Examining drug datasets

Final Report

Bella Said

Introduction

Right now with the current opioid epidemic it's crucial that we learn and understand the data that is being collected about drugs so that we can attempt to find ways of lessening the number of people affected by this epidemic. First I looked at the number of overdose deaths in an attempt to find a classification model that would best predict this data value. While looking at drug overdoses in general, it's important to look at other aspects of business that keep the drug epidemic alive, such as those who are distributing the drugs. Using the second dataset that included distributor data, I wanted to visualize some of the relationships in the data as well as compare the classification model accuracy of that dataset to the accuracy of the model from the first dataset that calculated the number of drug overdose deaths.

Motivating Questions:

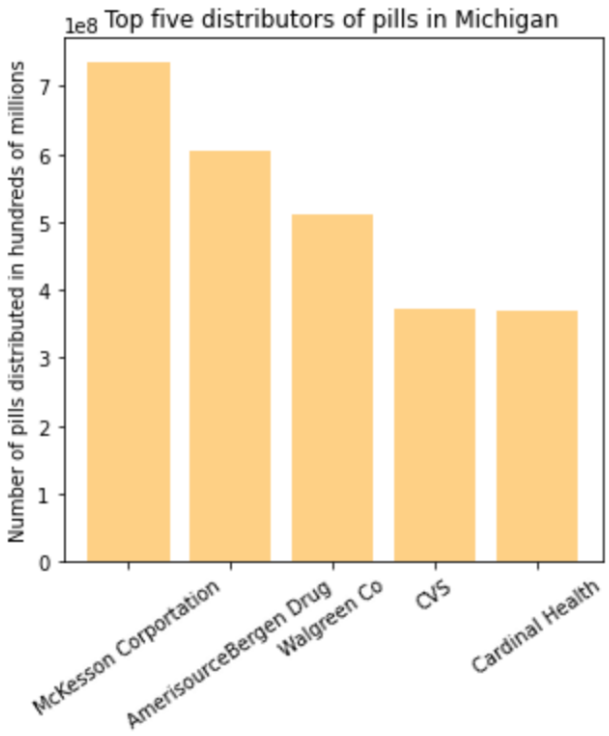
Question 1: What is the best model to predict the number of drug overdose deaths?

Question 2: What is the relationship between total records and total dosage unit?

Question 3: Will the distributor dataset's classification model have lower accuracy if the labels are balanced?

Related Work

In one of the articles that I used as a resource they had bar plots that represented the top five distributors, manufacturers, and pharmacies in regards to how many pills they distributed, manufactured, and were supplied respectively. I replicated these three bar plots that were shown in the article for the statewide data of Michigan using three different datasets; the distributor dataset, the manufacturer dataset, and the pharmacies dataset. Below is an image of the bar plot I made using the distributors dataset.



Of the three datasets that were utilized in the related work section, I decided to use only one of them and continued the rest of my project using the distributors data and another dataset from the CDC.

Datasets

I used two datasets for my project. The first dataset contains the drug overdose deaths each month by 12-month provisional month ending. This means that a data value for the month of April is the number of deaths from that month and the previous eleven months. This dataset includes data from the entire United States. It also includes month, year, and an indicator value that gives the potential cause of death. There are many indicator values ranging from all different kinds of drugs, so for my model creation I decided to focus just on the drug overdose death indicator, which can include many different kinds of drugs as the reason for the overdose.

The second dataset that I used included data only from the state of Michigan, and this dataset included data about the distributors of drugs from the year 2006-2014. It's important to look at this data because it might tell us a lot about where most of the drugs that are in the state are being distributed from, and it's important because many drug users begin by misusing prescription

drugs. For my model creation I focused on the total dosage unit column which had the data values for the number of pills distributed from a certain distributor.

Methods

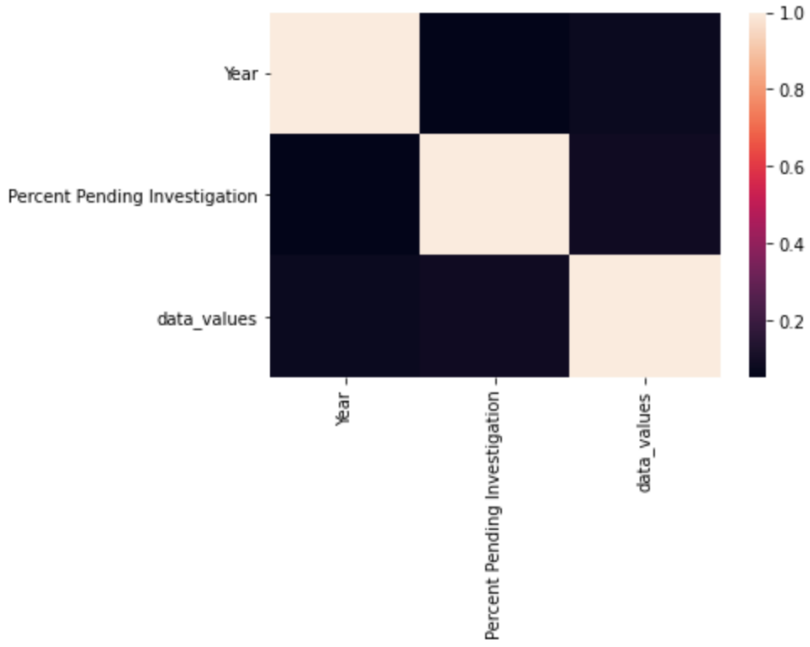
For my methods I decided to use a K Nearest Neighbors classification model and a Random Forest model for my two model creation questions. The reason I chose to do Random Forest is because it is a useful model that does not lose accuracy if there is missing data in the dataset, so I thought this would be a beneficial model to use in case there was some missing data still present in the cleaned dataset. The reason I chose KNN was because the features in my drug overdose dataset did not seem to be highly correlated with one another, and a KNN classification model works well with non-linear data, so I figured this might be a good model to utilize considering the datasets I'm working with.

Results & Discussion

Question 1: What is the best model to predict the number of drug overdose deaths?

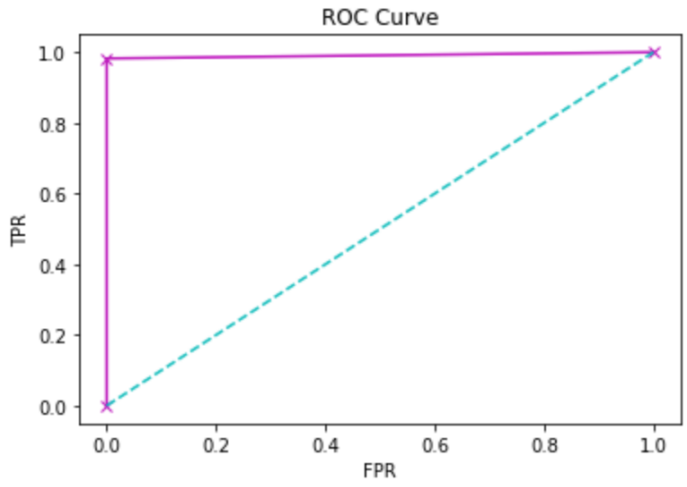
The first question I attempted to address was to find the best model to predict the number of drug overdose deaths. This question dealt with the drug overdose dataset and I used both KNN and Random Forest models to try and best predict the number of drug overdose deaths. Initially I used a heatmap to be able to visualize the correlations present in the

dataset. Below is an image of the heatmap.



It's clear in the image above that the features do not have much of a correlation with one another at all. This hints towards the idea that none of these features would be significant in the models and therefore we might end up with some low accuracy scores. However in doing the KNN model, we ended up with an extremely high accuracy rate. Below is an image of the ROC curve I made that shows the relationship between the false positive rate and the true positive rate.

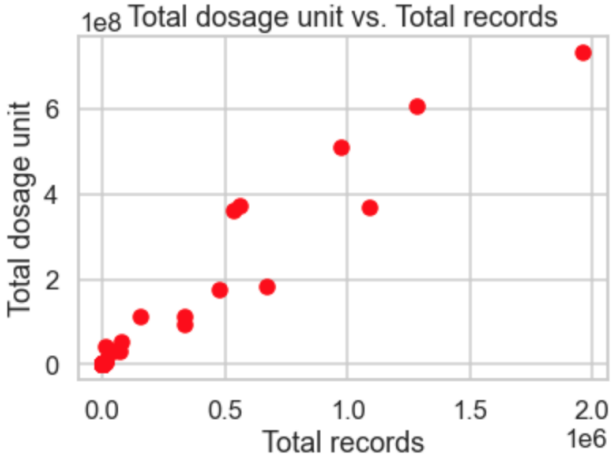
```
auc: 0.9911504424778761
```



As we can see in the plot above the ROC curve for the KNN model has an AUC of .99 which is almost 100% accurate. With the data I was using I was very surprised that this was the result, especially since the features did not seem to have high correlations with one another as discussed previously. This unusually high accuracy led me to believe that the dataset might have unbalanced labels, which can lead to overfitting when training the model. I decided to then check the amount of labels and see if they were balanced. In doing so I realized that the labels were heavily unbalanced and this caused overfitting towards the label that was significantly more present in the data. Since the model overfit to the label, we ended up with a very high accuracy score, which doesn't properly represent the classification model. When doing the random forest classification the model accuracy was also extremely high and had an AUC value of 1.0. This means that the best model for predicting the number of drug overdose deaths was the random forest classifier, however these models are not truly accurate due to the unbalanced nature of the labels.

Question 2: What is the relationship between total records and total dosage unit?

For this next question, I focused on the distributors dataset. I wanted to see if there was any visible relationship between the total records values in the data and the total dosage unit or the amount of pills being distributed. By creating a scatter plot with total records on the x-axis and total dosage unit on the y-axis I was able to visualize their relationship.



← why this is a important or interesting result?

As shown in the figure above, as total records increase the number of pills distributed, or total dosage unit also increases. When using a heatmap total records and total dosage unit had a pretty high correlation of around .97 with one another, so these features certainly seem significant. This

doesn't really change anything in terms of potential models, since correlation does not equal causation we cannot make any direct claims about their relationship, but this helps us visualize more of what is happening within the dataset.

Question 3: Will the distributor dataset's classification model have lower accuracy if the labels are balanced?

I wanted to try another model creation where I had balanced labels and see what the difference was with the accuracy. I decided to try and predict the total dosage unit of the distributors dataset using both a KNN model and a Random Forest classifier to see how the accuracy compared to the drug OD model's accuracy. This time when making my model I have an equal number of both types of labels, so there was no worry about unbalanced labels. This time, I got an accuracy score of .97 for both KNN and Random Forest so neither one was the 'better' model, however this accuracy score is lower than the accuracy score for the drug OD dataset. Another reason these models have such high accuracies even with balanced labels might be due to the size of the distributors dataset. The distributors dataset only had 147 rows while the drug OD dataset had over 2000, so the smaller dataset might lead the models to have a higher accuracy than expected.

Conclusion and Future Work

This project succeeded in answering the motivating questions, while also bringing up important instances of balanced and unbalanced labels, as well as dataset size. The drug OD dataset was a good size, however it had unbalanced labels, which led to an overfitted model. The distributor data had a small dataset but balanced labels, which might have also led to an overfitted model. Overall much was learned about drug deaths and how many drugs are being distributed in Michigan. Future work could include studying the other datasets from the Washington Post article such as the pharmacy and manufacturer datasets. It could also be beneficial to break up the number of drug overdose deaths by drug category to further explore which drugs are at the forefront of the epidemic.

References

1. “VSRR Provisional Drug Overdose Death Counts.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, data.cdc.gov/NCHS/VSRR-Provisional-Drug-Overdose-Death-Counts/xkb8-kh2a/.

2. “Drilling into the DEA's Pain Pill Database.” The Washington Post, WP Company, 21 July 2019,www.washingtonpost.com/graphics/2019/investigations/dea-pain-pill-database/#download-resources.