Machine Learning Engineer Nanodegree

Capstone Proposal

Jenny Lee
June 16, 2017

Background

The number of overdose deaths from substance abuse in the US was more than 33,000 in 2015, up from 19,000 in 2014. Prescription opiate drugs, especially methadone, oxycontin, and hydrocodone, are believed to have played a significant role in this public health crisis sweeping the US. Prescription opioid abuse, misuse, and dependence as a public health hazard is a daily phenomenon now; according to Centers for Disease Control and Prevention (CDC)[1], over 1,000 people are treated in emergency department everyday for misusing prescription opioid drugs. In addition to the health issues arising from overdose, opioid epidemic requires significant economic resources from cities and state governments for emergency call response and policing. The estimated total costs of US opioid epidemic reaches over 78 billion dollars.[2]

The major source of diverted opioids is physician prescription[3]. However, opioids prescription to patients with acute pain and patients with chronic pain requires a careful distinction. As Opioid is regarded as one of the most effective drugs for the acute pain management, limiting its use for patients who are in urgent need of pain control, post surgical status, cancer patients, and other health crisis, would not only be inhumane but also defeat its intended purpose. On the other hand, use of opioids for chronic non-malignant pain control has remained controversial for decades[4] and requires closer look in regards to the current opioids health crisis.

Problem Statement

[1] https://www.cdc.gov/drugoverdose/data/overdose.html
[2] "The Economic Burden of Prescription Opioid Overdose, Abuse, and Dependence in the United States, 2013," C. Florence et al. (2016)
[3] "Prescription opioid abuse: problems and responses," Compton WM et al. Prev Med (2015)
[4] "Health care reform in the United States: radical surgery needed now more than ever." Manchikanti L, Pain Physician. (2008)

This study attempts to build a predictive model of likelihood of a healthcare provider prescribing opioids drugs to patients with chronic pain. More specifically, we will identify correlated features of non-opioid drugs prescription history with opioid prescription. In addition, we will distinguish gender, specialty, and location that are more highly correlated to the prolonged use of a long term (more than 12 weeks) supply of opioids.

Datasets and Inputs

A total of four datasets, *detailed data, provider summary, national drug summary table, and state drug summary table*, were obtained from the web page of the Centers for Medicare and Medicaid Services (CMS)[5]. The *detailed data* contains the information on prescription drugs prescribed by individual health care providers and paid for under the Medicare Part D Prescription Drug Program in year of 2015. It includes the detailed prescription information such as the brand drug name, the number of patients who filled the drug more than ten times, the aggregate number of day's supply for which the prescription drug was dispensed.

The dataset *provider summary table* contains the demographics of the individual prescribers (n=1,102,268). Briefly, it includes the National provider identification number, name, gender, address, medical credential, specialty, and medicare enrollment status. It also includes the summary of the abstracted clinical data such as the number of total claims from the prescriber, total day's supply of all prescription drugs, total claims of antibiotic drugs, total claims of high-risk medication (HRM) drugs, total claims of antipsychotic drugs, and more importantly, the number of patients treated with opioid prescriptions, total days supply of opioids, and the ratio of opioid prescription to non-opioid prescription.

The rest two datasets are *national drug summary table* and *state drug summary table.* It lists the prescription drug names, whether the drug is categorized as antibiotics (n=78), opioids (n=29), antipsychotic (n=28), HRM (n=68), or others (n=951) and the number of prescribers of that drug grouped by nation and state, respectively. For more detailed information on how the dataset were collected, please refer to the CMS's webpage.

Solution Statement

We start by merging two datasets, *detailed data* and *provider summary,* to get combined features of providers' personal information and prescription history of drugs (sorted by its generic names, n=1154). Two new attributes, *op_avg_day_supply* and *op_longer*, are added. *Op_avg_day_supply* is the aggregate number of days supply

[5]https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/PartD2015.html

divided by total of patients for which opioids drugs were dispensed. *Op_longer* is labeled 1 if the provider has *Op_avg_day_supply* greater than 84 days (12 weeks), and 0 if less than 84 days. Then we train supervised classification models according to the *op_longer* labels to solve this large scale nonlinear problem.

Benchmark Model

This study was inspired by an article from kaggle[6], which used randomly selected subset of the dataset with samples (n=25,000) and features (n=250). In order to classify opioid prescribers, they built a gradient boosted classification tree ensemble model, and reported result of accuracy of 0.82, precision of 0.88, recall of 0.81, and F1 score of 0.84.

In this study, however, we are classifying prescribers who supply opioids in longer term (longer than 12 weeks) versus short term (less than 12 weeks). With naive predictor, that predicts all samples fall into long term label, the resulted accuracy is 0.13 and $f_{0.5}$ is 0.44.

Evaluation Metrics

Two evaluation metrics, accuracy and F-beta score, will measure the performance of our supervised classification models.

1. Accuracy

$$Accuracy = \frac{TP + TN}{FP + FN + TP + TN}$$

As two classes in the dataset are imbalanced and we want to put more weights on the recall than the precision, we will additionally incorporate F-beta score with beta equal to 0.5.

2. F- beta score with $\beta = 0.5$

$$F_{\beta} = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall}$$

Project Design

[6] https://www.kaggle.com/apryor6/us-opiate-prescriptions/kernels

This projects will consists of four parts: building automated input pipeline, exploratory data analysis, train/select/finetune models via scikit learn, and extension to the neural networks in TensorFlow.

Building automated input pipeline for minibatch learning

As the *detailed data* contains more than 25 million instances and its volume is larger than 3 GB, it easily takes up the memory of local machine, especially when we combine it with *prescriber summary* dataset to make a wide table. As such, we will utilize pandas' TextFileReader module and read in the instances by small chunks (size=100,000), which naturally lead us to use minibatch learning (online learning). In order to make sure of seamless and reproducible inputs, we automate the data input pipeline via data cleaning, feature selection, and feature scaling.

Exploratory data analysis

In order to gain insights, we explore the data by visualizing each attribute on some randomly selected batches. The t-distributed stochastic neighbor embedding is fitted to discover the dataset.

Train/select/finetune models via scikit learn

Scikit learn provides a handful of online learning algorithms. We will compare linear support vector machine via **stochastic gradient descent, logistic regression via stochastic gradient descent, perceptron, multinomial naive Bayes, and passive-aggressive algorithm** with default parameters. We will choose the one with highest $F_{0.5}$ score and fine tune the parameters via grid search.

Extend to neural networks

The large trove of given data, prepared in mini batches, containing complex nonlinear inner workings inspired us to consider extending to neural networks. We will move the same problem to the realm of TensorFlow and test if deep learning algorithms give improved performance.