Dylan Mather

4/26/2021

CMSE 381

Opioid Drugs

Background:

There is a rising trend in accidental death by overdose in the United states and part of that problem is some medical professionals are prescribing more than 10 opioids in a year. This heavily depends on what their specialty is and what other drugs they are prescribing. Some of these professionals seem to be justified in their higher rate of opioids they are prescribing based on their field of work however that does not change the fact that the death toll caused by overdoses from these drugs are rising.

My main goal for this project is try an predict if a medial professional is likely to prescribe opioids more than 10 times in a year based off the other drugs they use, their specialty and the deaths per capita from drug overdose in the state they practice. I hope to identify which professions pose the greatest risk to the rising overdose epidemic as well which states need stricter policies to help begin to fix this problem we are facing in our country.

In order to answer this question, I took my data from three data sets from kaggle. The main data set is the prescriber data which has information about each medical professionals' specialty, credentials, gender, state in which they practice, and a long list of drugs they prescribed in that year. The opioids data set is simply a list of opioid drug names, and the overdoses has information about how many deaths were in each state as well as the population of that state. I used the two smaller data sets to split the prescriber data into two categories, one that has only opioid drugs and one that has all the rest. Additionally, I added the data from the overdose data to the main data, also adding a deaths per captia column so I could more fairly compare death in states with large differences in population sizes.

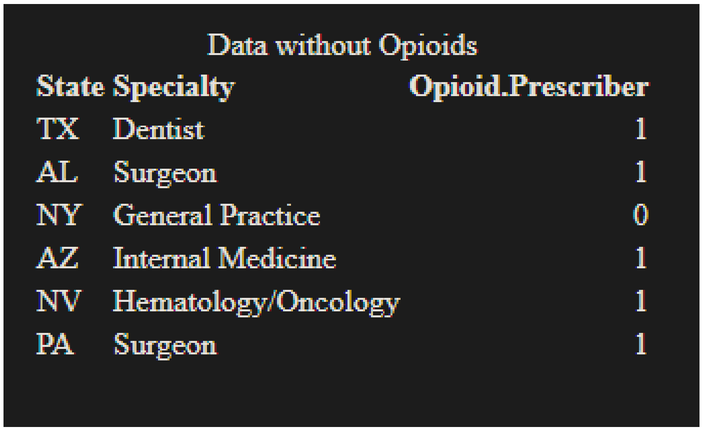
Related Project

https://www.kaggle.com/apryor6/detecting-frequent-opioid-prescription

This link will take you to the R project made by the person who posted this data on kaggle. His way of cleaning the data was very good but I also saw some ways in which I wanted to change it. I wanted to make the two seperate data sets with opioids and without as well as adding the additional layer of deaths per capita. In his project he used a boosting method that takes very long to run and gets a decent accuracy, but I decided to use a logistic regression, lasso and ridge regression models instead since they are less computationally expensive and I found they produced similar results.

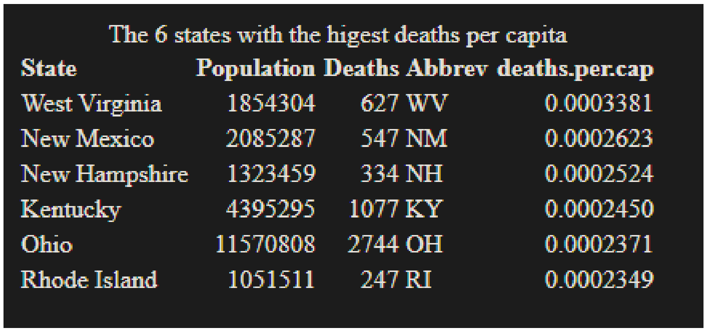
Data

The heads of these two data sets show the cleaned data with all the added dimensions I wanted to add. They do not show all the drugs included to save space on this report.



Data without Opioids

State	Specialty	Opioid.Prescriber
TX	Dentist	1
AL	Surgeon	1
NY	General Practice	0
AZ	Internal Medicine	1
NV	Hematology/Oncology	1
PA	Surgeon	1

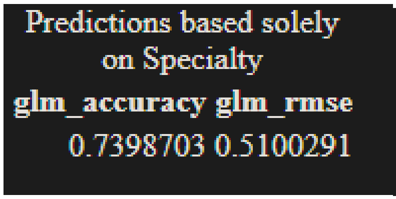


The 6 states with the higest deaths per capita

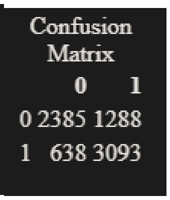
State	Population	Deaths	Abbrev	deaths.per.cap
West Virginia	1854304	627	WV	0.0003381
New Mexico	2085287	547	NM	0.0002623
New Hampshire	1323459	334	NH	0.0002524
Kentucky	4395295	1077	KY	0.0002450
Ohio	11570808	2744	OH	0.0002371
Rhode Island	1051511	247	RI	0.0002349

GLM Regression

I chose to use GLM regression for this project since It does well with data with large dimensionality. This function makes a linear model that is perfect for classification problems. I first fitted my data to predict the Prescriber column based on Specialty since I expected this to be one of the most important dimensions when predicting this value. When assessing the summary of the fit, some of the different specialties were more influential than the others, however I was surprised to see that Family Practice was included in that list. I think this is a red flag for this practice since its the odd one out of the other specialties since most of the others are in pain management or surgery. The accuracy of this model was decent, having a 73% prediction rate which was worse than the project on kaggle.

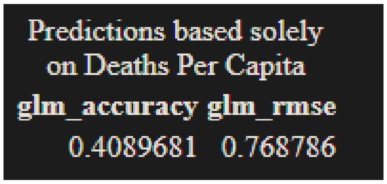


```
Predictions based solely
      on Specialty
glm_accuracy glm_rmse
     0.7398703 0.5100291
```

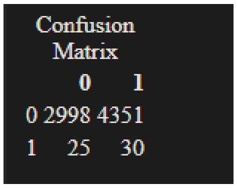


```
Confusion
  Matrix
     0    1
0 2385 1288
1  638 3093
```

I also wanted to look at how well my new column would predict the prescriber column. When I ran this fit it did not work well at all, having an accuracy significantly lower than 50% which would be the prediction rate if we simply flipped a coin. After looking the confusion matrix for this fit I can see that the model predicts most of the time that the doctor did not prescribe ten opioids in one year.

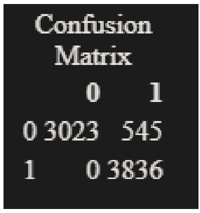


```
Predictions based solely
  on Deaths Per Capita
glm_accuracy glm_rmse
     0.4089681   0.768786
```

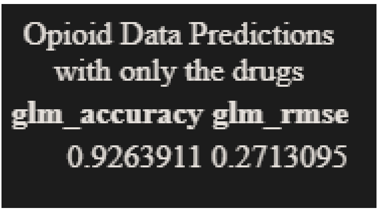


```
Confusion
  Matrix
     0    1
0 2998 4351
1   25   30
```

The last fit to predict the Opioid.Prescriber column I did for the data that included the opioid drugs had the best accuracy of the project. I fitted the data with only using the drug columns and got a 92% prediction accuracy. This high number could be due to the fact that the prescriber column is whether a medical professional prescribed more than 10 opioids in a year.

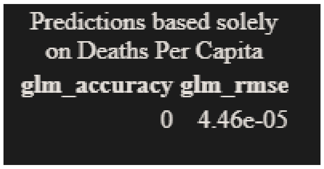


```
Confusion
  Matrix
     0    1
0 3023  545
1    0 3836
```



```
Opioid Data Predictions
   with only the drugs
glm_accuracy glm_rmse
     0.9263911 0.2713095
```

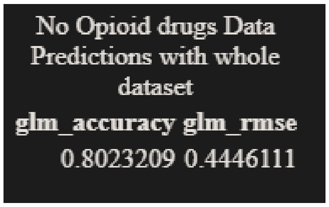
The last fit I did with the data that included the opioids and deaths per capita was to use the data to predict the deaths per capita. I could not easily get accuracy for deaths per cap estimate since I would need to pick how much to round the data which would change its accuracy, so I only show RMSE. One of the reasons I could have such a low RMSE for calculated deaths per capita is that they are the same for every state. I hoped to counteract this problem by only predicting it based off of specialty and the drug columns. Based on the RMSE value I calculated I can conclude that this model was very successful at predicting the deaths per capita so I think adding this column to the data was a good decision overall.



Predictions based solely on Deaths Per Capita

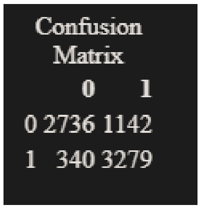
glm_accuracy	glm_rmse
0	4.46e-05

The first fit I did with the data that did not include the opioid drugs was with the whole data set and I achieved a prediction accuracy of 80%. It is not surprising that this model is not as accurate as the one with the opioid drugs, but I was still impressed with how well it did. When looking at the next model where I only used the non-opioid drugs, I can see that it has a similar prediction accuracy as the fit with only the specialty column. This tells me that both the Specialty column and the drugs columns have decent predictions on their own but produce a much more accurate fit when joined together.



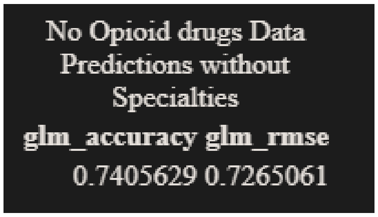
No Opioid drugs Data Predictions with whole dataset

glm_accuracy	glm_rmse
0.8023209	0.4446111



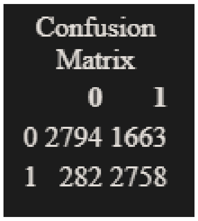
Confusion Matrix

	0	1
0	2736	1142
1	340	3279



No Opioid drugs Data Predictions without Specialties

glm_accuracy	glm_rmse
0.7405629	0.7265061

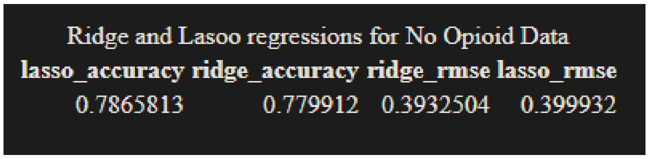


Confusion Matrix

	0	1
0	2794	1663
1	282	2758

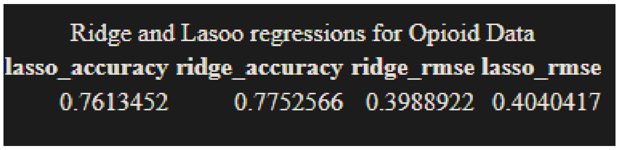
Lasso and Ridge Regression

For this section of the project, I did a ridge and lasso regression for each of the data sets to see if there was a difference in choosing the data with the opioid drugs or the one without them. I also wanted to see whether lasso or ridge worked better to see if there were many columns that had a large affect, meaning ridge regression would perform better or the opposite. I found that for both data sets ridge gave a higher prediction accuracy meaning that no one column had an overwhelming affect when predicting the prescriber column.



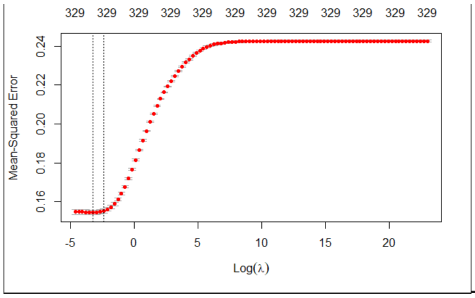
Ridge and Lasoo regressions for No Opioid Data

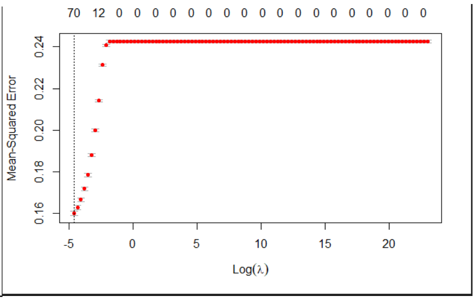
lasso_accuracy	ridge_accuracy	ridge_rmse	lasso_rmse
0.7865813	0.779912	0.3932504	0.399932



Ridge and Lasoo regressions for Opioid Data

lasso_accuracy	ridge_accuracy	ridge_rmse	lasso_rmse
0.7613452	0.7752566	0.3988922	0.4040417





These graphs helped me choose my gamma parameter for each of the lasso and ridge regression prediction models.

Results

Over the course of this project, I found that the GLM model performed very well. Even though I did not do as well as a job cleaning the data as the project on kaggle I had prediction accuracy that were not much worse. I was surprised at how well my model did a predicting the deaths per capita, having the lowest RMSE by far than the rest of the project. If I had more time to continue this project I would like to see If I could clean my data differently so, get better accuracy and predict different aspects of the data than solely the Opioid prescriber. I think that we can use data like this to help find the root of the opioid epidemic and hopefully help put an end to it.