Capstone 1: Milestone Report

Rika Yatchak

February 16, 2019

0 Introduction

In this project, I examine the Kaggle Opioid Prescriber Dataset for my Springboard Capstone project. To view the code used to generate the results in this report, see the corresponding Git repository.

In this report, we will

1. Define a guiding question for the project.
2. Identify hypothetical clients.
3. Describe the dataset and the cleaning/wrangling steps needed to prepare for analysis.
4. List other potential datasets that could be used.
5. Explain the findings.

1 Project Question

Can we predict drug prescriptions based on provider specialty?

2 Hypothetical Client

The hypothetical client is a drug company which would like to optimize their supply of drugs in various regions based on information about how many providers with various specialties practice in the region.

3 Dataset Description and Preparation

The Kaggle Opioid Prescriber Dataset dataset was uploaded to Kaggle in 2016. The dataset was compiled by AJ Pryor and is sourced from cms.gov. There are three different files:

- `opioids.csv` – Drug and generic names for 113 different opioids.
- `overdoses.csv` – Number of opioid-related overdose deaths in different states (no territories!) in the US in 2014.
- `prescribers.csv` – Prescriber information for 20,000 different prescribers in the US. Includes credentials, gender, state, specialty, and then the number of prescriptions given in 2014 for the 250 most commonly prescribed drugs. Finally, the boolean column `Opioid.Prescriber` indicates whether each provider prescribes opioids or not.

Pryor wrote an R script to generate `prescribers.csv`. This publically-available script could be used to generate more data, if desired.

Since the dataset originates from a Kaggle challenge, it is in very good condition: the csv files are easily read into pandas dataframes using `pd.read_csv()`. There are no missing values. The two processing steps I needed to undertake were: identifying outliers, and string processing of opioid names. Both of these tasks were completed relatively quickly. See the details of the steps I took here.

4 Other Datasets

The dataset Pryor provides on Kaggle is only a cleaned subset of the available data on cms.gov: one possibility to obtain more data would be to run the script Pryor made available to obtain more prescriber data.

For more information on opioid deaths (perhaps the drug company manufactures opioids and would like to contribute to the effort to end the opioid epidemic), we could apply for access to the National Death Index: access is only available to health and medical investigators and can take up to 3 months to be approved.

5 Findings

5.1 Insights from EDA

The full exploratory data analysis is available as a Jupyter notebook. In the exploratory data analysis portion of this project, I set out to explore the following questions:

Question 1: *Is the opioid overdose rate the same everywhere, or does it depend on the state? Are there US Census Bureau-defined regions or divisions that are associated with a higher rate of overdose?*

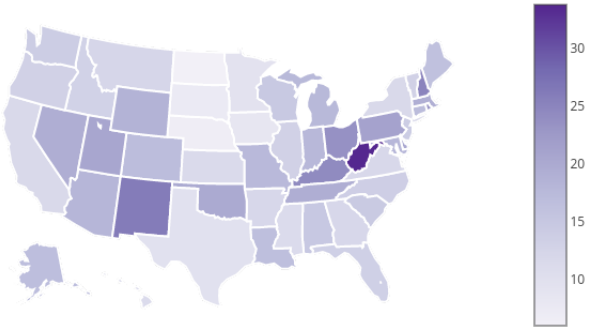

Figure 1: Geographic distribution of 2013 opioid overdose deaths in the US.

The opioid overdose death rate depends quite a bit on the state. We observed that, in general, rural states tended to have a higher than average opioid overdose death rate in 2013.

Question 2: *Is there really an opioid crisis?*

The answer to this question is an emphatic "yes" in the case of the 2013 overdose death numbers. There is evidence to show that opioid overdose deaths were more common than the national average suicide rate in many states: in other words, opioid overdose deaths appear to be one of the leading causes of death in certain US states.

Question 3 *Which variables available in the dataset influence overdose rate?*

There appears to be a positive correlation between state unemployment rates and opioid overdose death rates. Further analysis on this point is available in the inferential statistics section of the project.

Question 4 *Can we characterize provider specialties based on the drugs they prescribe in a year?*

TSNE visualizations showed that it is possible to distinguish between different specialist providers (such as optometrists and psychiatrists) based on the number and type of drugs they prescribed in 2013. This provides evidence that prediction of providers based on their drug prescriptions should be possible. Moreover, guessing it should be possible to predict prescription rates of certain drugs based on provider specialty.

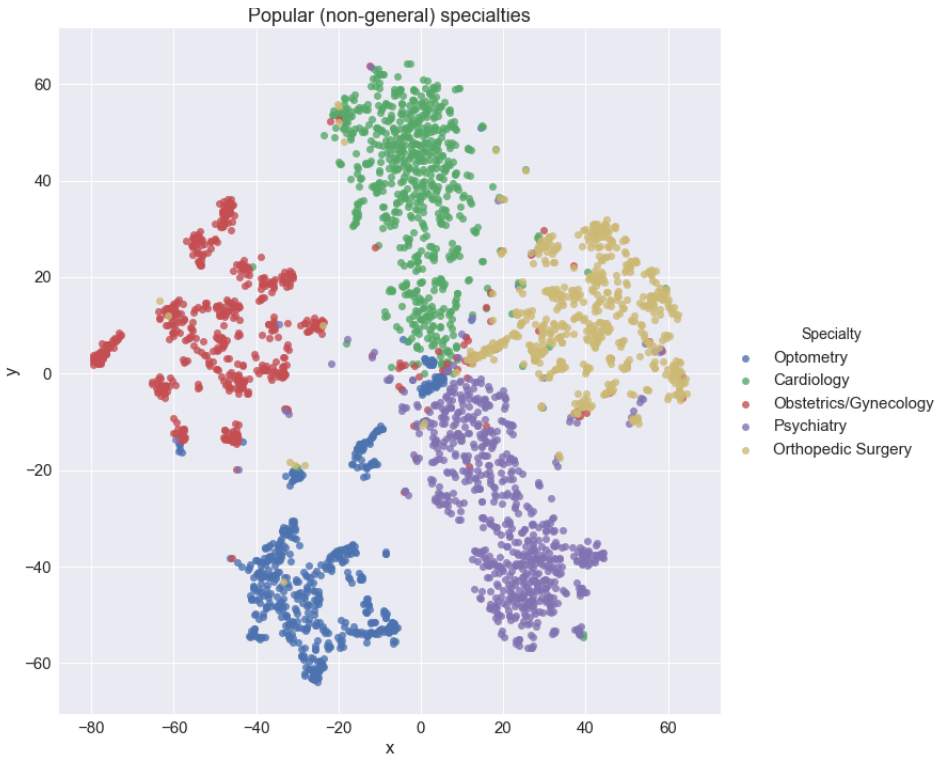

Figure 2: Sample TSNE visualization: specialist prescribers distinguished by their drug prescriptions.

5.2 Inferential Statistics

A separate report on the inferential statistics tasks I performed related to this project are available here, and the code for all statistical tests performed are available in the Data Story Jupyter notebook. A statistically significant correlation between state overdose death rate and state unemployment rate was observed. In contrast, there was no significant correlation between state overdose death rate and state population. Finally, statistical tests performed on the average numbers of drugs prescribed by different types of healthcare providers showed that, at least for some types of more specialized providers, there is an observable and statistically significant difference in prescription rate for certain drugs.

5.3 Prediction

The code for the prediction tasks performed are available in a separate Jupyter notebook.

5.3.1 Predicting Drug Prescription Based on Specialty

After isolating the most well-represented specialties in the dataset, we set about trying to predict drug prescription numbers based on prescriber specialty, opioid prescriber status, and total number of drugs types prescribed in a year. I created a linear model for each of the 20 most prescribed drugs in 2013 (within our dataset), and scored these models based on the coefficient of determination $R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$, where SS_{res} is the residual sum of squares and SS_{tot} is the total sum of squares.

For the best-performing drug Lisinopril, plotting the residuals indicated a strong relationship between the residuals (error of the prediction) and the target variable, which is not ideal for linear regression. After applying transformations to the prescriber Lisinopril prescriptions in per year as well as the total number of drugs prescribed by each provider, performance of the linear regression model improved from a R^2 score of .725 to a score of ¡bestscorehere¿. Moreover, the residual plot indicated that the residues are distributed much more randomly, which is desirable.

5.3.2 Predicting Specialty Based on Drug Prescriptions

For the selected specialties of Psychiatry, Cardiology, Obstetrics/Gynecology, Orthopedic Surgery, and Optometry, I used the number and types of drugs prescribed by each specialty to create a logistic regression model to predict the specialty of the provider. Performance taking only the 10 most popular drugs was poor, but using all drugs prescriptions recorded in the dataset resulted in a mean accuracy score of .96. Using a one-versus-rest implementation of the logistic regression model, I was able to calculate the ROC scores for each one-vs-rest logistic regression subproblem to provide further insight into the performance of the model. The AUC for each subproblem was .98 or greater, indicating that overall performance is excellent.

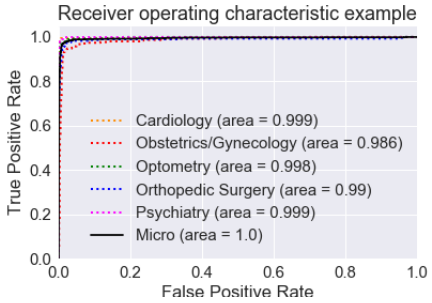

Figure 3: ROC curves for one-vs-rest subproblems