PREDICTIVE MODELING OF FRAUDULENT HEALTHCARE PROVIDERS

Julia Sheriff | Springboard, Data Science Career Track

CAPSTONE PROJECT 2

HOW CAN WE PREDICT POTENTIALLY FRAUDULENT MEDICARE PROVIDERS?

- Source: Medicare Data:
  - Beneficiary Claims
  - Outpatient Claims
  - Inpatient Claims
  - Provider Potential Fraud Status

BENEFICIARY DATA

- 138,556 BENEFICIARIES
- 27 FEATURES:
  - Demographics
  - Chronic health conditions
  - Insurance
- DATA WRANGLING:
  - Created binary variables for gender, death, and chronic health conditions
  - Removed strings from beneficiary id

OUTPATIENT CLAIMS

- 517,737 OUTPATIENT CLAIMS
- 25 FEATURES:
  - Claim Information (Location, Date, Time, Beneficiary, Provider)
  - Charges and Payments
  - Diagnosis and Procedure Codes
  - Physicians
- DATA WRANGLING:
  - Dropped procedure codes 4-6 due to missing data
  - Time variables converted to datetime objects
  - String parsing to remove redundant letters in provider id, beneficiary id, claim id, and physician id.

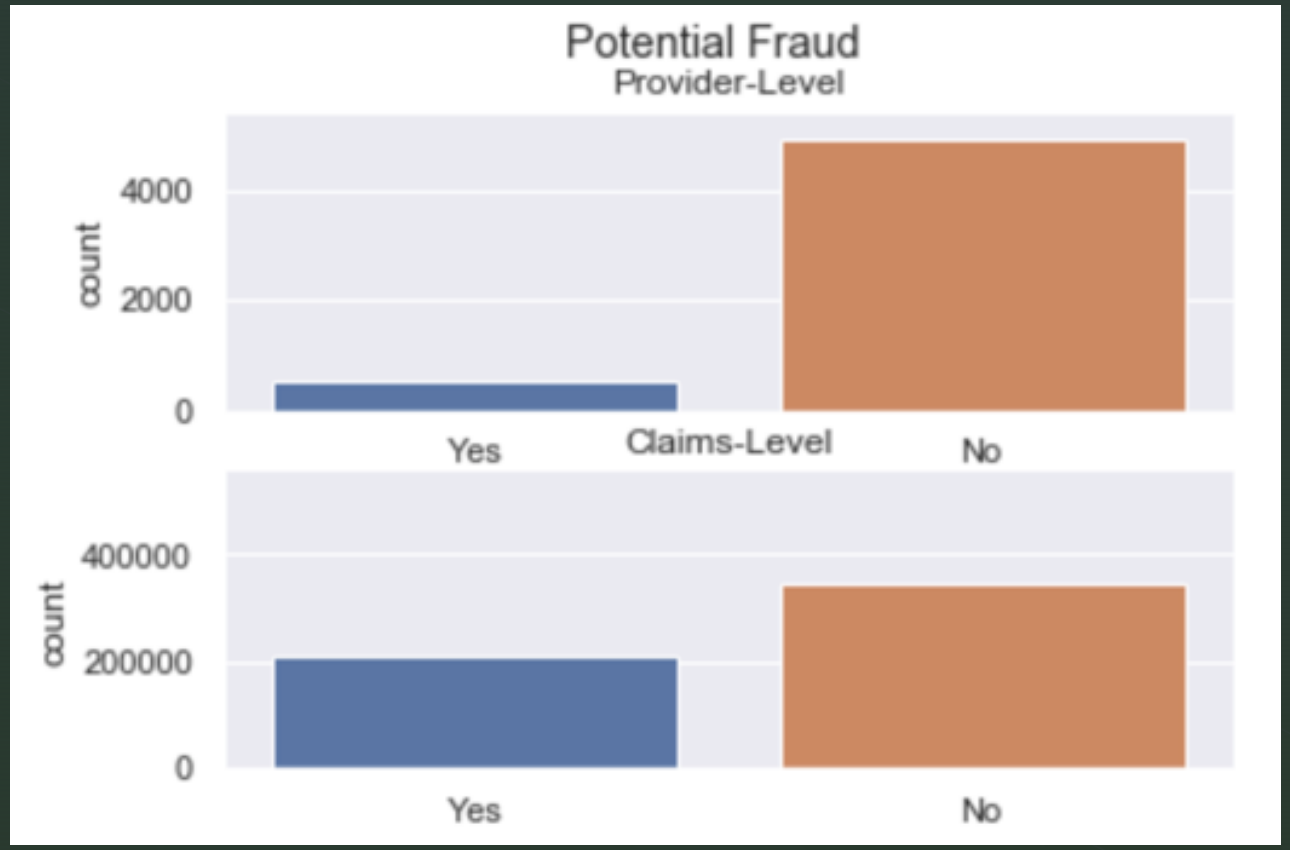
INPATIENT CLAIMS

- 40,474 INPATIENT CLAIMS
- 29 FEATURES:
  - Included all features from outpatient claims
  - Additional features:
    - Admission and Discharge Date
    - Diagnosis Group Code
- DATA WRANGLING:
  - New variable: duration
  - Dropped procedure codes 4-6 due to large quantities of missing data
  - Time variables converted to datetime objects
  - String parsing to remove redundant letters in provider id, beneficiary id, claim id, and physician id.

AGGREGATED DATAFRAME

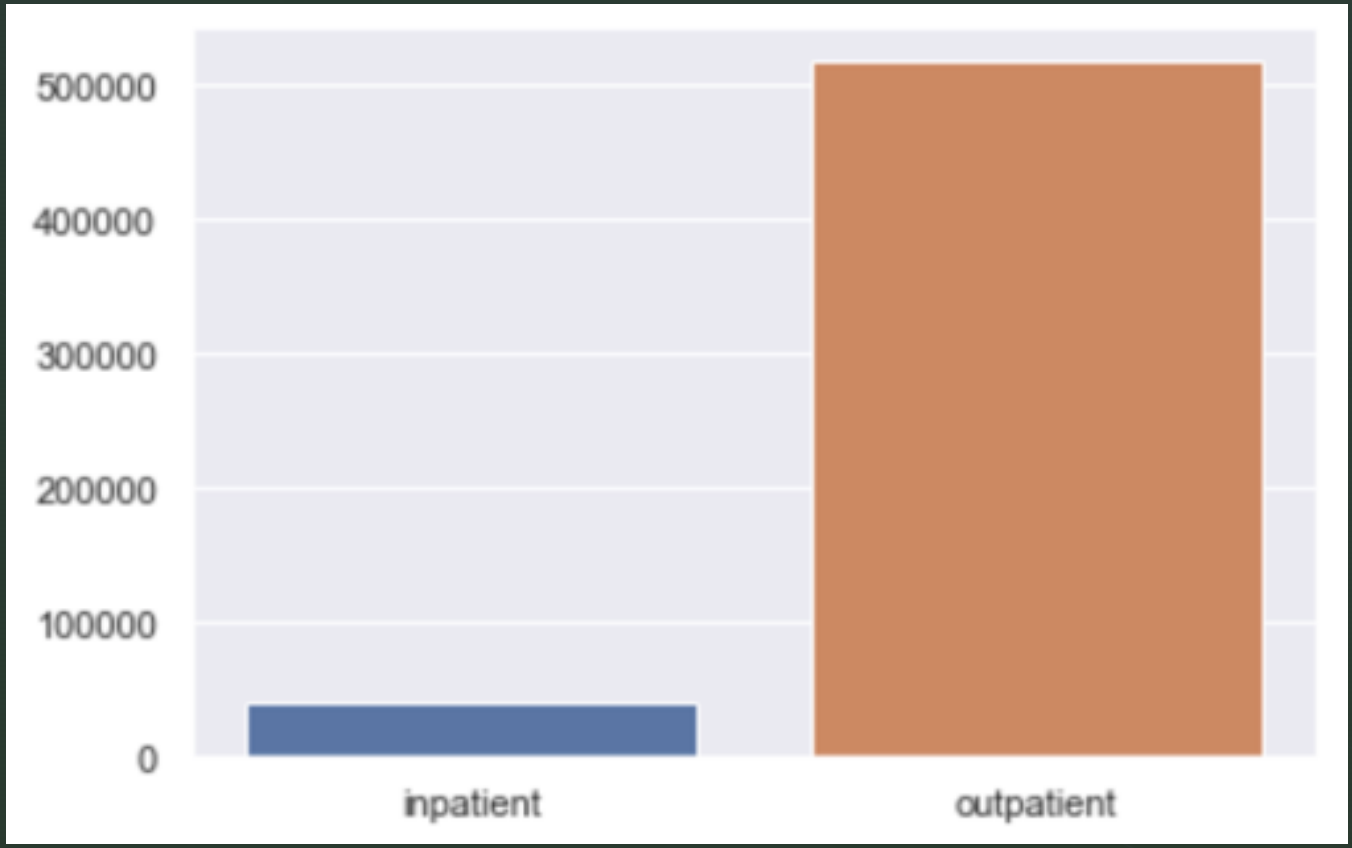
- 5,410 PROVIDERS
- 168,888 FEATURES:
  - Measures of central tendency of other variables
  - Distributions of demographic variables
  - Diversity in number of physicians per provider
  - Maximum mean insurance claim amount reimbursed for various diagnostic codes
  - Counts per procedure and diagnostic codes
- Used beneficiary and provider to merge all datasets

CLASS IMBALANCE: FRAUD ON THE PROVIDER AND CLAIMS LEVEL



A few providers are responsible for the majority of fraudulent claims.

IMBALANCE: INPATIENT VERSUS OUTPATIENT CLAIMS



The claims data used to describe providers is more influenced by outpatient claims.

LOGISTIC REGRESSOR

Test Classification Report:

	precision	recall	f1-score
0	0.97	0.90	0.93
1	0.46	0.75	0.57
accuracy			0.88

- The best model with ridge penalty and the best model with lasso penalty had the same class 1 recall and accuracy.

RANDOM FOREST CLASSIFIER

Test Classification Report:

	precision	recall	f1-score
0	0.98	0.86	0.91
1	0.37	0.81	0.51
accuracy			0.85

- 6% improvement in class 1 recall from the best logistic regression model.
- 3% decrease in accuracy from the best logistic regression model.

TOP PERFORMING MODEL: EXTREME GRADIENT BOOSTED (XGB) TREE CLASSIFIER

Test Classification Report:

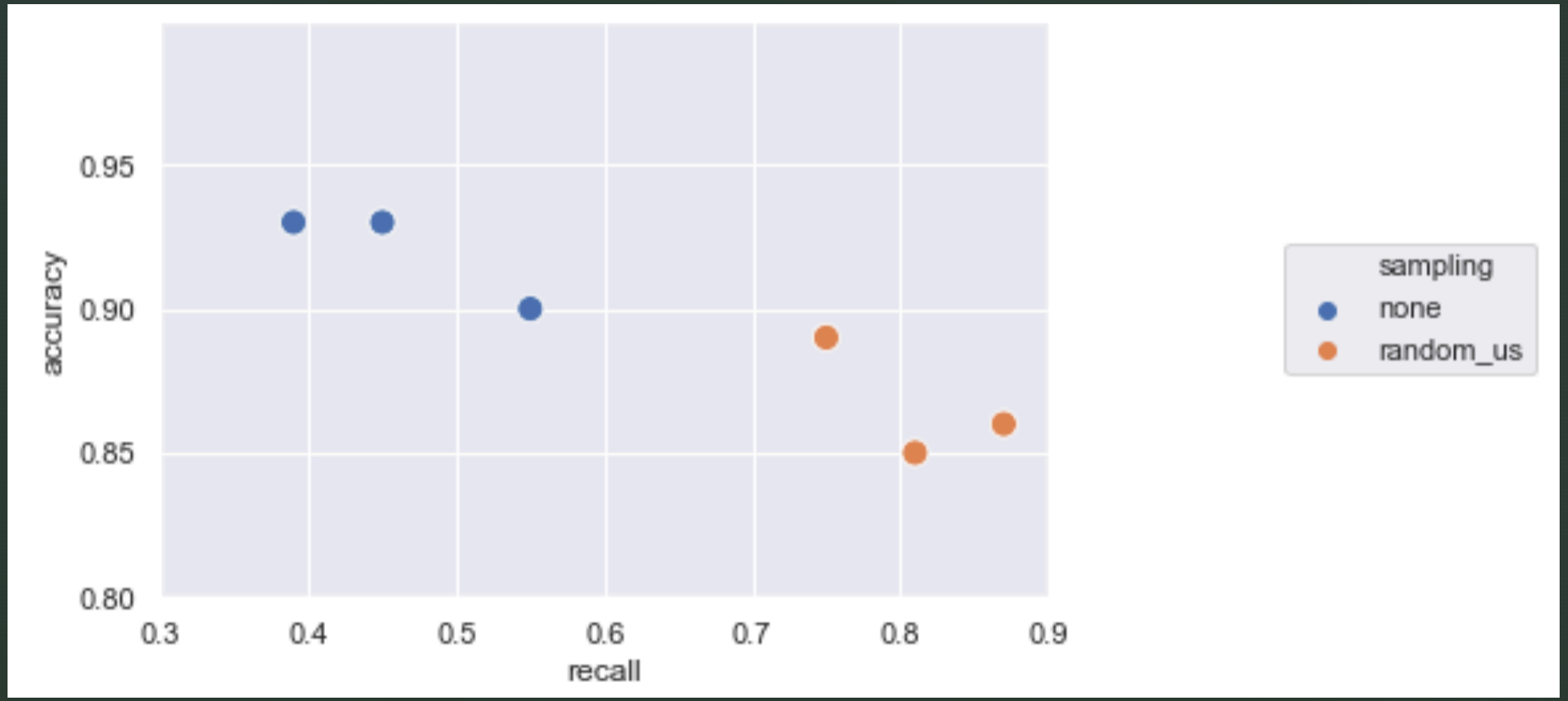
	precision	recall	f1-score
0	0.98	0.86	0.92
1	0.39	0.87	0.54
accuracy			0.86

- 6% improvement in recall from the best random forest model.
- 2% decrease in accuracy from the best logistic regression model.

MODELING NOTE: RANDOM UNDER-SAMPLING

- How it works: The algorithm takes the minority class (potentially fraudulent providers) and resamples with replacement until the size of the resampled class matches the size of the majority class
- Improved Class 1 recall across all models and decreased accuracy.
- The large gains in recall were beneficial in identifying more fraudulent providers

EFFECT OF RANDOM UNDERSAMPLING:



Recall is higher and accuracy is lower when implementing undersampling on the same models. Accuracy-recall tradeoffs were greater in XGB and random forest models than in logistic models.

XGB Feature Importances:

	Variable	Importance
4432	diag_41401	0.066810
4365	diag_40390	0.044962
5691	diag_5990	0.028702
4527	diag_4280	0.024929
4446	diag_4149	0.020869

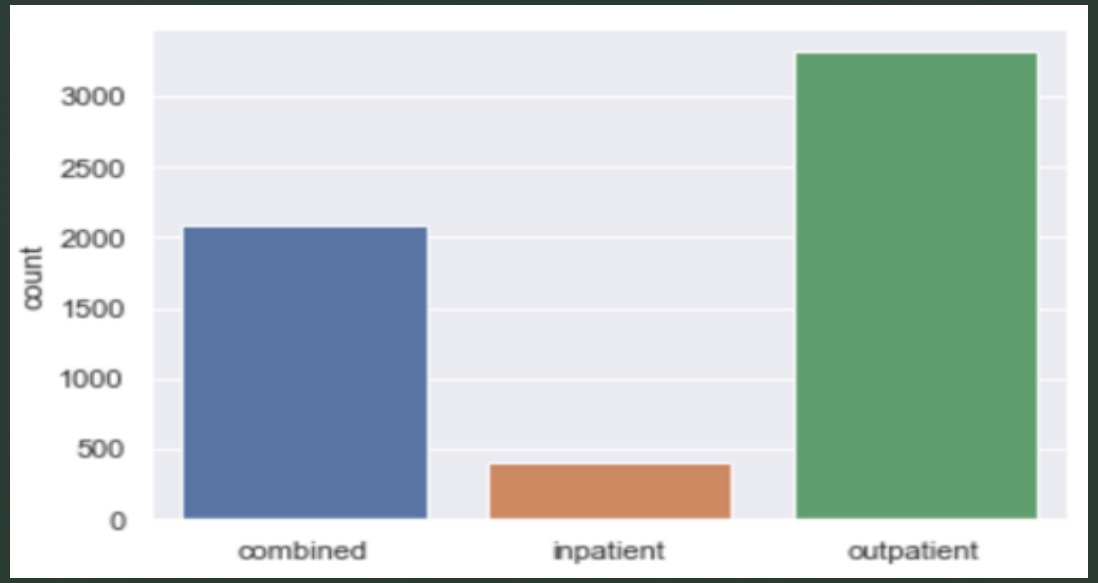
SUMMARY

- Mean Insurance Claim Amount Reimbursed was important in predicting potential fraud across all models
- Mean duration was one of the most important features in random forest models and XGB models.
- Counts of the following ICD-9 Diagnostic Codes indicated potential fraud in random forest and XGB models:
  - Coronary atherosclerosis of native coronary artery (41401)
  - Hypertensive chronic kidney disease, unspecified, with chronic kidney disease stage I through stage IV, or unspecified (40390)
  - Urinary tract infection, site not specified (5990)

CONCLUSIONS

- The top-performing model was an XGB Trees classifier with a Class 1 recall of 87% and accuracy of 86% in predicting potential provider fraud.
- The model with higher accuracy and the least comparative loss in recall was a Logistic Regressor with a recall of 75% and accuracy of 89%.
- Both models significantly improve the ability to predict provider fraud. A model using random selection to predict potential fraud has 9.31% Class 1 recall.

FUTURE WORK



Inpatient versus Outpatient Providers

- Modeling inpatient providers and outpatient providers separately could lead to new insights.
- Collapsing beneficiary and claims data to the provider level in different ways could yield alternative results due to different features.

RECOMMENDATIONS FOR THE CLIENT

- Having more data on inpatient claims and providers could assist in improving model performance with inpatient providers.
- Having additional data about health claims, particularly on the provider level, would be helping in improving model performance. For example:
  - Provider charges per claim
  - Provider operational costs
  - Provider date established
  - Provider legal history
  - Provider client reviews