Identifying Opioid Prescribers

Lauren Belknap
NetID: belknapl

Project Category: General Classification Machine Learning
https://www.kaggle.com/apryor6/us-opiate-prescriptions?select=prescriber-info.csv
https://github.com/fivethirtyeight/data/tree/master/drug-use-by-age

L
83

Final Report

1. Introduction

Drug use is a large problem in America. Trying to determine whether or not a doctor will prescribe opioids can be a huge help in trying to curve this issue. Opioids are amongst the top drugs that are abused. It is said that approximately over 10 million people in America 12 years or older misused opioids in the past year. Within that 10 million, 9.7 million people misused prescription pain relievers [1]. Deciding which doctors prescribe these opioids can help put an end to the opioid epidemic. During this report we will also take a look at an article that breaks down how and why certain drugs get abused more than others. Using methods such as Logistic Regression, Classification Trees, Linear Discrimination Analysis, and K Nearest Neighbors we will attempt to try and accurately predict whether or not a doctor will prescribe opioids.

2. Related Work

An existing article by *fiftythirtynine* called "How Baby Boomers Get High" focuses on Americans ages 50-64 who have used drugs in 2012. The top two drugs used were marjuana and pain relievers. In the article it states that Baby Boomers use less drugs compared to younger generations. If we look specifically at painkillers 2.52% of people aged 50-64 had used painkillers within the past year which is the second most used drug. Although this is a large category of drugs. An additional question was asked about specifically which drug. About 0.36 percent of those 50-64 year olds said that they had used OxyContin compared to 1.72 percent of

20 year olds. This is still considered an addictive drug which is a cause of concern. Baby Boomers are more likely to die from an overdose than any other generations according to Richard Miech, a professor at the University of Michigan. Research has said that the reason drug use is different for older generations is because of the reason it is being used. For older people drugs are used more for chronic pain relief, loss, isolation, and effects on the brain [2]. Regardless, drugs can be addictive for any age. Taking into consideration how addictive painkillers can be, I wanted to take a look at how we might be able to predict which doctors might be writing these prescriptions.

3. Datasets

Two datasets were used during this report. The first dataset was used to recreate the plot created by *fiftythirtynine*. In this dataset, there are a number of predictors including 13 different drugs and 17 different age groups. The second dataset was used in order to try and classify whether or not a doctor will prescribe opioids. This dataset has 256 columns. The last column is called "Opioid.Prescriber" which is classified as 0 or 1 indicating whether or not a unique physician has prescribed opioids more than 10 times a year. There are 25,000 rows which indicate unique prescribers. With this dataset 4 columns were strings which were cleaned by finding the unique names with the predictor and changing them to an integer. There were no missing values so the data was ready to be used.

4. Recreating Article Plot

In order to recreate the plot within the *fiftythirtynine* article a few things need to be done. First we just need the row for ages 50-64. Then for that row specifically we need to filter just the columns that were used for the plot in the article. Once that was done it was turned into a new dataframe for simplicity's sake. Using the ggplot2 library a bar graph was created with labels.

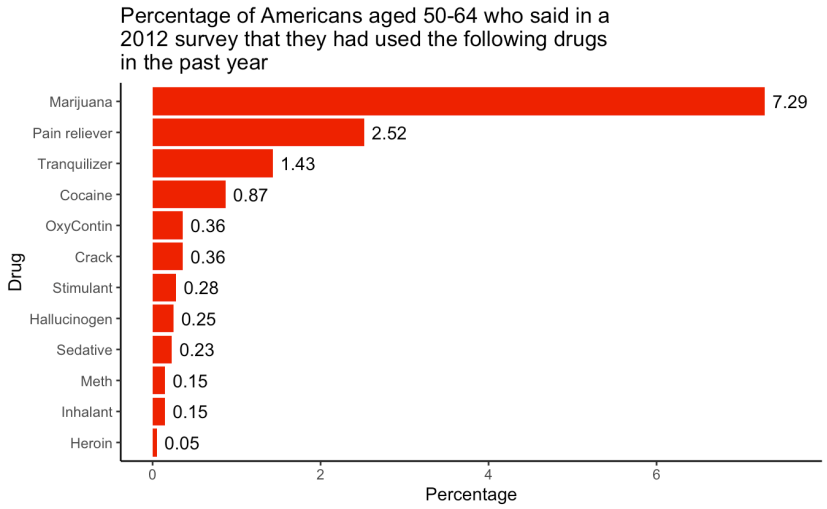

Figure 1: Recreated plot from Baby Boomer Article

I was able to almost get an identical plot to that of the *fiftythirtnine* Article. This figure is usefully in grasping which drugs are used more than others. A drawback to this article is we don't quite get enough information. For example how many people were surveyed, where did they get this information from, and what can we do about this. In my model I'm attempting to try and predict something that can be beneficial to the issue at hand.

5. Methods

Since this was a classification dataset only certain methods can be applied. For example I decided to use logistic regression instead of linear regression because again this is a classification problem. Specifically the response variable, "Opioid.Prescriber" is binary meaning that there are only two classifications. This was also the case when using tree based methods. A Tree Classifier method was used rather than regression because of the binary response variable and so on for the other methods. I chose these methods because I felt that they would suit the dataset the best and could potentially give the most accurate predictions.

5.1 Logistic Regression

The first method I used was a generalized Linear Regression model. I started first by using the full model, so all of the predictors except for NPI. This is because NPI is just the identification of the doctor and will not help with the model. After this was complete I had printed the summary of the model to determine which variables were statistically significant. It was found that about 41 variables were considered statistically significant when alpha is less than or equal to 0.05. Some of these variables include Levofloxacin, Levothyroxine Sodium, Prednisone, and Finasteride just to name a few. Now if we create a new Logistic Regression model with just these variables we end up with a worse error rate. This I believe is the case because even though these are the most statistically significant the other variables still play an important role because there are so many kinds of drugs. It was found that the test error rate for the full model was about 8% compared to the limited model which had an error rate of 26%. I also decided to plot the area under the curve for visualization using the informative value library.

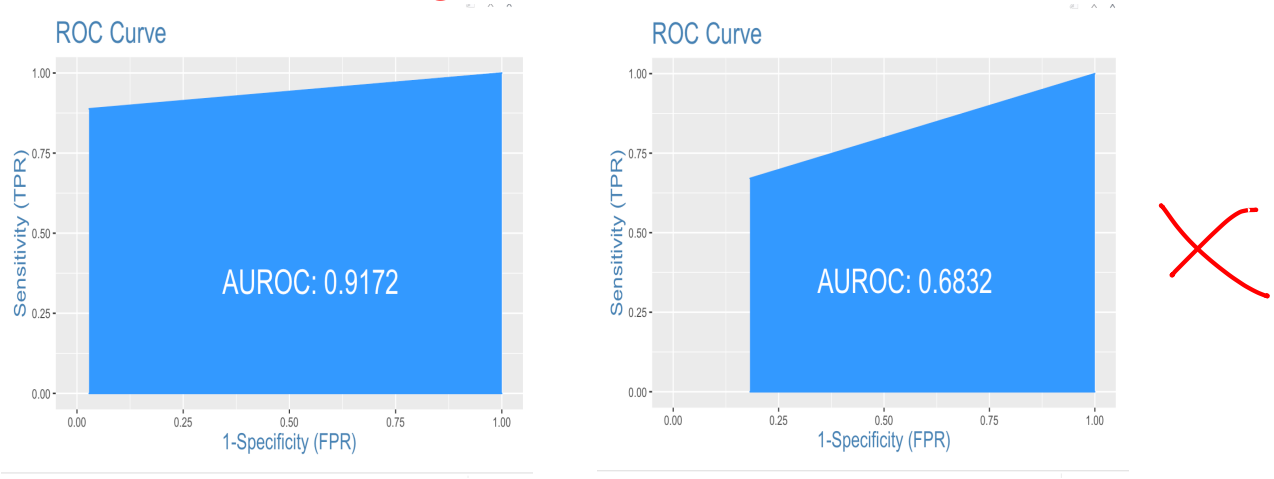

Figure 2: AUC Comparison between full and reduced logistic regression models

As we can see from the figure the area under the curve for the full model is much bigger than the model with only the statistically significant features. Thus we will continue our modeling methods using all variables.

5. 2 K Nearest Neighbors

The next model I wanted to create was a KNN. I wanted to create a loop for k neighbors one through one hundred. This loop will determine the accuracy for each neighbor in order to find which neighbor is the best. This is done by calculating the correct predictions over the total number of observations for each neighbor. The accuracy was then plotted to visualize the accuracy found for each neighbor.

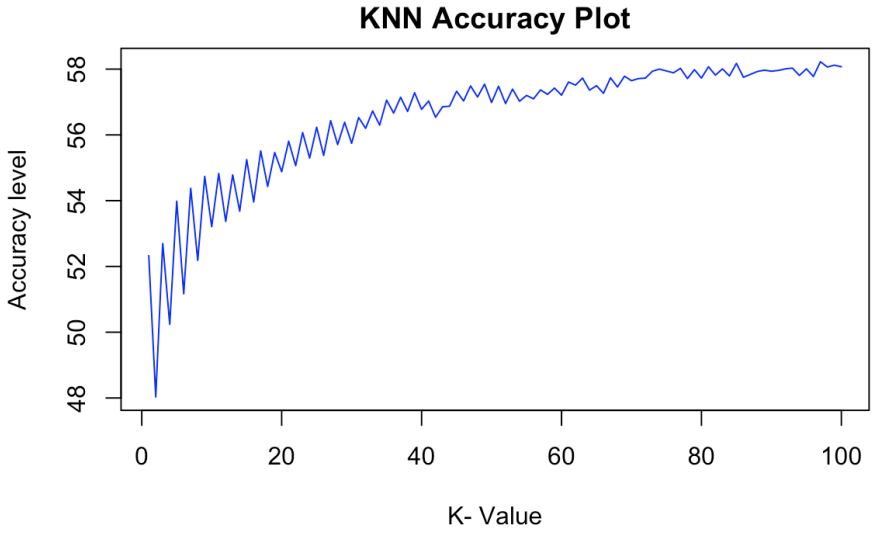

Figure 3: KNN plot to Determine Accuracy of each Neighbor

It was found that 97 neighbors gave the best prediction accuracy. This accuracy being 58.2 percent. This is not good at all considering that there are only two classes with random selection our accuracy would be 50 percent. This is just slightly above. Thus so far the logistic regression model has performed the best.

5.3 Tree Based Methods

The next method I decided to use is a Tree Classifier. In order to create my tree I first took the "Opioid Prescriber" column and turned it into a string column to not get the tree confused. Also it makes the tree look better. So, if the opioid prescriber was 1 meaning that they had prescribed opioids then it would now be "Yes" and "No" for 0. This column was then added to the end of the dataset. Now we are ready to make the tree. It was found that 6 terminal nodes is the best. I then plotted the tree for visualization

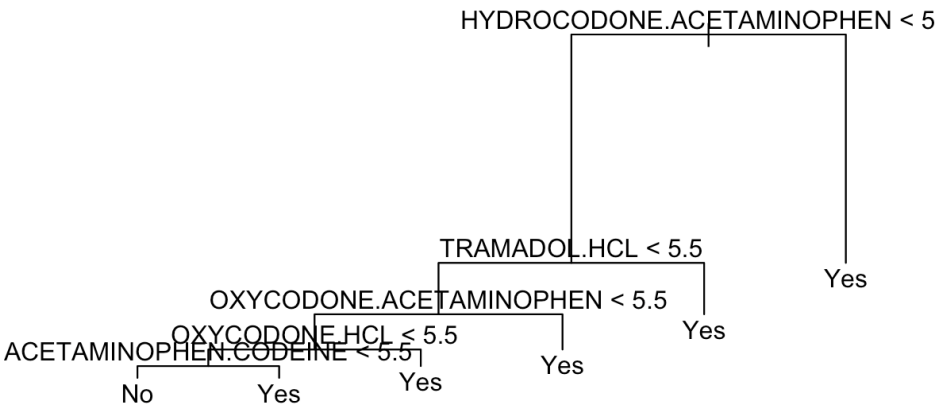

Figure 4: Tree plot for the dataset

It was found that the error rate for the tree based method was about .0768. Which makes it slightly better than the logistic regression model. I then pruned the tree in order to determine if we can get an even better error rate. However when pruned we get the same number of terminal nodes and thus we can accept our tree.

5.4 Linear Discrimination Analysis

The last model I decided to try was a Linear Discrimination Analysis. LDA uses Bayes' Theorem to estimate probabilities. This method is similar to the other models in the way of how you create it in R. Using "Opioid.Prescriber" as our response variable with all predictors. I then

created a confusion matrix using the predictions found from the LDA model. It was found that the error rate was about .205. This is worse than the tree method. Hence we won't use this model for the best classification.

6. Discussion and Results

Confusion matrix, area under the curve, and the error rate were used to calculate the performance of the models. The training set included a sample of half of the data, and the testing dataset was the other half. The results from the model can be seen in the table below.

Model	Test Error Rate
Full Logistic Regression	0.077
Reduced Logistic Regression	0.268
K Nearest Neighbors	0.418
Tree Based Methods	.0768
Linear Discrimination Analysis	0.205

We can see that the model that did the worst was KNN. The error rate for this is quite terrible. The next worst model was the reduced Logistic Regression. This model I used just the variables that were considered statistically significant however this had made worse than previously because we were getting rid of too many variables. In third place was Linear Discrimination analysis this preformed average compared to the others. The top two models were the Logistic Regression model with all features and the tree model. These models performed almost identically in test error rate. They both had about an accuracy of 92 percent, which is very good. The model I would recommend would be tree based method because it did slightly better than the Logistic Regression. However they are both good.

7. Conclusion and Future Work

This project was intended to try to most accurately predict whether or not a doctor will prescribe opioids. This was done through using predictors such as drug type, gender, state, specifically, and credentials. Based on these variables I attempted to use models such as Logistic Regression, K Nearest Neighbors, Tree Classifiers, and Linear Discrimination Analysis to get the most accurate results. It was found that the tree classifier performed the best based on the test error rate. The level of accuracy obtained by some of these models is promising for practical use of trying to find and predict whether or not a doctor is abusing power when prescribing opioids.

In the future I would like to try and tune models further for example doing some diagnostics on the logistic regression model to try and get the accuracy even higher. I would also try using other methods such as Bagging, Random Forest, and Subset Selection. I would also like to find possibly other datasets that could be a good addition to tuning the models. That could give us even more training and testing data, as well as, even more predictors to use when creating the models. We can try and answer even more questions like what speciality of doctors prescribe the most amount of opioids? Hopefully these models can be applied in real situations that can help fight the drug epidemic.

References

[1] (DCD), Digital Communications Division. “Opioid Crisis Statistics.” *HHS.gov*, Https://Plus.google.com/+HHS, www.hhs.gov/opioids/about-the-epidemic/opioid-crisis-statistics/index.html.

[2] Barry-Jester, Anna Maria, and Andrew Flowers. “How Baby Boomers Get High.” *FiveThirtyEight*,FiveThirtyEight,23Apr.2015, fivethirtyeight.com/features/how-baby-boomers-get-high/.

[3] James, Gareth, et al. *An Introduction to Statistical Learning: with Applications in R*. Springer.