Emily Lang
April 18, 2021
CMSE 381

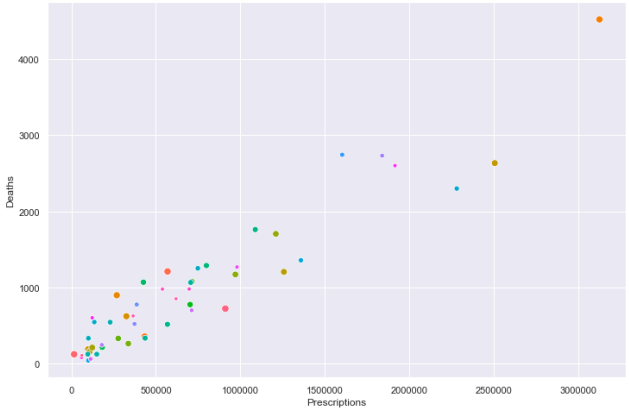
Opioid Use and Prescribers

1. Introduction

The abuse of opioids leading to opioid addiction has been an issue for many years. Although opioids are highly addictive, it is not uncommon for doctors to prescribe opioids for pain management, which can often lead to dependency and abuse. This project uses three different datasets, all from the year 2014, involving the use of opioids in different states, the amount of opioid use and overdoses in a state, and the different medical professionals that prescribe opiods. The goal of this project is to analyze the role that different features play in the use of opioids, which will first be examined in the related works. This project also aims to identify if there is a relationship between the amount of prescriptions made and the amount of opioid related deaths in a year. Another goal of the project is to determine if there is a relationship between an individual's credentials and whether or not that individual will be classified as an opioid prescriber.

2. The Dataset

The three datasets used for this project, found from kaggle[1], all contain different features relating to opioid use, and all the information is from the year 2014. The first set, “opioids.csv”, describes different opioids and their generic names. The next set, “prescriber_info.csv”, involves mostly information on the medical professionals that can prescribe opioids, providing information such as their credentials, specialty and whether or not they are classified as an “opioid prescriber”, which would mean they prescribe opioids 10 or more times a year. Finally, the last dataset “overdoses.csv” provides information about the amount of opioid users in each state, as well as the number of opioid related deaths in that state and the state's population. Different features were used from all of these sets in order to run the analysis. The following figure describes the information found in the prescribers data set and the overdoses set.



[1] https://www.kaggle.com/apryor6/us-opiate-prescriptions?select=opioids.csv

From these datasets, a strong linear relationship can be seen between the amount of prescriptions and the number of opioid related deaths. These features are used in the following analysis, as well as using "opioid prescriber", the binary variable which classifies a medical professional as an opioid prescriber or not, and "credentials" in order to determine if an individual's credentials correlate to if they are an opioid prescriber.

3. Related Work

A project by Avery Dunn[2] aimed to explore the relationship between an individual prescribing opioids and different features such as the state they practice in, their gender, and their specialty. To be labeled as an "opioid prescriber", one has to have prescribed opioids more than 10 times a year, as mentioned above. Dunn used the value counts of the states and those labeled as opioid prescribers in order to analyze the amount of prescribers in each state. From there, they selected the state of Alabama to take a closer examination, beginning by comparing the amount of prescriptions made in one year as well as how many opioid related deaths occured that year. Of the 344 medical professionals that could prescribe opioids that year, 244 were classified as opioid prescribers. It was found that, with the population of Alabama being about 4,833,722 for the year 2014, there were about 911,474 prescriptions written that year, as well as 723 deaths. Therefore it could be determined that of the people that were prescribed opiods that year, about 0.08% of those people had opioid related deaths. This test was run again, but this time using the state that was found to have the largest number of opioid prescriptions, which was California.

This maximum number of opioid deaths in a state in 2014 was 4521, from the California population of about 38,332,52. Of the 2,462 medical professionals that could write opioid prescriptions, 1462 were classified as opioid prescribers. Therefore it could be determined that about 0.14% of people that were prescribed opioids that year had died from opioids. The state with the highest percentage of opioid related deaths that year was Alaska with about 0.77% of deaths, totaling 124 from the population of 735,132.

The creator of this project mainly focused on the relationship between opioid use in a state and the total deaths of users, as well as looking at the number of prescribers. This project could be improved upon by looking at the total number of prescribers and searching for a correlation between one's credentials and whether or not they will be classified as an opioid prescriber. From this relationship, it could be tested if it is possible to predict if someone will be an opioid prescriber or not, based on their credentials. Additionally, we can explore the relationship between the number of opioid related deaths and the amount of prescriptions made in a state.

4. Methods- Regression

To begin, we will explore this relationship between the amount of prescriptions made and opioid related deaths, using the overdoses and prescribers datasets. A linear regression model can be used in order to analyze this relationship. A linear regression is a good choice of model for

[2] https://www.kaggle.com/averydunn/dreamland

this test because the goal is to examine the relationship between two numerical variables, neither of which being binary variables. Linear regression is also a good model choice because the goal is to predict a quantitative response, the amount of deaths, based on a singular predictor, the number of prescriptions made.

4.1 Classification

Next, for the question of whether predicting if a person will be considered an opioid prescriber or not based on their credentials, several different classification models can be created. It is beneficial to run several different models in order to find a model that may improve upon the accuracy of the other models, which can help to determine if this prediction can be made, and if so, how accurately it can be made. First, a logistic regression model can be applied, as this model aims to make predictions by using the probability that an individual will fall into the category of a prescriber, or not a prescriber. Next, K-nearest neighbors can be made using the same labeled training and testing data as the logistic regression. KNN is a supervised learning model which uses labeled training data to learn which features correspond to each class. KNN is a good choice of model because the algorithm can predict if an individual will be an opioid prescriber or not, using the information of whether or not the surrounding elements are prescribers. Lastly, a random forest model can be applied to create a number of bootstrapped samples, obtained with replacement. Random forest will create a decision tree for each sample, using the associated bootstrap sample as the training data, resulting in k different decision trees. Random forest is a good model choice as the classifiers from random forest are often less correlated than other methods.

5. Discussion and Results

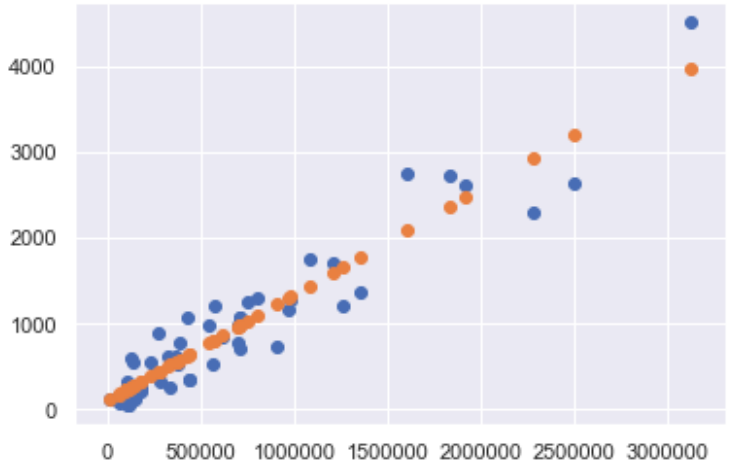
To begin, the linear regression model was created to analyze the relationship between the number of opioid related deaths in a year and the amount of opioid prescriptions made. The following is a summary of the models results.

OLS Regression Results

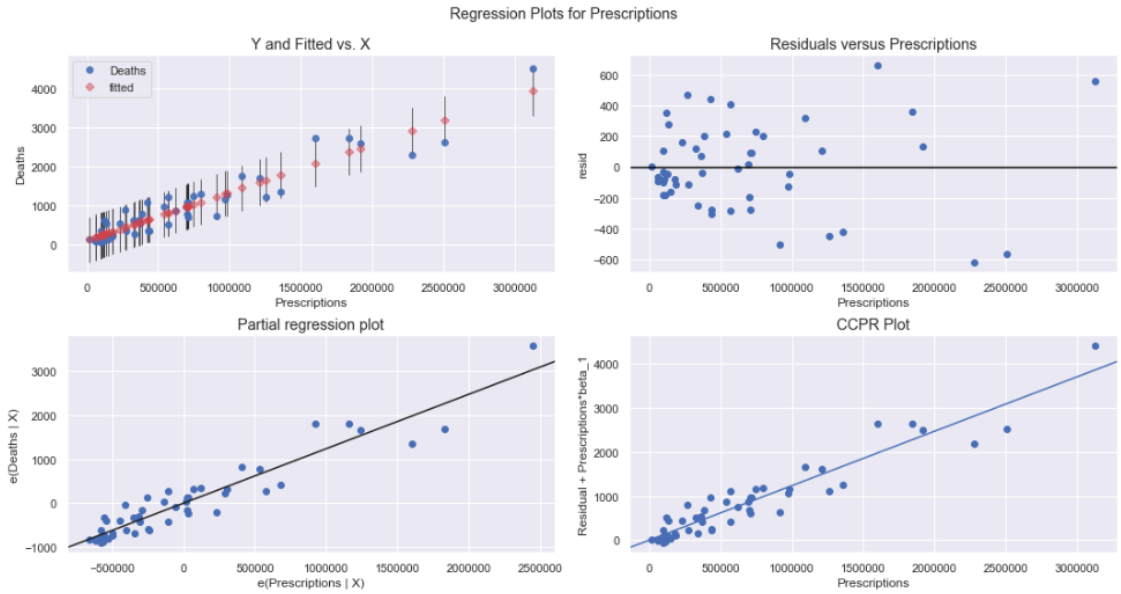
Dep. Variable:	Deaths	**R-squared:**	0.899
Model:	OLS	**Adj. R-squared:**	0.897
Method:	Least Squares	**F-statistic:**	429.0
Date:	Thu, 08 Apr 2021	**Prob (F-statistic):**	1.40e-25
Time:	14:48:08	**Log-Likelihood:**	-353.05
No. Observations:	50	**AIC:**	710.1
Df Residuals:	48	**BIC:**	713.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	101.0182	57.401	1.760	0.085	-14.394	216.431
Prescriptions	0.0012	5.96e-05	20.713	0.000	0.001	0.001

Omnibus:	0.033	**Durbin-Watson:**	1.545
Prob(Omnibus):	0.984	**Jarque-Bera (JB):**	0.123
Skew:	0.055	**Prob(JB):**	0.940
Kurtosis:	2.783	**Cond. No.**	1.36e+06



From the graph, it is clear that there is a strong linear relationship between the deaths and the number of prescriptions. Additionally, from the summary the R-squared value of .89 shows that this is in fact a strong positive relationship. The following graphs further explore the relationship between these variables.



Beginning in the top left, this graph shows that when using fitted values for the x variable, prescriptions, this strong positive correlation still exists. The bottom left graph displays the same relationship, with a natural log applied to both variables. The graph on the bottom right also shows this same relationship, with a beta value added to the y variable of deaths. Overall, these graphical and numerical representations expose that the number of opioid related deaths positively correlates with the number of opioid prescriptions made each year.

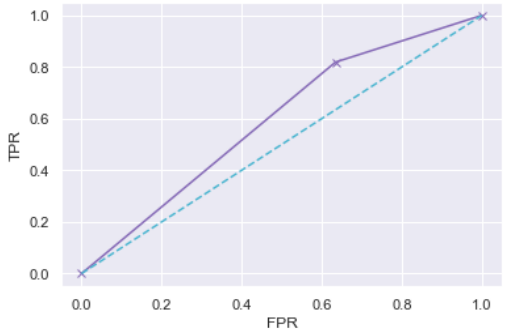
Next, for the classification model, the logistic regression model was created. The model aimed to predict the probability of being labeled an opioid prescriber using one's credentials as the predictor. The following is a summary of the models results.

```
Optimization terminated successfully.
         Current function value: 0.640155
         Iterations 9
                           Logit Regression Results
==============================================================================
Dep. Variable:      Opioid.Prescriber   No. Observations:                18177
Model:                          Logit   Df Residuals:                    18162
Method:                           MLE   Df Model:                           14
Date:                Thu, 08 Apr 2021   Pseudo R-squ.:                0.05567
Time:                        21:02:23   Log-Likelihood:                -11636.
converged:                       True   LL-Null:                       -12322.
Covariance Type:            nonrobust   LLR p-value:                1.861e-284
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.1603      0.201      0.799      0.424      -0.233       0.554
creds_ARNP    -0.7704      0.250     -3.079      0.002      -1.261      -0.280
creds_CRNP    -0.0222      0.274     -0.081      0.935      -0.559       0.514
creds_DDS     -0.6428      0.208     -3.093      0.002      -1.050      -0.235
creds_DMD     -0.3289      0.218     -1.510      0.131      -0.756       0.098
creds_DO       1.0726      0.213      5.031      0.000       0.655       1.491
creds_DPM      0.3061      0.236      1.298      0.194      -0.156       0.768
creds_FNP      0.5054      0.246      2.051      0.040       0.022       0.988
creds_MD       0.3736      0.202      1.853      0.064      -0.022       0.769
creds_MDPHD   -0.5748      0.292     -1.970      0.049      -1.147      -0.003
creds_NP      -0.0694      0.222     -0.312      0.755      -0.505       0.366
creds_OD      -5.0730      0.613     -8.273      0.000      -6.275      -3.871
creds_PA       0.5264      0.223      2.358      0.018       0.089       0.964
creds_PAC      0.6274      0.215      2.917      0.004       0.206       1.049
creds_other   -0.0858      0.208     -0.413      0.680      -0.493       0.321
==============================================================================
The accuracy of the model is 0.6448844884488449
```

From this summary it can be seen that the accuracy of the model is about 64.5%. Additionally, some of the features have high p-values, greater than 0.05, meaning they are not statistically significant indicators of the class and could potentially bring down the accuracy. However, these features are all professions and the goal is the prediction based on profession so it may not be wise to remove all these features. Although this is not a very high accuracy, this shows that the model was able to classify some professions as opioid prescribers or not, based on their credentials.

Next, the k-nearest neighbors model was run using the same training and testing datasets as the logistic regression, with the goal to make the same classification and potentially improve on the accuracy. The following results show the confusion matrix and ROC curve from the resulting KNN.

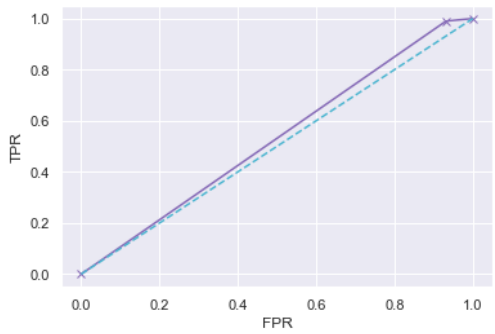
```
[[ 896 1555]
 [ 654 2955]]
0.6354785478547855
auc: 0.5921757214470963
```



From these metrics, it is clear that the k-nearest neighbors model was not a very strong predictor of if an individual will be classified as an opioid prescriber or not. The accuracy score shows about 63.5% correct predictions with the area under the ROC curve being only 0.59. Similarly to the logistic regression, the model does make some correct predictions, however the overall accuracy of the KNN is relatively low.

Finally, a random forest model was created with the same training and testing sets and the same goal as the KNN and the logistic regression models. The goal of using this model was to improve on the accuracy of the first two models, especially considering the random forest can decrease correlation between variables. The following shows the confusion matrix as well as the ROC curve from the resulting random forest model.

```
0.6173267326732673
[[ 167 2284]
 [  35 3574]]
auc: 0.529218738818668
```



Similar to the KNN, these metrics show that the random forest was not a strong classifier of being an opioid prescriber. The random forest shows an accuracy score of about 61.7%, the lowest score of the three models. Additionally, the area under the ROC curve is about .53, which is very close to only being as accurate as a guess. Given the results of these metrics, the random forest is not a strong model for making this prediction.

6. Conclusion

The first goal of this project was to determine the relationship between the number of opioid related deaths and the amount of opioid prescriptions made in a year. For this goal, a linear regression model was created which showed that there is in fact a strong linear correlation between these variables. This model showed that as the amount of prescriptions made in the year 2014 increased, so did the number of opioid related deaths.

The second goal of this project was to predict if a person would be classified as an opioid prescriber or not, based on their credentials. The models used to make this classification were logistic regression, k-nearest neighbor, and random forest. Overall, these models all had relatively low accuracy scores, as they were all between 60-65% accurate. The logistic regression did show the highest accuracy, of about 64.5%. It is likely that these models made many incorrect predictions due to the features in the data, and it can be concluded that one can not be placed into the class of being a prescriber or not solely based on their credentials. These models could potentially be improved if they were given different features on the individual that may correlate more strongly with the classification of whether or not one is on opioid prescriber.