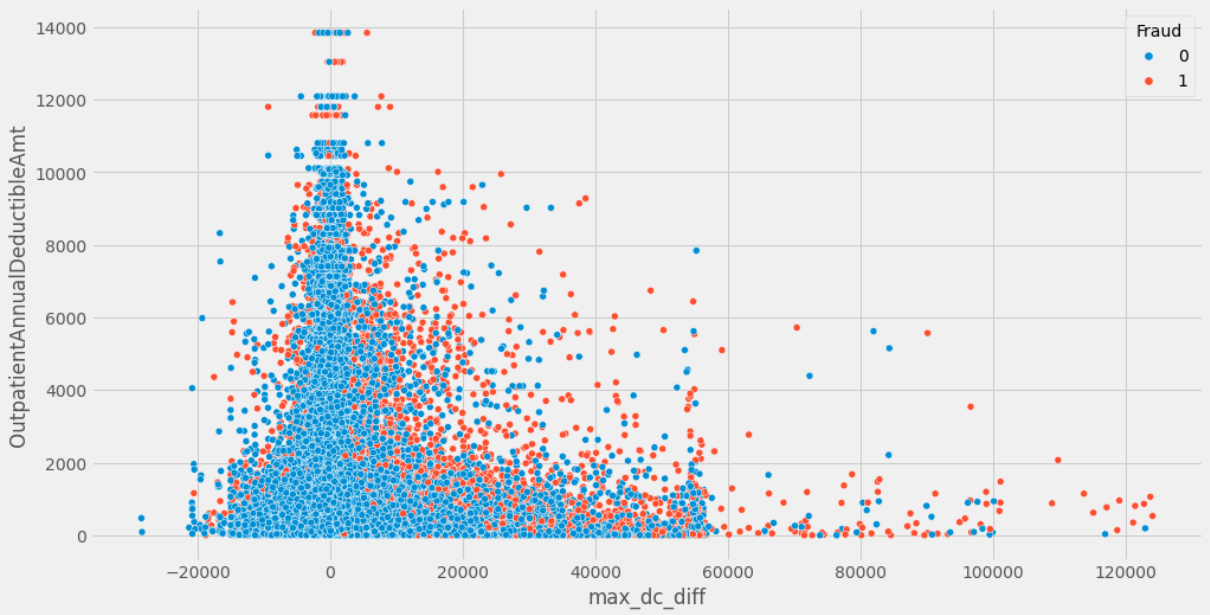
b. Hypothesis 2: Claims with high reimbursement amounts tend to be from fraudulent providers

This is based on our assumption that providers would want to fraud more from patients with low deductible amounts because insurance companies will have to pay more for these patients than patients with high deductible amounts. Hence, medical providers have more incentives to make fraudulent claims on low-deductible patients.



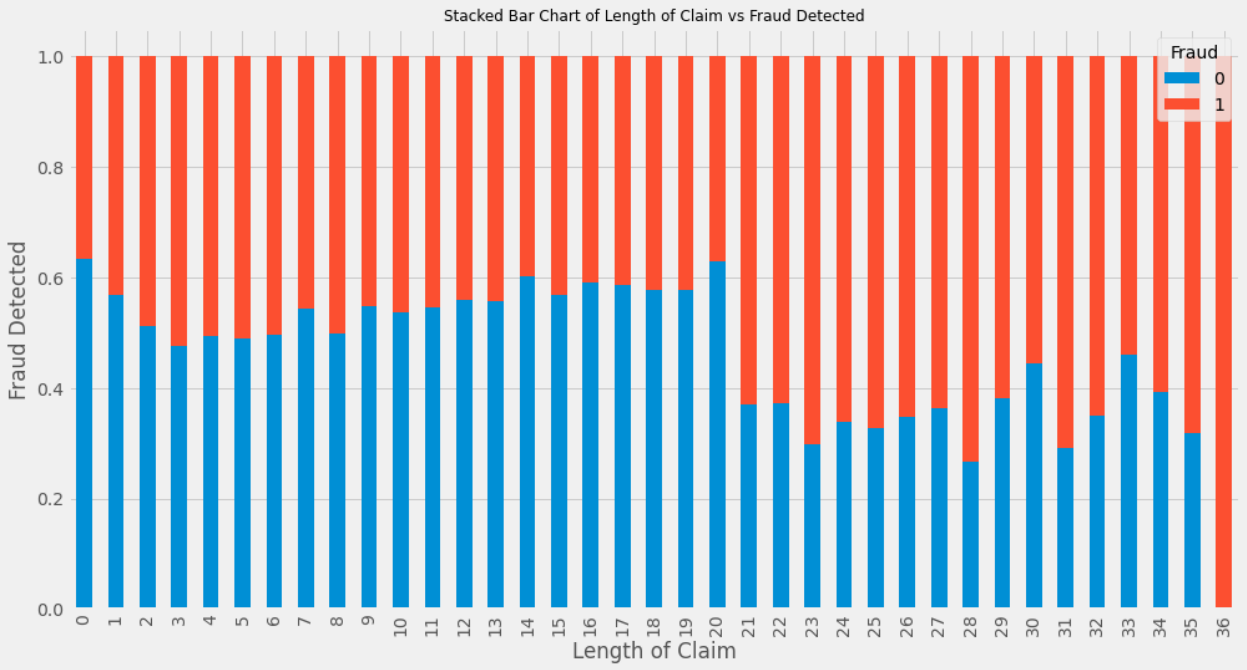
The scatter plot shows that this hypothesis is feasible. We can see that at any patient deductible amount, fraud providers' claims have higher reimbursement amount than non fraud providers' claims

c. Hypothesis 3: Claims with reimbursements amount having higher deviation from average costs for claims with similar diagnosis codes tend to be from fraudulent providers

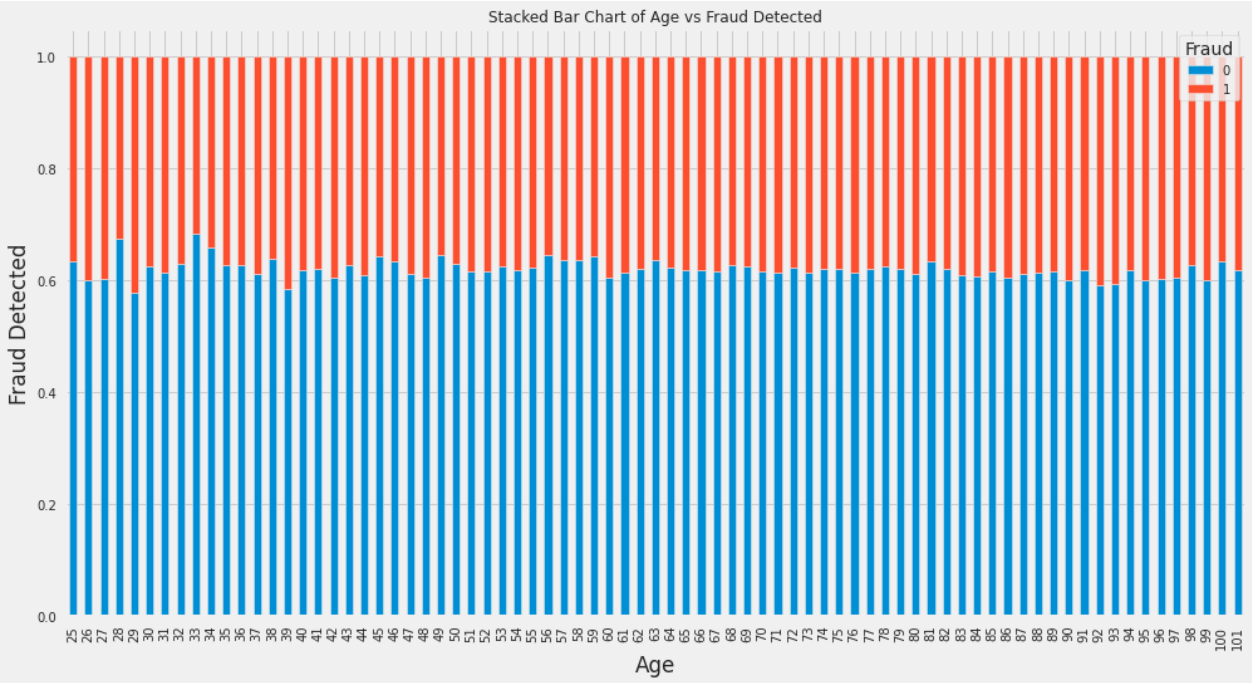


Scatter plot between the maximum deviation from average cost among claims with similar diagnosis codes and outpatient annual deductible amount shows that for the same patient deductible amount, claims from fraud providers tend to have larger deviation than claims from non-fraud providers.

d. Hypothesis 4: Providers may report longer length of service or length of hospitalization to claim more money



Stacked bar chart shows that length of claims, presumably representing length of medical service provided, is longer among fraud providers than non-fraud providers

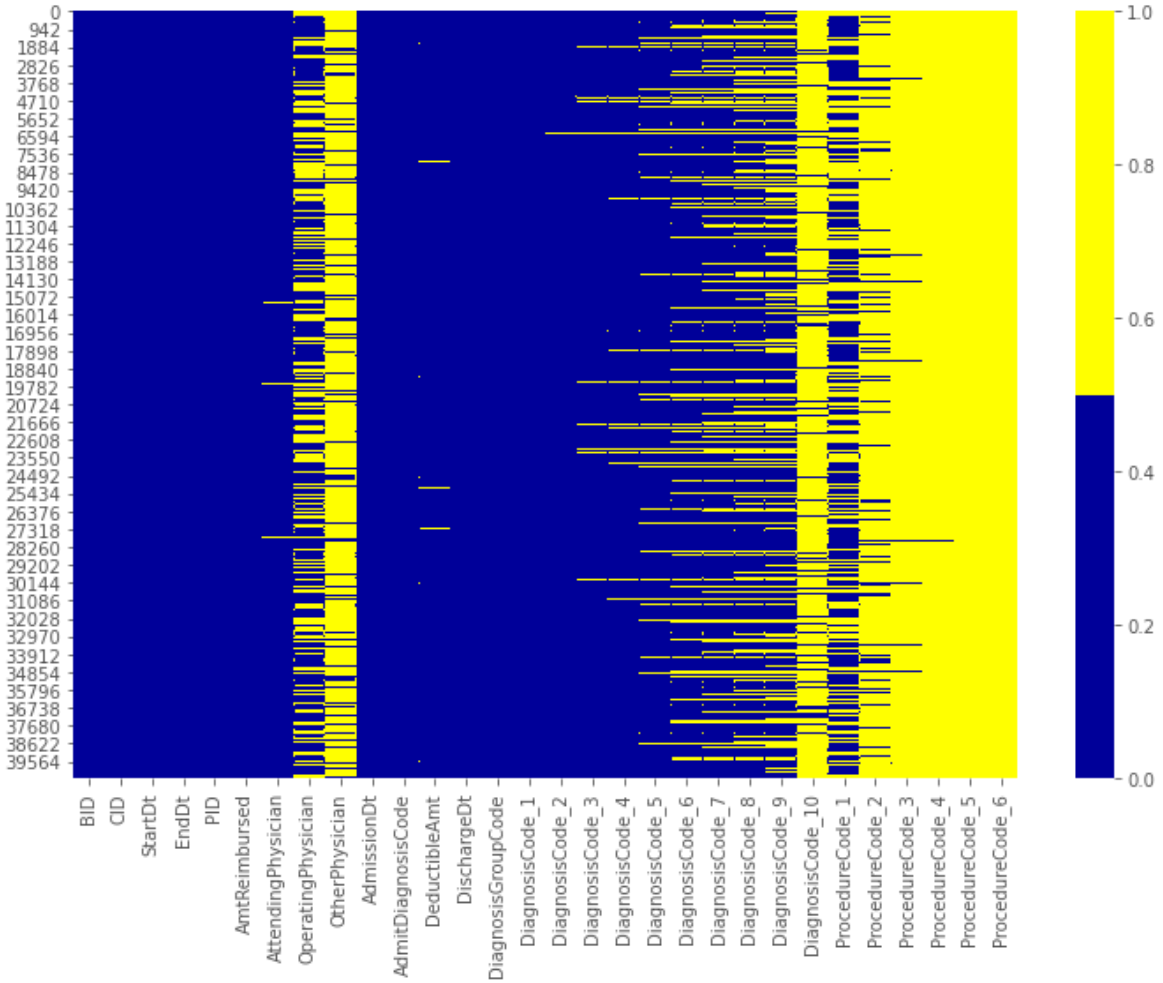


Meanwhile, patients' traits such as age and patients' pre-conditions (number of chronic conditions) don't seem to vary among fraud vs non-fraud providers.

3. Data Preparation

We concatenated the inpatient and outpatient datasets, merged it with the beneficiary dataset by Beneficiary ID (BID), then finally merged it with the providers dataset by Provider ID (PID) to get the 'master' dataset.

a) Missing data



The heat map above shows missing data (yellow) vs. non-missing data (blue). Since there is a lot of missing data in procedure code (PC) variables, we decided to replace these variables with a dummy variable indicating if the claim has or does not have a procedure code ('without PC': 0 or 'with PC': 1).

There is also a lot of missing data in some diagnosis code variables. Since reimbursement amounts depend on the codes filed, we wanted to create some variable that shows the impact of each diagnosis code on the total reimbursement amount. For each diagnosis code in the model, we calculated the deviation of AmtReimbursed (amount the provider is reimbursed) from average AmtReimbursed for all claims in the dataset with that diagnosis code, then found the maximum deviation amount for each claim. We called this variable max_dc_diff.

b) Irregular data (outliers)

We looked at descriptive statistics and distribution of each variables and decided to remove DeductibleAmt and NumOfMonths_PartACov and NumOfMonths_PartBCov because these variables seem have constant values

Minimum	0
5-th percentile	0
Q1	0
median	0
Q3	0
95-th percentile	1068
Maximum	1068
Range	1068
Interquartile range (IQR)	0

(Descriptive statistics for DeductibleAmt)

Minimum	0
5-th percentile	12
Q1	12
median	12
Q3	12
95-th percentile	12
Maximum	12
Range	12
Interquartile range (IQR)	0

(Descriptive statistics for NumOfMonths_PartACov and NumOfMonths_PartBCov)

A report of descriptive statistics of all variables in the model is included in the Appendix.

c) New variables

Based on the stated hypotheses we created new variables that we deem would affect fraudulent status:

- <len_claim = EndDt- StartDt> Length of claims - an estimation of length of service
- <len_claim = DischargeDt - AdmissionDt> Length of hospital stay
- <age_start = StartDt – DOB> Age of patient at start of healthcare service
- physician_count: No of physicians involved in claim
- chronic_count: No of chronic conditions patients have
- dc_count: No of diagnosis codes reported in claim
- inpatient_outpatient: Whether claim is inpatient (0) or outpatient (1)

d) Overview of the final ‘master’ dataset:

```
df.nunique()
```

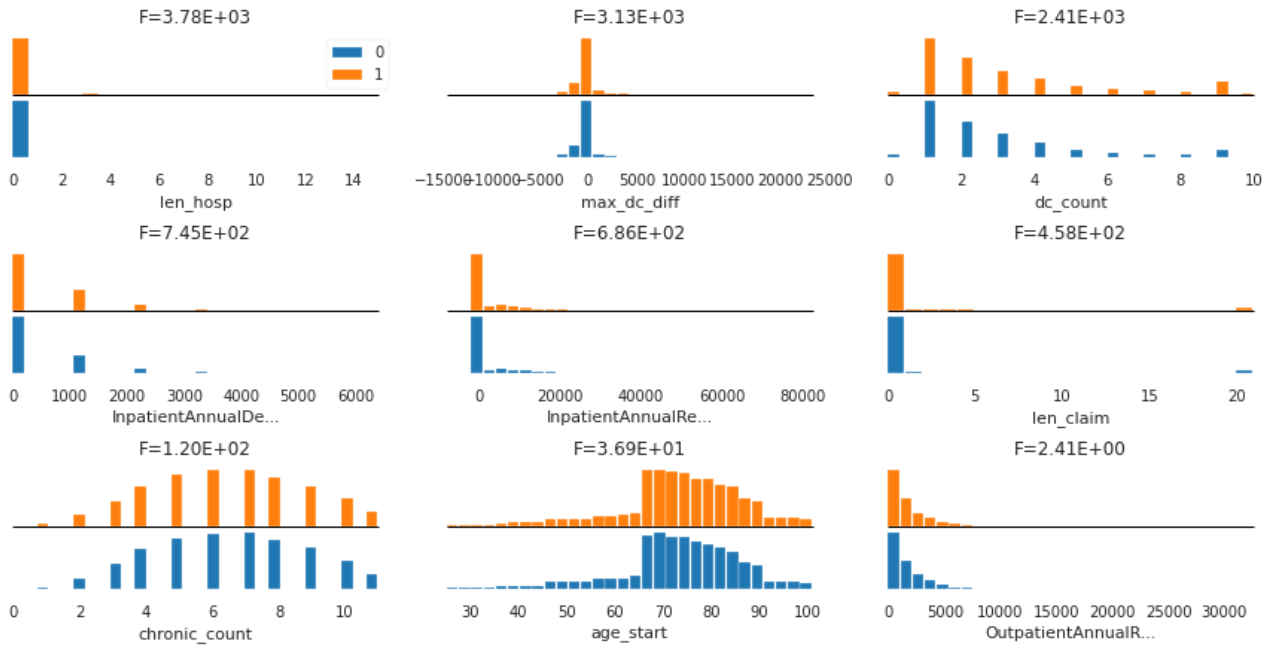
```
BID                                 138556
CID                                 558211
PID                                   5410
AmtReimbursed                          438
Procedure                                2
inpatient_outpatient                     2
Gender                                   2
RenalDisease                             2
Chronic_Alzheimer                        2
Chronic_Heartfailure                     2
Chronic_KidneyDisease                    2
Chronic_Cancer                           2
Chronic_ObstrPulmonary                   2
Chronic_Depression                       2
Chronic_Diabetes                         2
Chronic_IschemicHeart                    2
Chronic_Osteoporasis                     2
Chronic_rheumatoidarthritis              2
Chronic_stroke                           2
InpatientAnnualReimbursementAmt       3004
InpatientAnnualDeductibleAmt           147
OutpatientAnnualReimbursementAmt      2078
OutpatientAnnualDeductibleAmt          789
Fraud                                    2
len_claim                               37
len_hosp                                36
dc_count                                11
physician_count                          4
chronic_count                           12
max_dc_diff                          97169
age_start                            27085
dtype: int64
```

a) Unique values

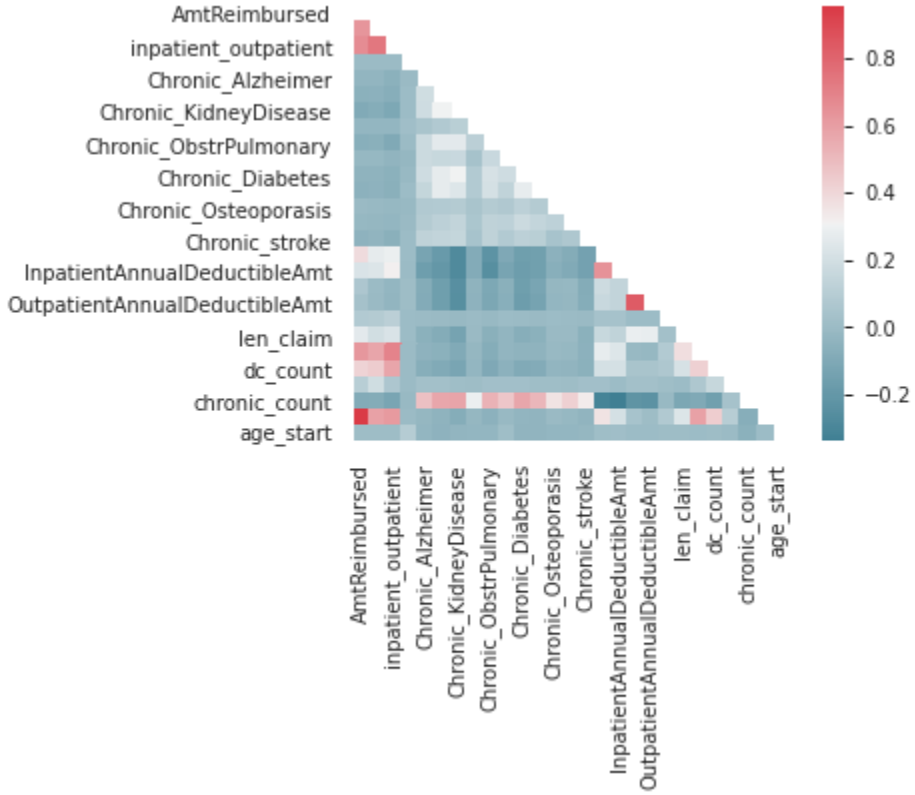
```
df.isnull().sum()/len(df)
```

```
BID                                 0.000000
CID                                 0.000000
PID                                 0.000000
AmtReimbursed                       0.000000
Procedure                           0.000000
inpatient_outpatient                0.000000
Gender                              0.000000
RenalDisease                        0.000000
Chronic_Alzheimer                   0.000000
Chronic_Heartfailure                0.000000
Chronic_KidneyDisease               0.000000
Chronic_Cancer                      0.000000
Chronic_ObstrPulmonary              0.000000
Chronic_Depression                  0.000000
Chronic_Diabetes                    0.000000
Chronic_IschemicHeart               0.000000
Chronic_Osteoporasis                0.000000
Chronic_rheumatoidarthritis         0.000000
Chronic_stroke                      0.000000
InpatientAnnualReimbursementAmt     0.000000
InpatientAnnualDeductibleAmt        0.000000
OutpatientAnnualReimbursementAmt    0.000000
OutpatientAnnualDeductibleAmt       0.000000
Fraud                               0.000000
len_claim                           0.000000
len_hosp                            0.000000
dc_count                            0.000000
physician_count                     0.000000
chronic_count                       0.000000
max_dc_diff                         0.018726
age_start                           0.000000
dtype: float64
```

b) % of missing values



Distribution of chosen variables is balanced.



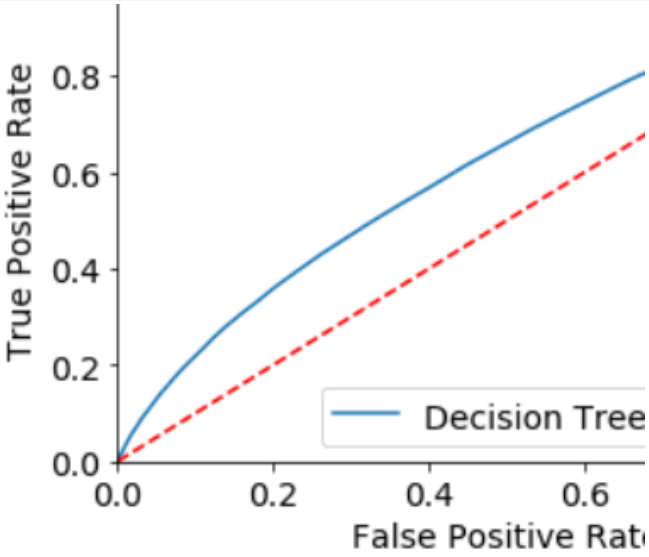
The correlation matrix (green shows positive and red shows negative relationship) shows no critical issue, but age_start seems to correlate highly with AmtReimbursed.

4. Modeling

The data that we have is from Medicare Insurance. The answer between yes/no for “whether the claim provider is fraudulent?” is a binary classification task. Therefore, we attempt to perform classification models in Python to detect fraud transactions. We compared different available classifiers including Decision Tree and RandomForest.

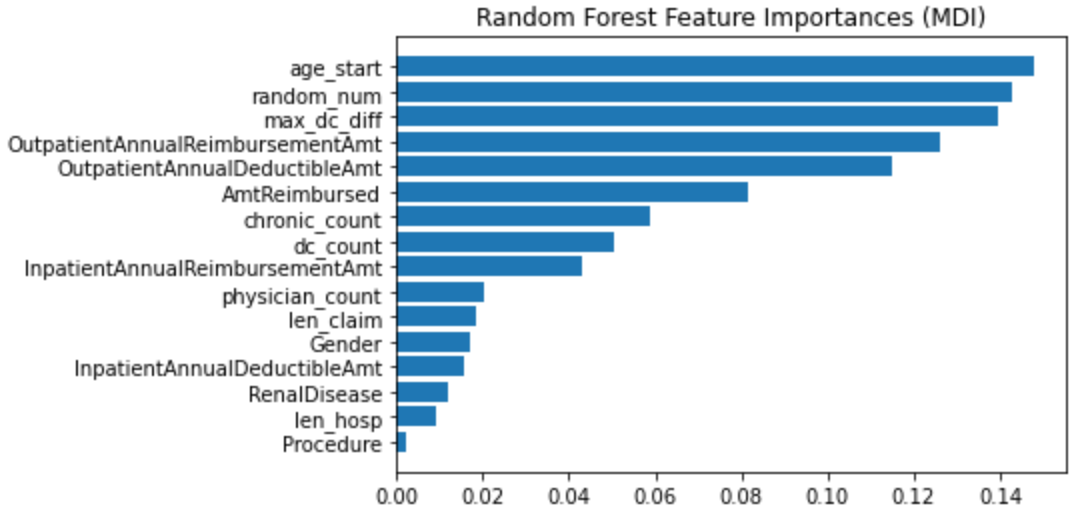
1. Decision Tree

	precision	recall	f1-score	support
0	0.66	0.88	0.76	103397
1	0.59	0.28	0.38	64067
accuracy			0.65	167464
macro avg	0.63	0.58	0.57	167464
weighted avg	0.64	0.65	0.61	167464

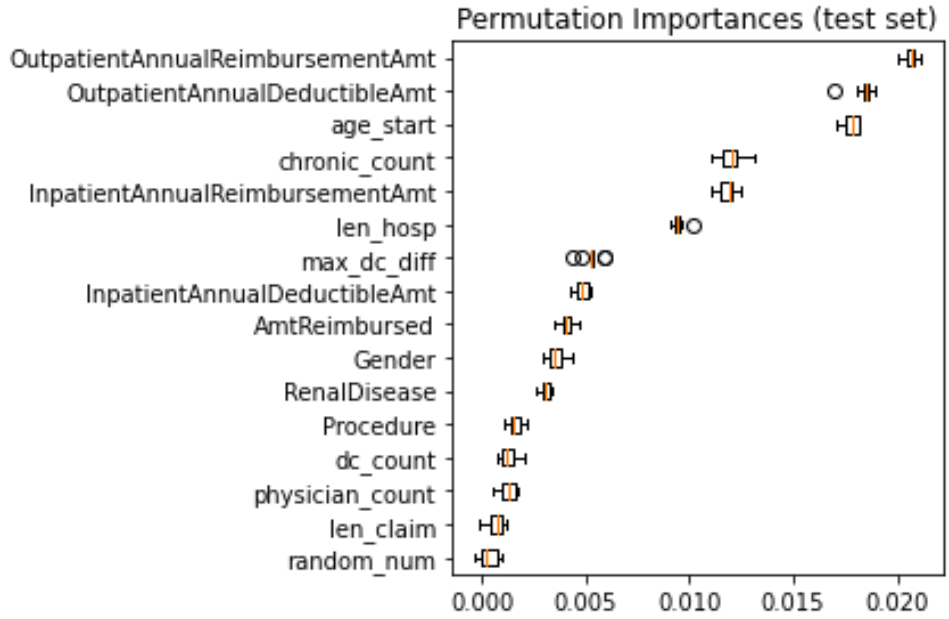


2. Random Forest

The impurity-based feature importance ranks the numerical features to be the most important features. As a result, the non-predictive random_num variable is ranked the second most important. Contrary to our hypothesis, the starting age of patients was ranked the most important feature to determine whether the claim is from a fraud or non-fraud provider. As expected, patient’s average annual deductible amount and claim’s amount reimbursed are important in identifying fraud/ non-fraud.



As an alternative, the permutation importances of rf are computed on a held out test set. This shows that the low cardinality categorical feature, outis the most important feature.
Also note that both random features have very low importances (close to 0) as expected.



C. Conclusion and Future Work

- We attempted to look at the most common diagnosis codes among claims of Fraud vs Non-fraud providers, and map it with ICD9/10 codes dictionary from the CMS to see potential outlier services on claims of Fraud providers. However, top 20 most frequent diagnosis codes are very similar among two groups (claims of Fraud vs. non Fraud providers):

Most common diagnosis/ procedure codes among fraud claims

Condition	ICD9/10	Proportion
Unspecified essential hypertension	4019	14.6%
Diabetes mellitus without mention of complication	25000	7.1%
Other and unspecified hyperlipidemia	2724	6.9%
Long-term (current) use of other medications	V5869	4.3%
Atrial fibrillation	42731	4.1%
Benign essential hypertension	4011	4.1%
Long-term (current) use of anticoagulants	V5861	3.6%
Unspecified acquired hypothyroidism	2449	3.4%
Pure hypercholesterolemia	2720	3.3%
Congestive heart failure, unspecified	4280	3.3%
Coronary atherosclerosis of native coronary artery	41401	3.1%
Esophageal reflux	53081	3.0%
Chronic airway obstruction, not elsewhere classified	496	2.7%
Anemia, unspecified	2859	2.6%

- For future work, we would like to use more Medicare claim data (i.e. Part D on Drug Coverage) to dive deeper into types of fraud including upcoding, provision of unnecessary procedures, and providing services with nurses and staff that should be provided by doctors.

- If we have more information on the physician's specialty, type of procedures performed and drug prescribed, we can try to predict the expected medical specialty of a physician based on type and count of procedures performed, then compare it with his/her actual specialty. For example, if the model predicts a physician as a dermatologist but his/her actual specialty is optometrist, he or she might be performing procedures indicating fraud
- Looking at historical data on providers who were identified as "Fraud" and understand their behaviors can be another way. Besides, we would like to implement other classifier algorithms.

References:

[1]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6181733/
[2]https://www.lifehealthpro.ca/news/ontarios-healthcare-system-losing-over-100-million-yearly-to-pharmacy-fraud-236893.aspx
[3]