Machine Learning Engineer Nanodegree Capstone Project

Prescription Opioid in the US[1]

Jenny Lee
jennylee.stat@gmail.com
July 12, 2017

Introduction

The number of overdose deaths from substance abuse in the US was more than 56,000 in 2016, up from 33,000 in 2015. 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)[2], over 1,000 people are treated in emergency department every day for misusing prescription opioid drugs. In addition to the health issues arising from overdose treatment, 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.[3]

The major source of diverted opioids is physician prescription[4]. 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, near end of life 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 (CNMP) control has remained controversial for decades,[5] and requires a closer look in regards to the current opioids health crisis.

---

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

This study demonstrates a predictive model of the likelihood of a healthcare provider prescribing opioids drugs to patients with chronic pain. More specifically, we used a variety of classification techniques to predict whether or not a healthcare provider dispense prescription opioids in long term (more than 84 days). In addition to predicting opioids prescription to chronic pain patients, we investigated features including non-opioid drugs prescription history, provider's gender, medical specialty, and location to understand the factors highly associated with the prolonged term supply of opioids drugs.

Materials and methods

Study population

Two datasets, *detailed data* and *provider summary*, were obtained from the web page of the Centers for Medicare and Medicaid Services (CMS)[6]. 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 the year of 2015. It includes the detailed prescription information including the generic drug names, the number of beneficiaries, and the aggregate number of day's supply for which the prescription drug was dispensed. The dataset includes a total of 1,701 unique generic drug names, including 36 opioids.

The dataset *provider summary table* contains the demographics of the individual prescribers. 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 including the number of total day's supply of all prescription drugs (mean = 77,996, SD = 153,926), total claims of antibiotic drugs (mean = 94.05, SD = 137.94), total claims of high risk medication (HRM) drugs (mean = 47.37, SD = 88.49), total claims of antipsychotic drugs (mean = 15.18, SD = 72.74), and more importantly, the number of patients treated with opioid prescriptions (mean = 37.97, SD = 60.09), total days supply of opioids (mean = 2724.59, SD = 9524.14), and the percentage of opioid prescription to non-opioid prescription (mean = 13.30, SD = 18.54).

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

Data preparation

We took the following steps to prepare the dataset. First, we added two new attributes `avg_op_day_supply` and `op_longer`. We defined the average days of opioid supply as the total days of opioid supply divided by total number of beneficiaries for which the opioids were dispensed to. We labeled `op_longer` as 1 if the health care provider's `avg_op_day_supply` is greater than 84 days. Second, we dropped drug name features if a number of annual prescribers were less than 1,000. Among the remaining 1154 drug names, all the flagged drugs - antibiotics, antipsychotics, and hrms drugs were included. Further, we sampled to select 300 more drugs based on its frequency of total prescriptions. After converting categorical features into dummy variables, the total number of feature was 535. Next, each numerical feature was log transformed, then min-max scaled to be ranged between zero and one. After data preparation, the total number of instances in the combined dataset was 630,900. Of these, 438,183 (69.44%) of health care providers prescribed opioids in 2015, and only 96,993 (15.37%) of them prescribed opioids longer than 84 days on average.

Mini batch learning (online 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 utilized pandas' TextFileReader object and read in the instances by small chunks, size of 100,000, which naturally led us to use minibatch learning (online learning). In order to make sure of seamless and reproducible inputs, we automated the data input pipeline via data cleaning, feature transformation, feature selection, and feature scaling. Preprocessed dataset was further splitted in fixed batch size of 256, and shuffled before being passed to the each of classifiers and deep neural network.

Machine learning algorithms

Five supervised classifiers candidates 1) stochastic gradient descent algorithm with logistic regression loss function, 2) stochastic gradient descent with linear support vector machine loss function, 3) perceptron, 4) naive Bayes multinomial classifier, 5) Passive-aggressive classifier support `partial_fit()` function in scikit learn API. We compared the runtime, validation accuracy, and validation F-score at its default settings to select the best classifiers.

Then we extended our analysis to deep neural network to test if deep learning algorithms improve the performance. The classification loss is defined as the sum of the cross entropy between true label and predicted label, averaged by the number of instances.

$$loss\ (w,\ b)\ =\ \frac{1}{N}\sum_i\sum_j (wx_i + b) * ln\hat{y}_j$$

Where w and b represents weight matrix and bias, and x and y are input instance and label, respectively. We passed the scaled features to the first hidden layer with 256 hidden units, followed by the second hidden layer with the 128 hidden units. We use the Adam optimizer with a learning rate of 0.00005 to train the deep neural net. Dropout with probability 0.5 is applied to both hidden layers.

Performance measure

Two evaluation metrics, accuracy and F-beta score, were used to measure the performance of the 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 additionally incorporate F-beta score with beta equal to 0.5.

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

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

With naive predictor, that predicts all samples fall into positive label, `op_longer`, the resulted accuracy is 0.1537 and F-score (beta = 0.5) is 0.4759. If naive predictor predicts no samples have positive labels, then the accuracy and F-score(beta = 0.5) were 0.8463 and 0.0000, respectively.

Exploratory data analysis

First three batches that are not included in the test set are explored to get a peek of the data structure and to gain insights. After concatenating and preprocessing the batches, the number of instances in the sampled dataset was 17,703.

1. Overdose deaths Vs. the ratio of long term opioid prescribers

The ratios of health care providers who prescribed opioids longer than 84 days to total number of healthcare providers are compared for each state. Montana was the highest (44.19%), followed by Wyoming (38.09%), Alaska (32.14%). South Dakota was the lowest, only 4.65% of providers prescribed opioids in long term, followed by Vermont (11.11%), and Mississippi (13.47%). In Figure 1, we plotted these ratios against the number of opioids related overdose deaths in 2014[7]. The size of the marker reflects the number of overdose death in the state. They seem to be somewhat linearly associated; States with higher overdose deaths per population tend to show higher ratios of long term opioid prescribers. This suggests that dispensing opioids in longer term might be related to the opioids health crisis.

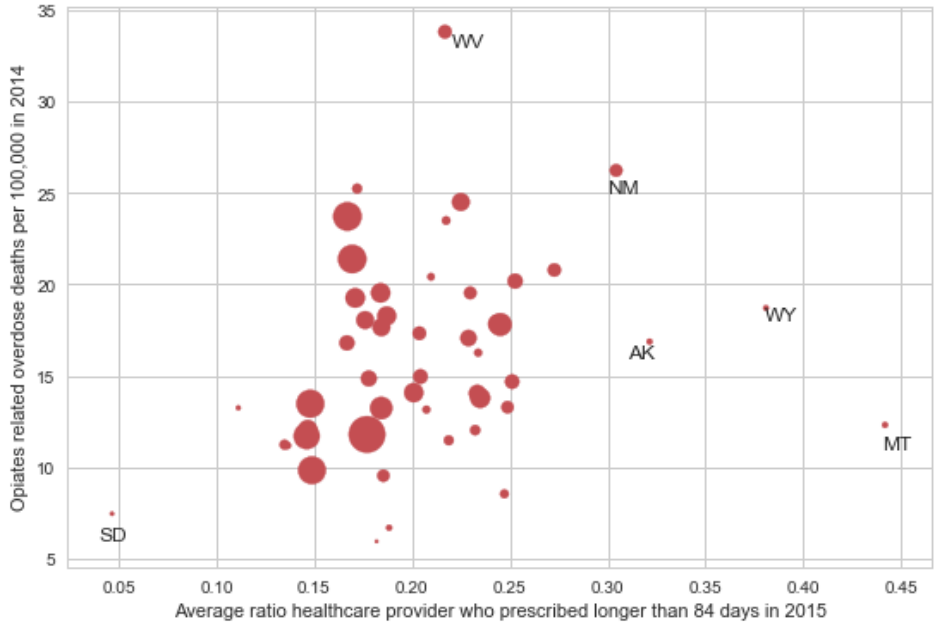

Figure 1. The number of opioids related overdose deaths is plotted against the ratio of long term opioid supplying prescribers to total prescribers. The size of marker reflects the number of overdose deaths in the state. They seemed to be roughly linearly correlated, suggesting that dispensing opioids in longer term might be related to the opioids health crisis.

2. Specialty Vs. average days of opioid supply

The boxplot in Appendix A Figure 1. shows the distribution for average days of opioids supply. Average days of opioid supply vary widely by individual prescriber's medical specialty; Pain management, rehabilitation, anesthesiology tend to supply opioid longer period of time compared to others. On the other hand, specialties that involve surgeries tend to have much shorter average days of opioid supply per patient. One thing to note in the plot is the fat right tails of the family practice, internal medicine, nurse

[7] https://www.kaggle.com/apryor6/us-opiate-prescriptions/downloads/overdoses.csv

practitioner, and physician assistant. These four require careful look as they, combined with anesthesiology (5.42%), take up about 72.96% of total opioid prescription in the US. Like other specialties in the dataset, they have right skewed data. They also show the longer fourth quantile and many outliers at extreme values. A good classifier trained on this data will be able to learn and explain how these outliers are distinct from the majority of the instances in the same specialty.

3. Opioids drugs vs `op_longer`

We categorized the dataset by medical specialty, and computed the correlation between the label (`op_longer`) and opioids drug names. The correlation matrix for most commonly prescribed opioids and most frequent specialty is visualized as a heatmap in **Figure 2.** Hydrocodone-acetaminophen, oxycodone HCl, oxycodone HCl acetaminophen, and tramadol seem to be most highly correlated with the label for most specialties. These four opioids are more likely to be prescribed in prolonged term. Of them, hydrocodone-acetaminophen prescribed in diagnostic radiologist showed the most strongly correlated to the label. Only handful of total 29 opioids are positively correlated with the label for most of specialties, while pain management, anesthesiology, physical medicine and rehabilitation used more varied opioids prescription.

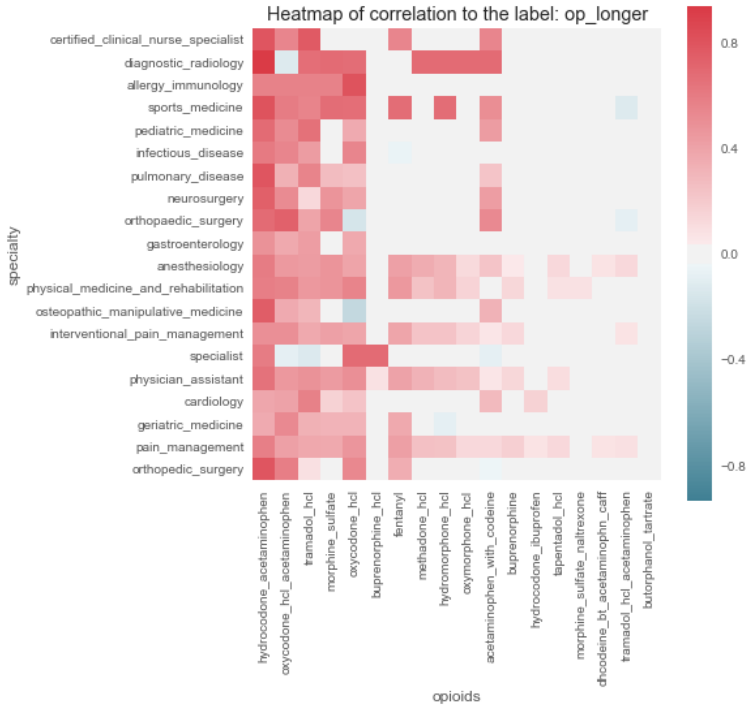

Figure 2. Training set is categorized by medical specialty and the correlation between the label and average days of supply for each opioids drugs are computed. The resulted correlation matrix is shown as a heatmap. For most specialties, only handful of opioids including hydrocodone acetaminophen, oxycodone HCl acetaminophen show positive correlation with the label.

Results

Training and evaluation with scikit learn classifiers

1. Runtime

The resulted prediction time and training time per 10,000 instances for five supervised classifiers are plotted in Figure 3. The naive Bayes multinomial classifier took the longest prediction time, and the longest training time, thus resulted the longest total time. The rest of four classifier took almost the same time, near 0.09 seconds per 10,000 instances in total time. However,it seems that all classifiers can train fairly quickly via `partial_fit()` and the time difference is insignificant. The slowest model, NB multinomial took only 7.18 seconds to train the whole instances in training set.

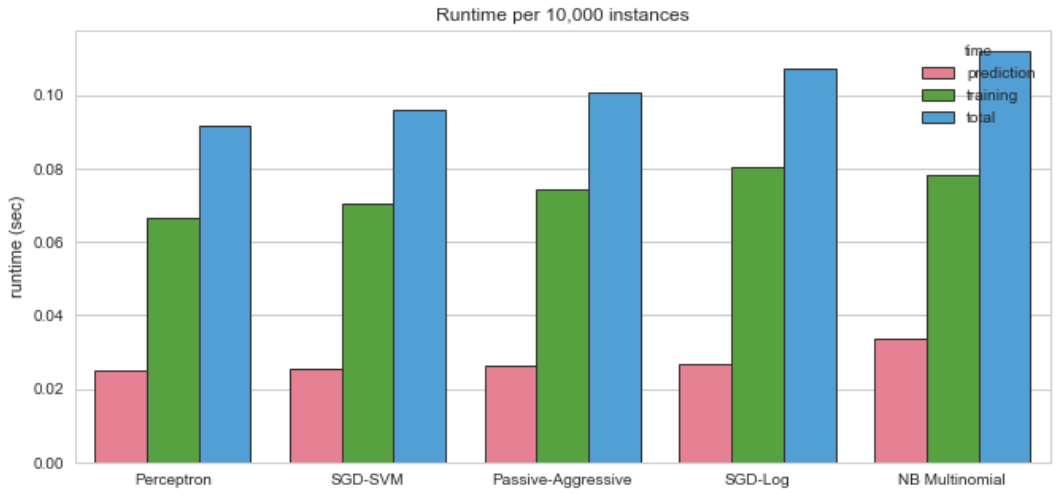

Figure 3. Runtime of five supervised classifiers at default setting are compared. Naive Bayes multinomial classifier was the slowest, taking about 0.11 seconds to process 10,000 instances. The rest four algorithms results were very similar for both training time and prediction time.

2. Accuracy and $F_{0.5}$ score

The accuracy score and F-score on training mini batches from five classifiers are shown in **Figure 4** below. For both accuracy and f-beta score, the SGD-SVM and SGD-logistic were top two classifiers on training dataset, while the naive-Bayes scored the lowest on both metrics. The SGD-Logistic performed highest on validation accuracy and validation f-beta score, closely followed by SGD-SVM. **Table 1** shows mean validation scores on {} validation batches.

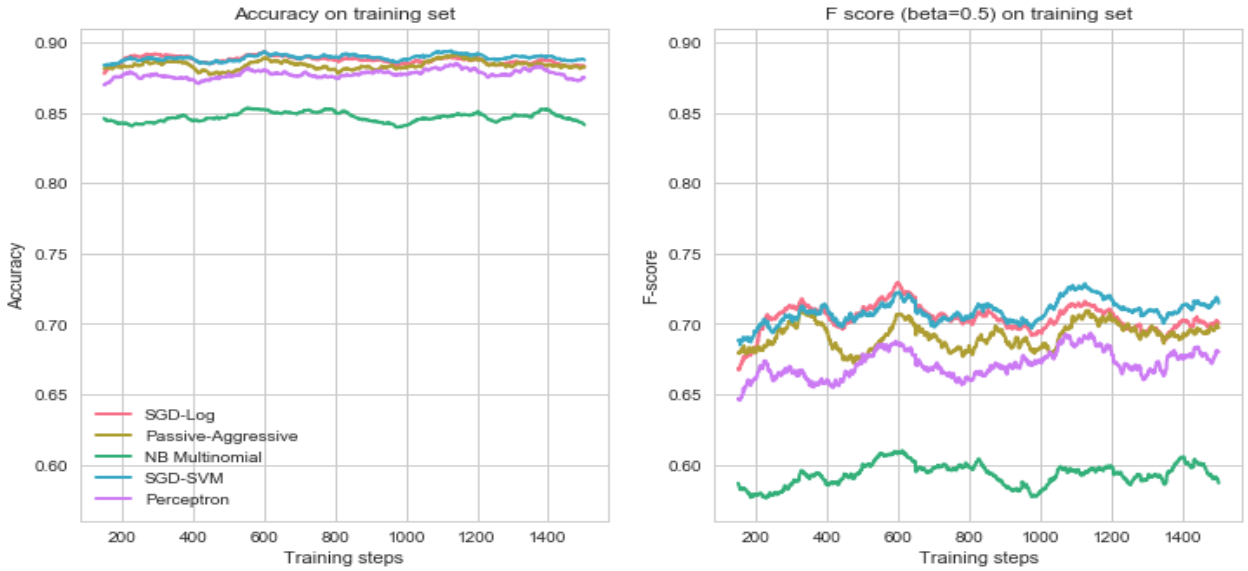

Figure 4. Accuracy and f-beta score (beta=0.5) are plotted against training steps (rolling window = 150). SGD-Logistic and SGD-SVM performed better than other classifiers on training dataset.

classifier	validation accuracy	validation F-score
SGD-Log	0.8693	0.6428
SGD-SVM	0.8584	0.6122
NB Multinomial	0.8395	0.5608
Perceptron	0.8311	0.5364
Passive-Aggressive	0.8317	0.5351

Table 1. Average accuracy and average F-score on validation set is shown above. SGD-logistic scored highest on both accuracy score and f-beta score, closely followed by SGD-SVM.

Fine tuning via grid search

We further compared SGD-SVM and SGD-Logistic by additionally searching for the hyper parameter space. The hyperparameters of elastic net ratio mixing parameter, regularization ratio, and learning rate are evaluated by grid search. Top 5 results from various hyper parameter profiles are shown in the table 2 below.

classifier	alpha	L1 ratio	Validation F-score
SGD-Log	0.0001	0.9	.7143
SGD-Log	0.0001	0.8	.7143
SGD-SVM	0.0001	0.8	.7080
SGD-Log	0.0001	0.0	.6977
SGD-Log	0.0001	0.2	.6977

Table 2. Grid search results of hyper parameter space. Top two models achieved the validation f-beta score(beta=0.5) of 0.7143.

Top two models had the same validation F-beta scores. After the optimal hyperparameters were chosen, the classifier again trained on both training set combined with validation set. The best and final model achieved an overall accuracy of 0.8752 and f-beta score of 0.6661 on the hold-out test set and 0.6661.

	Observed class	
	False	**True**
False	57,619	3,602
True	5,722	8,127

Table 3. Test set confusion matrix for the best classifier, SGD-logistic(alpha = 0.001, L1 ratio=0.9)

Neural Network

Accuracy and F-beta scores were validated every 50 batches by randomly selected 256 instances from hold out validation set. The results are visualized in **Figure 5** for total number of epochs of 50. This model achieved the accuracy of 0.8782 and f-beta score of 0.6711 on the hold out test batches, slightly improved than SGD-logistic (alpha = 0.9, learning rate = 0.0001).

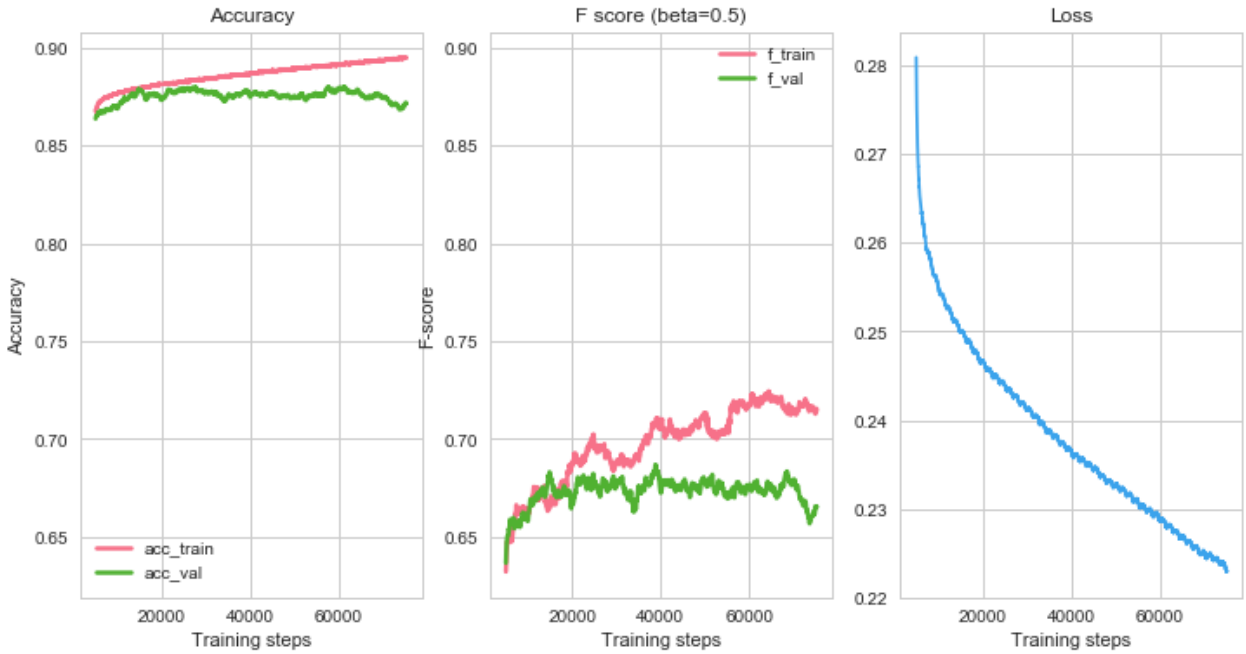

Figure 5. Accuracy and F-beta score results on both training set and validation set for n_epochs = 50.

	Observed class	
	False	**True**
False	57,589	3,632
True	5,473	8,376

Table. 4. Test set confusion matrix for deep neural network. This model achieved accuracy of 0.8782 and f-beta score of 0.6711.

Discussion

For our dataset, fine tuned DNN performed slightly better than the stochastic gradient descent - logistic classifier. The accuracy is not 'perfect' because we set the labels by categorizing continuous variable (average days of opioid supply) with an arbitrary threshold (84 days). Both model had higher false positive rate than false negative rate.

classifier	Avg test accuracy	Avg test F-score
SGD-Log	0.8751	0.6661
DNN	0.8782	0.6711

Table 5. Average test accuracy and F-beta score on hold out test set for final two models.

We forward pass the identity matrix to the DNN and softmax transformed to estimate probabilities of each feature. For SGD-Logistic classifier, `get_proba()` was utilized to compute probabilities. **Figure 6** shows the top 25 estimated probabilities of which a prescriber is a long term opioid supplier, if that feature is flagged 1 and others zero. Many features came up for both models. Carisoprodol, diazepam, alprazolam, gabapentin, pregabalin, hydroxychloroquine sulfate, cyclobenzaprine HCl, and tizanidine HCl were highly associated with the label based on both model. As we expected from the EDA, of medical specialties, pain management, interventional pain management, rheumatology, hematology, and hospice and palliative care were estimated the label is true with higher probability for both model. Among states, only Alaska came up within 25 for both models.

A limitation of the study is that the dataset only includes a snapshot of the total population as beneficiaries of Medicare are elderly population of age greater than 65.

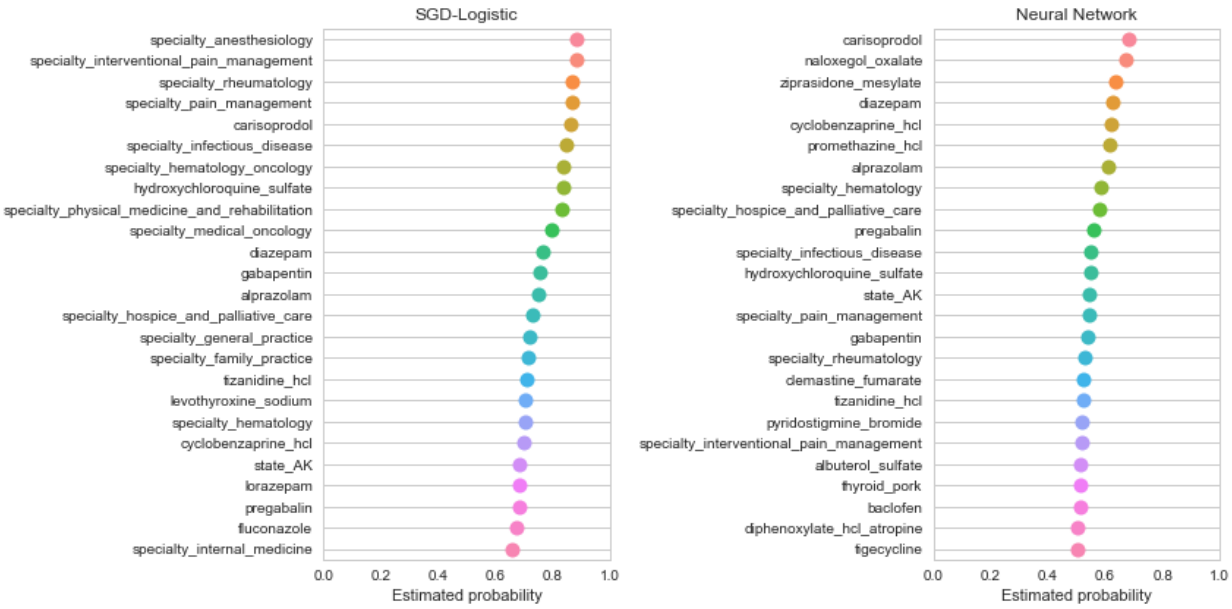

Figure 6. A comparison of estimated probabilities for two final models.

Appendix

A. Figures

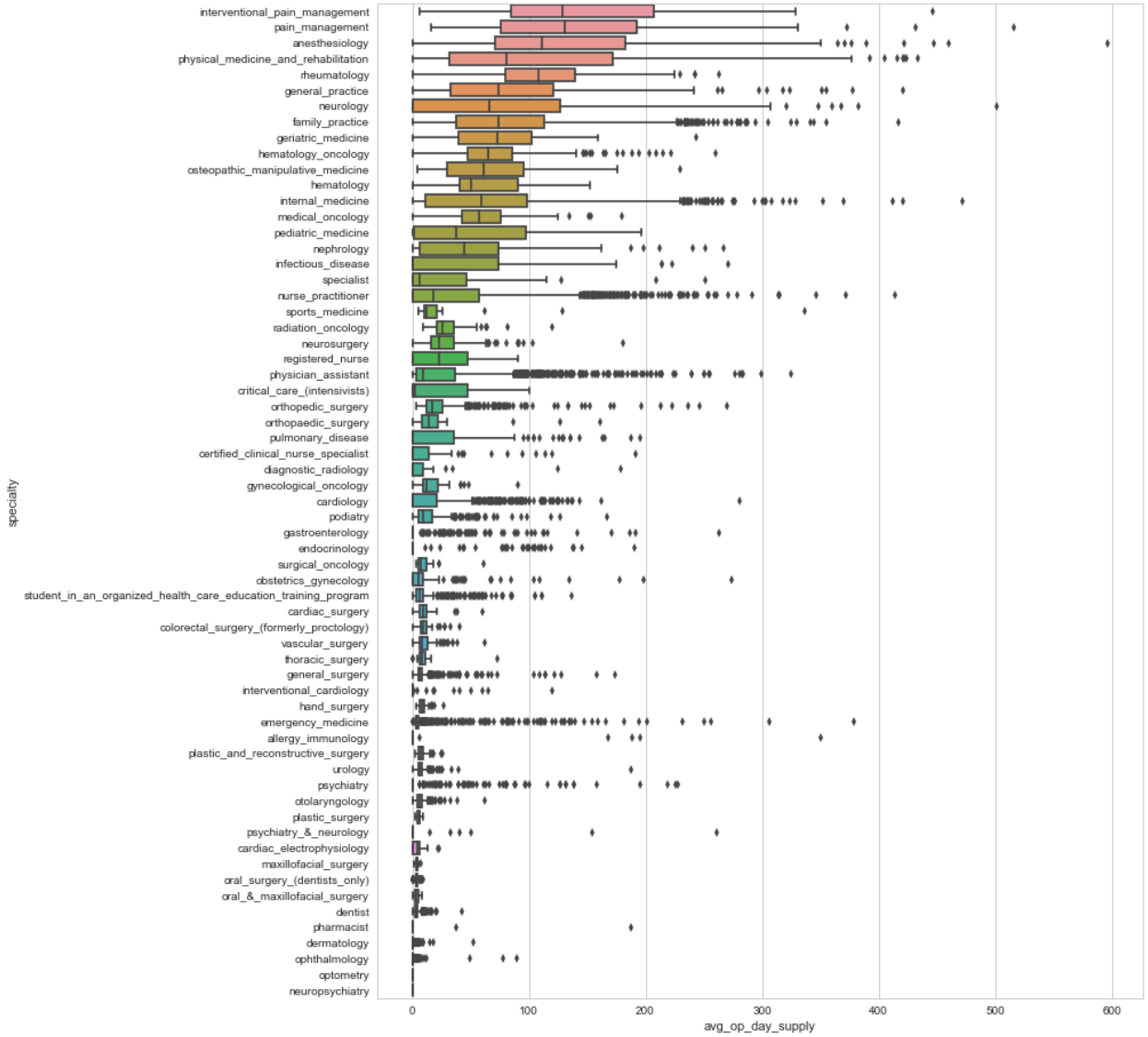

Figure 1. The sampled small dataset for exploratory analysis is grouped by medical specialty of each health care providers. The distributions for the average days of opioid supply of each medical specialty are compared as box plot shown above. Pain management, rehabilitation, and anesthesiology show longer term of average days of opioid supply, while specialties that involve surgeries tend to dispense opioids shorter term on overage.