Healthcare Provider Fraud

Group 11

Clara Tay Linn Qi (A0204413X)

Kok Ze Xuan (A0189596B)

Loo Hui Lin (A0203151B)

Wee Zhen Qi, Tarcius (A0190085E)

Wong Chung How, Brian (A0202022J)

Table of contents

01

Introduction

Problem Statement
Dataset Overview

Project timeline

Choosing a common pain point to target and learn more about the industry

Research

Making the data usable and digestible

Preprocessing

Finding the best model to predict fraudulent transactions

Evaluation

01

02

03

04

05

Data

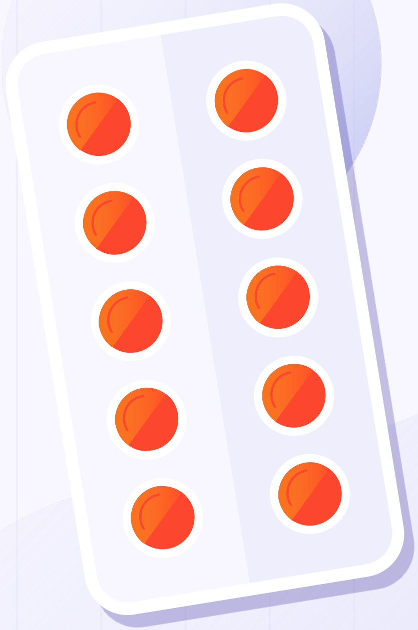
Finding an appropriate dataset

Modelling

Testing and comparing models

Medicare

- Healthcare insurance program
- Based in United States
- Target group: Senior Citizens (>65)



Problem Statement

Monetary loss

US$60 billion (2009) loss due to Medicare fraud

Increases premium

Insured are at the disadvantage of higher premiums

Wastage of resources

Resources are diverted from actual, truthful claims

Data source

United States Medicare claims in 2009 extracted from Kaggle

Datasets

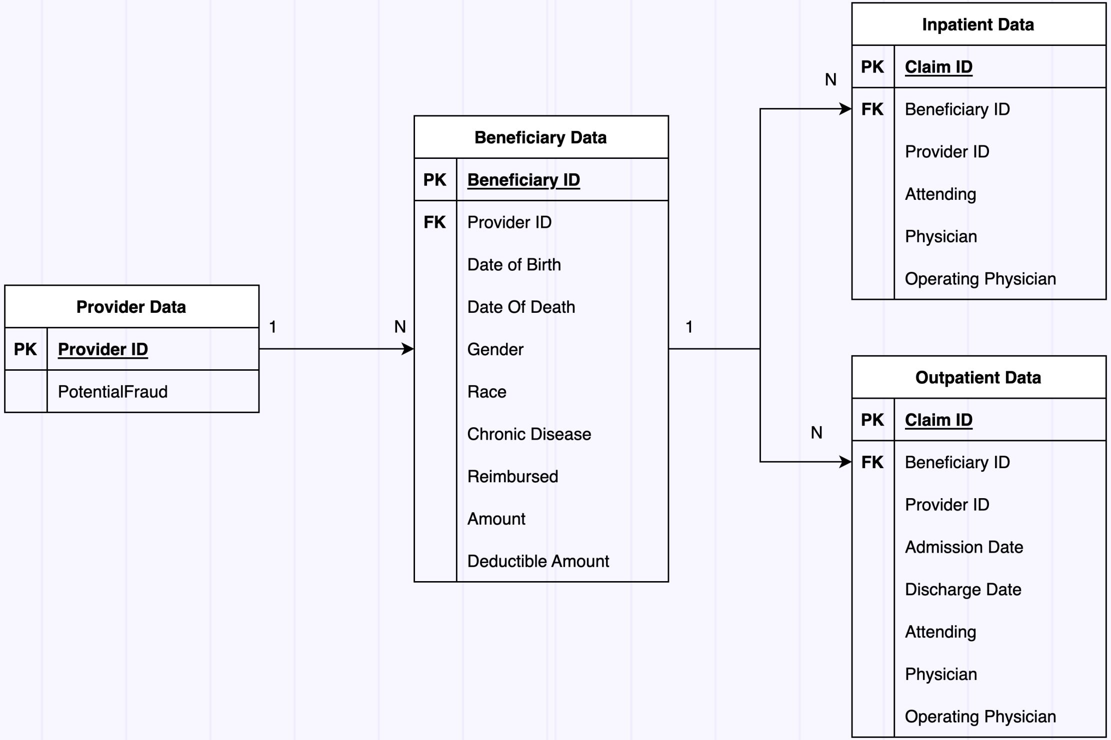
Beneficiary
- Demographics of beneficiary
- Reimbursement + Deductible amounts

Provider
- Provider ID
- Target variable ('PotentialFraud')

Inpatient
- Claims of patients admitted to the hospital

Outpatient
- Claims of patients not admitted to the hospital

Database Schema



02

Preprocessing

Exploratory Data Analysis
Feature Engineering

Overview of Beneficiary Data

Data Type	Variables
Object (4)	\`BeneID\`, \`DOB\`, \`DOD\`, \`RenalDiseaseIndicator\`,
Integer (21)	\`Gender\`, \`Race\`, \`State\`, \`County\`, \`NoOfMonths_PartACov\`, \`NoOfMonths_PartBCov\`, \`ChronicCond_Alzheimer\`, \`ChronicCond_Heartfailure\`, \`ChronicCond_KidneyDisease\`, \`ChronicCond_Cancer\`, \`ChronicCond_ObstrPulmonary\`, \`ChronicCond_Depression\`, \`ChronicCond_IschemicHeart\`, \`ChronicCond_Osteoporasis\`, \`ChronicCond_rheumatoidarthritis\`, \`ChronicCond_stroke\`, \`IPAnnualReimbursementAmt\`, IPAnnualDeductibleAmt\`, \`OPReimbursementAmt\`, \`OPDeductibleAmt\`

- Demographic information of the beneficiaries
- Identifies the chronic condition that they suffer from
- Inpatient and Outpatient Reimbursement and Deductible amounts

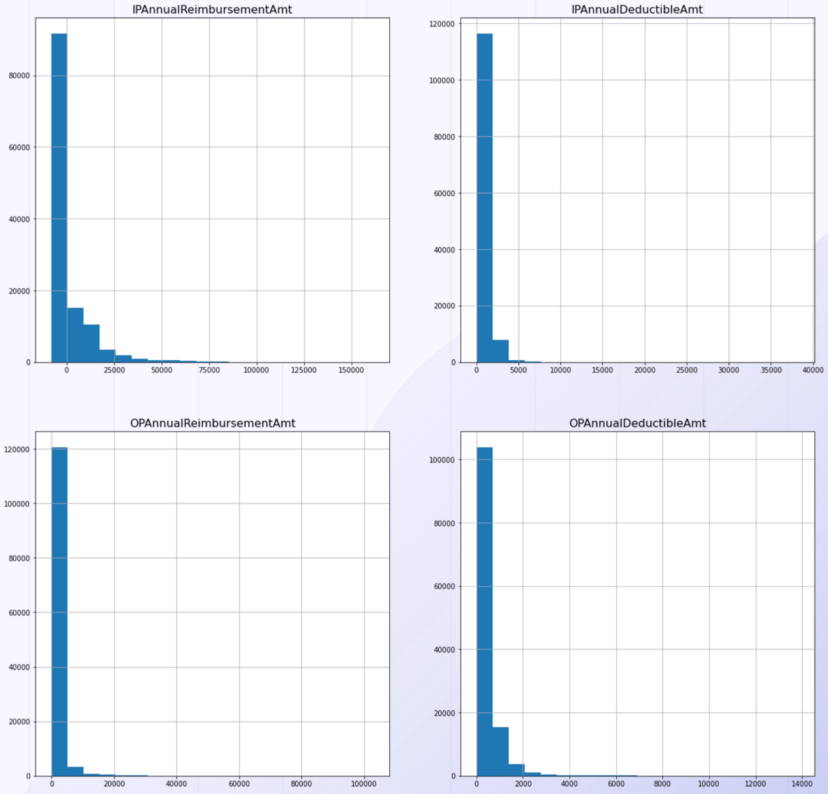
Reimbursement and Deductibles

Reimbursement

Medicare reimbursement is the amount which a **doctor or health facility receives** for providing medical services to a Medicare beneficiary.

Deductibles

Medicare deductible is the annual amount a person pays for covered healthcare services before their Medicare plan starts to pay.

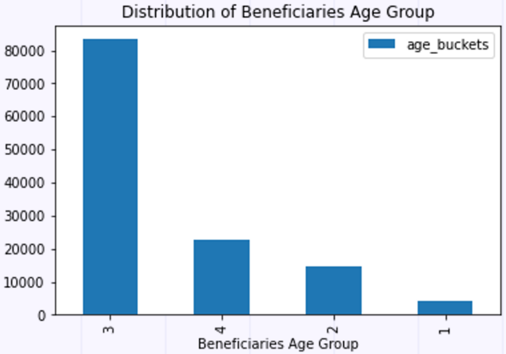


Demographics of Beneficiaries

01

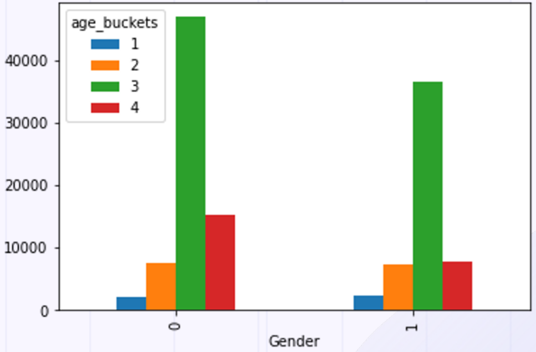
Generated Age Buckets

1. 19 – 44 years
2. 45 – 64 years
3. 65 – 84 years
4. 85 + years



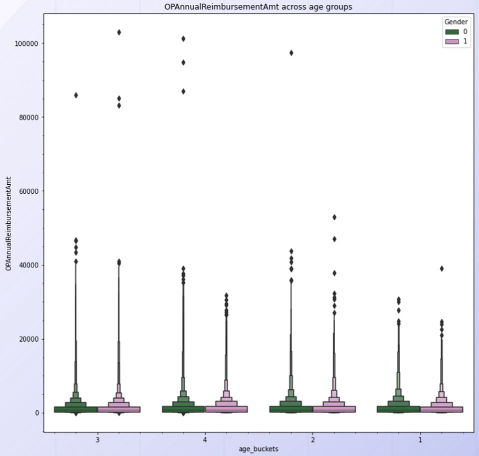
02

Explored for differences in age buckets and gender distribution of beneficiaries.



03

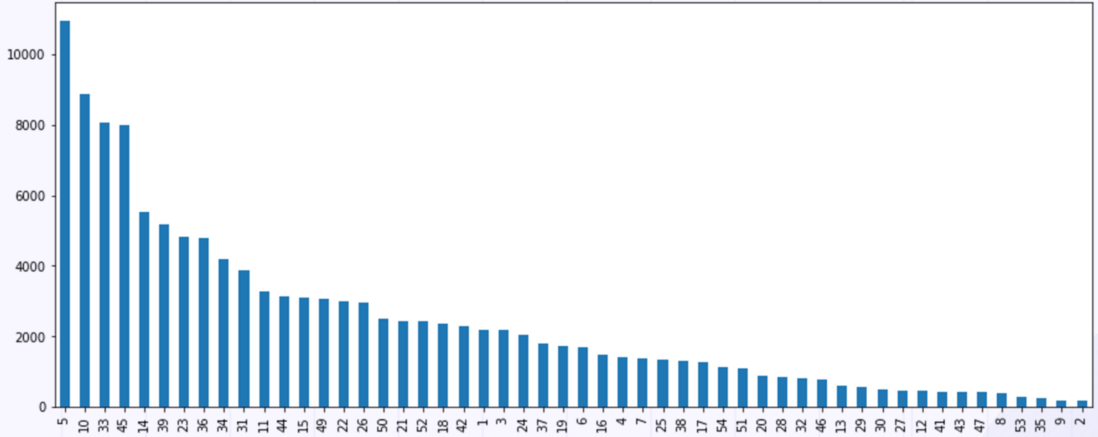
Explored for differences in reimbursement amounts across gender and age buckets.



Demographics of Beneficiaries

04

Explored State and County distribution of beneficiaries



05

Most prevalent conditions in each State

Top 3:

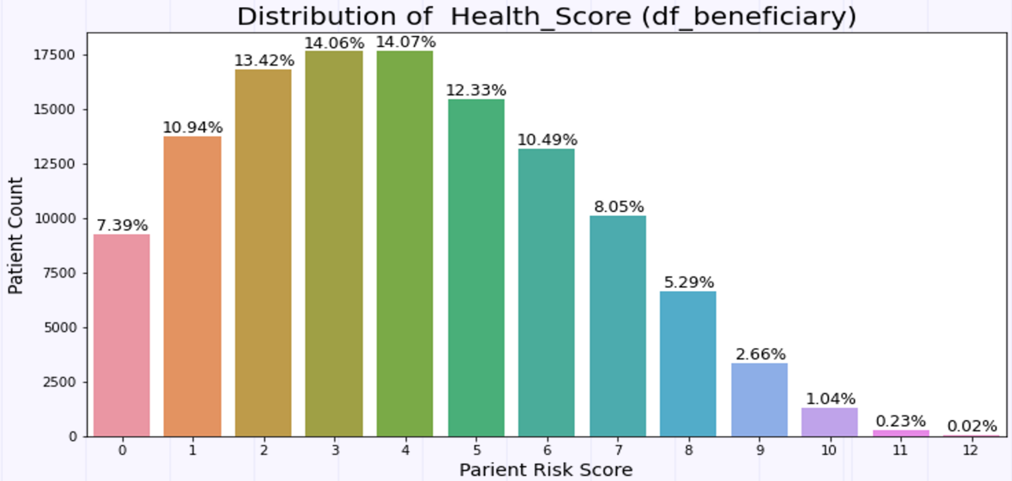
1. Stroke
2. Cancer
3. ObstrPulmonary (condition relating to the lungs)

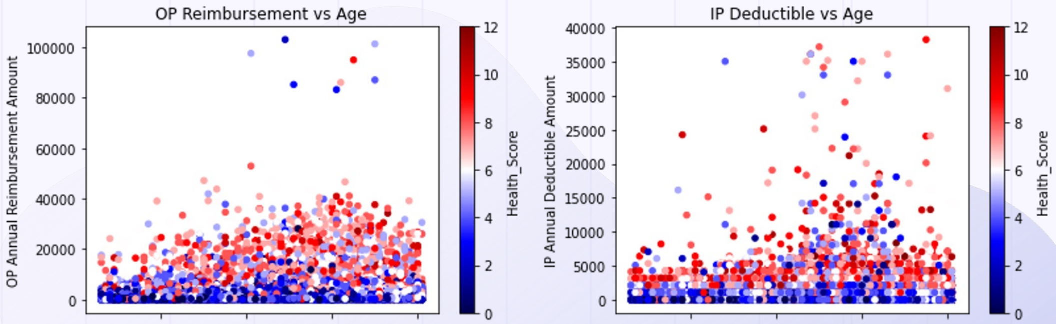
Correlation between Conditions

Should there be any correlation between conditions, it could help explain correlations between diagnosis codes and claim procedure codes in the Patient dataset.

However, there are **little to no correlation between the conditions**, other than Renal Disease and Kidney Disease, which both relates to the Kidney.

Patient Health Score





- Created a Patient Health Score which indicates the number of chronic conditions the beneficiary suffers from.
- The higher the score, the less healthy the person is.
- Explored possible relationships between the deductible and reimbursement amount, age and health score

Overview of Patient Data

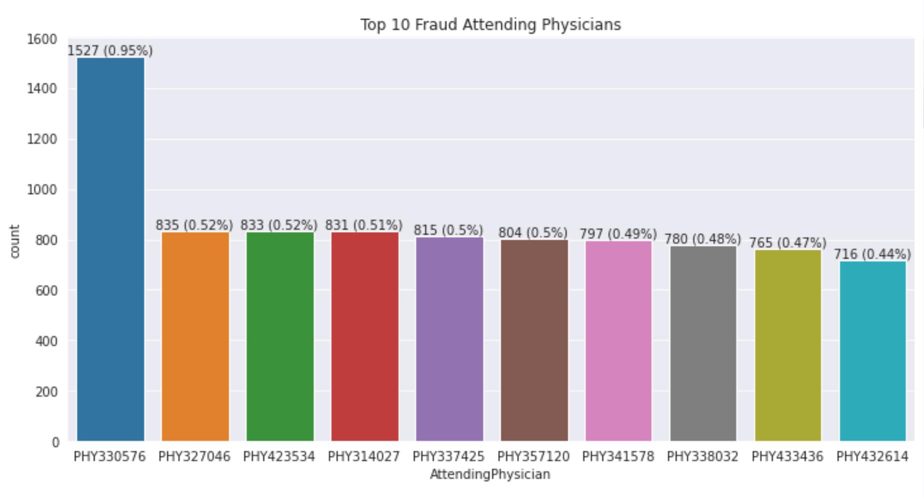
Data Type	Variables
Object (23)	\`BeneID\`, \`ClaimID\`, \`ClaimStartDt\`, \`ClaimEndDt\`, \`Provider\`, \`AttendingPhysician\`, \`OperatingPhysician\`, \`OtherPhysician\`, **\`AdmissionDt\`**, \`ClmAdmitDiagnosisCode\`, **\`DischargeDt\`, \`DiagnosisGroupCode\`,** \`ClmDiagnosisCode_1\`, \`ClmDiagnosisCode_2\`, \`ClmDiagnosisCode_3\`, \`ClmDiagnosisCode_4\`, \`ClmDiagnosisCode_5\`, \`ClmDiagnosisCode_6\`, \`ClmDiagnosisCode_7\`, \`ClmDiagnosisCode_8\`, \`ClmDiagnosisCode_9\`, \`ClmDiagnosisCode_10\`, \`source\`
Integer (1)	\`InscClaimAmtReimbursed\`
Float (7)	\`DeductibleAmtPaid\`, \`ClmProcedureCode_1\`, \`ClmProcedureCode_2\`, \`ClmProcedureCode_3\`, \`ClmProcedureCode_4\`, \`ClmProcedureCode_5\`, \`ClmProcedureCode_6\`

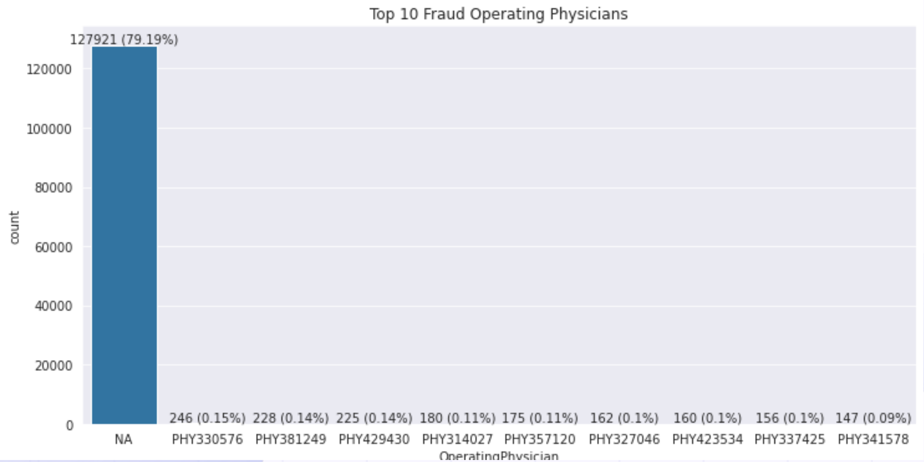
- Merged both inpatient and outpatient datasets
- Information on claims of inpatient and outpatients that were admitted to the hospital
- Highlighted are the additional columns from inpatient dataset

Physician ID

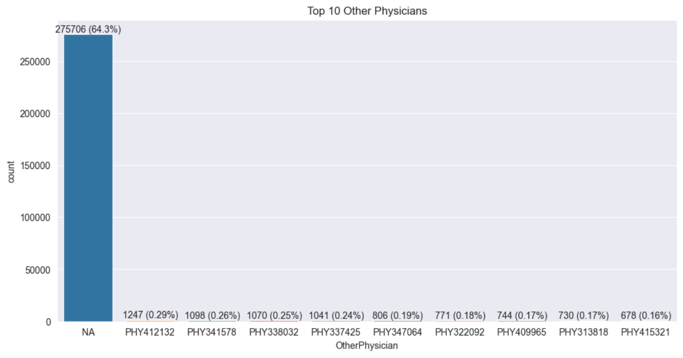
- Certain physician IDs only appeared in fraudulent claims which could be a good predictor

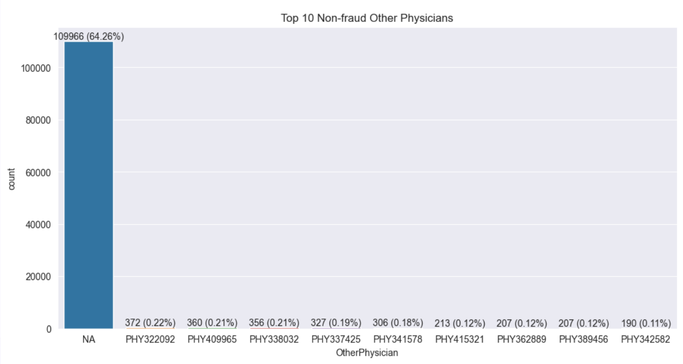
- Can be used to build a dictionary of fraudulent physicians which the claim approver can closely monitor

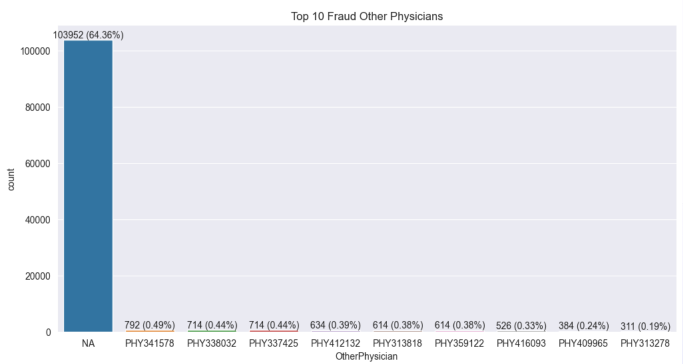




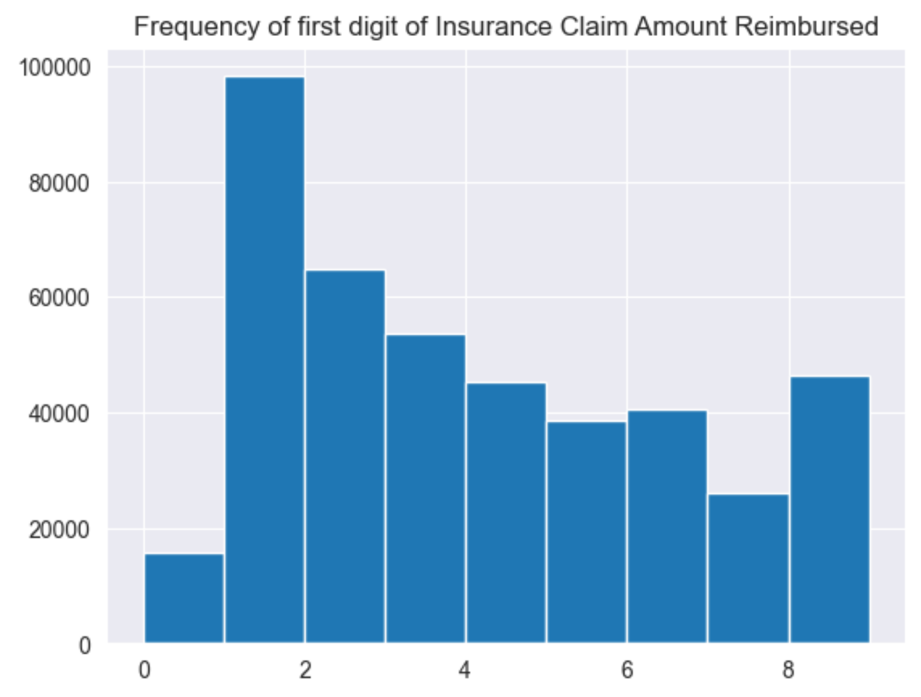
Physician ID





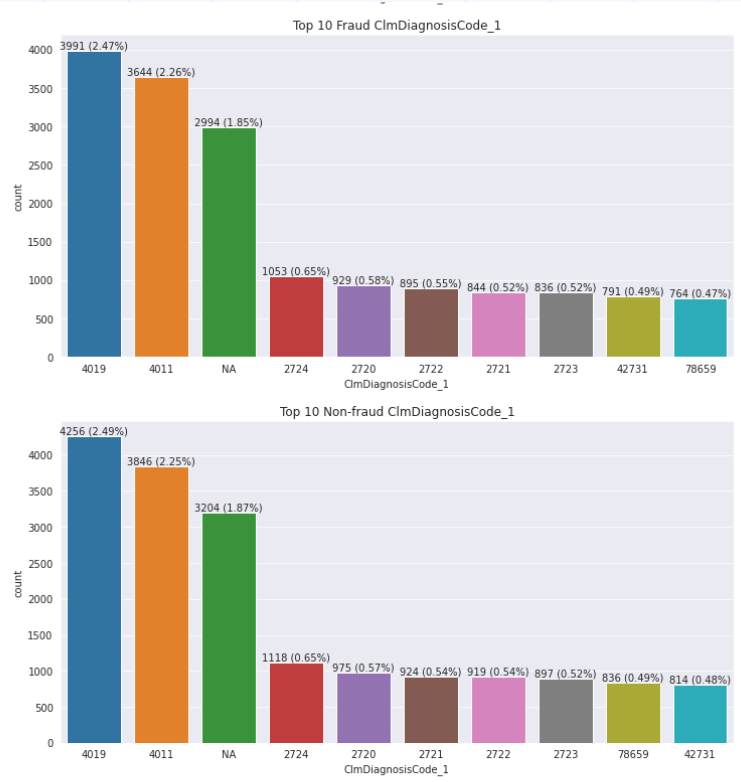


Insurance Claim Amount



- According to Benford's law, the larger digits have a smaller probability of occurring
- The last value should have the least number of cases
- More prevalent than the 4 preceding values

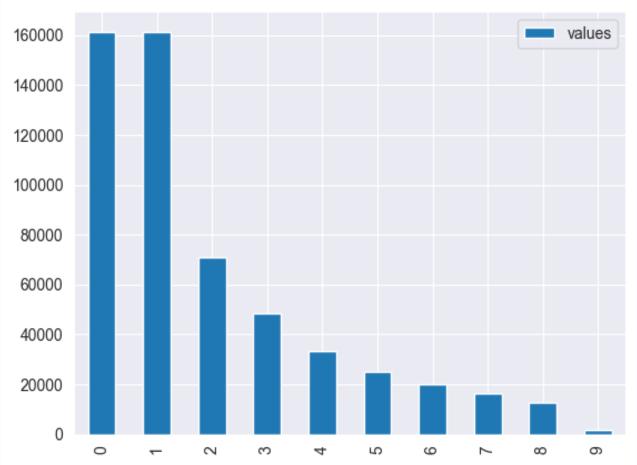
Diagnosis and Procedure Codes



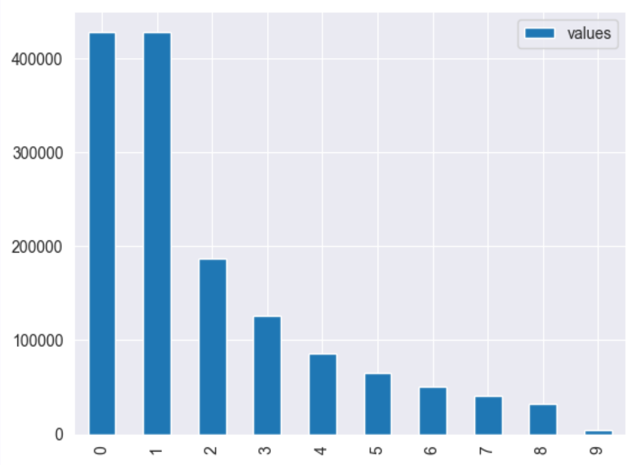
Example of one of the Fraud vs Non-Fraud plot

- We explored Diagnosis and Procedure Codes to see if there were some which would be more prevalent amongst the fraud cases.
- The distribution of codes are similar across both diagnosis and procedure codes.
- There were no significant differences in distribution patterns.

Diagnosis and Procedure Codes



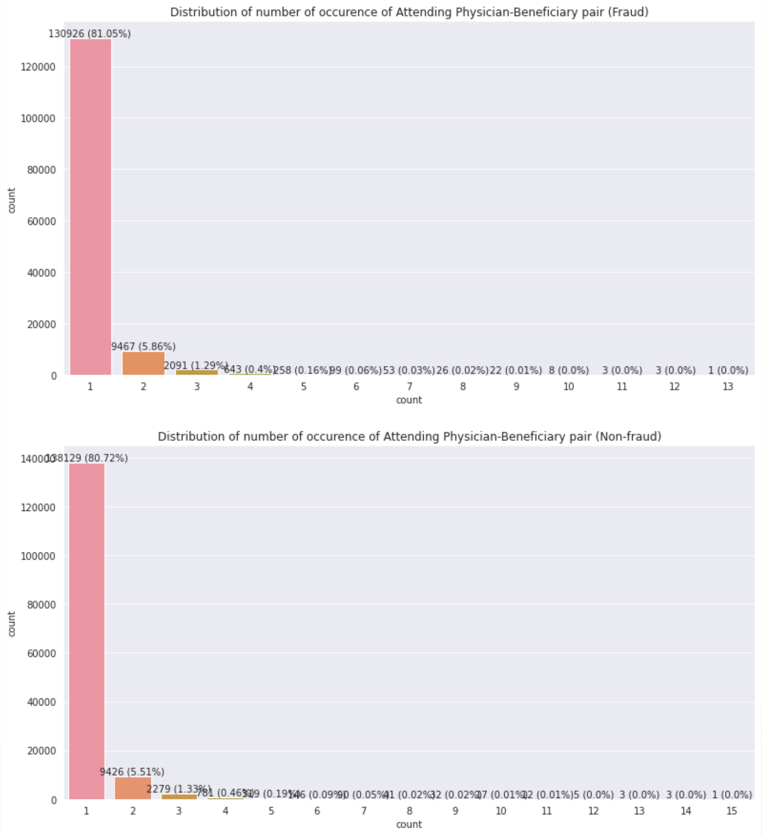
Non-Fraudulent Claims across Codes



Fraudulent Claims across Codes

- Distribution among across each code for fraudulent / non fraudulent cases were also were largely similar
- Codes may not be useful to predict fraud

Physician – Beneficiary Pair



Healthcare providers and beneficiaries work together to submit the Medicare claims.

Physician – Beneficiary pairs as it could be indicative of fraud.

However, there are no significant differences between the fraud and non-fraud pairs.

03

Modelling

Training
Evaluation

Model Comparison

Model	Advantages	Disadvantages
Logistic Regression	Easy to implement and interpret the results	Assumes linearity and not suitable for our high dimensionality dataset
Decision Tree	Easy to interpret, does not require data to be of a certain distribution	Unstable
Random Forest	More accurate than Decision Tree	Hard to interpret results
XGBoost	Parallel processing, faster computation time	Hard to interpret results
Neural Network	Can learn complex relationships which benefits our large dataset	Prone to overfitting

Evaluation

Model	Recall	F1
Logistic Regression	0.401	0.433
Random Forest	0.135	0.216
Decision Tree	0.272	0.362
XGBoost	0.314	0.406
Neural Networks	0.584	0.464

Integration of Model into Business Process

- Run submitted claims through the model
- Select claims that have been predicted to be potential fraud
- Send these claims to domain experts for further evaluation
- Reduce workload of domain experts with reduced cases to evaluate

04

Conclusion

Limitations
Future Extensions

Limitations

Non-semantic features

ClmDiagnosisCode & ClmProcedureCode (numerically coded)

Relationship between the features provides useful information
(eg. Diagnosis shows no kidney issues but Procedure includes dialysis)

Lack of semantic information thus missing out on potentially highly predictive feature

Limitations

Computational Power

Reduce number of categories for encoding hence losing features and information

Limited hyperparameter tuning
→ Reduce number of epochs
→ Tuning only selected hyperparameters

Future Extensions

Custom Ensemble Methods

Train a selection of different base models

↓

Use their predictions to get a new dataset with targets

↓

Train a meta-model on this new dataset

Future Extensions

Application to related problems

Providers and beneficiaries often work together to commit the fraud

↓

Understand interactions between these parties in a graph-based approach

↓

Predict potential beneficiary frauds by inferring from their interactions

Thank you

Please keep this slide for attribution