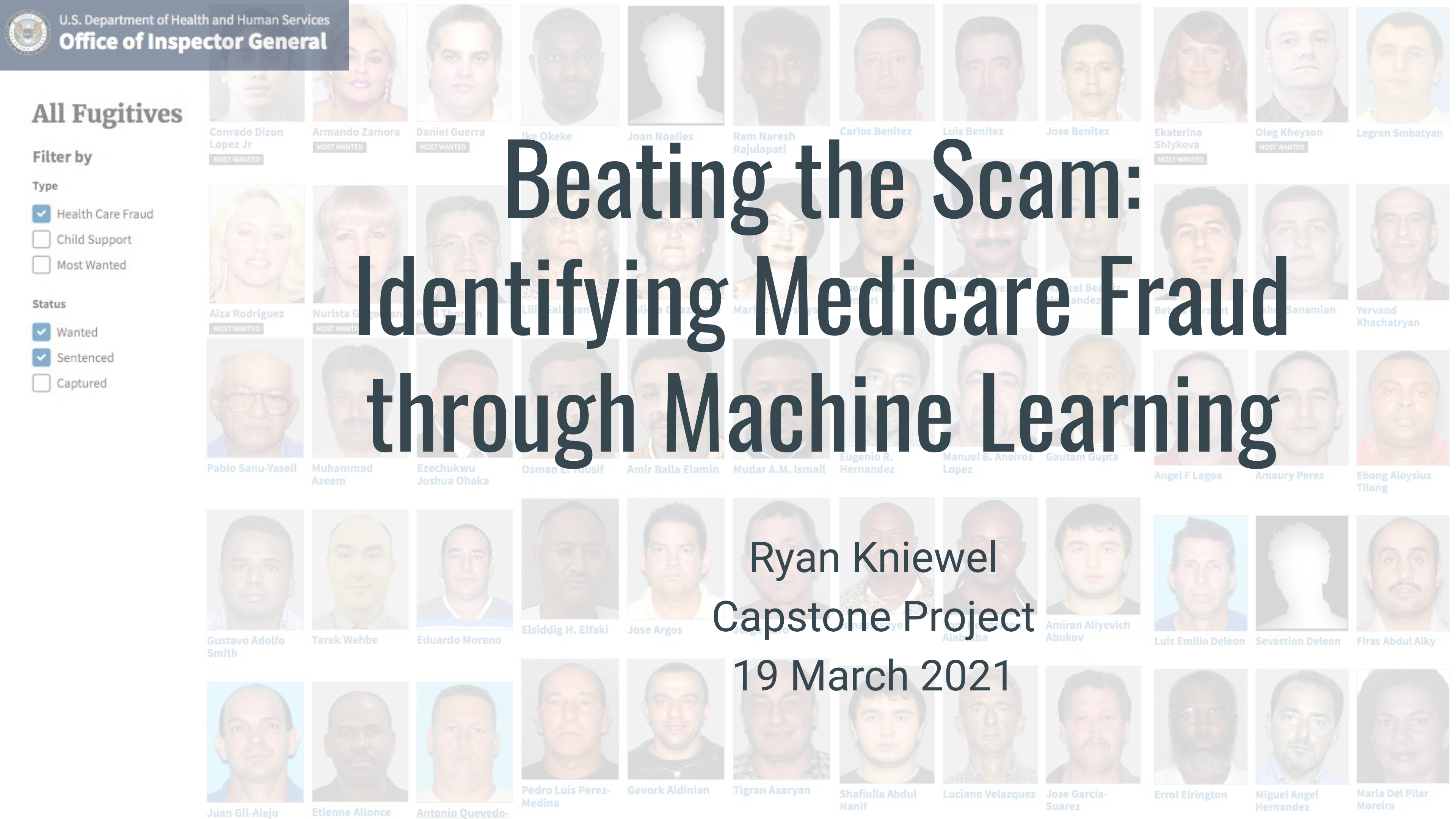


Beating the Scam: Identifying Medicare Fraud through Machine Learning

Ryan Kniewel

Capstone Project

19 March 2021

Medicare Fraud Schemes

Losses due to Medicare fraud in the United States are estimated to inflate public health expenditures from between 3 and 10%

This means that up to $300 billion is misappropriated from Medicare patients into the hands of criminals on an annual basis

National Health Care Anti-Fraud Association and FBI Financial Crimes Report 2009

Medicare Fraud Schemes

Losses due to Medicare fraud in the United States are estimated to inflate public health expenditures from between 3 and 10%

This means that up to $300 billion is misappropriated from Medicare patients into the hands of criminals on an annual basis

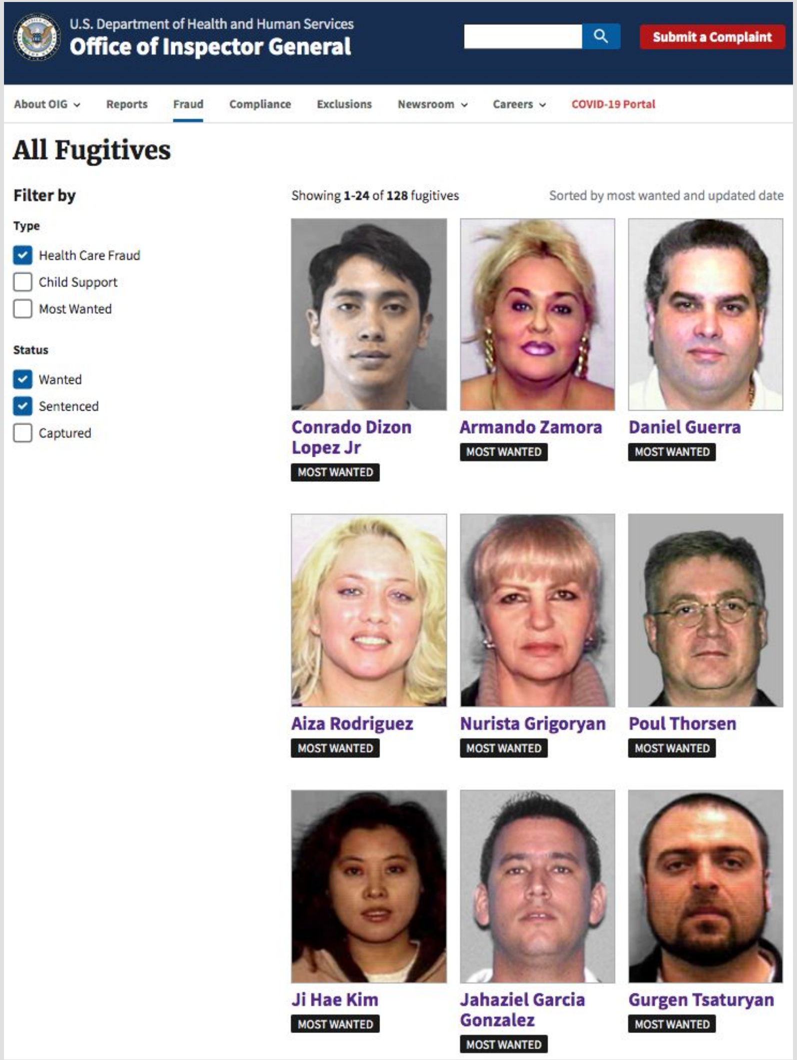
- Duplicate Claims
- Kickbacks
- Billing for Services not Rendered
- Upcoding of Services
- Unbundling
- Excessive or Unnecessary Services

National Health Care Anti-Fraud Association and FBI Financial Crimes Report 2009

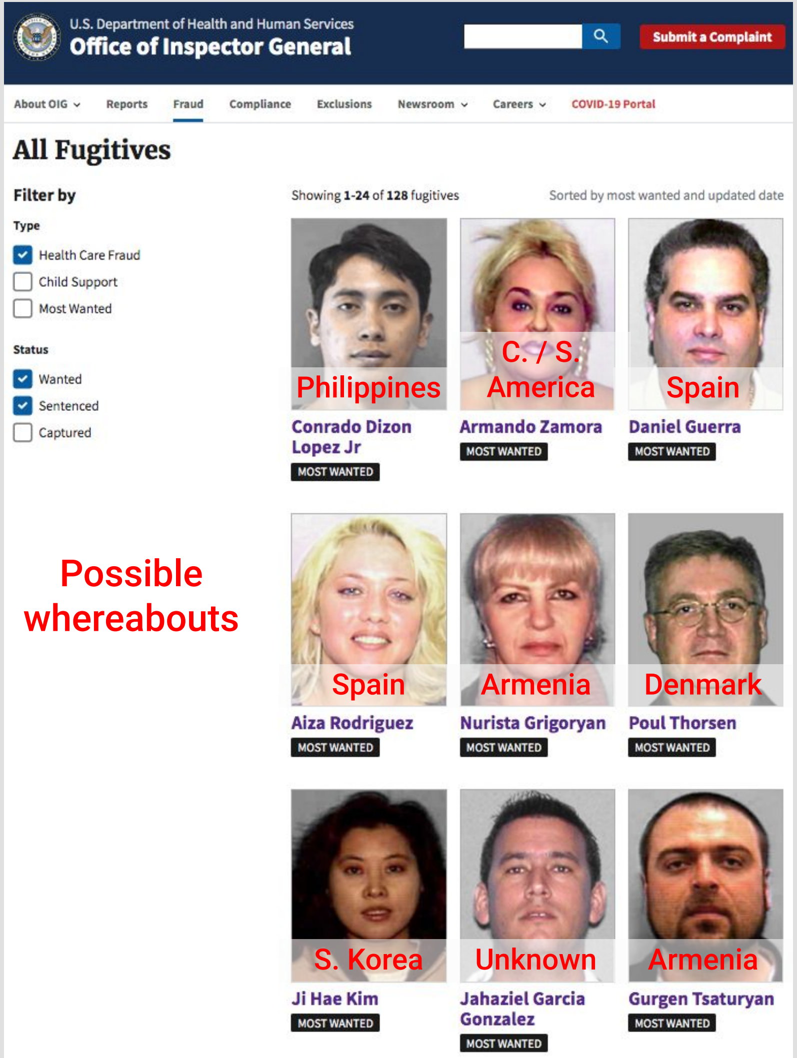
Involvement by International Criminal Enterprises

“ ... the U.S. Department of Justice announced the indictments of 102 members of an Armenian crime syndicate, accusing them of participating in massive health care fraud schemes. The gang operated a health clinic in Miami that paid individuals to refer “patients” of staged accidents … “for treatments that were either not medically necessary or were not provided.”

The Armenians allegedly had ties to Russian and Eastern European crime gangs as well as the Mexican Mafia.”



US HHS Office of Inspector General Website



US HHS Office of Inspector General Website

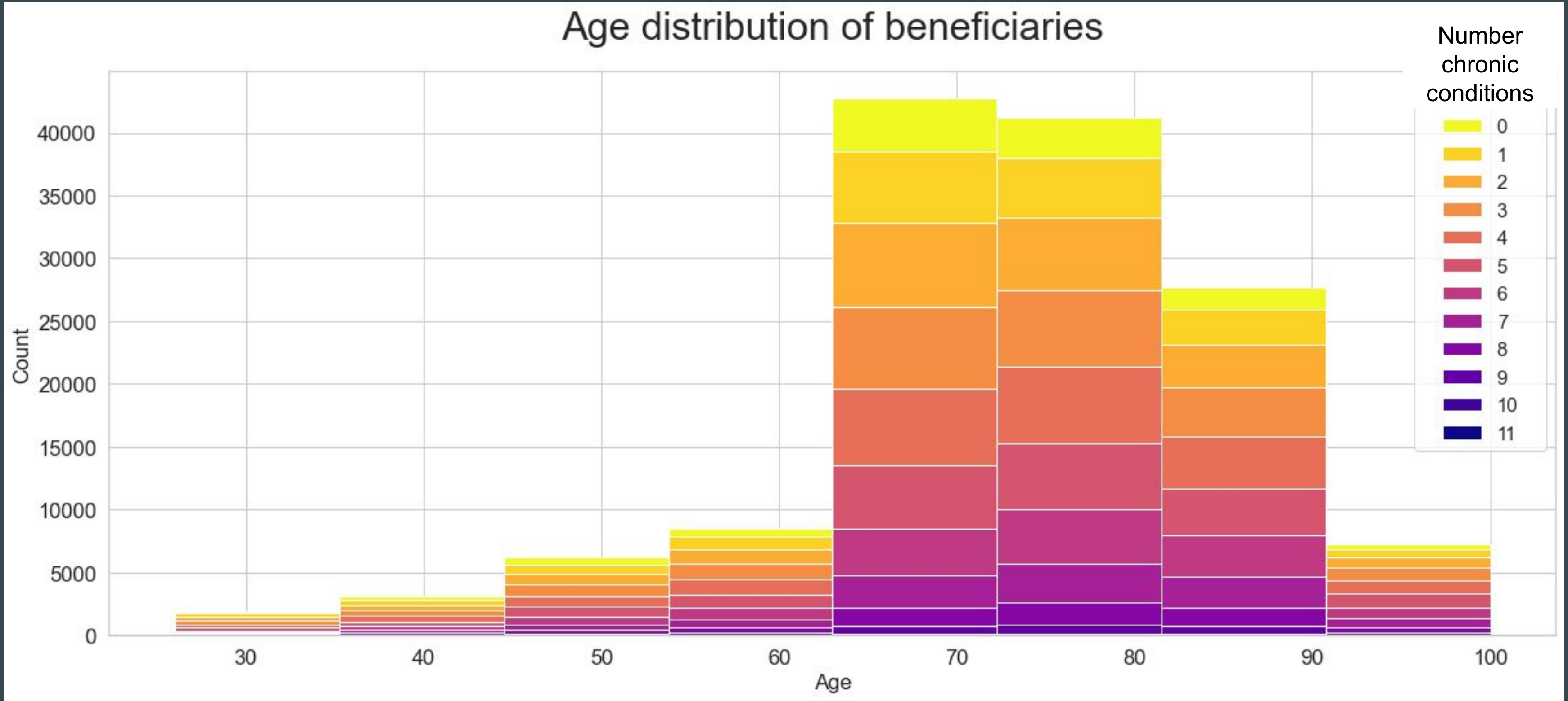
Sample of Medicare Claims Data from 2009

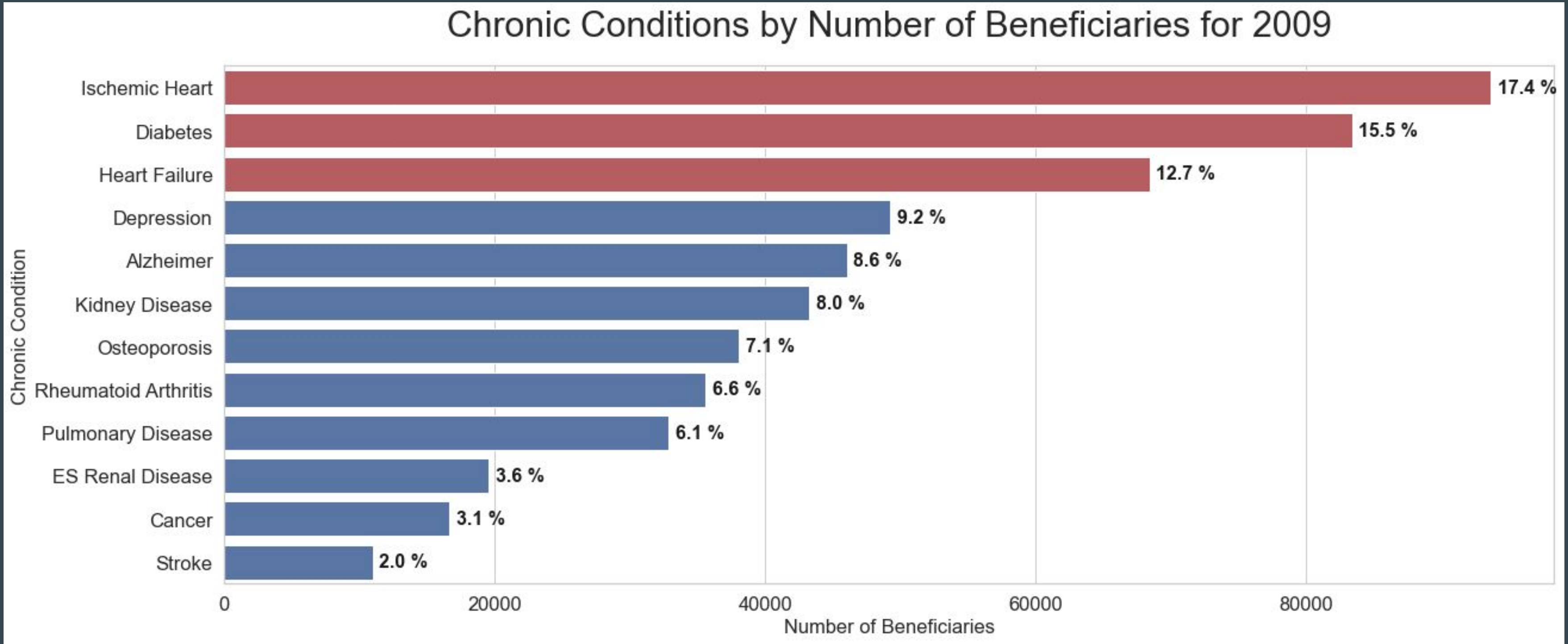
Data in four parts

- Beneficiary data is encoded for gender, race, state, county, chronic conditions
  - Encoding key is available in the CMS Record Data Dictionary 2009
- Claims data separate for inpatient and outpatient treatments
- Potentially fraudulent providers data
  - Not clear what potentially fraudulent signifies

Center for Medicaid Services, Research Data Distribution Center LDS Denominator Record Data Dictionary 2009

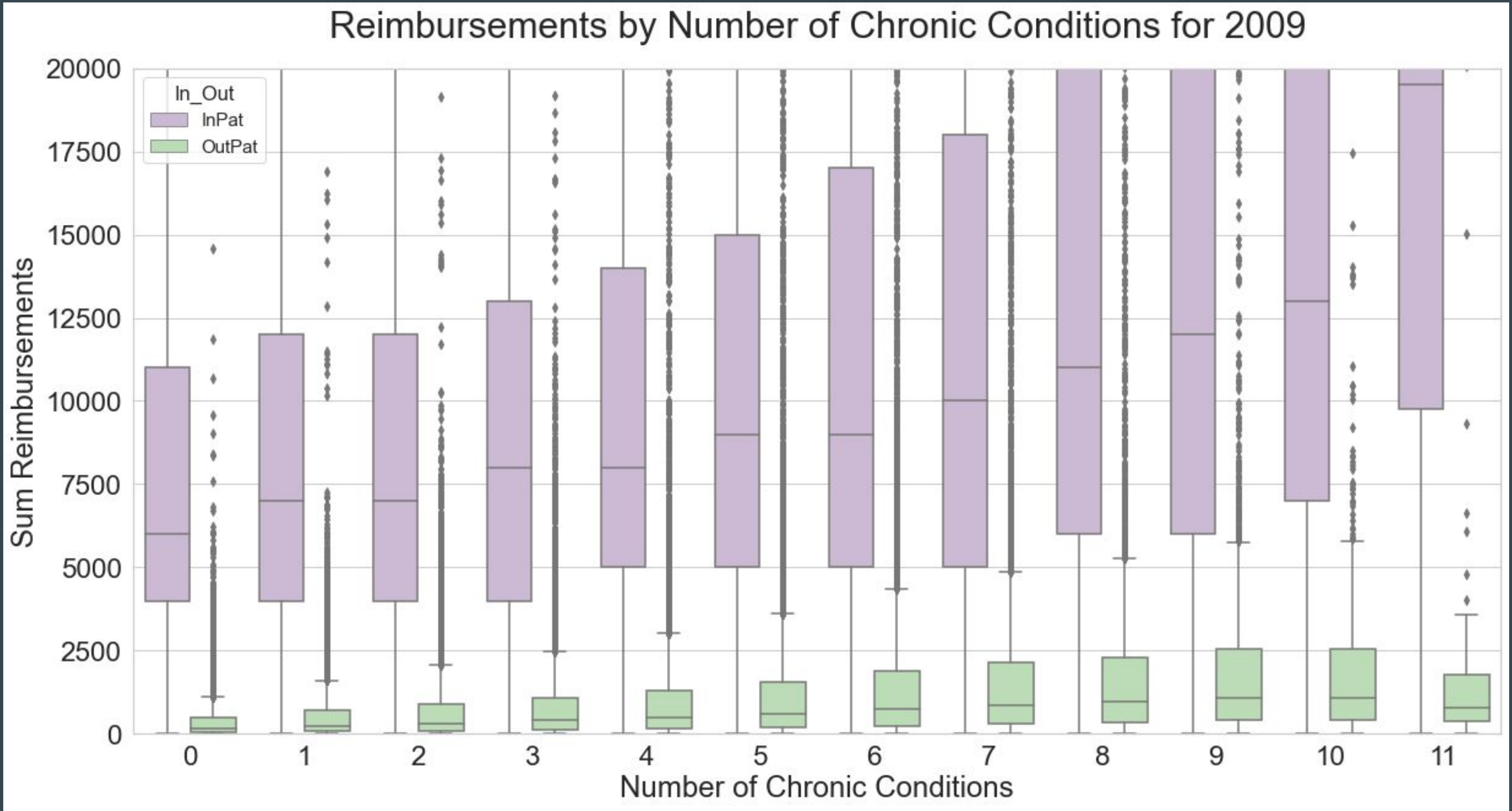
Demographics of Beneficiaries


Age distribution of beneficiaries
Count
Age
40000
35000
30000
25000
20000
15000
10000
5000
0
30
40
50
60
70
80
90
100
Number chronic conditions
0
1
2
3
4
5
6
7
8
9
10
11

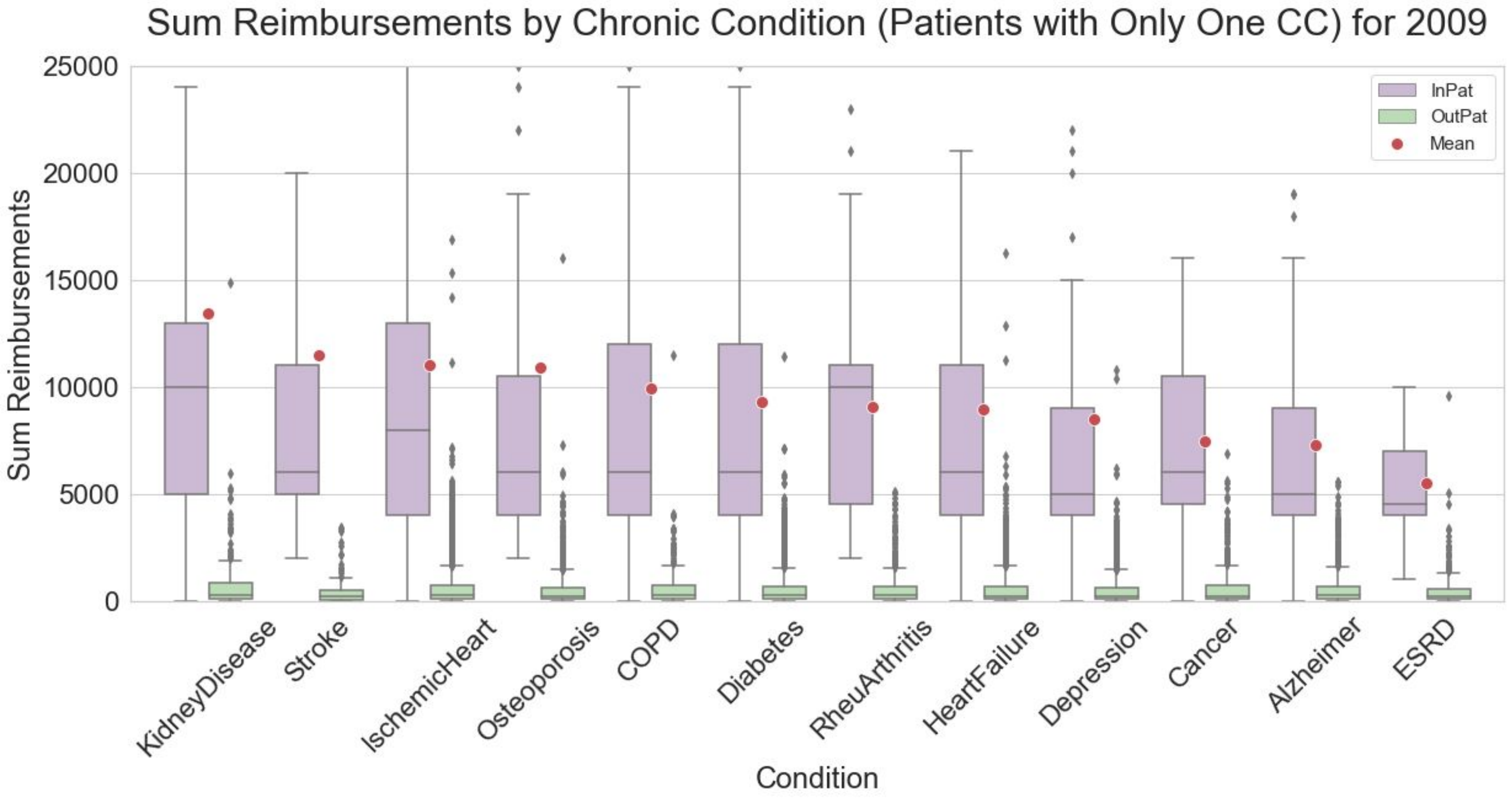


Chronic Conditions by Number of Beneficiaries for 2009
Chronic Condition
Ischemic Heart
Diabetes
Heart Failure
Depression
Alzheimer
Kidney Disease
Osteoporosis
Rheumatoid Arthritis
Pulmonary Disease
ES Renal Disease
Cancer
Stroke
17.4 %
15.5 %
12.7 %
9.2 %
8.6 %
8.0 %
7.1 %
6.6 %
6.1 %
3.6 %
3.1 %
2.0 %
0
20000
40000
60000
80000
Number of Beneficiaries

Reimbursements


Reimbursements by Number of Chronic Conditions for 2009
In_Out
InPat
OutPat
Sum Reimbursements
20000
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
3
4
5
6
7
8
9
10
11
Number of Chronic Conditions

Chronic Conditions

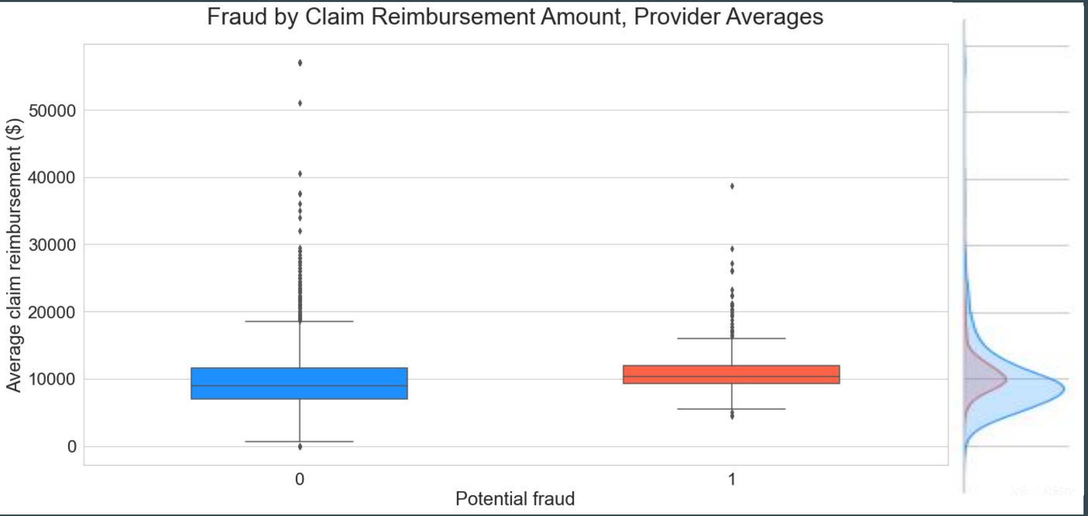

Sum Reimbursements by Chronic Condition (Patients with Only One CC) for 2009
InPat
OutPat
Mean
Sum Reimbursements
25000
20000
15000
10000
5000
0
KidneyDisease
Stroke
IschemicHeart
Osteoporosis
COPD
Diabetes
RheuArthritis
HeartFailure
Depression
Cancer
Alzheimer
ESRD
Condition

Provider-Focused EDA

Inpatient claims only

Two sample T-test:

p-value = 1.2×10^{-4}

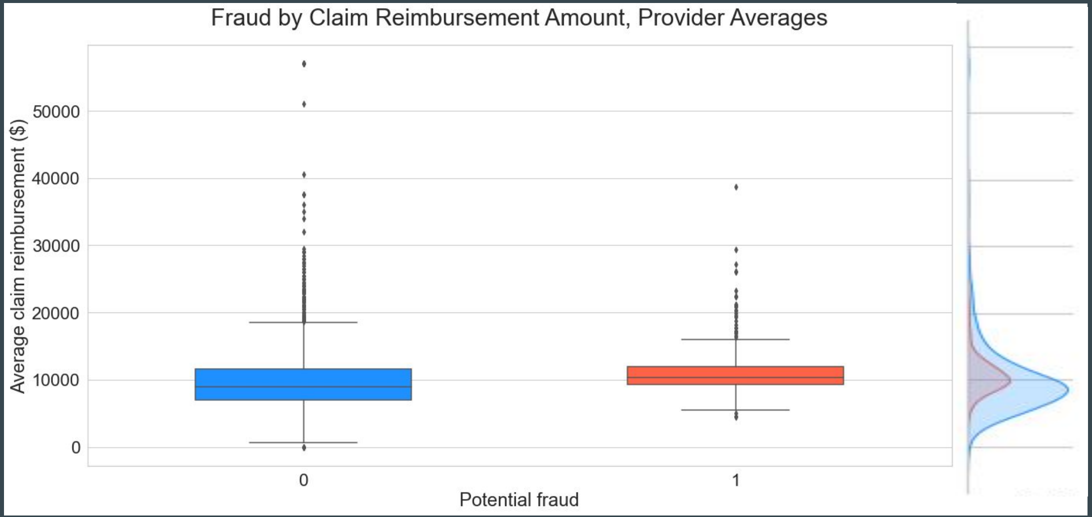


Levene’s test for equal variances:

p-value = 7.3×10^{-9}

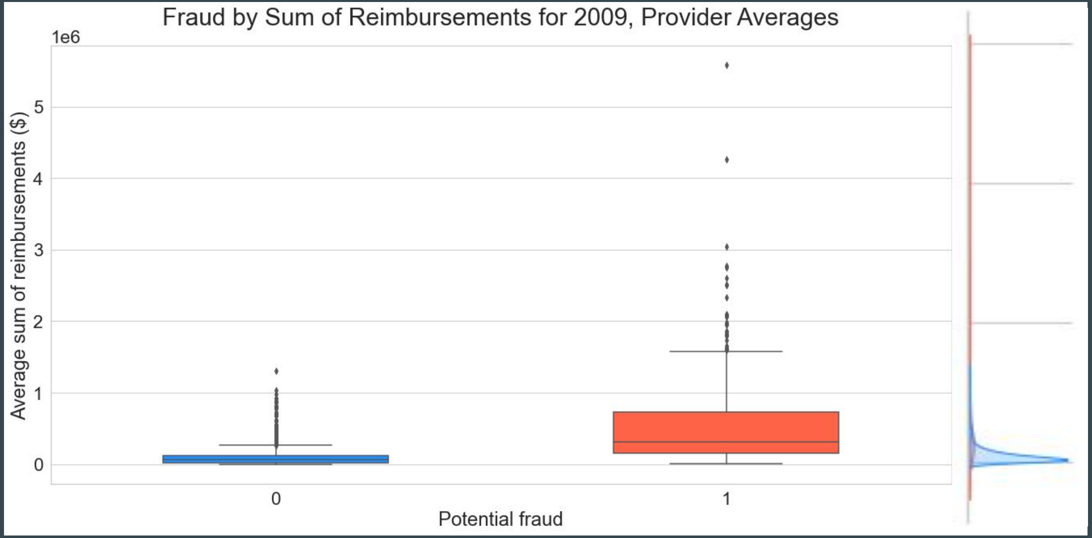
Two sample T-test:

p-value = 1.2×10^{-4}



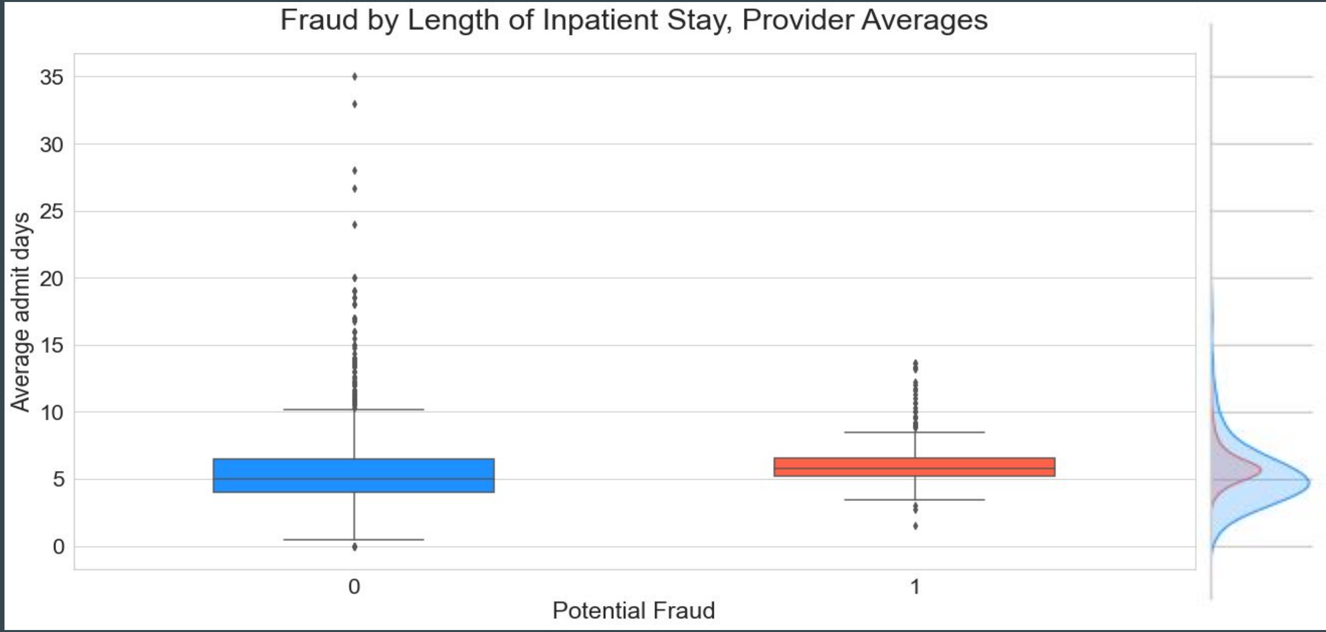
Levene's test for equal variances:

p-value = 7.3×10^{-9}



p-value = 1.7×10^{-138}

p-value = 4.1×10^{-95}



Two sample T-test:

p-value = $1.2\text{x}10^{-4}$

Levene's test for equal variances:

p-value = $1.8\text{x}10^{-13}$

Supervised Machine Learning to Predict Fraud

Outline

DATA PREPARATION	1. Data pre-processing 2. Feature engineering
MODELING	3. Establish important scoring metrics 4. Model selection 5. Select Data Scaling Method 6. Assess need for balancing 7. Feature importance 8. Hyperparameter tuning

Supervised Machine Learning to Predict Fraud

DATA PREPARATION

1. Data pre-processing
2. Feature engineering

Total reimbursements:

INPATIENT 40474 claims totaling $408,297,020

OUTPATIENT 517737 claims totaling $148,246,120

Total potentially fraudulent providers:

INPATIENT 440 of 2092 providers (slightly imbalanced)

OUTPATIENT 462 of 5012 providers (imbalanced)

Feature Engineering focusing on providers: Inpatient claims

Drop or convert and drop: 4

(eg. DOB converted to Age and dropped)

New beneficiary-based features: 8

(ClaimDuration , AdmitDays, DOD to binary, Sum number of diag or proc codes to new features, Sum number of physicians, Reimbursement amount per day, Number chronic illnesses)

New provider-based features: 27

(Num unique beneficiaries per provider, Num unique claims per provider, Sum reimbursements per provider, Number unique Admit, Group, Diag, Proc codes by provider)

Feature Engineering focusing on providers: Inpatient claims

Convert to fraction

DOD (deceased binary) to fraction

Fraction not white

M/F gender ratio

Fraction of beneficiaries with each chronic illness

Sum

Unique Counties

Unique ClmAdmitDiagCodes

Unique DiagGroupCodes

Unique DiagCodes

Unique ProcCodes

SUMMARY STATISTICS BY PROVIDER

Mean and MAD

InscClaimAmtReimbursed

ReimbPerDayAdmit

ClaimStartDt

AdmissionDt

ClaimDuration

AdmitDays

Number chronic illnesses

Age

Mean and MAD

NoOfMonths_PartACov

NoOfMonths_PartBCov

IPAnnualReimbursementAmt

IPAnnualDeductibleAmt

Number Physicians

Number DiagCodes

Number ProcCodes

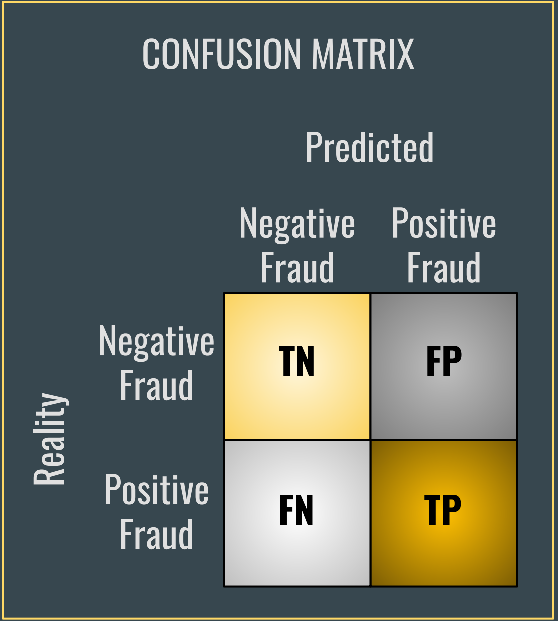
Feature Engineering focusing on providers: Inpatient claims

51 FEATURES,

ALL CONTINUOUS

Supervised Machine Learning to Predict Fraud

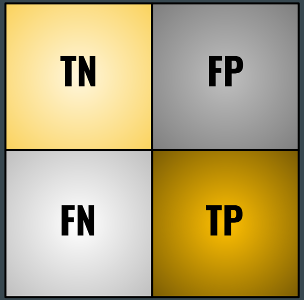
- Consider that no model is expected to predict fraud with 100% accuracy
  - Fraud predictions fall into two categories:
    - True Positive (TP)
    - False Positive (FP)
  - Similarly, for the fraud negative class:
    - True Negative (TN)
    - False Negative (FN)



Values to Direct Scoring Metric Selection

Number of Medicare fraud defendants investigated in 2009: 1786

45% of these led to charges



US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Values to Direct Scoring Metric Selection

Number of Medicare fraud defendants investigated in 2009:
1786
45% of these led to charges

Average cost per investigation:
$260,000
Cost of a False Positive

TN	FP
FN	TP

US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Values to Direct Scoring Metric Selection

Number of Medicare fraud defendants investigated in 2009:
1786
45% of these led to charges

Average cost per investigation:
$260,000
Cost of a False Positive

TN	FP
FN	TP

Maximum possible,
> number of FP

Average settlement per investigation leading to charges:
$1,451,000
Benefit of a True Positive

US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Values to Direct Scoring Metric Selection

Number of Medicare fraud defendants investigated in 2009: 1786
45% of these led to charges

TN	FP
FN	TP

Maximum possible,
> number of FP

Average cost per investigation: $260,000
Cost of a False Positive

Fiscally justifiable if 1 in 5 investigations lead to a conviction

Average settlement per investigation leading to charges: $1,451,000
Benefit of a True Positive

US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Values to Direct Scoring Metric Selection

Number of Medicare fraud defendants investigated in 2009: 1786

45% of these led to charges

Average cost per investigation: $260,000

Cost of a False Positive

Fiscally justifiable if 1 in 5 investigations lead to a conviction

TN	FP
FN	TP

Average loss from uncharged fraud: $548,000

Cost of a False Negative

Minimum possible

Maximum possible,
> number of FP

Average settlement per investigation leading to charges: $1,451,000

Benefit of a True Positive

US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Values to Direct Scoring Metric Selection

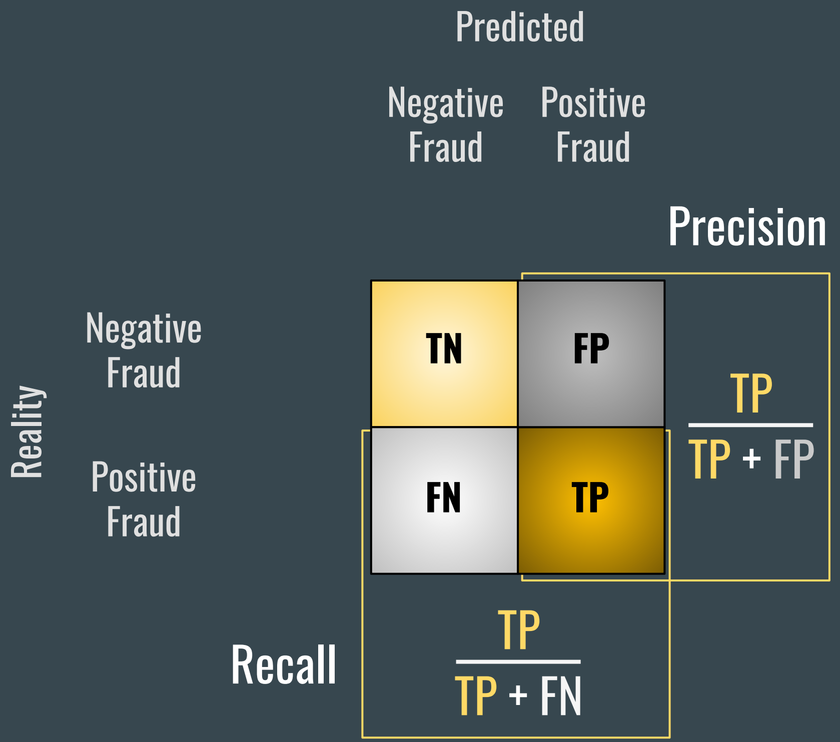
MODELING 3. Establish important scoring metrics

Precision vs. Recall

F_2 score:

Reduced precision

Increased recall



Supervised Machine Learning to Predict Fraud

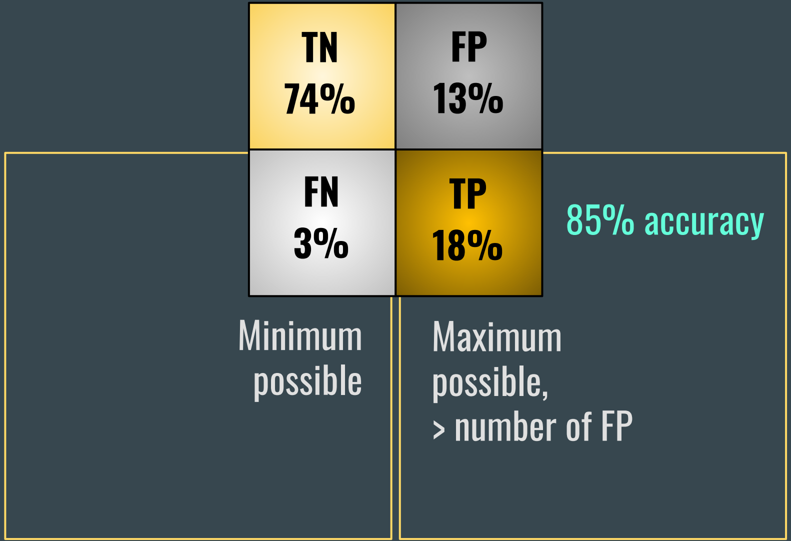
MODELING 4. Model selection

- Baseline screen of 14 different models (stratified cross validation)
    a. Linear algorithms: Logistic Regression, Linear Discriminant Analysis, Naive Bayes
    b. Nonlinear algorithms: Decision Trees, K-Nearest Neighbors, Support Vector Classifiers
    c. Ensemble algorithms: Bagging Classifier, Random Forest, Extra Trees, Gradient Boosting
- Select for highest possible recall scores
- Best: Gradient Boosting Classifier

Gradient Boosting Classifier to Predict Fraud

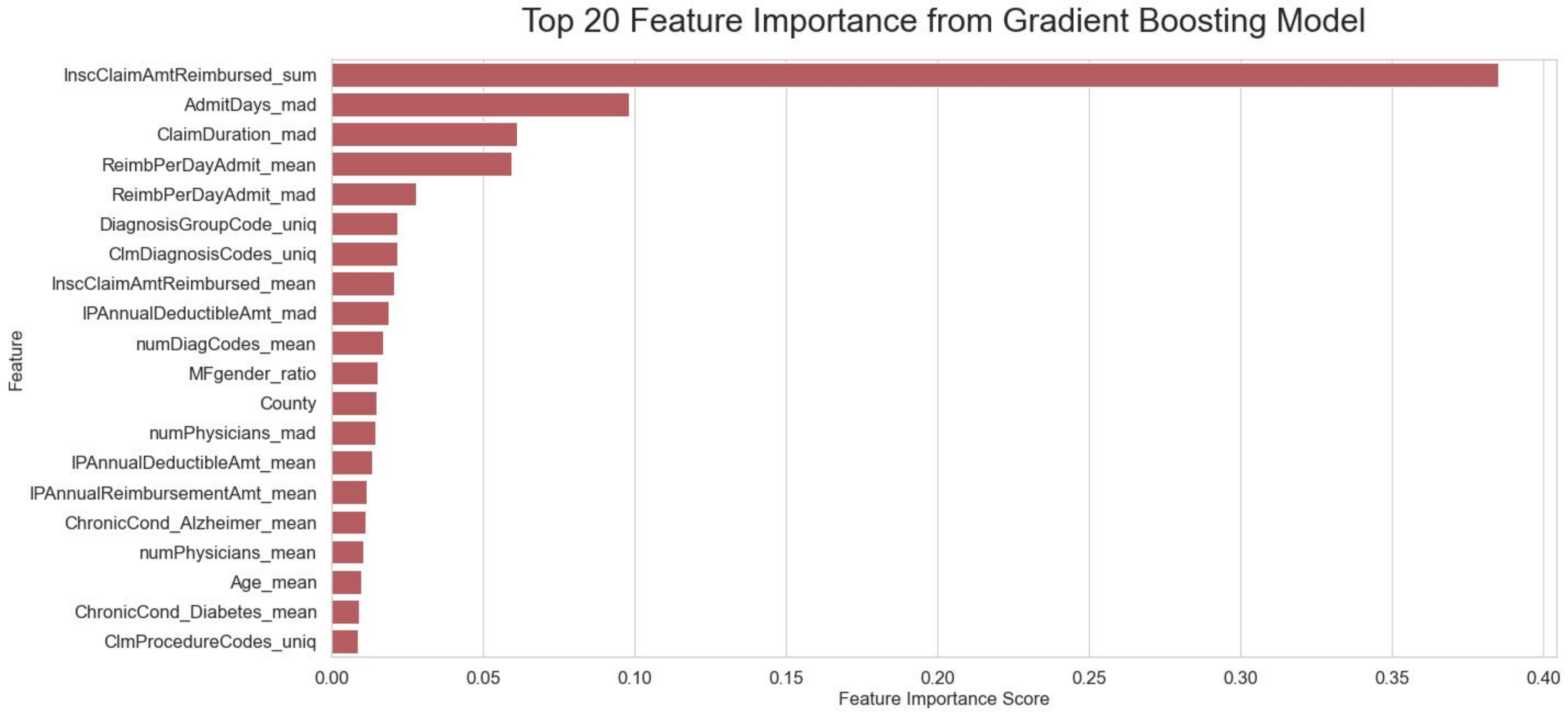
MODELING	5. Select data scaling method 6. Assess need for balancing

- Gradient Boosting Classifier
- Data balanced by undersampling
- Train performance:
  - Recall: 0.983
  - Precision: 0.983
- Test performance:
  - Recall: 0.852
  - Precision: 0.573



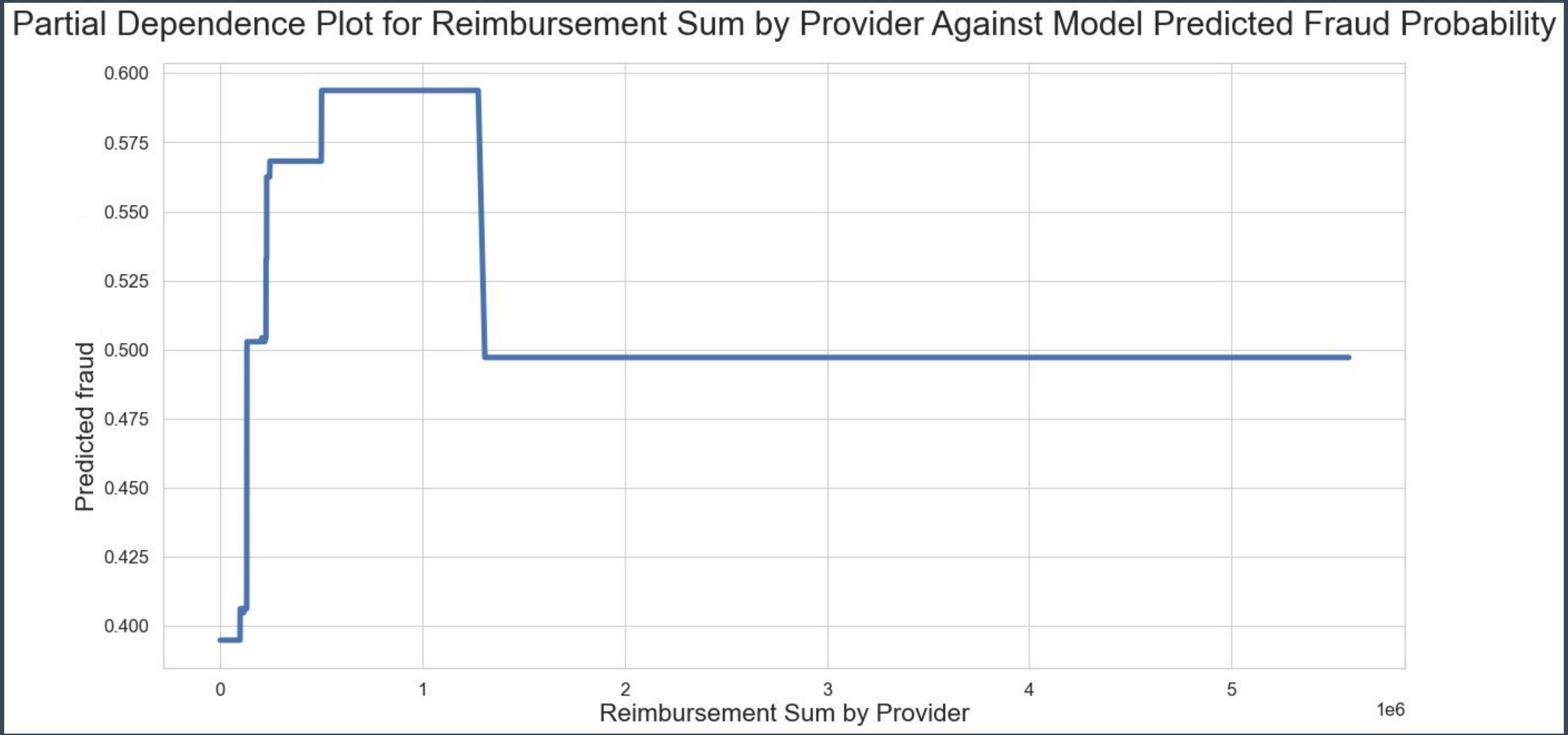
Gradient Boosting Classifier to Predict Fraud

MODELING 7. Feature importance



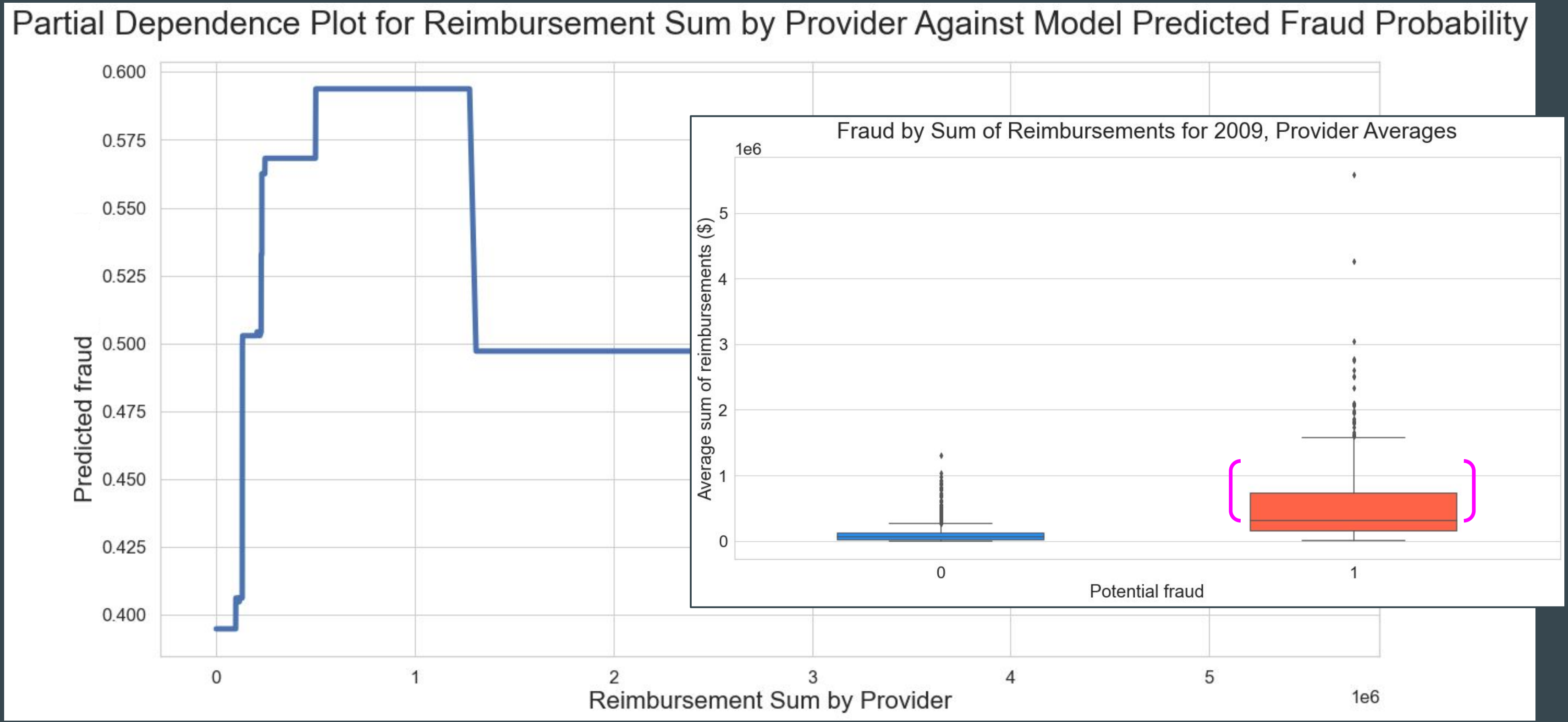
Gradient Boosting Classifier to Predict Fraud

MODELING 7. Feature importance



Gradient Boosting Classifier to Predict Fraud

MODELING 7. Feature importance



Gradient Boosting Classifier to Predict Fraud

Analysis of Top 8 Features from Gradient Boosting Model

1. Amount reimbursed (sum for provider)
2. Admit days (MAD for provider)
3. Claim duration (MAD for provider)
4. Reimbursement per day admitted (mean for provider)
5. Reimbursement per day admitted (MAD for provider)
6. Number of unique diagnosis group codes (for provider)
7. Number of unique clinical diagnosis codes (for provider)
8. Amount reimbursed (mean for provider)

Actionable Insights from the Model

Revenue saved by optimizing model to minimize false negatives...

“Catch everything” model, and then work out the TP, FP details during subsequent investigations

Actionable Insights from the Model

Revenue saved by optimizing model to minimize false negatives...

“Catch everything” model, and then work out the TP, FP details during subsequent investigations

Applying this model to validation dataset and applying average values from Health Care Fraud and Abuse Control Program Report

US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Actionable Insights from the Model

Revenue saved by optimizing model to minimize false negatives...

“Catch everything” model, and then work out the TP, FP details during subsequent investigations

Applying this model to validation dataset and applying average values from Health Care Fraud and Abuse Control Program Report

- Investigation cost for positive class (TP + FP): 131 x $260,000 (avg. investigation) = $34 million

US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Actionable Insights from the Model

Revenue saved by optimizing model to minimize false negatives...

“Catch everything” model, and then work out the TP, FP details during subsequent investigations

Applying this model to validation dataset and applying average values from Health Care Fraud and Abuse Control Program Report

- Investigation cost for positive class (TP + FP): 131 x $260,000 (avg. investigation) = $34 million
- Loss to incorrectly classifying false negatives: 13 x $548,000 = $7.1 million

US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Actionable Insights from the Model

Revenue saved by optimizing model to minimize false negatives...

“Catch everything” model, and then work out the TP, FP details during subsequent investigations

Applying this model to validation dataset and applying average values from Health Care Fraud and Abuse Control Program Report

- Investigation cost for positive class (TP + FP): 131 x $260,000 (avg. investigation) = $34 million
- Loss to incorrectly classifying false negatives: 13 x $548,000 = $7.1 million
- Gain from correctly identifying true positives: 75 x $1,451,000 (avg. settlement) = $109 million

US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Actionable Insights from the Model

Revenue saved by optimizing model to minimize false negatives...

“Catch everything” model, and then work out the TP, FP details during subsequent investigations

Applying this model to validation dataset and applying average values from Health Care Fraud and Abuse Control Program Report

- Investigation cost for positive class (TP + FP): 131 x $260,000 (avg. investigation) = $34 million
- Loss to incorrectly classifying false negatives: 13 x $548,000 = $7.1 million
- Gain from correctly identifying true positives: 75 x $1,451,000 (avg. settlement) = $109 million
- Benefit of applying model (millions $): 109 - (34 + 7.1) = $68 million

US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Actionable Insights from the Model

Important benefit of using this machine learning model:

- Fewer false positives than using standard Health Care Fraud Abuse Control measures in place in 2009 which assigned 55% of positive class as false positives (considered to be investigations that did not lead to charges (983 / 1786))

US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Actionable Insights from the Model

Important benefit of using this machine learning model:

- Fewer false positives than using standard Health Care Fraud Abuse Control measures in place in 2009 which assigned 55% of positive class as false positives (considered to be investigations that did not lead to charges (983 / 1786))

- In contrast, this model assigns 44% of the positive class as false positives and can be re-tuned to further decrease the percentage of false positives as desired

US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Actionable Insights from the Model

Important benefit of using this machine learning model:

- Fewer false positives than using standard Health Care Fraud Abuse Control measures in place in 2009 which assigned 55% of positive class as false positives (considered to be investigations that did not lead to charges (983 / 1786))

- In contrast, this model assigns 44% of the positive class as false positives and can be re-tuned to further decrease the percentage of false positives as desired

- This decrease in false positives leads to an average savings of $30,000 per fraud investigation

US Office of the Inspector General, Health Care Fraud and Abuse Control Program Report FY 2009

Improving the Model

Considering the top important features, it is likely that the model can be improved with additional data

- Inpatient admission days: important features 2 - 5 are all related

Improving the Model

Considering the top important features, it is likely that the model can be improved with additional data

- Inpatient admission days: important features 2 - 5 are all related
  - Any additional information regarding admission duration could be helpful to bolster the model

Improving the Model

Considering the top important features, it is likely that the model can be improved with additional data

- Inpatient admission days: important features 2 - 5 are all related
    - Any additional information regarding admission duration could be helpful to bolster the model
    - For example, having information regarding the US average admission stay for the 13 chronic conditions present in the beneficiary data could be used to establish a baseline hospital stay for each.

Improving the Model

Considering the top important features, it is likely that the model can be improved with additional data

- Inpatient admission days: important features 2 - 5 are all related
  - Any additional information regarding admission duration could be helpful to bolster the model
  - For example, having information regarding the US average admission stay for the 13 chronic conditions present in the beneficiary data could be used to establish a baseline hospital stay for each.
- Location of the providers (this can only be inferred by the mode values of beneficiary's home state):
  - This could help to correlate provider state with states known to have higher or lower rates of Medicare fraud (available in records from the FBI and US Sentencing Commission).

Thank you

Inpatient vs. Outpatient Claim Data and Modeling

What about fraud for the outpatient claims?

This model focused solely on the inpatient data

An additional model could be created to detect fraud in the outpatient claims data

Inpatient vs. Outpatient Claim Data and Modeling

What about fraud for the outpatient claims?

This model focused solely on the inpatient data

An additional model could be created to detect fraud in the outpatient claims data

Keeping the two separate was decided due to the obvious differences in the two datasets:

- The magnitude and range of the reimbursement amounts between the inpatient and outpatient claims are vastly different
- The outpatient data has three fewer features, and importantly lacks admission duration information (which is not relevant for outpatient treatment)
- Four of the top 8 important features from the inpatient model relate to admission duration, suggesting that a combined model may not be suitable to apply to both datasets

Sample of Medicare claims data from 2009

- 2009 was first year the implementation of electronic health records for Medicare started transitioning (due to American Recovery and Reinvestment Act)

- Additionally, in this period there was a transition between two medical coding standards, International Statistical Classification of Diseases and Related Health Problems (ICD)-9 and ICD-10 and is reflected in the confusing mix of codes used

- Therefore, one would expect there to be various problems with the quality of the data and whether providers flagged as potentially fraudulent were actually committing fraud or merely just committing errors in the record-keeping process

Center for Medicaid Services, Electronic Health Records Provider Fact Sheet 2015