# Predicting Opioid Drug Prescriptions among Prescribing Specialists

*Jimmy Smart*

## 1. Introduction and Data background

Accidental death by fatal drug overdose is a rising trend in the United States. What can we do to help?

Below is a dataframe that I will be using to dive into the opioid crisis as I wrangle through and provide visuals that pertain to the csv dataframes I have available.

The 1st one, opioid.csv, lists the known opioid drug names and their generic names. The Overdoses.csv contains data provided from the 50 states in regards to their population and opioid related deaths. Finally, the Prescriber-info.csv provides data regarding a specified prescriber. their gender, state, and speciality are provided, as well as the drugs they prescribed to their patients. Opioids and non opioid drugs.

This dataset contains summaries of prescription records for 250 common opioid and non-opioid drugs written by nearly 25,000 unique licensed medical professionals in 2014 in the United States for citizens covered under Class D Medicare as well as some metadata about the doctors themselves. This is only a small subset of data that was sourced from a much larger file: cms.gov.

The data was acquired from Kaggle: U.S. Opiate Prescriptions/Overdoses by Alan “AJ” Pryor, Ph.D. The full dataset contains almost 24 million prescription instances in long format. In the Kaggle form, the data has already been previously cleaned and compiled here in a format with 1 row per prescriber and limited the approximately 1 million total unique prescribers down to 25,000 to keep it manageable.

## 2. Key for reading the dataset

NPI – unique National Provider Identifier number
Gender - (M/F)
State - US State by abbreviation
Credentials - set of initials indicative of medical degree
Specialty - description of type of medicinal practice (109 unique specialties)
Opioid.Prescriber - a boolean label indicating whether or not that individual prescribed opiate drugs more than 10 times in the year

## 3. Regression Analysis

To better simplify our dataset we'll drop the list of drug names as our main concern to the total amount of opioid and non opioid prescriptions prescribed by our specialists. We'll look at the correlation between the sum of Opioids and non opioids along with population, states and our specialists.


| | Gender | State | Specialty | Population | Opioid_Prescriber |
|---|---|---|---|---|---|
| **Gender** | 1.000000 | -0.009391 | -0.144758 | 0.014512 | 0.067798 |
| **State** | -0.009391 | 1.000000 | 0.015114 | -0.262505 | 0.000096 |
| **Specialty** | -0.144758 | 0.015114 | 1.000000 | 0.000354 | -0.062143 |
| **Population** | 0.014512 | -0.262505 | 0.000354 | 1.000000 | -0.038426 |
| **Opioid_Prescriber** | 0.067798 | 0.000096 | -0.062143 | -0.038426 | 1.000000 |

We can see the correlations our Opioid_Prescriber variable has with the other variables within our dataset. The correlations aren't super strong.

## 4. Ordinary Least Squares Regression Analysis

For our model we went with Opioid_Prescriber as our dependent variable and added Gender, State and Speciality variables. As the image below shows, our R-squared score of 0.338 isn't as high as we would have liked. The variables we used are categorical variables, meaning we must turn the values into numerical values (0 or 1) using C().

| | | | |
|---|---|---|---|
| **Dep. Variable:** | Opioid_Prescriber | **R-squared:** | 0.338 |
| **Model:** | OLS | **Adj. R-squared:** | 0.334 |
| **Method:** | Least Squares | **F-statistic:** | 79.92 |
| **Date:** | Mon, 04 May 2020 | **Prob (F-statistic):** | 0.00 |
| **Time:** | 10:00:16 | **Log-Likelihood:** | -12485. |
| **No. Observations:** | 24759 | **AIC:** | 2.529e+04 |
| **Df Residuals:** | 24601 | **BIC:** | 2.657e+04 |
| **Df Model:** | 157 | | |
| **Covariance Type:** | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| **Intercept** | 0.5862 | 0.241 | 2.432 | 0.015 | 0.114 | 1.059 |

To explain:

**Adj. R-squared** indicates that 33.8% of Opioid Prescriptions can be explained by our predictor variables.

The **regression coefficient (coef)** represents the change in the dependent variable resulting from a one unit change in the predictor variable, all other variables being held constant. In our model, a one unit increase in our variables increases our index.

The **standard error** measures the accuracy of our variables coefficient by estimating the variation of the coefficient if the same test were run on a different sample of our population. Our standard error(s) are low and therefore appears accurate.

The **p-value** means the probability of our coef decrease in Opioid_Prescriptoins due to a one unit increase in our variables is 0.01%, assuming there is no relationship between the two variables. A low p-value indicates that the results are statistically significant, that is in general the p-value is less than 0.05.

## 5. Random Forest Regression

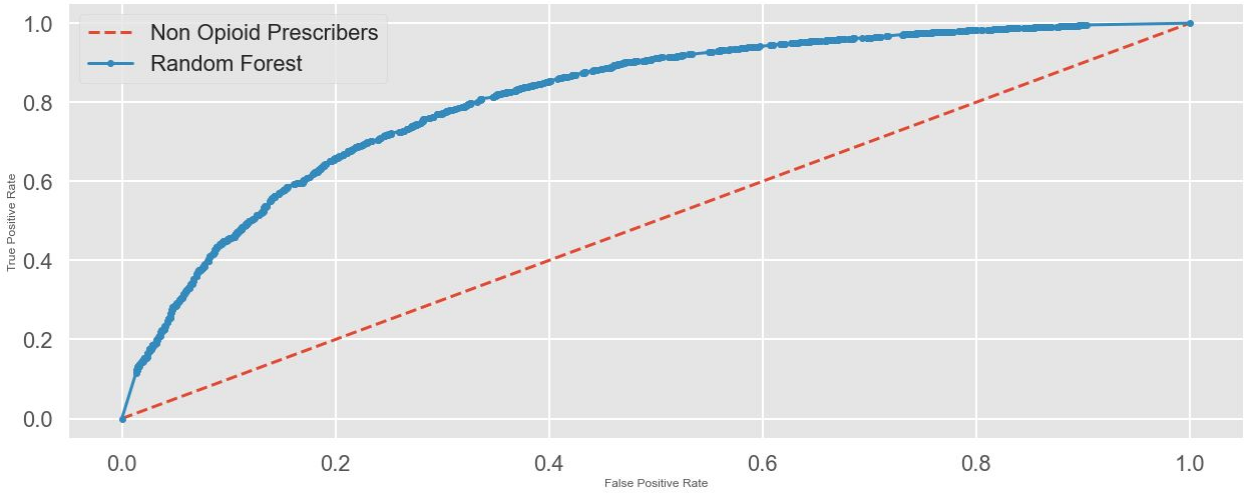
The Random Forest model produced a low accuracy R-squared score of 0.273 and produced a good accuracy a score: 0.7427551

| | Feature | importance | rank |
|---|---|---|---|
| **74** | Specialty_Family Practice | 0.106276 | 1 |
| **0** | Population | 0.070633 | 2 |
| **71** | Specialty_Emergency Medicine | 0.066525 | 3 |
| **91** | Specialty_Internal Medicine | 0.063420 | 4 |
| **119** | Specialty_Orthopedic Surgery | 0.048439 | 5 |

We can see that Family Practice, Emergency Medicine, Internal Medicine and Orthopedic Surgery specialists came in at the highest importance. Population is also a factor.

Our **Baseline Test (default parameters)** produced:

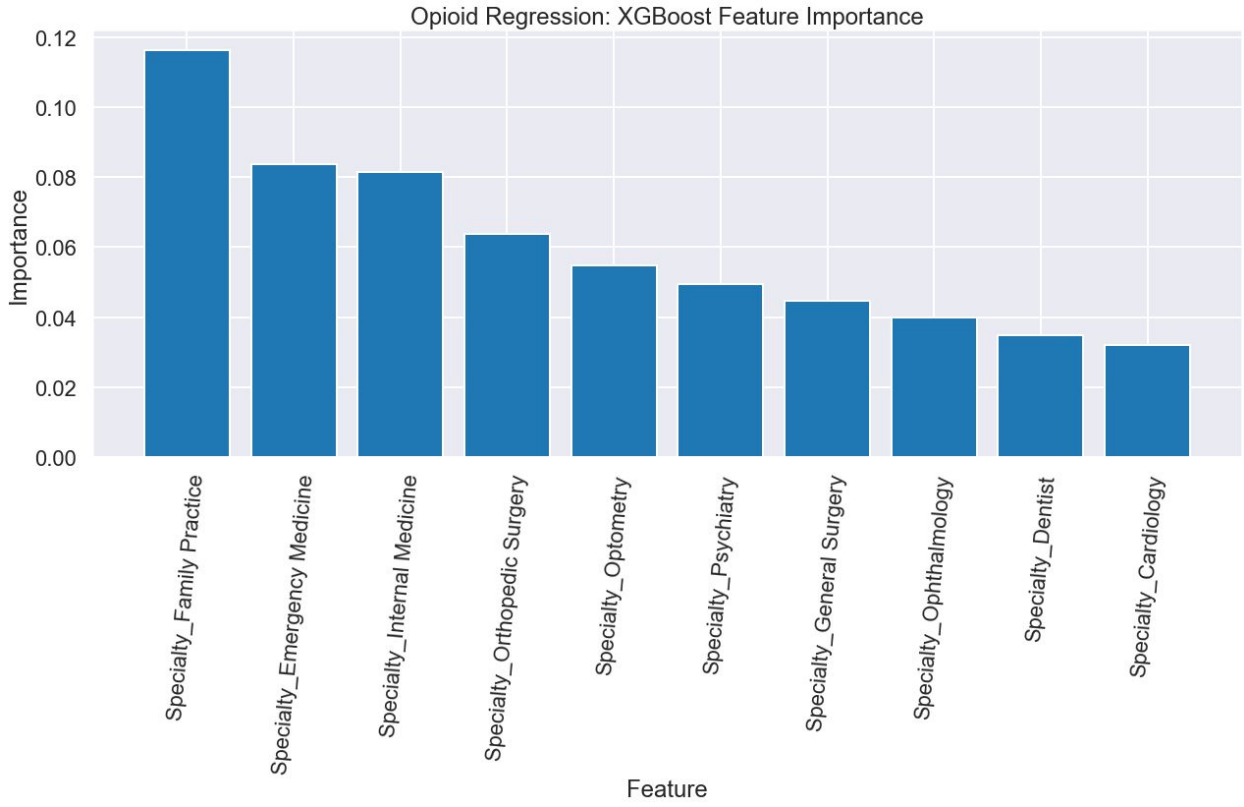
| Model | Class | Precision | Recall | F1-score | Support | Accuracy |
|---|---|---|---|---|---|---|
| Random Forest Classifier | 0 (less than 10 prescriptions) | 0.68 | 0.70 | 0.69 | 3084 | 0.74 |
| | 1 (10 or more opioid prescriptions) | 0.79 | 0.77 | 0.78 | 4344 | |




We created our model based on our best performing variables with our best performing model with a good accuracy rate. The values (Precision, Recall and F1-Score) could be higher (1) but are still good.

## 6. XGBoost

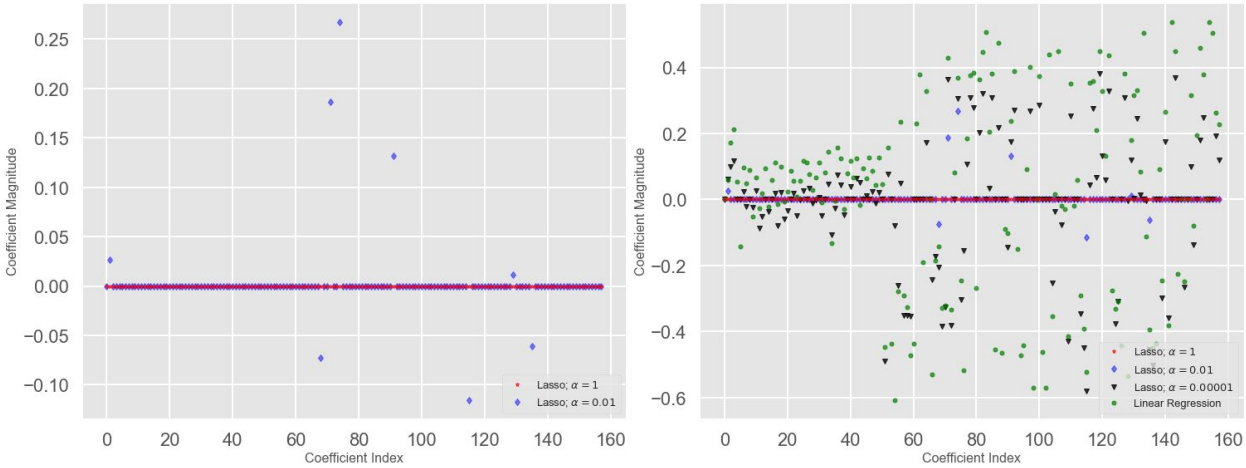
Our XGBoost model produced a test score of 0.247 and a training score of .253.

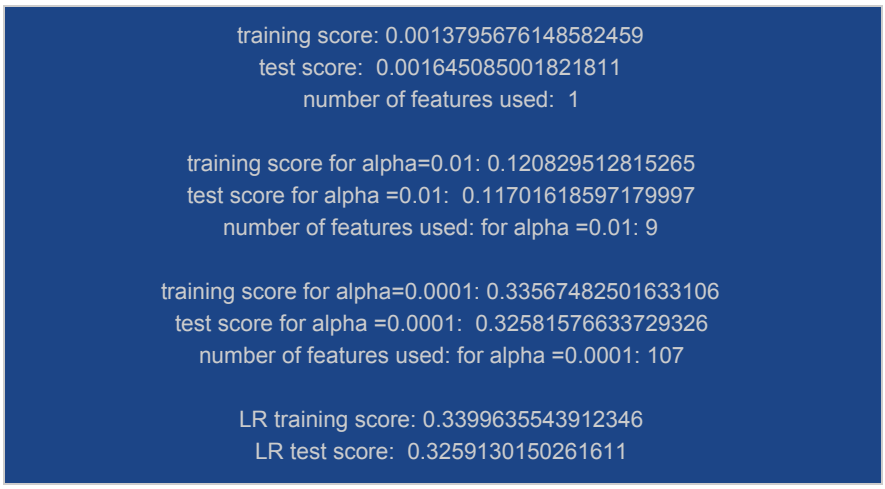



Just like in previously with our Random Forest model, this model we also can see that Family Practice, Emergency Medicine, Internal Medicine and Orthopedic Surgery specialists came in at the highest importance. Population is also a factor.

## 7. Lasso

We chose Lasso regression since it not only helps in reducing over-fitting but it can help us in feature selection.






The default value of regularization parameter in Lasso regression (given by α) is 1.

With this, out of the features our dataset, only 1 feature is used (non zero value of the coefficient).

Both training and test score (with only 1 features) are low; conclude that the model is under-fitting the dataset. Reduce this under-fitting by reducing alpha and increasing number of iterations. Now α = 0.01, non-zero features= 9, training and test score increases.

For alpha= 1, we can see most of the coefficients are zero or nearly zero, which is not the case for alpha=0.01. Further reduce α =0.0001, non-zero features = 107. Training and test scores are similar to basic linear regression case. In the right panel of figure, for α = 0.0001, coefficients for Lasso regression and linear regression show close resemblance.

## 8. Comparing Models

Our models weren't especially impressive but offered valuable information. Adding more variables to our OLS model only caused our r-squares results to raise slightly. Random Forest and XGBoost offered our highest accuracy along with insightful information in forms of providing us an insight of which specialists had a higher chance of prescribing opioid prescriptions.

| Name | r-squared |
|---|---|
| OLS Regression | 0.337779 |
| Random Forest | 0.272954 |
| XGBoost | 0.247265 |
| Lasso | 0.339964 |

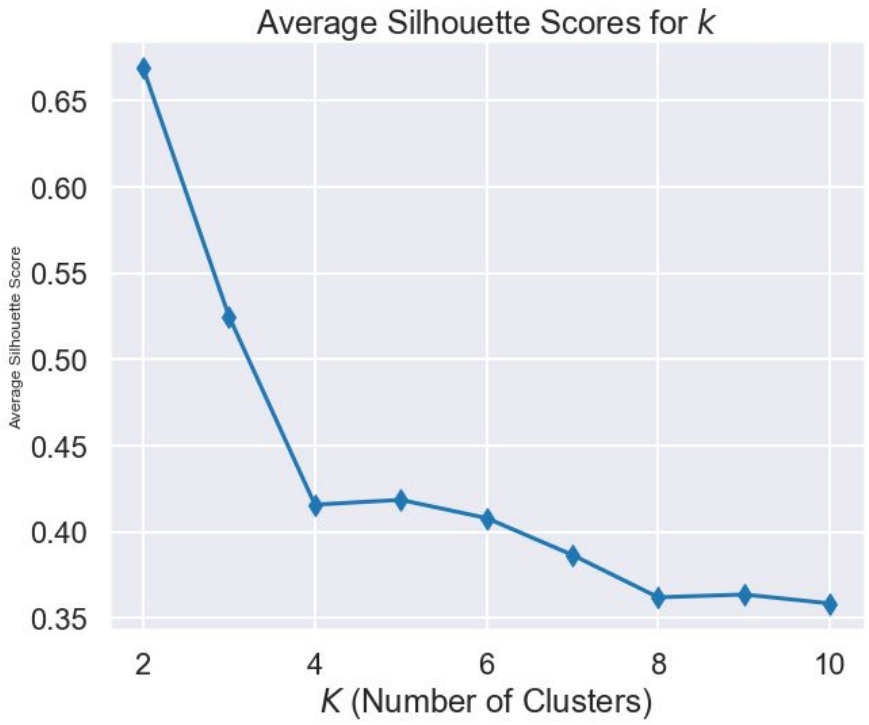
## 9. Clustering

We created two versions for our clustering models:

**Clustering with Specialty and SumOpi columns (SumOpi = total sum of opioids prescribed per specialists)**

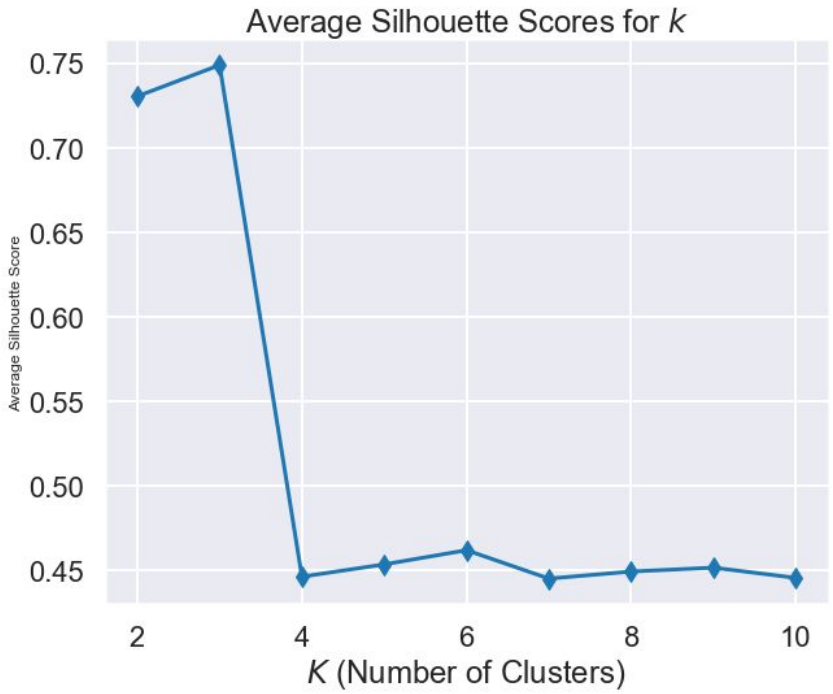
*and*

**Clustering with Specialty and NonOpi columns (NonOpi = total sum of non opioids prescribed per each specialists)**




Silhouette method (SumOpi model) suggests that the $K$ value of 2 is the best value to choose since it has the highest score. At over 0.65 it is a reasonable structure. The rest of the clusters having scores under 0.50 means their structures are weak. In the future, we should probably increase our test range.

**In comparison:**




Silhouette method (NonOpi model) suggests that the $K$ value of 3 is the best value to choose since it has the highest score. At 0.75 it is a strong structure. Cluster 2 is also strong. The rest of the clusters having scores under 0.50 means their structures are weak. In the future, we should probably increase our test range.

## 10. Summary

Our models did not score as high as we would have liked but we were able to see some benefits from them. Certain types of specialists and locations did show that they were factors when predicting opioids prescriptions. More tuning and reworking with the categorical variables will be needed to better tune the models.

## 11. Next Steps

Further modeling and hyperparameter tuning can be done. Additional data would also be needed to better improve the model