Predicting Fraud in Medicare Data

Matvei Shaposhnikov

Datasets Used:

CMS Medicare Physician- by Provider (https://data.cms.gov/provider-summary-by-type-of-service/medicare-physician-other-practitioners/medicare-physician-other-practitioners-by-provider)

OIG LEIE Exclusions (https://oig.hhs.gov/exclusions/exclusions_list.asp)

Main Methods:

Python EDA, Linear Regression, Random Forest, XGBoost

Main Goal:

Create a model agencies like OIG can use to detect Medicare Fraud quickly and efficiently

Motivation & Questions

Context: **Every year our government loses millions to medicare fraud**

Why: **This will help Americans by improving Medicare spending/funding**

How: **Understanding and using existing fraud data to predict fraud in CMS data**

Q1. **What factors are most associated with fraud**

Q2. **Can I build a model that predicts with <u>useful</u> accuracy**

Q3. **How competitive would a linear regression be with a forest based predictive model**

Data (pre-EDA)

CMS	OIG
Time per.: **2022-2025**	Time per.: **2022-2025**
Size: **~1M rows (per year), 108 col**	Size: **8274 rows, 17 col**
Key var.: **NPI, Type, Specialty, Funding, Beneficiaries**	Key Var.: **Exclusion date, State, Exclusion type**

Data (post-EDA)

CMS	OIG
Time per.: **2022-2025**	Time per.: **2022-2025**
Size: **12,478 rows (per year), 101 col**	Size: **7723 rows, 13 col**
Key var.: **NPI, Type, Specialty, Funding, Beneficiaries**	Key Var.: **Exclusion date, State, Exclusion type**

EDA Tools

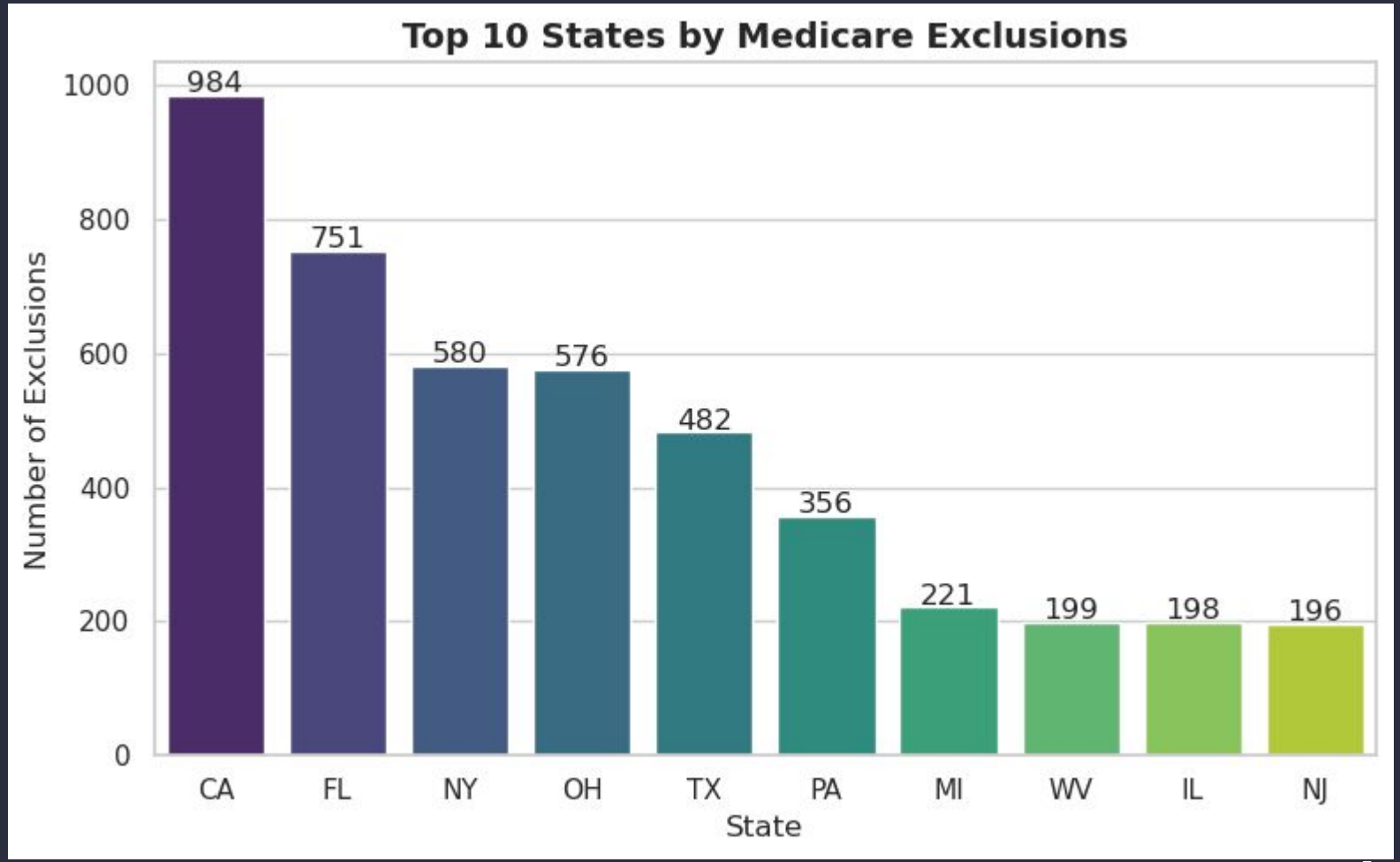
01. **Pandas, numpy, matplotlib, seaborn**

02. **Sampling/Filtering N/A rows**

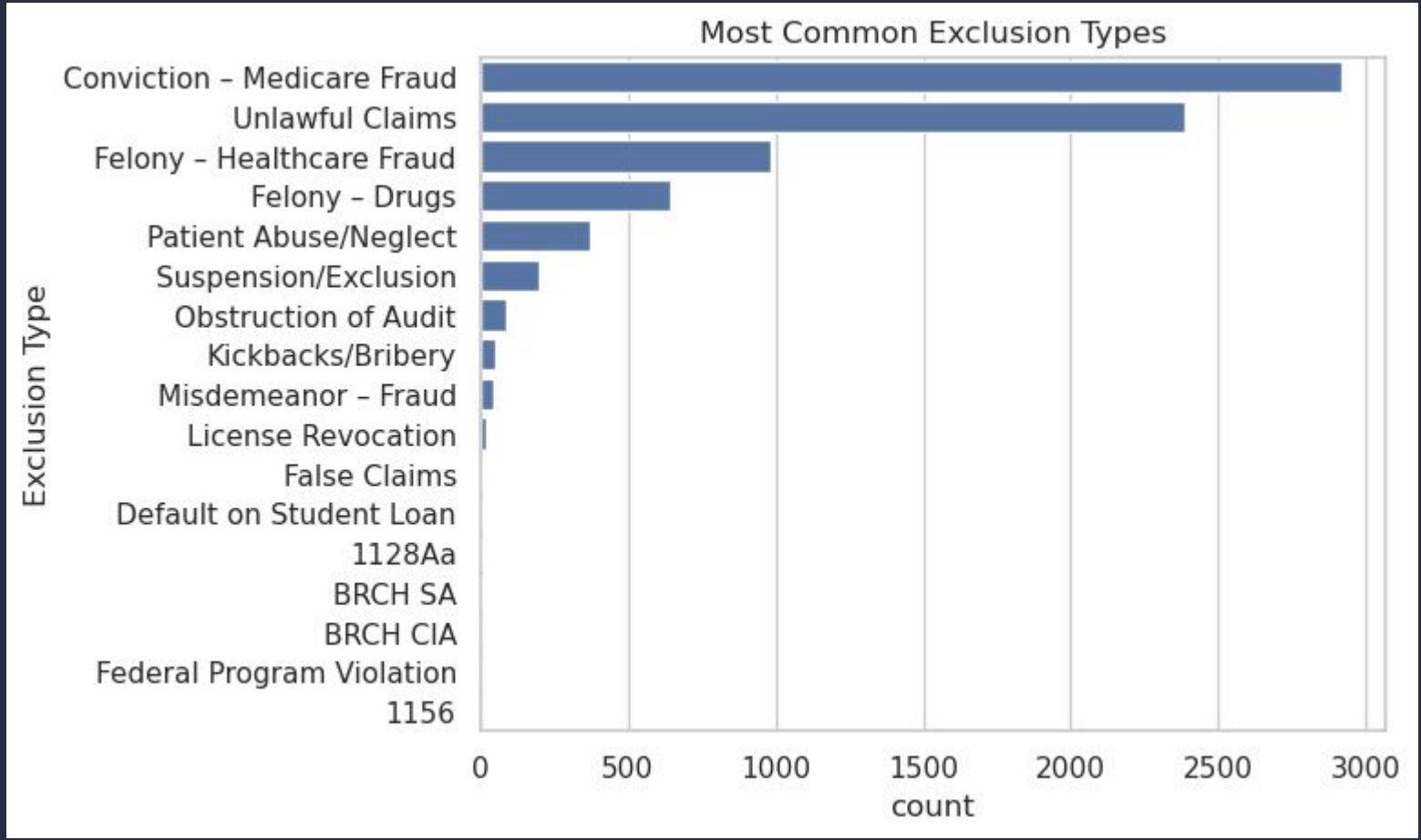
03. **Merging/Aggregating data**

04. **Creating visualizations**

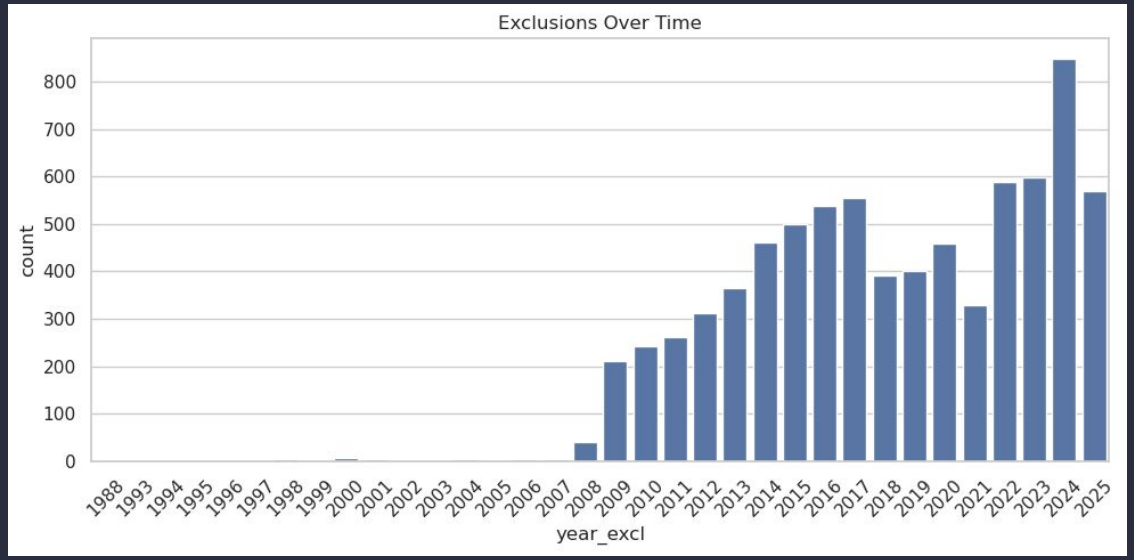
OIG

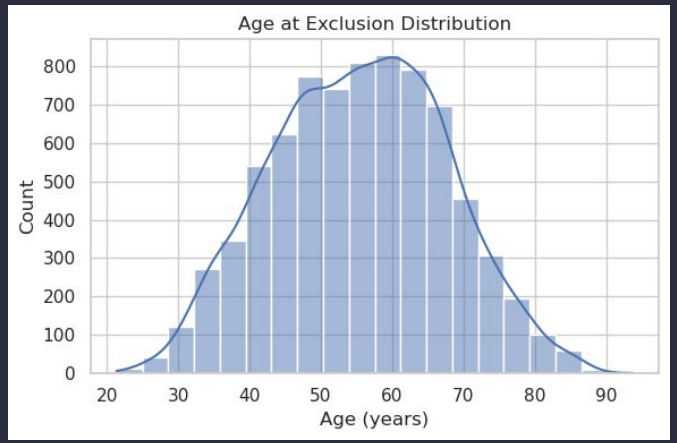

Top 10 States by Medicare Exclusions
Number of Exclusions
0
200
400
600
800
1000
984
751
580
576
482
356
221
199
198
196
CA
FL
NY
OH
TX
PA
MI
WV
IL
NJ
State

OIG


Most Common Exclusion Types
Conviction – Medicare Fraud
Unlawful Claims
Felony – Healthcare Fraud
Felony – Drugs
Patient Abuse/Neglect
Suspension/Exclusion
Obstruction of Audit
Kickbacks/Bribery
Misdemeanor – Fraud
License Revocation
False Claims
Default on Student Loan
1128Aa
BRCH SA
BRCH CIA
Federal Program Violation
1156
Exclusion Type
0
500
1000
1500
2000
2500
3000
count

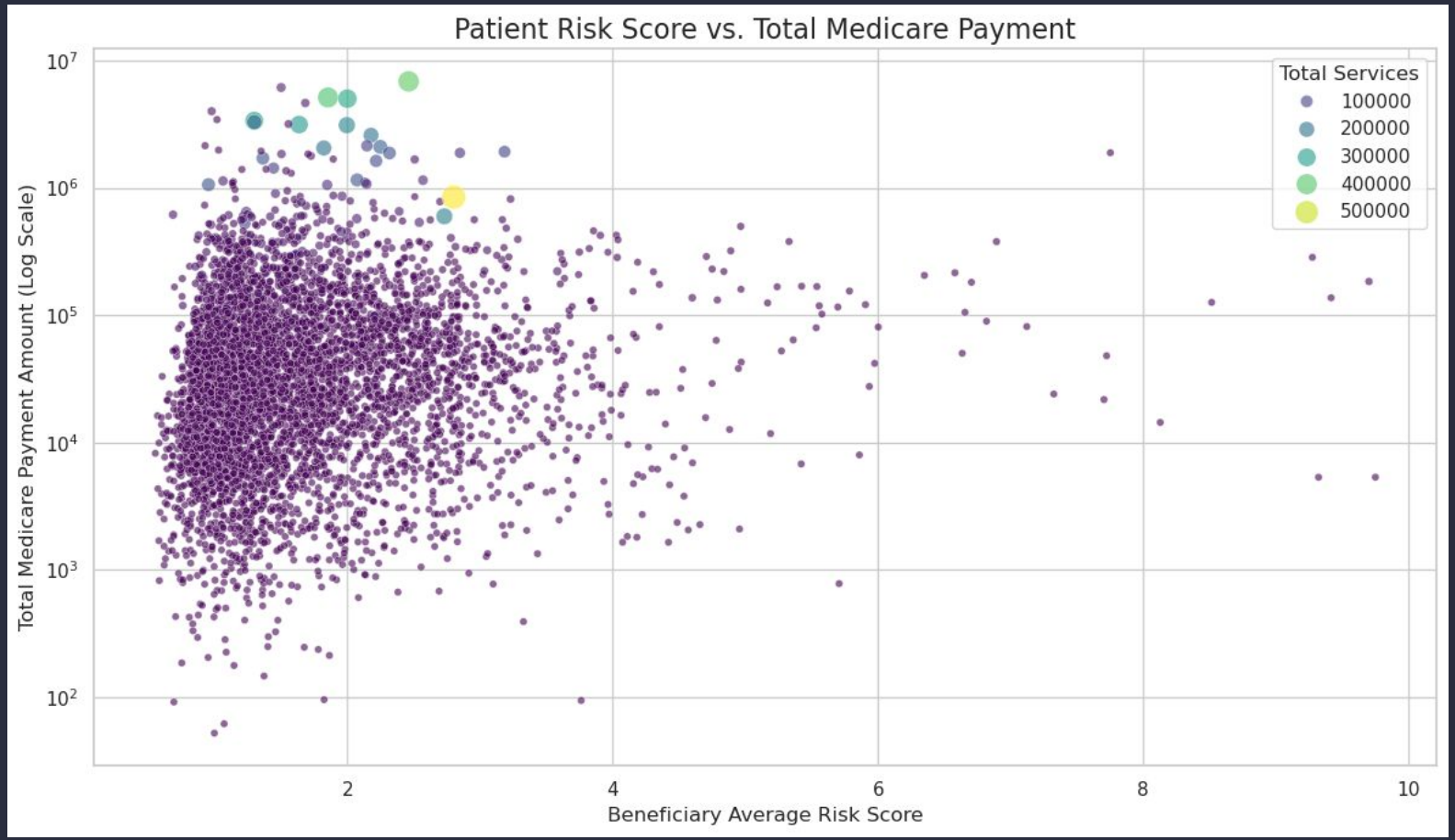
OIG





Age at exclusion is misleading, the prosecution takes years, and excluding could take up to a decade, many providers get excluded much later than they should.

CMS

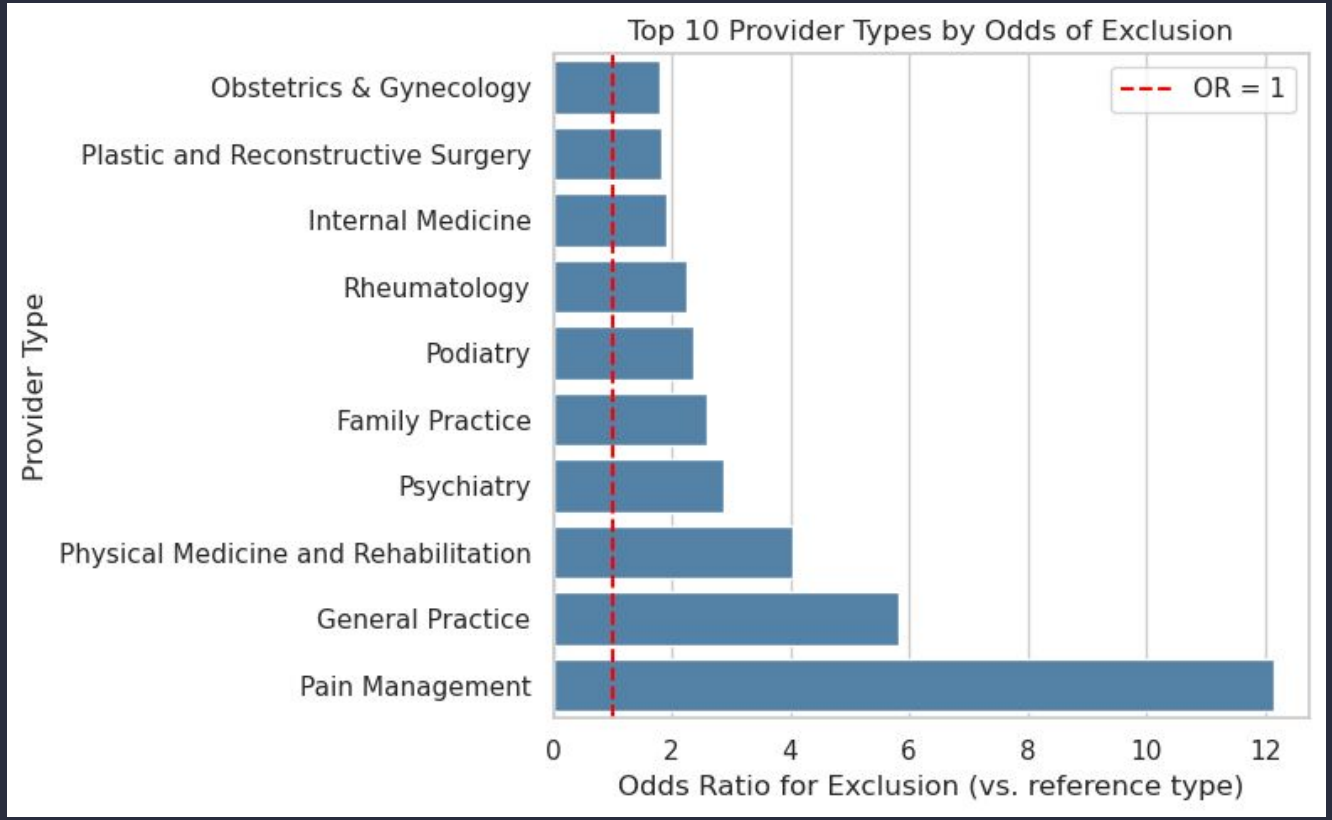

Patient Risk Score vs. Total Medicare Payment
Total Medicare Payment Amount (Log Scale)
10^7
10^6
10^5
10^4
10^3
10^2
2
4
6
8
10
Beneficiary Average Risk Score
Total Services
100000
200000
300000
400000
500000

Modeling (Linear Reg.)

Using a python libraries statsmodels and scikit-learn I constructed two Linear models to predict exclusion odds for providers by:

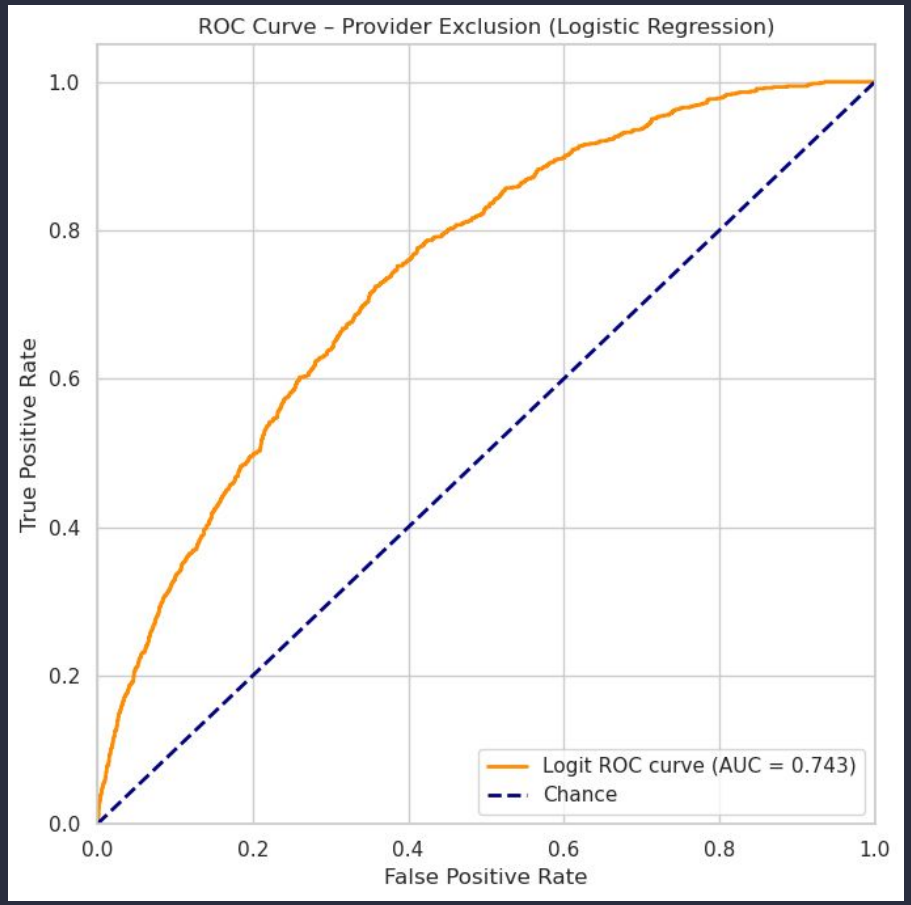
- Merging both (filtered and aggregated) datasets
- Using statsmodels to create a provider type linear regression
- Calculating Log odds/Odds ratios for each provider specialty
- Using statsmodels and scikit-learn to create an advanced LR to predict exclusions based on specialty, funding, and state

Provider fraud odds compared to reference group (OLM)



Formula used: exclusion odds ~ provider type

Regression results for predicting excluded providers



Formula used:
Excluded ~ provider type + state + log(funding received)

Sensitivity=
(True Positives+False Negatives)/True Positives

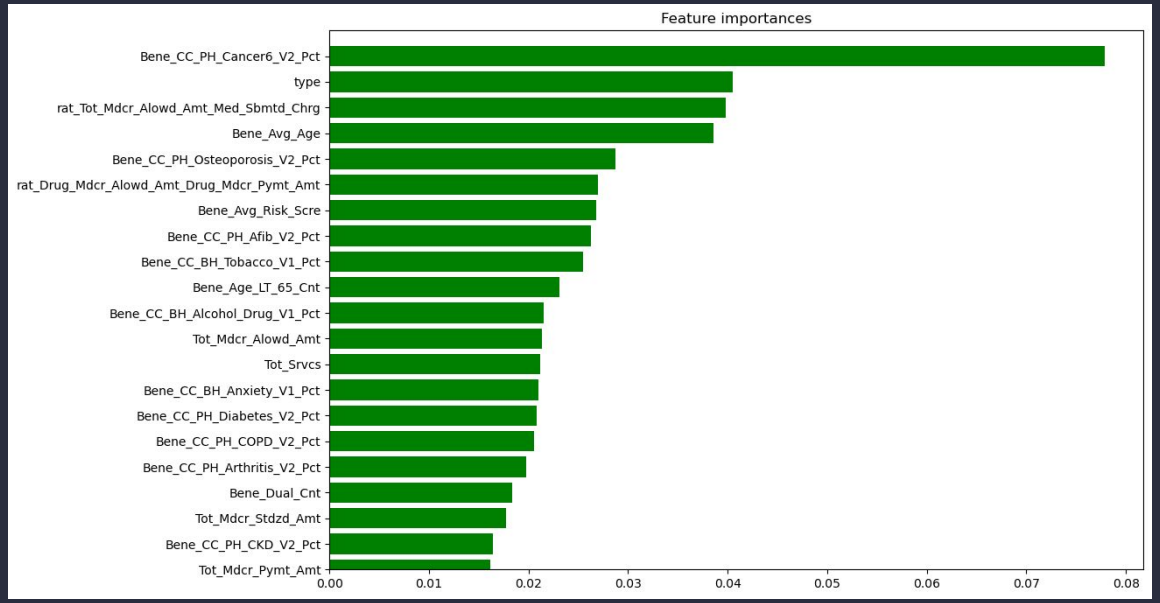
(1 - Specificity) =
(False Positives+True Negatives)/False Positives

AUC= (Sensitivity+Specificity)/2

Modeling (Predictive Models)

Using machine learning libraries (scikit-learn, RandomForest, and xgboost) I constructed two predictive models using:

- Feature engineering/rebalancing/importance
- Classifying and encoding
- Training using 'test train split' method

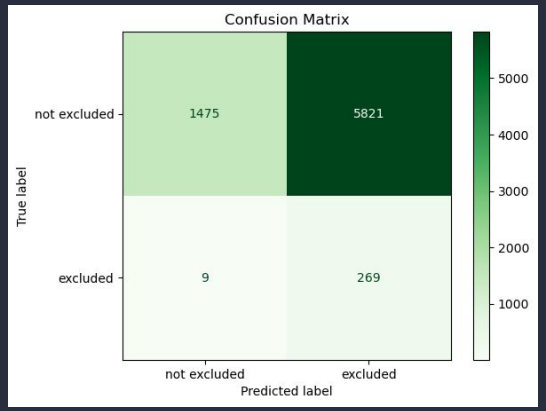


While training the models, I created this barplot of the most predictive features.

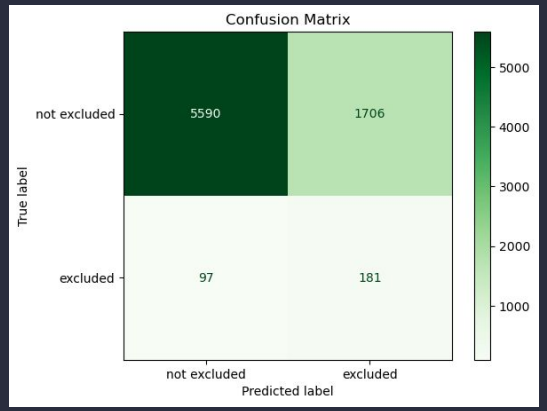
Using this graph I can retrain the model based on however many features I wish.

All that needs to be done is isolating the top most important features from the features set

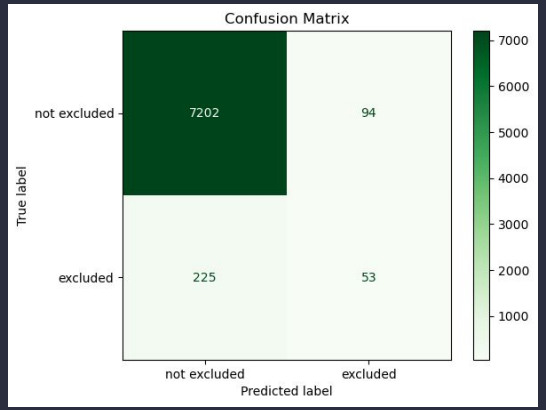
Probability filtering = 0.2



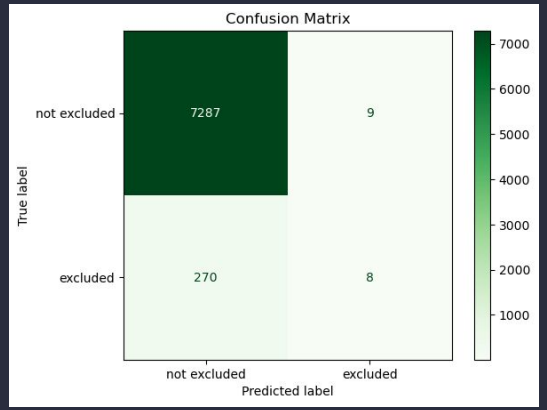
Probability filtering = 0.4

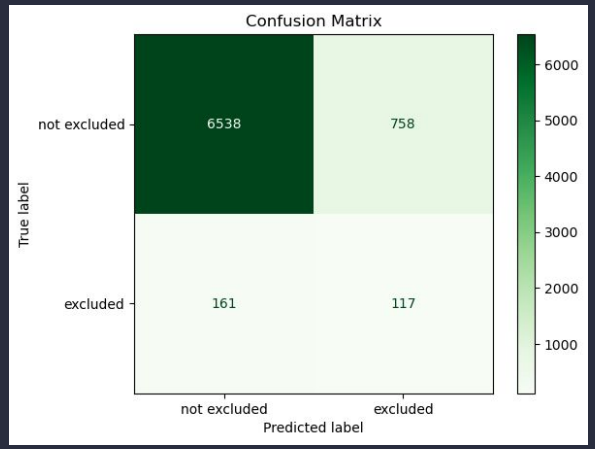


Probability filtering = 0.6

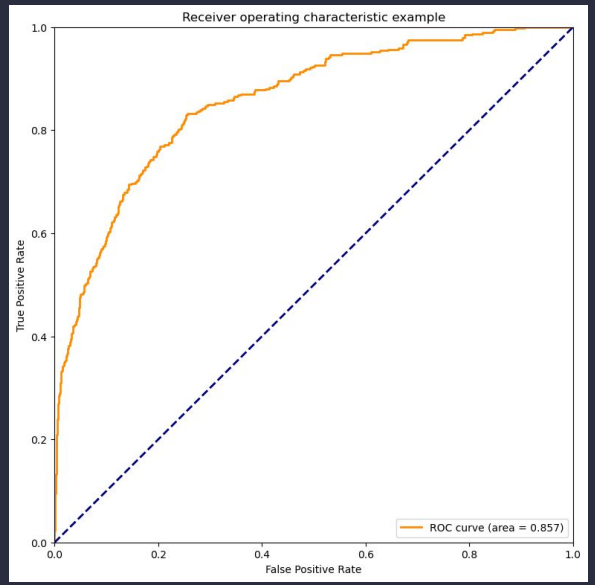


Probability filtering = 0.8



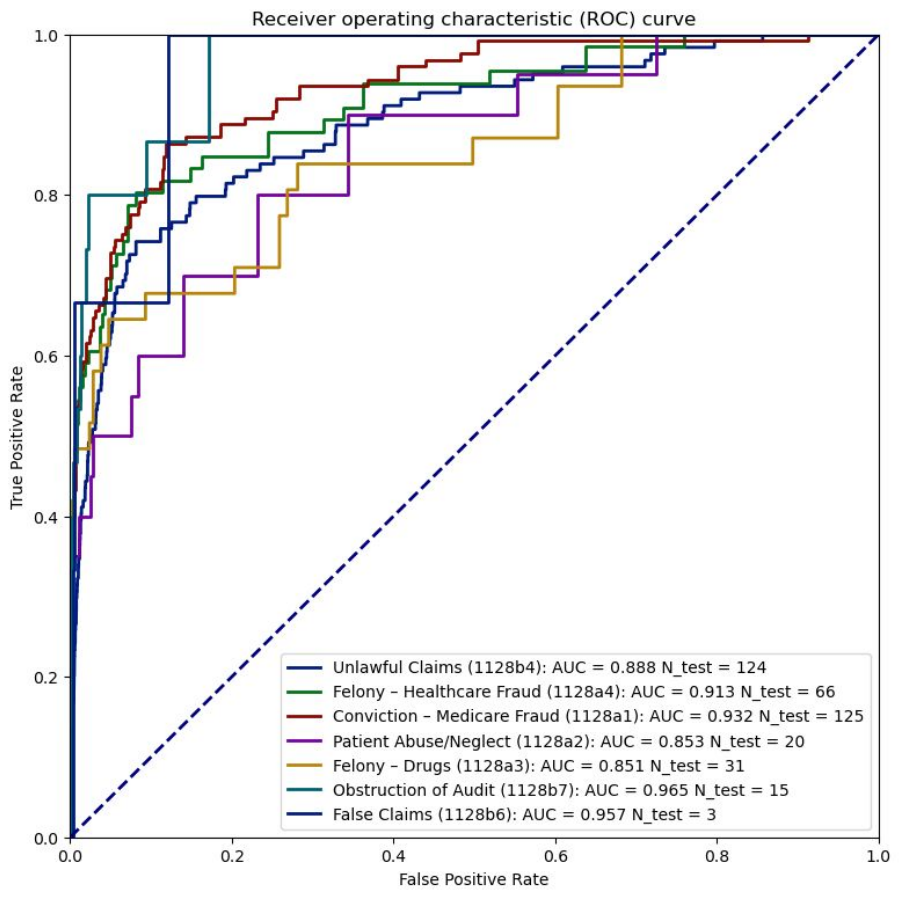


Using probability filtering = 0.55



The ROC curve shows a consistent range of 0.8-0.85 AUC which classifies the model as a good predictor

Using XGBoost I went further to predict by exclusion type, and the model can predict fraud with an average AUC of 0.90



Conclusion

Both the Linear Regression and Predictive model serve their own purposes but can work together to make an efficient solution to the medicare fraud issue.

- The linear regression can be used in tandem with the predictive model
- The robust and reliable structure of the Linear regression can be used in Prosecution and audits
- The Predictive model will be used to sift through large datasets to create a list of potential criminals to investigate
- Together they create an efficient method to find, audit, and prosecute any providers who file fraudulent claims.

Possible Next Steps:

- Refine the model further, using math to find the perfect probability filter and push the model to minimize false positives/negatives
- Create additional models tuned for different business purposes
- Create a decision process that will sort data based on prediction made by the model

Acknowledgements

Prof. Lori Perine - Knowledge and skills provided guided me through this project
My Father - Essential advice and methods provided from a senior data analyst
CMS data analysis team - Provided key context on data allowing deep insightful analysis

Thank you for listening!

Any questions you have are welcome now.

Link to my DATA 205 Github repository:
https://github.com/matsha2266/DATA205