Opioid overdoses

Devarshi Mehta | CMSE 381 Final Project | Spring Semester 2021

**Introduction:**

For this following project, I have selected to work with Opioid datasets. The package I have used in the project was found at Kaggle.com. This dataset contains summaries of prescription records for 250 common opioid and non-opioid drugs written by 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 a small subset of data that was sourced from cms.gov. The full dataset contains almost 24 million prescription instances in long format. The dataset has been cleaned and compiled. This dataset here is 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. There are 3 files in the whole dataset: The main data is in prescriber-info.csv. There is also opioids.csv that contains the names of all opioid drugs included in the data and overdoses.csv that contains information on opioid related drug overdose fatalities.

The data consists of the following characteristics for each prescriber:

1. NPI – unique National Provider Identifier number

2. Gender - (M/F)

3. State - U.S. State by abbreviation

4. Credentials - set of initials indicative of medical degree

5. Specialty - description of type of medicinal practice A long list of drugs with numeric values indicating the total number of prescriptions written for the year by that individual.

6. Opioid. Prescriber - a Boolean label indicating whether that individual prescribed opiate drugs more than 10 times in the year.

**Questions:**

There are 3 main questions that I would like to address. As we are all familiar that Opioid usage in United States has been a big problem and I strongly believe that there is not enough attention given to this problem. Following are the questions:

1. Where are the opioid prescriptions usually prescribed at and what drugs are correlated to the number of prescriptions prescribed?
2. After finding out the correlation of which drugs are usually prescribed in the prescriptions, I will find out that where are usually sold from? Also, to help us understand we will find out which gender is exposed to opioids more and predict the future prescriptions of opioids using predictive tools.
3. Our third question is to find out if there is any relation between number of deaths and population of each state?

**Methods and Results:**

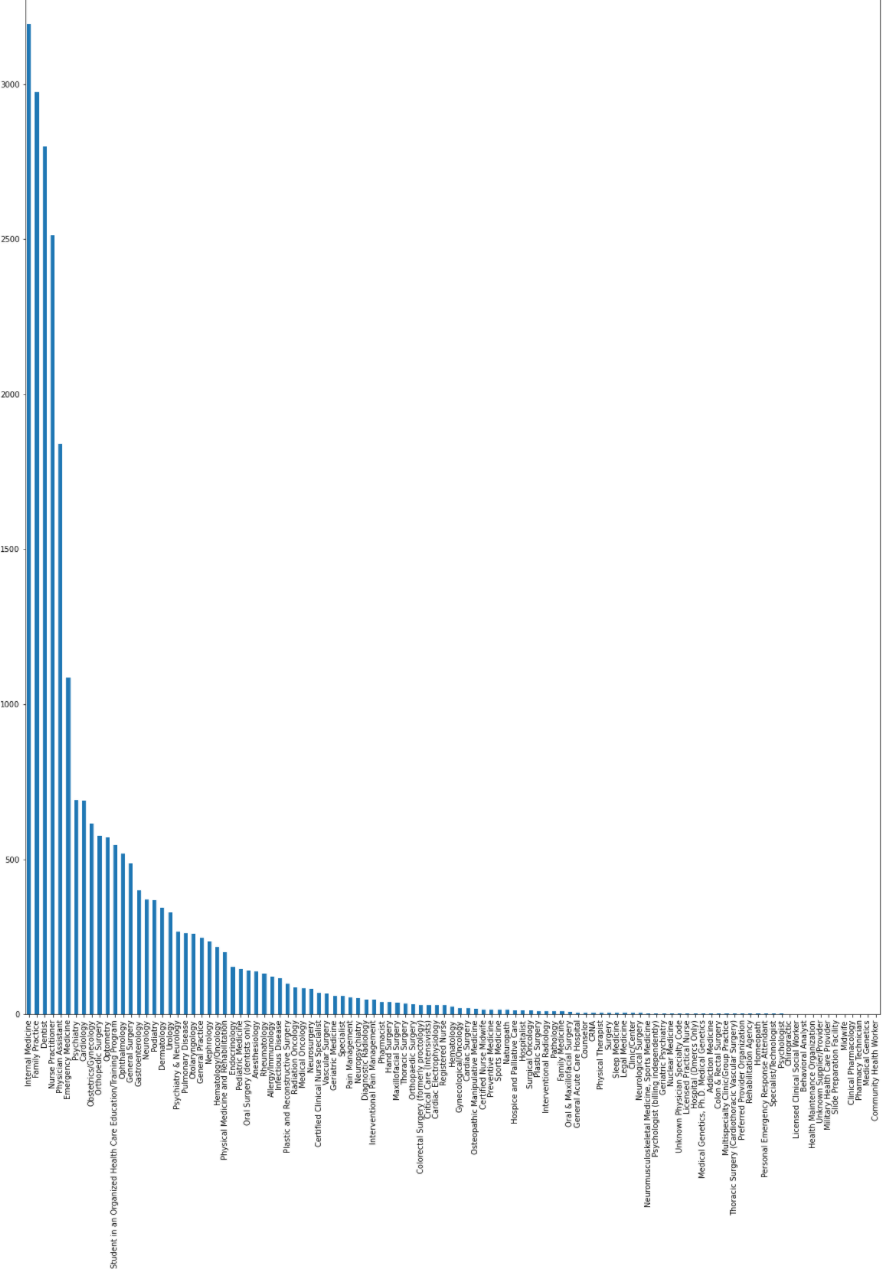
To answer our questions, we will use the help of Kaggle code that was provided to me with the dataset and for question 3, I will use my own knowledge to answer that question.

*Note: The first two questions are done using Kaggle code and I do not own any of this code.*

To answer the first question, we will be using the prescriber-info file which has the prescription information and has all the main information of the dataset.

To begin, we plotted out the specialty feature which showed us the number of prescriptions from each prescriber. The result from the plot shows us each prescriber and where was it prescribed at?

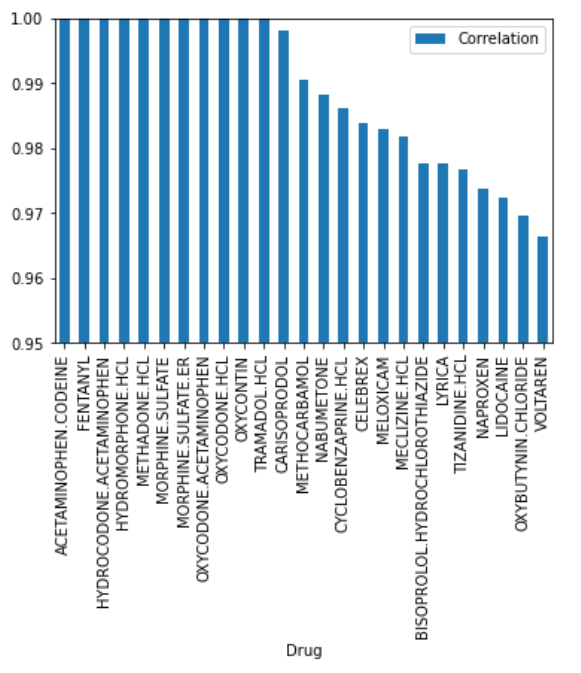
The plot shows that Internal medicine, Family Practice, Dentist, Nurse Practitioner and Physician assistant were the ones that prescribed the opioids most. This would make sense since this is were most of the public would get their control substance prescriptions from. Also do bear in mind, these are few of the most visited places from general public which plays a big part in the plot.



There are also other prescribes from other places that have prescribed opioids, but the prescribers mentioned above are the top ones in the list. Now that we know where people get prescribed opioids from, we move on to answer, which opioids are usually prescribed?

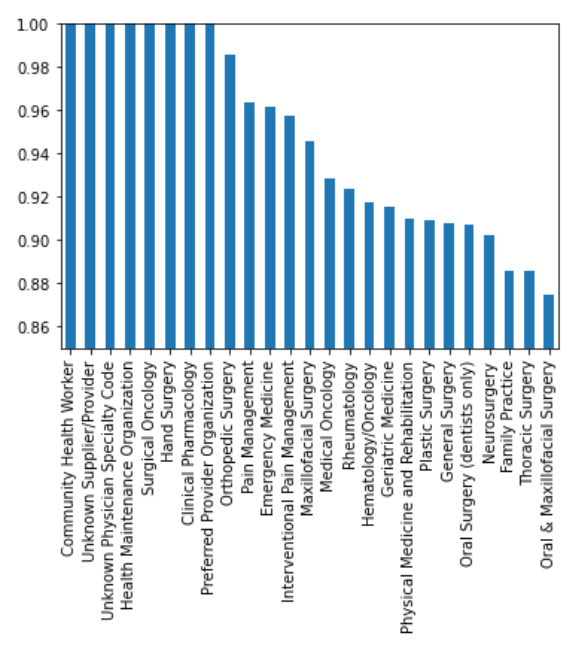


This is the list showing us the most the prescribed opioids in the country. We decided to run a correlation test which basically took in the data from the prescriber dataset and for every prescription generated the name of medication prescribed per prescription will be added to our model. The correlation test is a great approach for this since we must find out the most prescribed opioids and this test will do just that and that too for every prescription. The closer our correlation is to 1, the more prescribed that drug will be. Now using that knowledge, we noticed that Acetaminophen with codeine, Fentanyl, Hydrocodone Acetaminophen, Hydromorphone HCL, Methadone HCL, Morphine Sulfate, Morphine Sulfate Extended Release, Oxycodone HCL, Oxycontin and Tramadol were showed 1.00 for the correlation meaning for almost every prescription wrote for opioid, these medications were the most written for. One of the main reasons why these drugs would be scheduled is for pain relief. People believe in taking these meds when they suffer following their procedures or any incident but often get in the addiction of the drug which is hard to control considering there are no other alternatives.

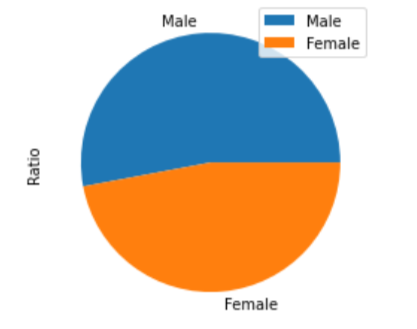


Now that we know the correlation and the drug mainly prescribed, we will turn our attention to the places these medications are sold at and who is highly exposed to it?

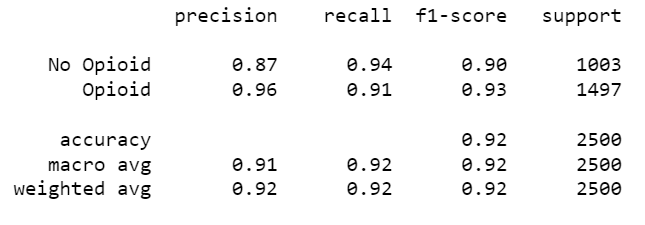
We decided to use the knowledge from the prescriber file again to find out the places the prescriptions were released from and plot it out. For this plot, we will use the correlation again to plot out the places the opioids were sold mainly. The higher the correlation means that most of the prescriptions sold were sold from that following place.



The results show that community health worker, unknown supplier, unknown physician, health maintenance organization, surgical oncology, hand surgery, clinical pharmacology and preferred provider organization were the places most prescriptions were released from. The most shocking result from this plot was Oral Surgery was quite low than other places considering it was one of the top places to prescribe the opioids (results from plot 1). This could be harmful because the prescription can be mishandled and can cause big issues. Now that we know the places they were prescribed at, we will plot out the ratio between male and female for their correlation with opioids.

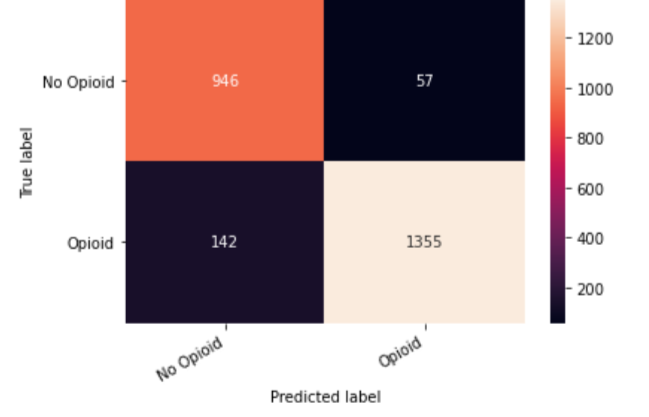


As we see that more male population is exposed to opioids than female population. Now that we know answers to few of our questions, we will find out that using 10% of data as testing data for our model, what will be the chances of opioids getting prescribed again. We will use bootstrap method for this as this method will help us not to overfit our results from the predictive model.

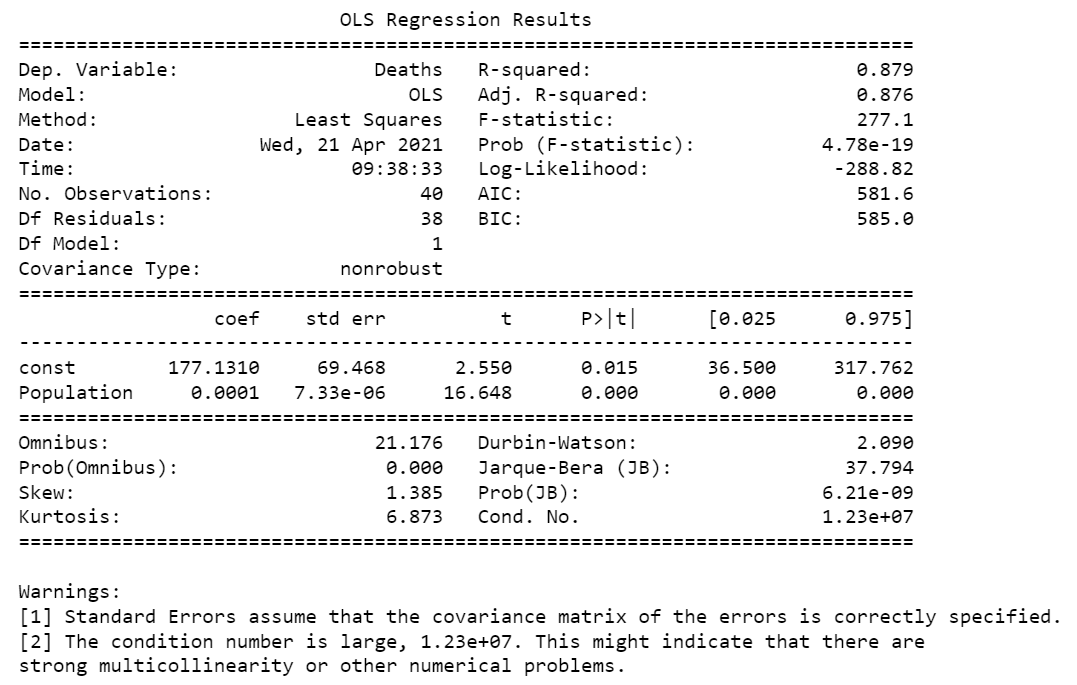


This model shows that precision is very good for opioids meaning the model did good predicting the positive results. Recall will show us the data recalled from the positive outputs from the original data, f1 score is the harmonic mean of precision and recall which is also very good for the model. Support will show the number of true results for each class. The accuracy of our model is at 92% which is amazing and would make sense because we used 90% as training data for the model.

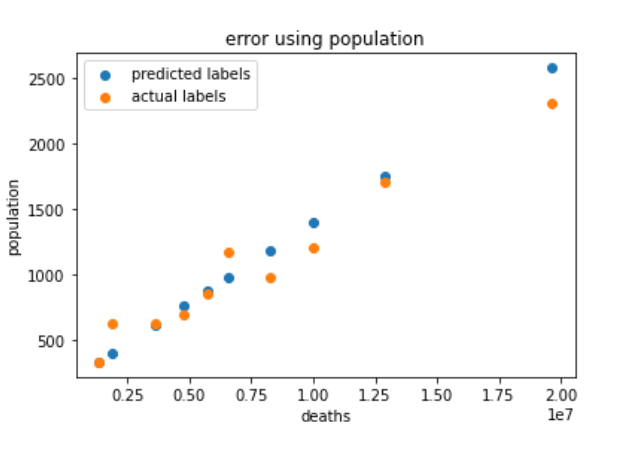
Next, we will use random forest method to print out a confusion matrix to show us the true positive/negative and false positive/negative. Random forest is great way to do this as it will not be bias towards highly correlated opioids, but instead will create diverse results from the random subset features it creates.



Now to answer our final question, that is if there are more deaths linked to more population. To answer this, I used Ordinary Least Squares (OLS) method to find out the significant of population as feature for the deaths caused by overdoses. OLS is one of the most common way used in machine learning to print us out the summary of our features and its significance to the feature we are testing it with. The method focuses on differences between observed values and the predicted values from the model we will build. Also, we will use the overdoses file which will give us the number of deaths reported per state with each states’ population.



The OLS results does show what we assumed which was the population and its significance on the number of deaths. But to test out our predictive model, we will plot out our results that was generated by our predictive model and compare it to the actual data.



Our model did fantastic for predicting the number of deaths per population compared to the observed data from the actual dataset.

**Conclusion:**

After successful results from each of our tests, there are multiple observations we can conclude from this project. Starting with, based on the number of prescriptions written per prescriber in this country, there should strictly rules and regulations to manage the problem better. Not enough attention is given to the problem and that is the reason why the number of deaths from opioid overdose keeps rising. As we saw from one of our tests, that there more than multiple opioids that had positive correlation to the number of prescriptions wrote in this country meaning the more prescriptions wrote, the more common it will be for those medications to be prescribed, which again will lead to mor deaths. Next tests showed us that based on where the medications are sold, one of the big issues found out was at Dental Office. Maybe, if there is a way that they can release the medication from there instead of sending it to other pharmacies which may help us prevent deaths. Our predictive model did predict that the correlation may go even higher for future prescriptions if steps are not taken towards it. Our project also found out that if population is higher for a state, it will have more reported deaths which would make sense, but the model was so accurate that it indicates that if we want to stop the overdoses or even try to reduce the number of overdoses, we should target bigger populated states. At the end, we designed a predictive model that showed that our predictive labels calculated the number of deaths per population very accurately and can be used for future references to prevent overdoses death.

At the end, I would like to share that the reason why I chose to show the relation between population and number of deaths is because it was not answered in any of the Kaggle code or online and one of the most important steps for us to take to prevent further overdoses and the deaths related to it, we need to see if population matters because if it does matter we should focus on areas with higher population to make a drastic difference. My model was also able to predict labels that were very accurate and aligned almost perfectly with the actual labels from the dataset, proving this can be one of good ways to solve the problem of overdoses and its deaths.

**References:**

<https://www.kaggle.com/apryor6/us-opiate-prescriptions?select=overdoses.csv> (The link to where the dataset and its information was collected)

<https://www.kaggle.com/dylanbui/notebookc8d4c9ee87> (Link to the code I used for question 1 and 2)

Question 3 was done using prior knowledge from other courses on campus and in CMSE 381.