Haojun Yang

CMSE 381

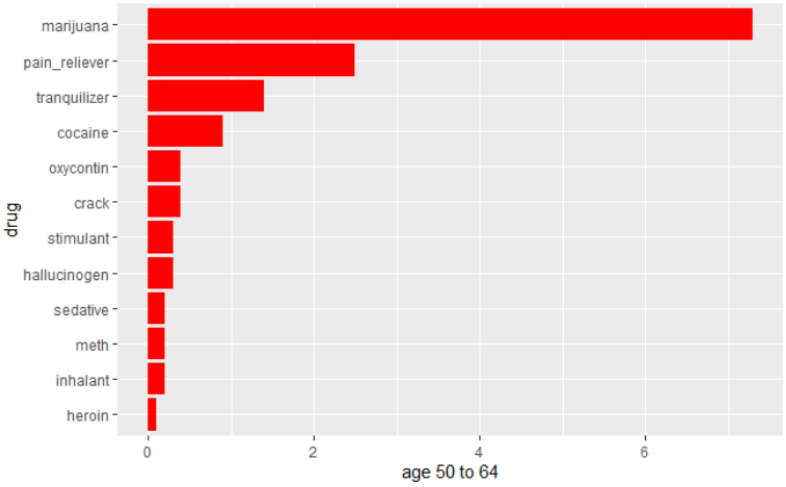
Dr. Xie

2021.04.16

Final Project

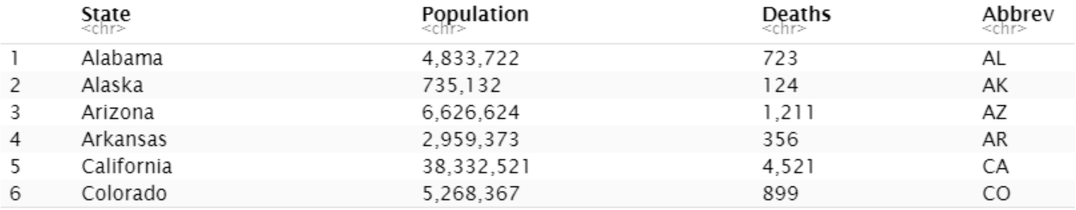
Introduction:

Drug using and over dosing is getting more and more popular nowadays among all ages. Between the age 50 to 64, 7.3% of the people have used marijuana. That's almost 5 percent higher than pain reliever which is the second place (excluding alcohol). Because of this "baby boomer" of drug using, overdose happens more often now. Opioid prescription, among all drugs, is one of the big cause of drug addiction. Thus, this study, we will mainly focus on what's the most important drug feature in predicting whether the prescriber is an opioid prescriber or not among family practice in the U.S., and what's the best model for predicting accuracy.

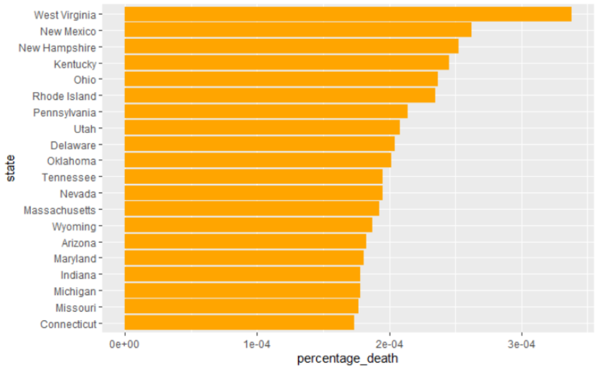


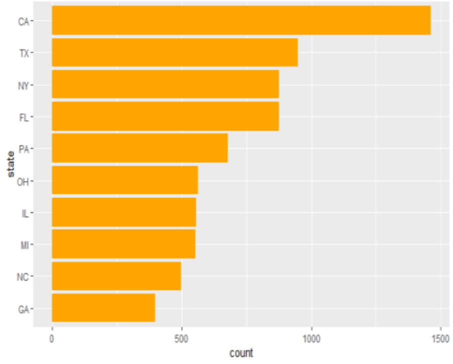
Data:

Two datasets were used in doing this project, they are “prescriber-info” and “overdoses”. The overdose dataset mainly focuses on population and death by overdose by state. To find out the percentage of deaths per state, a “percentage of death” column was calculated by using “Deaths” to divide by “Population” and added to the dataset. After plotting out the bar plot (plot on the left), we can see West Virginia has the highest percentage, followed by New Mexico and New Hampshire as second and third. The “Prescriber-info” dataset also has a state column showing which state the opioid prescribers from. From the bar plot (the plot on the right we can tell California has the most amount of opioid prescribers, followed by Texas and New York as second and third. Three of the top 10 opioid prescriber states appeared to be top 20 overdose states in the U.S. Family practice prescribes the most amount of opioid prescriptions, followed by internal medicine and nurse practitioners as second and third (plots shown below on the left). And among all credentials, M.D. and D.O. prescribes the most amount of opioid prescriptions (plots

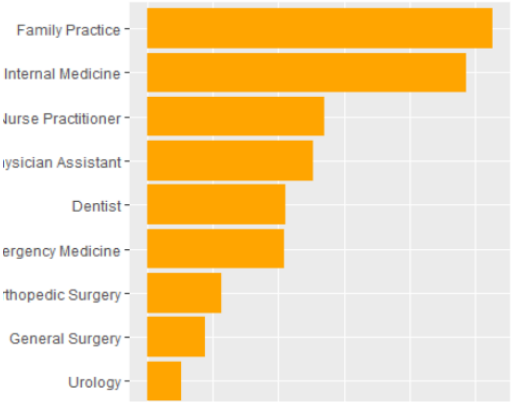


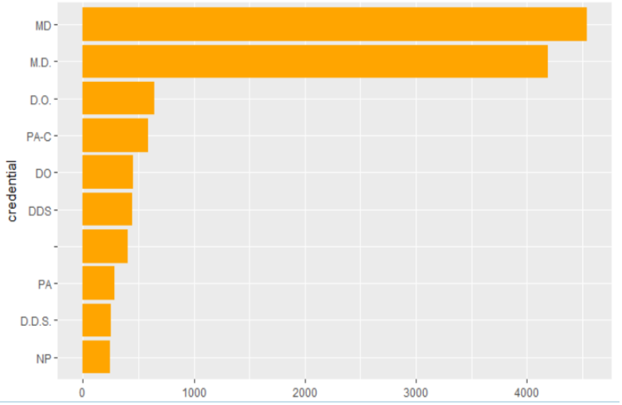
	State <chr>	Population <chr>	Deaths <chr>	Abbrev <chr>
1	Alabama	4,833,722	723	AL
2	Alaska	735,132	124	AK
3	Arizona	6,626,624	1,211	AZ
4	Arkansas	2,959,373	356	AR
5	California	38,332,521	4,521	CA
6	Colorado	5,268,367	899	CO





shown below on the right). “Prescriber-info” is a 25000x256 dataset. The first five





columns are “ID”, “Gender”, “State”, “Credentials”, and “Specialty” of the prescriber. The next 250 columns are drug info prescribed by the prescriber. The last column is a factor column indicates whether the prescriber is a opioid prescriber or not.

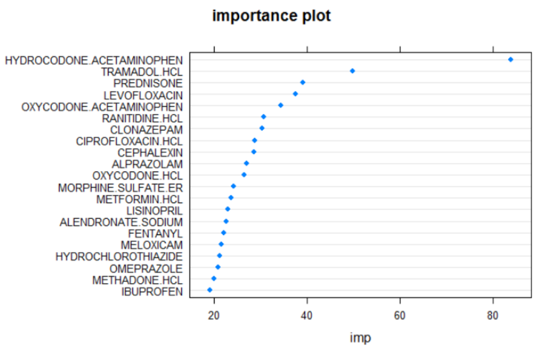
Modeling:

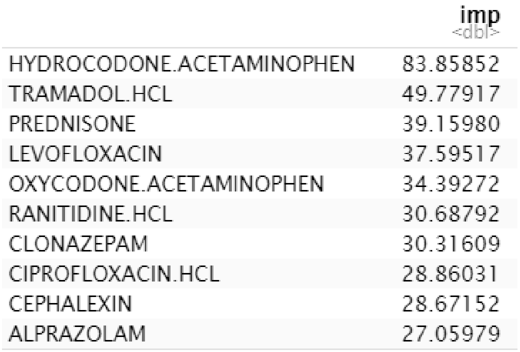
This section will be divided into 4 sections, “Tree Based Method”, “SVM”, “Logistic Regression”, and “KNN”. The dataset is divided 0.75 and 0.25 for training and testing set for each model.

1. Tree Based Method

Tree based method is a very powerful model to use when it comes to classification. The regular tree based method is no precise at all while doing predicting, but random forest and boosting method, which were both used in this project, are some very robust models that give precise prediction. To fit the random forest model, 1000 trees were “grown”, and 16 features were chosen randomly to be predictors for each tree. The most important variable for this model is Hydrocodone

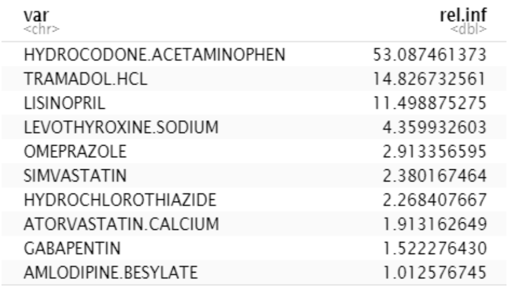
Acetaminphen, which has around 83.86% importance on this model.



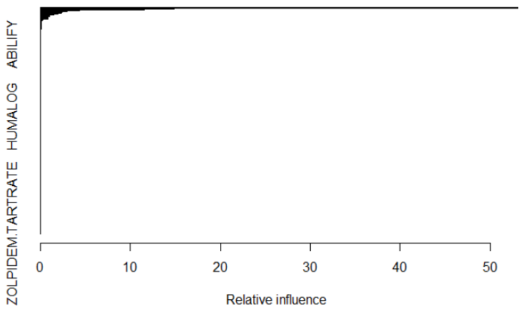


	imp <dbl>
HYDROCODONE.ACETAMINOPHEN	83.85852
TRAMADOL.HCL	49.77917
PREDNISONE	39.15980
LEVOFLOXACIN	37.59517
OXYCODONE.ACETAMINOPHEN	34.39272
RANITIDINE.HCL	30.68792
CLONAZEPAM	30.31609
CIPROFLOXACIN.HCL	28.86031
CEPHALEXIN	28.67152
ALPRAZOLAM	27.05979

TramaDol HCL and Lisinopril have the second and third strongest influences on the model. The prediction error on the testing set is 0.0614. Boosting method is a bit different to the random forest model. Boosting method "grows" small trees and builds the next tree based on the previous tree's residual. I "grew" 1000 trees and set the interaction depth as 1. Hypdrocodone Acetaminphen still stood out to be have the



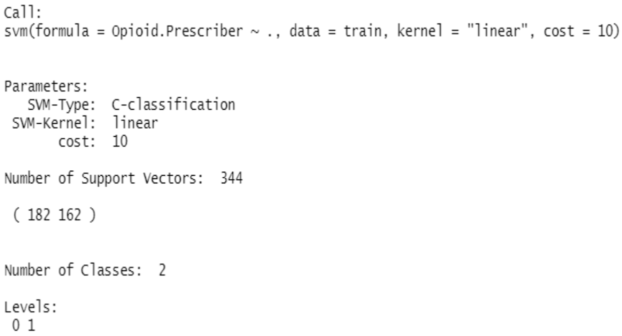
var <chr>	rel.inf <dbl>
HYDROCODONE.ACETAMINOPHEN	53.087461373
TRAMADOL.HCL	14.826732561
LISINOPRIL	11.498875275
LEVOTHYROXINE.SODIUM	4.359932603
OMEPRAZOLE	2.913356595
SIMVASTATIN	2.380167464
HYDROCHLOROTHIAZIDE	2.268407667
ATORVASTATIN.CALCIUM	1.913162649
GABAPENTIN	1.522276430
AMLODIPINE.BESYLATE	1.012576745



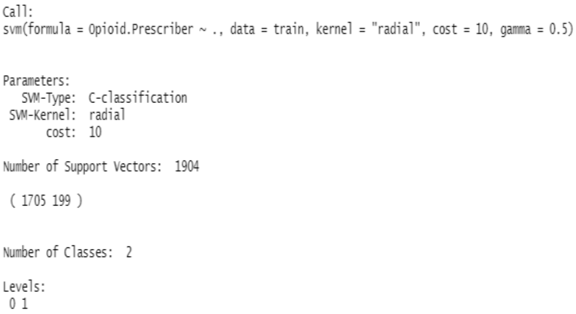
strongest influence on the model with 53.09% relative influence. Tramadol HCL and Lisinopril are still the second and third place. Boosting method also has a lower prediction error of 0.053. Those two models answered our first question. Hydrocodone Acetaminophen, Tramadol HCL, and Lisinopril are the 3 most important drug features for predicting whether or not the prescriber is a opioid prescriber.

2. SVM

SVM is a strong model for predicting small and medium size dataset like “Prescriber-info”. The advantage of SVM is the “kernel trick” and calculating model in higher dimension to separate the data better. I used three kernels, linear, radial, and polynomial. However, after doing cross validation, polynomial chose degree of 1 as the best choice, which is the same as linear. Thus, we will combine those two together. For linear kernel, cross validation chose 10 to be the cost. This model uses 344 support vectors to separate two classes. For radial kernel, cross validation chose cost to be 10 and gamma to be 0.5 for the model. 1904 support vectors were used to set the boundaries. Linear and radial kernel both did a good job on prediction. Linear had a test error of 0.089 and radial had a test error of 0.067. However, Boosting method and random forest still have a better prediction accuracy so far.



```
Call:
svm(formula = Opioid.Prescriber ~ ., data = train, kernel = "linear", cost = 10)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  10

Number of Support Vectors:  344

 ( 182 162 )


Number of Classes:  2

Levels:
 0 1
```

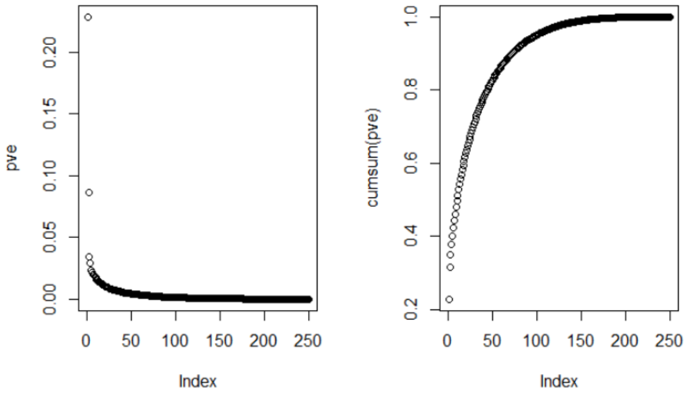


```
Call:
svm(formula = Opioid.Prescriber ~ ., data = train, kernel = "radial", cost = 10, gamma = 0.5)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  10

Number of Support Vectors:  1904

 ( 1705 199 )


Number of Classes:  2

Levels:
 0 1
```

3. KNN and PCA

KNN is a classic classification model. This model uses its neighbor’s average to determine whether the class should be predicted as 1 or 0. However, KNN is not great when the dimension of the data is high like this one. This is where PCA comes in handy. Principle components analysis does a great job on reducing dimensions. After using PCA, we reduced the dimension down to 45 predictors (45 components explains

80% of the variance). Using cross validation with KNN can determine that when k = 8, KNN has the smallest prediction error of 0.12. However, boosting and random forest is still better than 0.12.



4. Logistic Regression

Logistic regression is basically a linear regression model but for classification. It uses linear regression to calculate the probability of class being picked. Logistic regression did the worst job for this problem, it had the highest prediction error of 0.154. According to SVM, the radial kernel has a lower error rate than linear kernel, which means this data is not a linear data. This explains why logistic regression didn’t do as good of a job as the other models.

Conclusion:

This project helped us to learn that Hydrocodone Acetaminophen, Tramadol HCL, and Lisinopril are the 3 biggest influence on predicting whether a prescriber is an opioid prescriber or not, and the best model to predict this data is boosting method and random forest.