Final Project

Caitlyn Locke

4/6/2021

The problem

Part of understanding a problem is looking at where it began. An opioid epidemic is sweeping across the United States, and medical professionals have a part to play in it. Whether justified or not, doctors and other medical professionals are prescribing opioids to their patients. The question is how much and do their specific specialties play a role in it?

The Prescriber data set will be the main data set used for this project; it contains demographic information on the Prescribers such as gender, state and specialty. It also contains separate columns for over 200 different drugs that tells us how many times that doctor prescribed that specific drug. The Opioids data set is used to clean the Prescriber data set and it contains the non-generic and generic drug names for some common opioid drugs. This project aims to use these data sets to create predictive models that can be used to predict the likelihood of an individual prescribing an opioid drug more than ten times a year. Having such a model can help identify where policy and regulations may need to be changed in order to fight this epidemic.

Related Work

Found here: https://www.kaggle.com/apryor6/detecting-frequent-opioid-prescription (https://www.kaggle.com/apryor6/detecting-frequent-opioid-prescription)

The link above will direct you to the code made by the person who posted the data sets, being used for this project, to Kaggle. In his code he only uses a single method for creating models to make predictions. He also removes and changes quite a bit of the data from the Prescriber data set to make his modeling. In this project I will expand his methods by using a boosting method similar to what he did and using a logistic regression method. In attempt to remove less information from the data sets and improve accuracy, I will also clean my data somewhat differently than he did.

The Data

Below are tables that contain information from the data sets that I will be using in this project. For the Prescriber data set, I have only include six out of 256 columns, since all of the other columns are about specific drugs and I wanted to save some space in the output. Also in effort to save output space only the first 6 rows of the data set are being shown. The second table, on the Opioids data set, contains all of the data sets columns, but again only the first six rows to save space.

Prescriber Data Set

	NPI	Gender	State	Credentials	Specialty	Opioid.Prescriber

NPI	Gender	State	Credentials	Specialty	Opioid.Prescriber
1710982582	M	TX	DDS	Dentist	1
1245278100	F	AL	MD	General Surgery	1
1427182161	F	NY	M.D.	General Practice	0
1669567541	M	AZ	MD	Internal Medicine	1
1679650949	M	NV	M.D.	Hematology/Oncology	1
1548580897	M	PA	DO	General Surgery	1

Opioids Data Set

Drug.Name	Generic.Name
ABSTRAL	FENTANYL CITRATE
ACETAMINOPHEN-CODEINE	ACETAMINOPHEN WITH CODEINE
ACTIQ	FENTANYL CITRATE
ASCOMP WITH CODEINE	CODEINE/BUTALBITAL/ASA/CAFFEIN
ASPIRIN-CAFFEINE-DIHYDROCODEIN	DIHYDROCODEINE/ASPIRIN/CAFFEIN
AVINZA	MORPHINE SULFATE

Summary

The summary function in R, gives us some key statistical information about each variable (column) in a data set. I have only included the summary for the six columns seen previously for the Prescriber data set, in effort to save space. Since the Opioids data set only contains names, or character vectors, I did not include its summary output, as it would give no additional information about the data. A similar looking output can be seen below for the Gender, State, Credentials and Specialty variables from the Prescriber data set.

Prescriber Summary Output

NPI	Gender	State	Credentials	Specialty	Opioid.Prescriber
Min. :1.003e+09	Length:25000	Length:25000	Length:25000	Length:25000	Min. :0.0000
1st Qu.:1.245e+09	Class :character	Class :character	Class :character	Class :character	1st Qu.:0.0000
Median :1.498e+09	Mode :character	Mode :character	Mode :character	Mode :character	Median :1.0000
Mean :1.498e+09	NA	NA	NA	NA	Mean :0.5875

NPI	Gender	State	Credentials	Specialty	Opioid.Prescriber
3rd Qu.:1.740e+09	NA	NA	NA	NA	3rd Qu.:1.0000
Max. :1.993e+09	NA	NA	NA	NA	Max. :1.0000

My Models

Before I can work with the data, I must clean it so that I have only the information needed to form predictions. It will also help reduce the overall complexity of my models. I have removed the gender and credentials variables from the data set and I have performed other cleaning methods. See my code file to learn more about how I cleaned the data.

Before making any predictive models the data set needs to be split into training and testing sets, that way the model accuracy can be tested. I made a training data set by randomly selecting 20000 of the rows from the Prescriber data set. The testing data set was then made from the remaining observations left in the Prescriber data set. For this project I will make a boosting model and a logistic regression model to predict whether a doctor will prescribe opioids more than ten times in a year. I have included the first six rows and non-drug name columns for both the testing and training sets below.

Training Data Set

State	Specialty	Opioid.Prescriber
SC	Ophthalmology	0
OH	Dermatology	0
PA	Internal Medicine	1
NC	Physician Assistant	1
NE	Student in an Organized Health Care Education/Training Program	0
KY	Dentist	1

Testing Data Set

State	Specialty	Opioid.Prescriber
TX	Dentist	1
NV	Hematology/Oncology	1
NH	Family Practice	1
OH	Cardiology	0
IA	Nurse Practitioner	0
MT	Internal Medicine	1

Boosting Model

A boosting model is a type of random forest model. It builds many small decision trees to make a prediction about the outcome. Each decision tree is based off of the residuals from the previous decision tree. This process will continue until a certain number, chosen by the model maker, of trees have been made. The boosting model gives us an advantage by listing which variables are the most influential to the model. For my model the most influential variable is the Specialty variable, making up about 89% of the relative influence. Thus confirming that the specialty of the medical professional does play a role in the amount of times that individual prescribes an opioid drug.

Logistic Model

A logistic regression model is a linear model that is restricted to be between zero and one and is used for classification problems, such as the one in this project.

Results

Boosting Model

Accuracy

Accuracy
0.7480962

Boosting Model Confusion Matrix

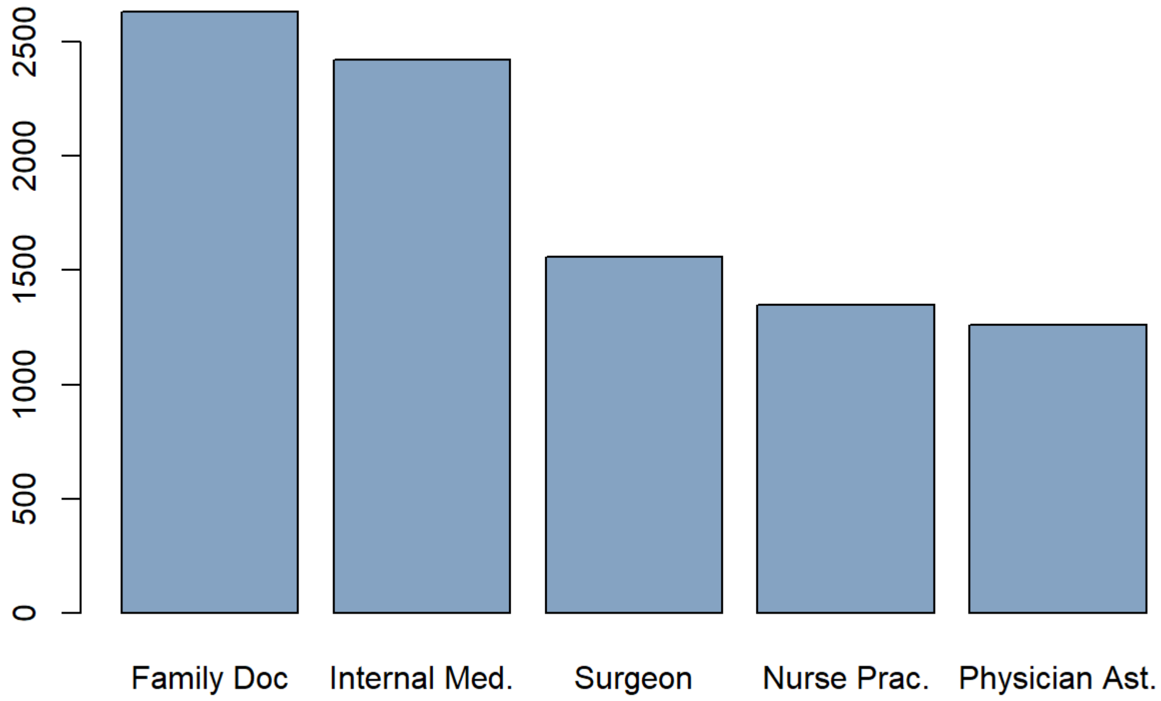
	0	1
0	1599	798
1	459	2134

The columns correspond to values from the Prescriber data set, and the rows correspond to the predicted values of the model. Zero represents those who do not prescribe opioids ten or more times in a year, and number one represents those who do. When the row value matches the column value (ex. row = zero, column = zero), this show that the models prediction was correct.

Logistic Regression

Opioid Prescribers Based on Specialty

Below I have created a box plot that shows the top five speciatlties that most frequently prescribe opioids more than ten times a year. I have also provided a table for easier interpretability of the data. The columns of the table correspond to a specific specialty and the row is the total number of medical professionals that prescribe opioids more than ten times a year. In order to make all the labels appear on the plot, I had to shorten the names of the variables.



Top Five Opioid Prescriber Specialties

	Family Practice	Internal Medicine	Surgeon	Nurse Practitioner	Physician Assistant
x	2633	2424	1559	1350	1261

Accuracy

Accuracy
0.808016

Logistic Regression Model Confusion Matrix

	0	1
0	1850	750
1	208	2182

The columns correspond to values from the Prescriber data set, and the rows correspond to the predicted values of the model. Zero represents those who do not prescribe opioids ten or more times in a year, and number one represents those who do. When the row value matches the column value (ex. row = zero, column = zero), this show that the models prediction was correct.

Discussions and Conclusions

This project attempts to test how changing the model and changing the data cleaning method affects the accuracy of the predictions. Unfortunately changing the way the data was cleaned, even by a little, greatly reduced the accuracy of the model. For my boosting model, I got an accuracy score of about 75%, whereas the previous coder had an accuracy score of about 82%. This 8% decrease in accuracy shows that his data cleaning is superior to mine. While my modeling did not perform better than his, I did get a better accuracy score of about 80% for my logistic regression model. This suggests that with his cleaned data and my logistic regression model, it would out perform his boosting model, thus making it the model choice for this project.

The accuracy scores for my models can be confirmed by the confusion matrices above. When analyzing these matrices, it is obvious that the logistic regression model has a larger number of variables correctly predicted, further showing that it is a superior model.

Looking the bar graph above, one will see the specialties: Family Practice, Internal Medicine, Surgeon, Nurse Practitioner and Physician Assistant. Since opioids are a type of pain killer, it makes sense that surgeon is among the top five specialties. However, I am surprised to see that Surgeon was not the top, and that Family Practice was. While family practice is a medical specialty that works with people of all ages, and thus all different kinds of needs, it does seem that the frequency is too high. Perhaps regulating how and how much family practitioners can prescribe opioids is a great place to start fighting this epidemic.