Identifying Healthcare Fraud

With Machine Learning

By Zheng Zhang

DATASET & OVERVIEW

- Healthcare provider fraud is one of the biggest problems facing Medicare, and it contributes to the total Medicare spending growth. Healthcare fraud is an organized crime which usually involves peers of providers, physicians, beneficiaries acting together to make fraudulent claims.
- We want to identify the potentially fraudulent providers based on the claims that they filed. In addition, we will also discover important variables helpful in detecting the behaviour of those providers.
- The datasets include provider data, beneficiary data, inpatient claims data and outpatient claims data. It can be found here: https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis

The Value Proposition

- The ability to accurately identify the fraudulent claims and the perpetrating healthcare providers are highly critical as fraud losses within US healthcare system cost US taxpayers tens of billions a year, could be as high as $100 billion based on some estimates (United States Department of Justice).

- We want to discourage providers to engage in fraudulent activities.
- We want to save public and private healthcare insurers fund so the savings can be passed to the subscribers.

PROJECT FLOW

1. **Loading Data**
   a. Kaggle.com
2. **Exploration**
   a. Cleaning data
   b. Beneficiary demographics/medical condition
   c. Claim distribution
   d. Different claim distributions by race, medical conditions
   e. Different claim distributions by provider's class
3. **Modelling I**
   a. Baseline tests
   b. Imbalanced data issue
4. **Modelling II**
   a. Balanced sampling
   b. Confusion matrix
5. **Conclusion**
   a. Summary, further optimization

LOADING DATA

LOADING DATA

- The datasets were downloaded from Kaggle’s website, which include provider data, beneficiary data, inpatient claims data and outpatient claims data. There are separate training dataset and test dataset. Only training datasets were used in this project.
- There are 5410 providers, 138,556 beneficiaries, 40,474 inpatient and 517,737 outpatient claims.
- The first five rows of the inpatient claim data:

	BeneID	ClaimID	ClaimStartDt	ClaimEndDt	Provider	InscClaimAmtReimbursed	AttendingPhysician
0	BENE11001	CLM46614	2009-04-12	2009-04-18	PRV55912	26000	PHY390922
1	BENE11001	CLM66048	2009-08-31	2009-09-02	PRV55907	5000	PHY318495
2	BENE11001	CLM68358	2009-09-17	2009-09-20	PRV56046	5000	PHY372395
3	BENE11011	CLM38412	2009-02-14	2009-02-22	PRV52405	5000	PHY369659
4	BENE11014	CLM63689	2009-08-13	2009-08-30	PRV56614	10000	PHY379376

EXPLORATION

CLEANING DATA

- Missing data:

  Missing entries in the variable “date of death” in the beneficiary data are expected.

  Missing entries on physician ID and diagnosis/procedure code variables were left as they were.

- Outlier detection:

  Positive outliers are expected, however, negative values in the claim deductible and amount are not expected, in those cases, the negative values were changed to 0.

Beneficiary Demographics and Medical Conditions

1. Gender: 43% are men
2. Race: 84% white, will be dichotomized for further analysis
3. Renal Disease: 14%
4. From 52 states and 314 counties
5. Baseline medical condition:

    a. Alzheimer: 33.2%

    b. Heart failure: 49.4%

    c. Kidney disease: 31.2%

    d. Cancer: 12.0%

    e. Obstructive-Pulmonary: 23.7%

    f. Depression: 35.6%

    g. Diabetes: 60.2%

    h. Ischemic heart disease: 67.6%

    i. Osteoporosis: 27.5%

    j. Rheumatoid Arthritis: 25.7%

    k. Stroke: 7.9%

Inpatient/Outpatient Claim Data

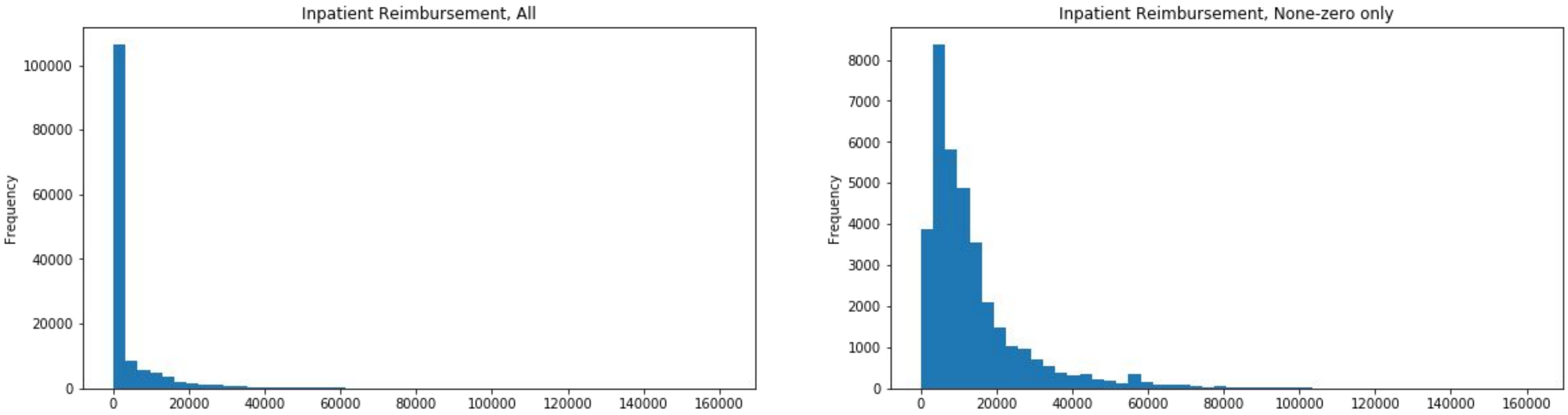
Inpatient claims:

- A total of 31,289 beneficiaries had inpatient claims
- A total of 40,474 claims from 2092 providers
- Maximum claims/beneficiary: 8
- Maximum claims/provider: 516

Outpatient claims

- A total of 133,980 beneficiaries had outpatient claims
- A total of 517,737 claims from 5012 providers
- Maximum claims/beneficiary: 29
- Maximum claims/provider: 8240

Inpatient annual claim reimbursement per beneficiary

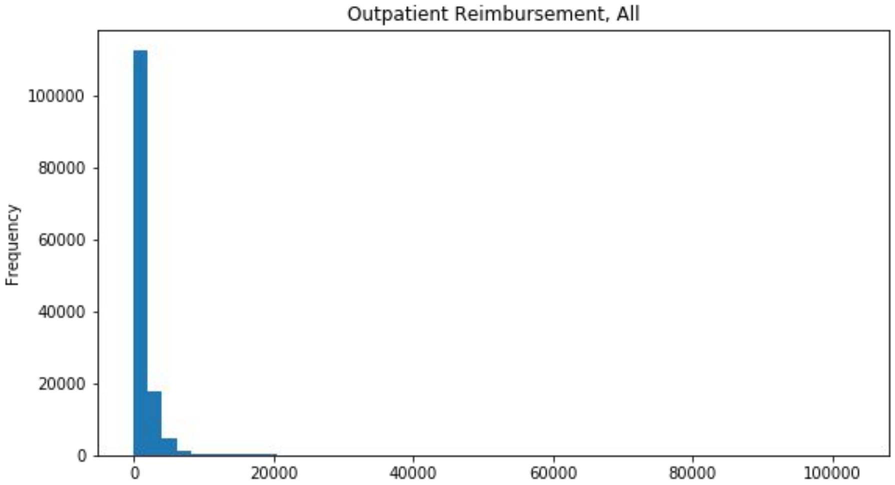


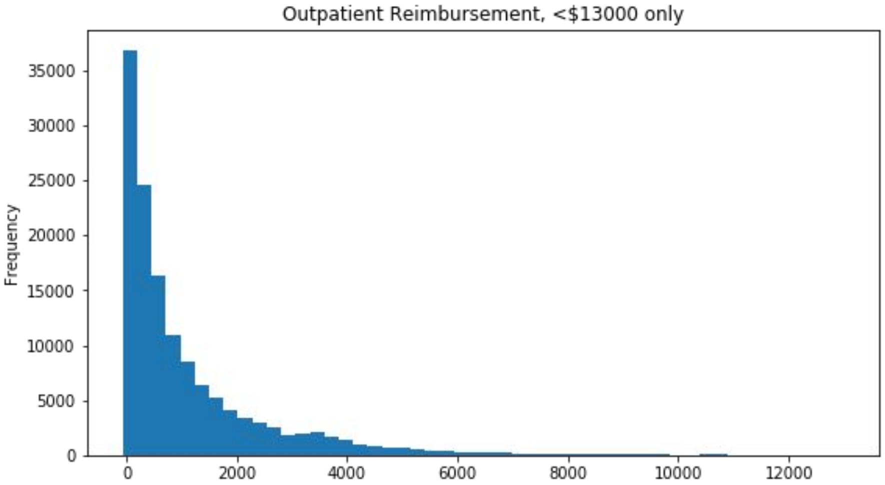
Inpatient Reimbursement, All

Inpatient Reimbursement, None-zero only

- Only 23% (31,289) beneficiaries have inpatient claims, so there are 77% at zero (left panel).
- Of those who had inpatient claims, the claim amount is highly skewed (right panel). The mean claim amount is $3660, SD is $9569, the maximum claim amount is $161,470.

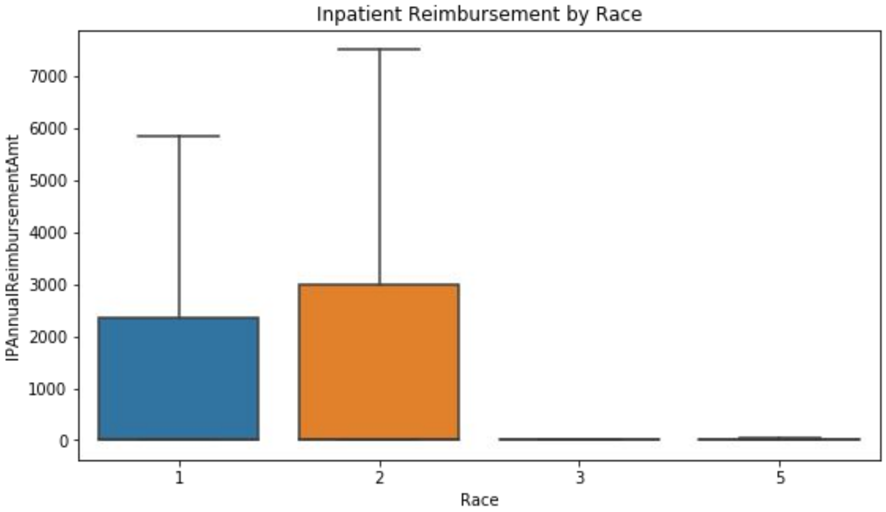
Outpatient annual claim reimbursement per beneficiary

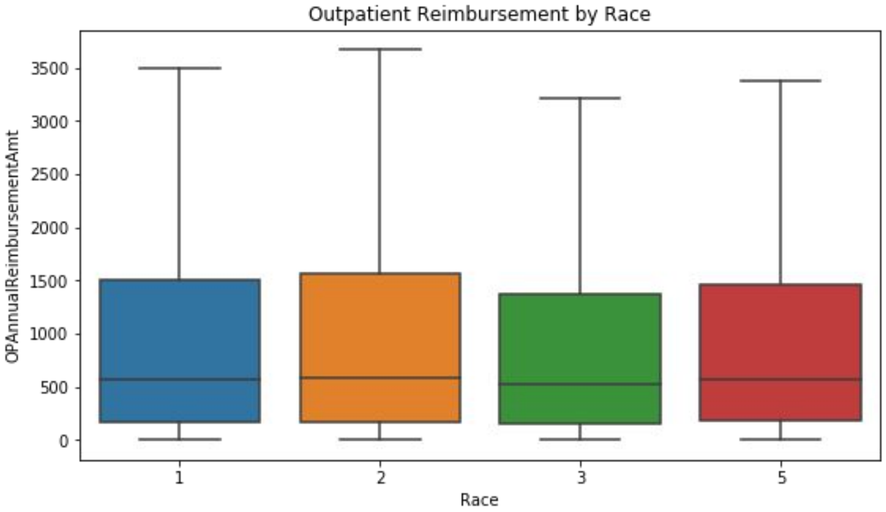




- 97% beneficiaries had outpatient claims.
- The distribution of the outpatient claims is also skewed, with mean claims $1298, median claim $570 and the maximum claim amount is $102,960.

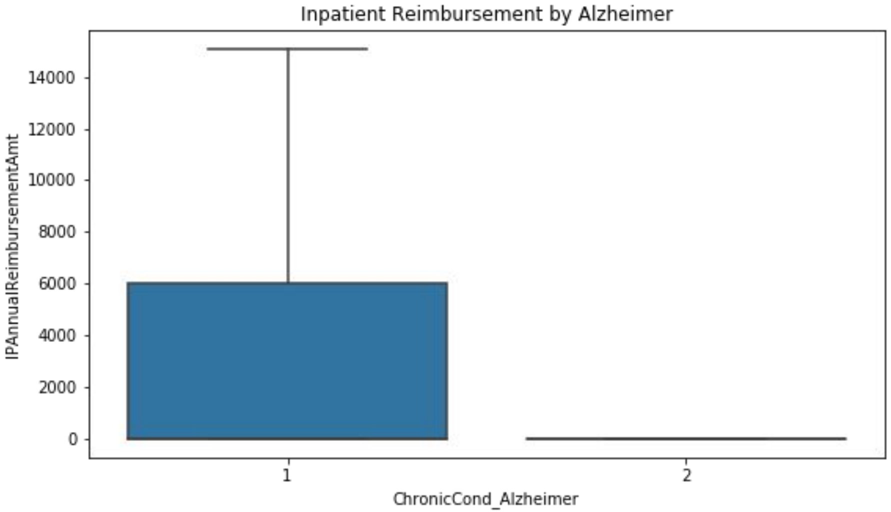
Claim amount by race

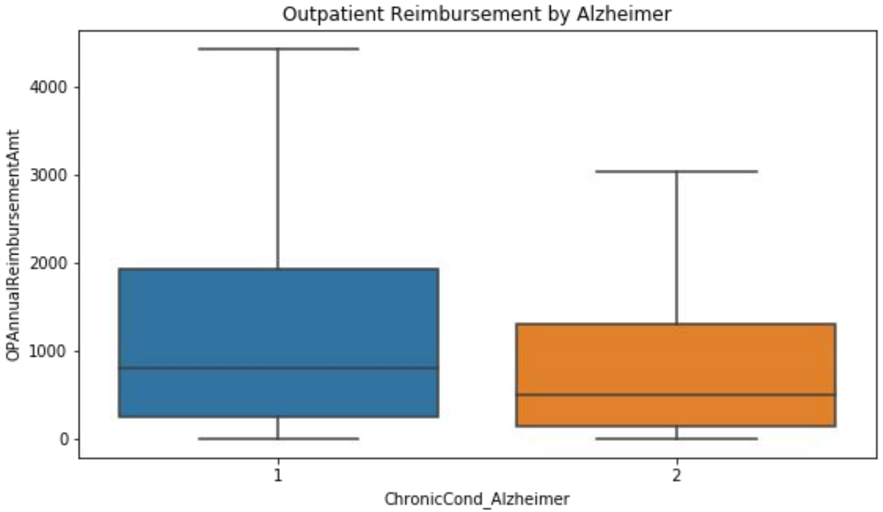




- Groups 3 and 5 have almost no inpatient claim.
- No significant difference observed across racial groups for outpatient claims.

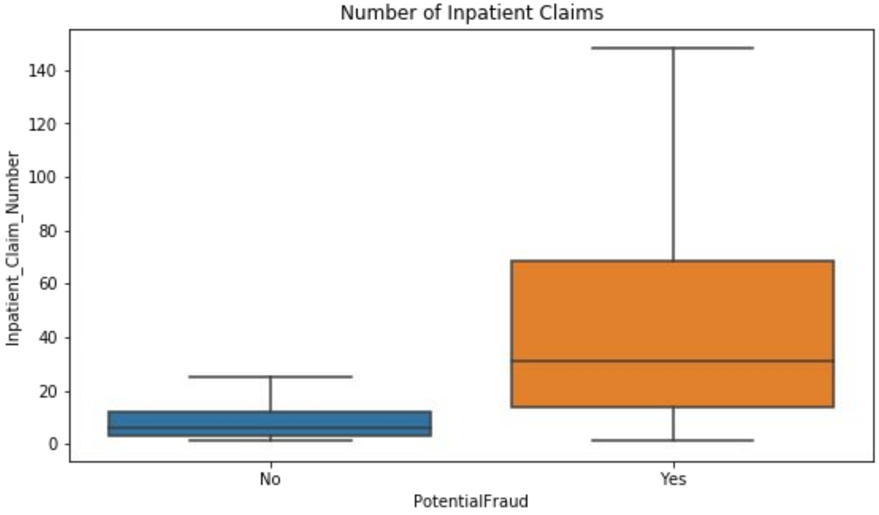
Claims amount by chronic medical conditions

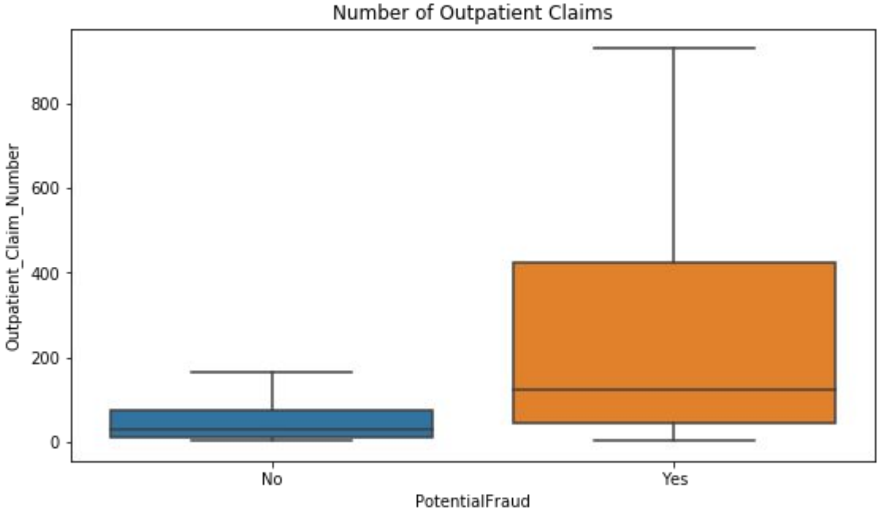




- people with Alzheimer’s incur almost all inpatient claims. They tend to have higher outpatient claim amounts.
- The pattern is the same for other chronic medical conditions.

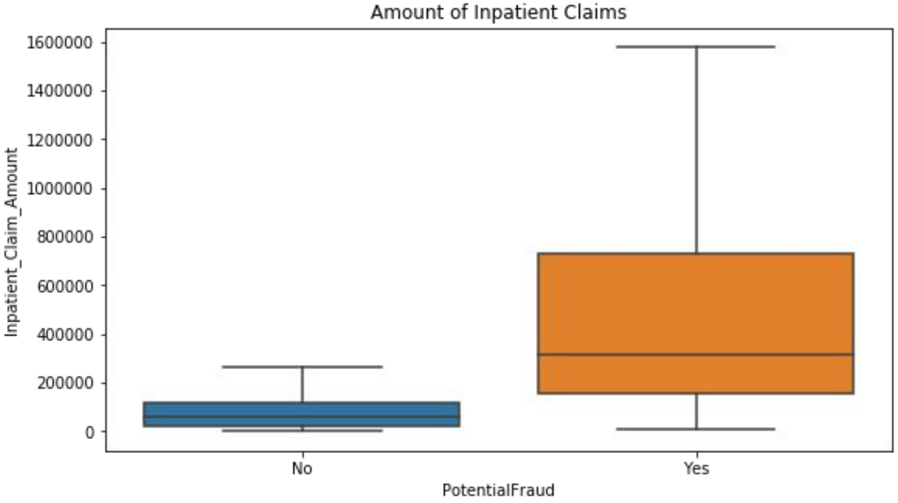
Fraudulent vs. Honest Providers: # of claims

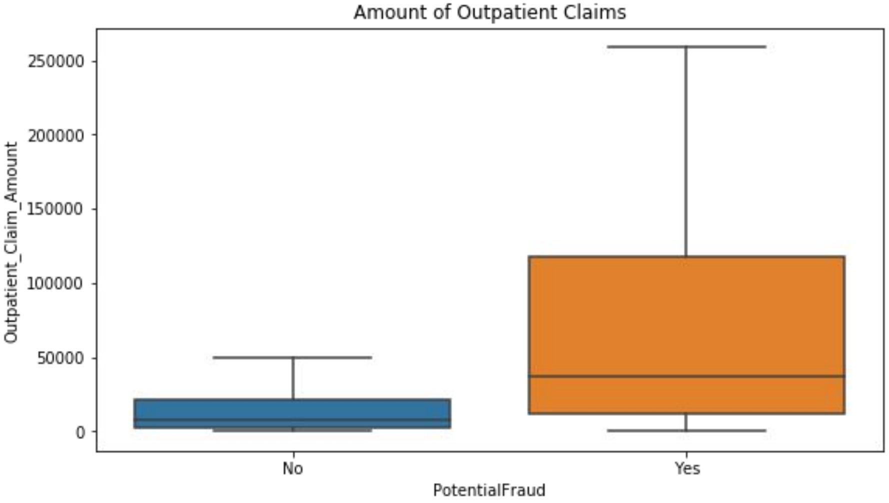




- The potentially fraudulent providers had substantially more claims submitted.

Fraudulent vs. Honest Providers: reimbursement





- The potentially fraudulent providers had substantially more claims submitted and had been reimbursed substantially more.

MODELLING I

MODELLING I

- Split the dataset into a 70% training set and 30% test set, stratified by the fraud status.

- Some algorithms have hyperparameters, where GridSearch has been used to identify the best choice for such hyperparameters (hyperparameter tuning) by using 5-fold cross-validation.

- All models are run with the same random state when applicable
- The sampling of the training data did not consider data imbalance.

BASELINE RESULTS (simple sampling)

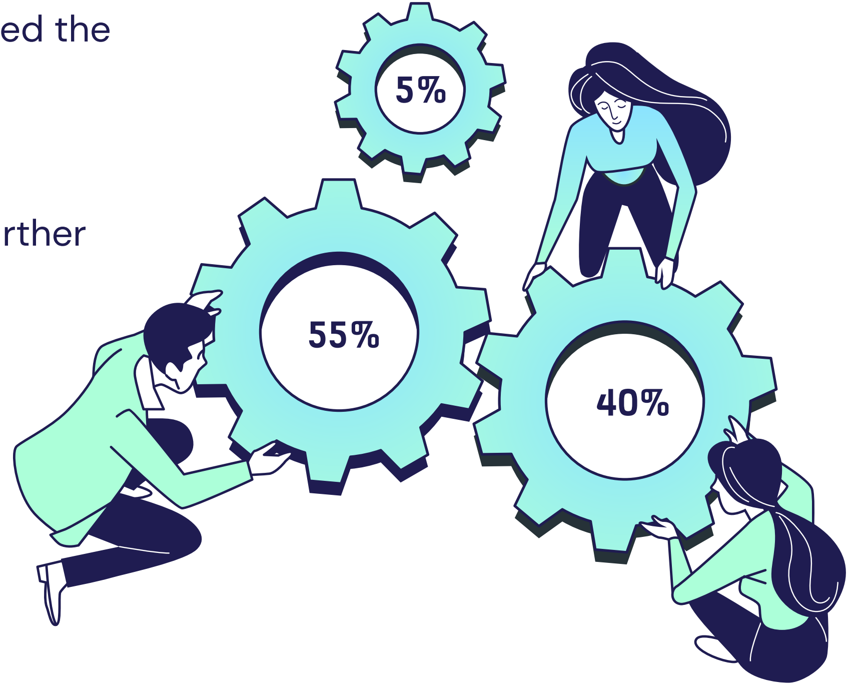
Model	Sensitivity	Specificity	Misclassification rate	Tuning parameter
Random Forest	0.49	0.97	0.077	N=51
Logistic Regression	0.46	0.98	0.065	
K-Nearest Neighbor	0.47	0.98	0.069	K=8

- Logistic regression is the best overall model, but all three had low sensitivity despite high accuracy rate.

MODELLING II

MODELLING II

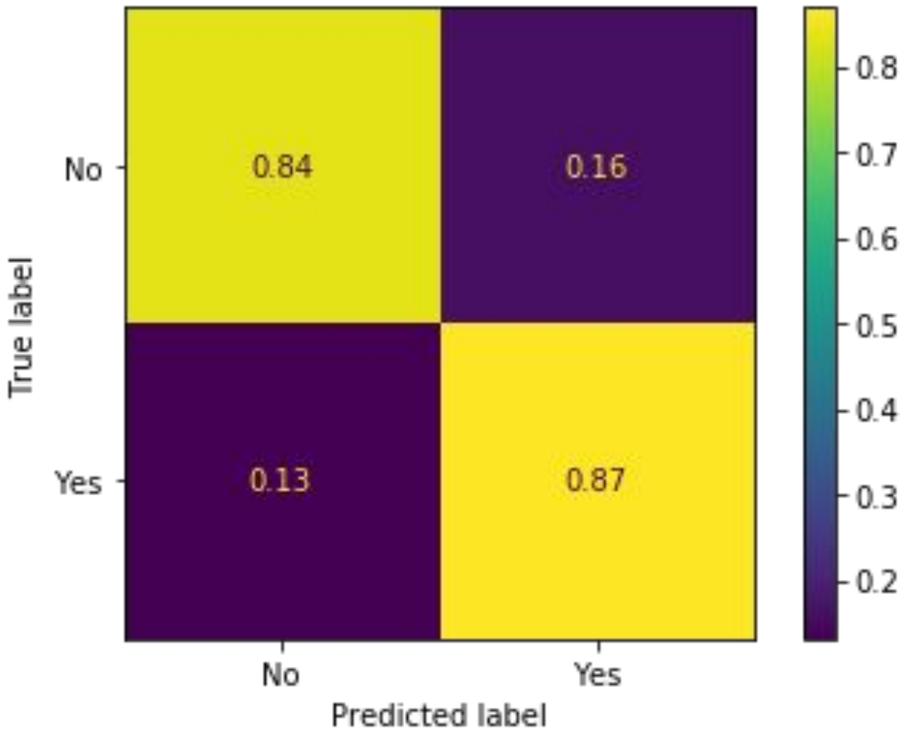
- In this phase of modelling, we have adjusted the sampling algorithm of the training data to balance the two classes
- We'll also look at confusion matrices to further analyze performance



MODEL RESULTS II (Balanced Sampling)

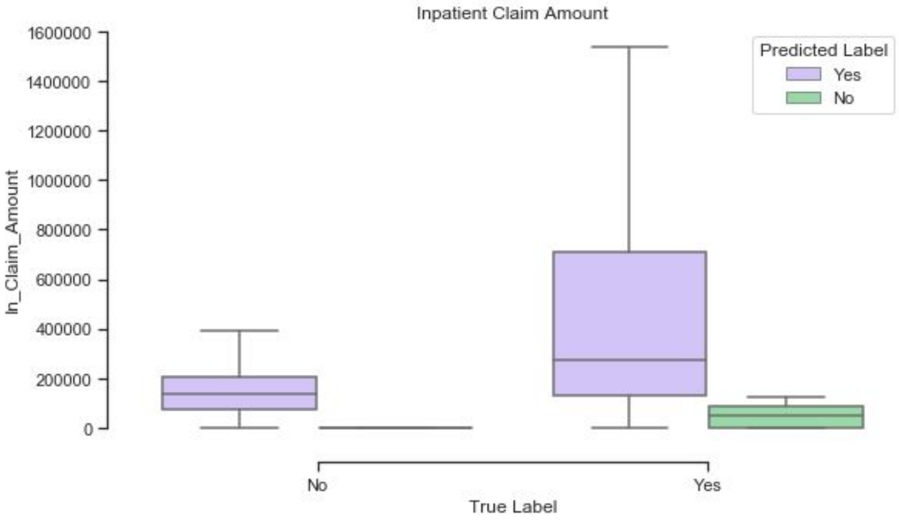
Model	Sensitivity	Specificity	Accuracy	Tuning parameter
Logistic Regression with oversampling	0.78	0.91	0.90	
Balanced Random Forest	0.87	0.84	0.84	N=81
Balanced Bagging	0.77	0.88	0.87	
Rusboosting	0.86	0.83	0.83	N=73
Adaboosting	0.86	0.85	0.85	

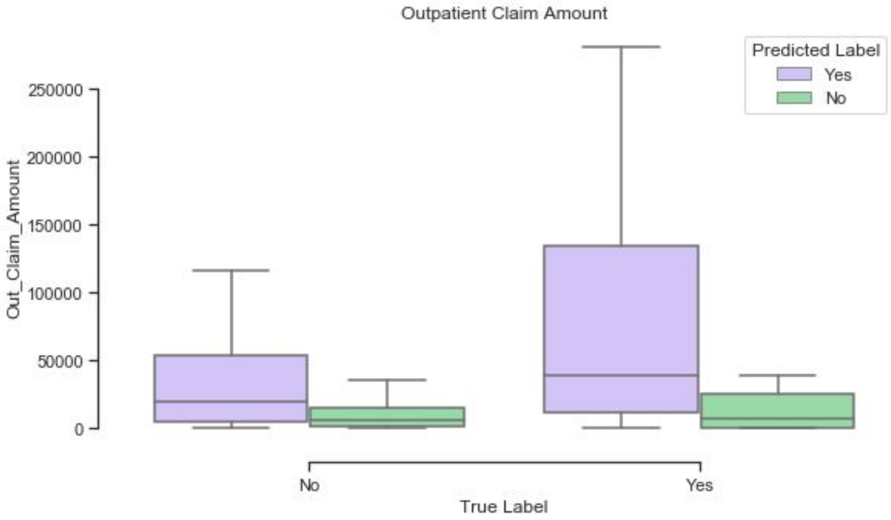
CONFUSION MATRIX (Balanced Random Forest)



- Correctly labelling 87% of fraudulent providers

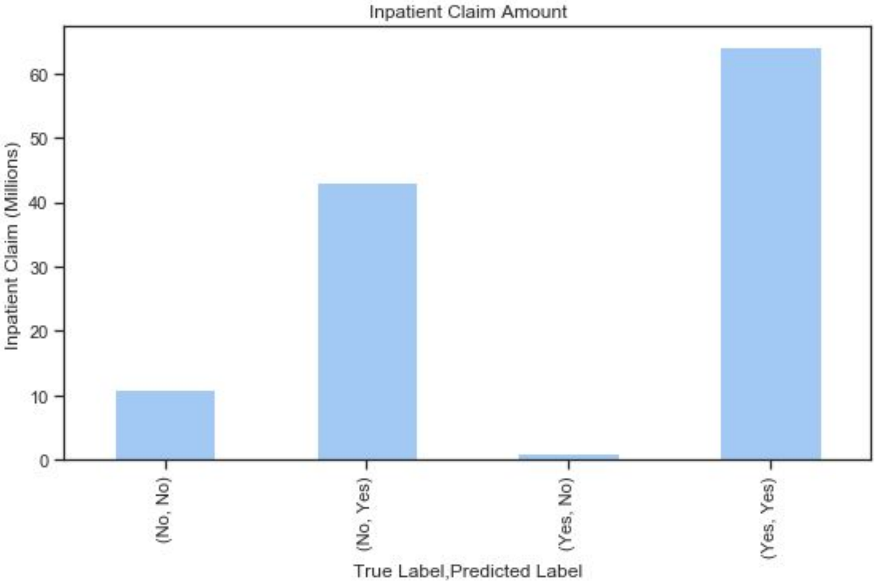
Predicted vs. True Label: Claim Reimbursement

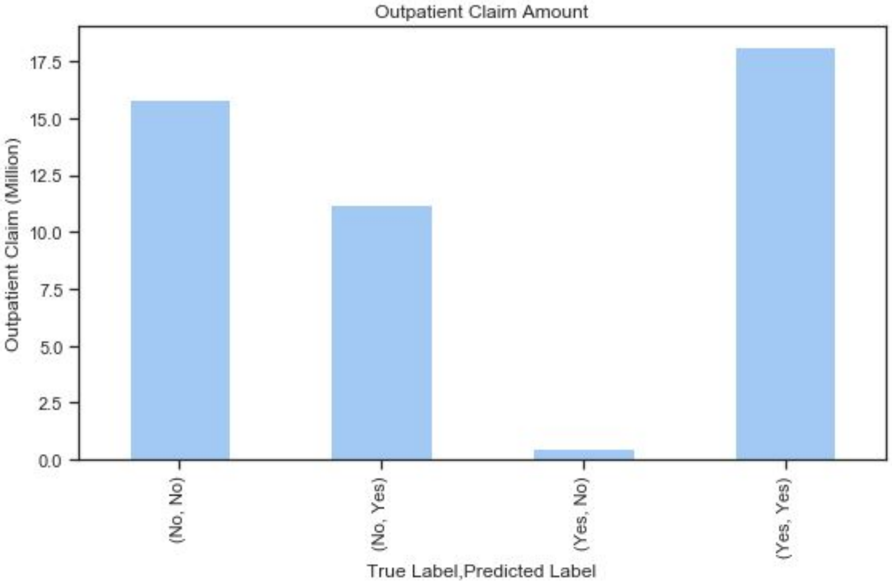




- Balanced random forest has identified 87% of fraudulent providers, whose average claims were $485,161(median $276,200) for inpatient and $137,259 (median $38,480) for outpatient claims.

Predicted vs. True Label: Total Claim Reimbursement





- Balanced random forest has identified 87% of fraudulent providers, who accounted for > 98% ($64 million) fraudulent inpatient and > 97% ($18 million) outpatient claims. Our model could be employed to identify the providers who account for over 97% fraud and save Medicare over $82 million dollars.

CONCLUSION

SUMMARY

- The main challenge of this project was the imbalance of the data, as less than 10% providers are labelled as potentially fraudulent. Ignoring this fact and using simple resampling method cause poor performance on identifying the most important (but minority) class of fraudulent providers.
- Balanced random forest had achieved highest sensitivity at 87% in the test set.
- By using balanced random forest, we were able to identify 132 out of 152 fraudulent providers, who had submitted a total of $64 million inpatient and $18 million outpatient claims, whose average annual claim amount per provider was $137,259 for outpatient and $485,161 for inpatient claims.
- Our model could be employed to identify the providers who account for over 97% fraudulent amount and save Medicare over $82 million dollars.

FURTHER OPTIMIZATION

- A Priori decisions on the cost of false positives and false negatives will help determine the optimal prediction model.
- Random forest and logistic regression were top models for this project, but other unexplored model may be better.
- The diagnosis codes were not used as they introduced a lot of noise but will be explored more in the future.
- We would like to do more in-depth parameter tuning