Foundations of Data Science
Project Report

A Predictive Modelling and Analysis
On
Causes of Death by Opioid Consumption

Proposed By:
Akshay Tambe (apt321@nyu.edu)
Aditya Bhatt (apb462@nyu.edu)
Team: The Cavalry

NYU | TANDON SCHOOL OF ENGINEERING

Project Source Code

Github Link: https://github.com/akshaytambe/Opioid-Prescription-Predictive-Analysis

Problem Statement & Motivation

The 21st Century runs on statistical analysis focussing mainly on social causes. This motivates us to go for healthcare data as “Analysis in Healthcare” will definitely contribute for Social Good. From the healthcare community article, we found that there is an increase in the national consumption of heroin and prescription painkillers.

A class of controlled pain-management drugs that contain natural/synthetic chemicals based on morphine which is the active component of opium. These drugs are often called as opioids which effectively mimic pain-relieving chemicals that the body produces naturally. We wish to mitigate the problem of Drug Overdoses by uncovering trends and latent features in prescription data and overdose deaths data with respect to different regions.

Problem:

Opioid Prescription helps to treat moderate to severe pain but it also leads to addiction and hence, people misuse it by consuming it at a higher rate. An overdose of Opioids leads to death.

1. How to improve ways the opioids are prescribed?
2. How to reduce the death toll due to a drug overdose in North America?

Goal: Detect opium components in the data and predict prescribers with opioid prescriptions which may lead to drug addiction.

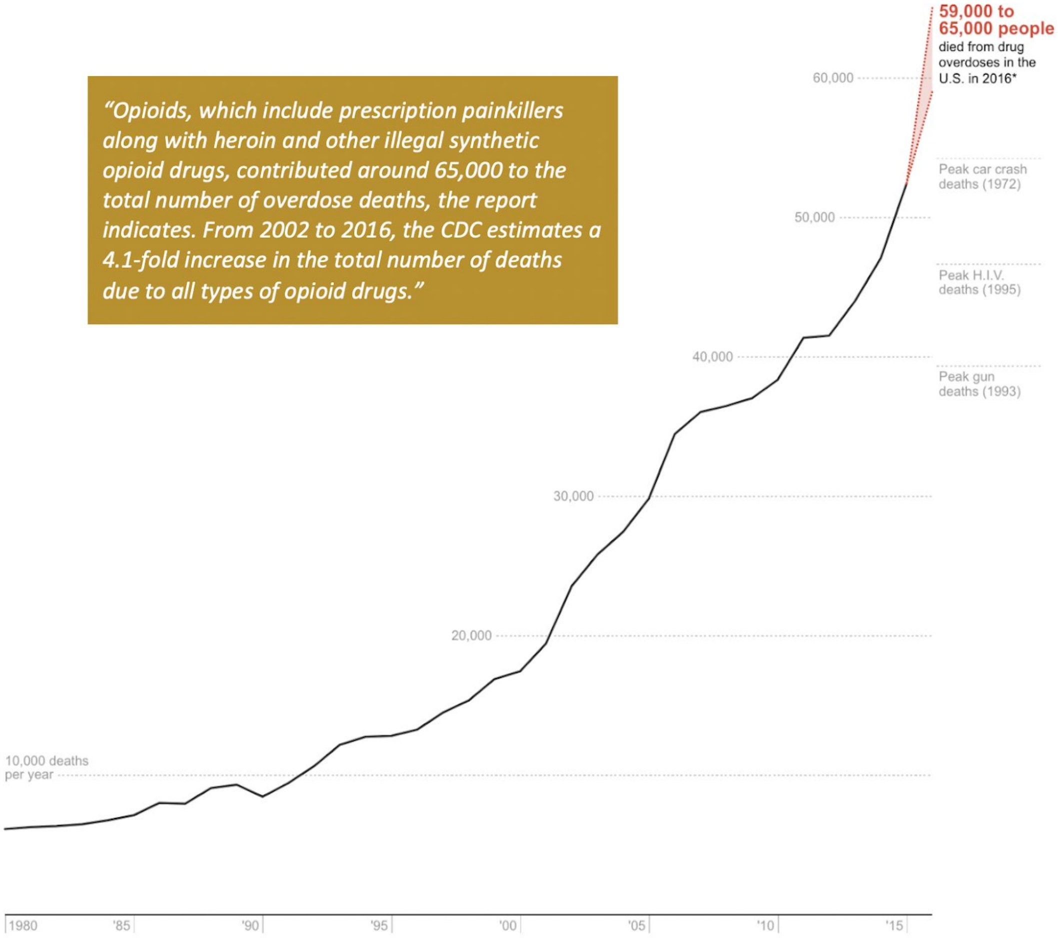
Data: Prescription data with drug components, medicare domain, opium drug list, state-wise overdose deaths.

Target Variable: Opioid Prescriber or not?

Impact: Model is able to predict prescriber with opioid prescription and individual chemicals and features of high importance causing opioid addiction.

Background

There numerous publications on this that take the initiative to find ways for the government to intervene at the right time, in order to prevent Opiate-related drug overdoses.



> *"Opioids, which include prescription painkillers along with heroin and other illegal synthetic opioid drugs, contributed around 65,000 to the total number of overdose deaths, the report indicates. From 2002 to 2016, the CDC estimates a 4.1-fold increase in the total number of deaths due to all types of opioid drugs."*

CNN Report and Video to explain *"Why opioids are addictive?"*
https://www.cnn.com/2018/08/16/health/us-overdose-death-report-cdc/index.html

Publications:

1. Stat Article - Discusses a disastrous tally of deaths caused by drugs.
2. NY Times on Drug Deaths in America - Discusses growing death cases caused by opioids.

We have also worked with an employee intern who works at *"The United States Pharmacopeia"* which is an organization responsible for standardizing medical products all over the world and is working on a similar research. They have a compendium of drug information which will be helpful for us for fetching interesting insights from the dataset. It was helpful for us to get more insights into chemicals and prescription procedures that may relate to drug addiction.

Data Collection

To conduct this study, we source data available from publicly accessible datasets provided.

1. **Information on drug prescriber (Dataset with Target Variable for Prediction):** (https://www.cms.gov/ :*“Medicare Provider Utilization and Payment Data: Part D Prescriber”)*

   This dataset contains information of different drug prescribers which drug prescription data and the predictor variable which gives the likelihood of a given doctor being a significant prescriber of opiates.

Attribute	Description	Data Type	Attribute Type
NPI	Unique National Provider Identifier Number	Numerical	Categorical
Gender	Gender of the Subject (M/F)	String	Categorical
State	U.S. State by abbreviation	String	Categorical
Credentials	Set of initials indicative of medical degree	String	Categorical
Specialty	Description of type of medicinal practice	String	Categorical
Drug_List	A long list of drugs with numeric values indicating the total number of prescriptions written for the year by that individual	Numeric	Quantitative
Opioid.Prescriber	A boolean label indicating whether or not that individual prescribed opiate drugs more than 10 times in the year	Boolean	Categorical

2. **List of Drugs Classified as Opiates (External Dataset used in feature engineering to find chemicals which leads to opioid addiction)** (https://www.cdc.gov/drugoverdose/resources/data.html - “Center of Disease Control and Prevention)

   This dataset contains the list of chemicals and their generic drug name which are classified as opiates. This information is useful for us to find features in *Dataset 1* leading to opiate addiction.

Attribute	Description	Data Type	Attribute Type
Drug Name	Name of the drug classified as opioid	String	Categorical
Generic Name	Generic Name of the drug	String	Categorical

3. **Data on Overdose Deaths (Used in Descriptive Analysis):**
(https://www.data.gov/ - *"Accidental and Drug Related Deaths"*)

This dataset contains a listing of accidental death associated with the drug overdose in all states in USA. This information was useful in comparison with *Dataset 1* in descriptive analysis.

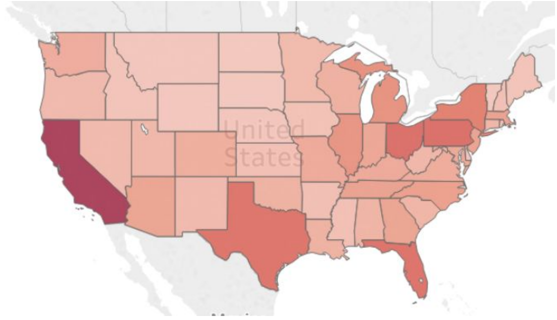
Attribute	**Description**	**Data Type**	**Attribute Type**
State	Name of the State in US	String	Categorical
Population	Population of the state	String	Quantitative
Deaths	Accidental number of deaths associated with the drug overdose	String	Quantitative
Abbrev	Abbreviation of the State	String	Categorical

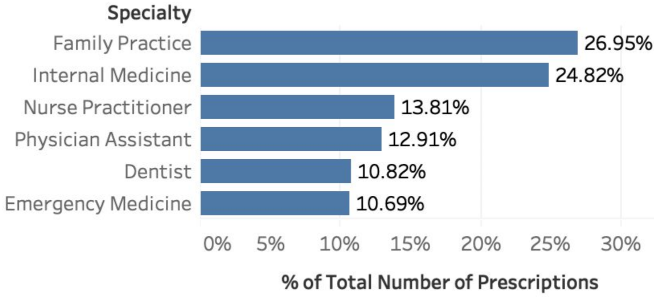
Data Preprocessing

Performed Data Cleaning:

- Missing Values found only in *Credentials* Column from *Drug Prescriber Data* - Replaced Missing Values with String Value. (Eg. NaN value replaced to 'UNKNOWN' in Credentials column).
- Mismatch in the number of states in *Drug Prescriber and Overdose Data* - Cleaned up States from *Prescriber Data* to match the list of states in *Overdose Death Data.*
  (Eg. States with abbreviations *AE, ZZ, AA, PR, GU, VI* from Overdose Data for further comparison with Prescription Data)
- Quantitative Value stored as a String in *Overdose Data* - Removed commas from String values to convert it into numerical type. (Eg. 4,833,722 replaced to 4833722 in Population and Deaths column).
- Remove Spacing and Special Characters from Drug Name and Replace it with '.' in *Drug List Data* for comparison with *Prescription Data*.

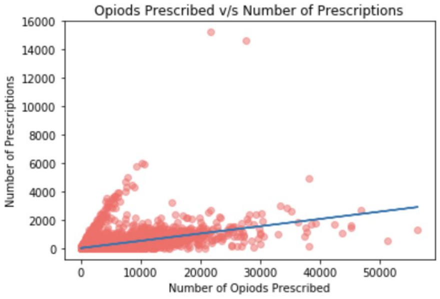
Data Exploration

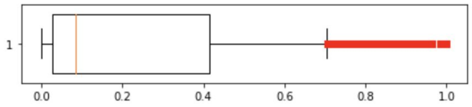


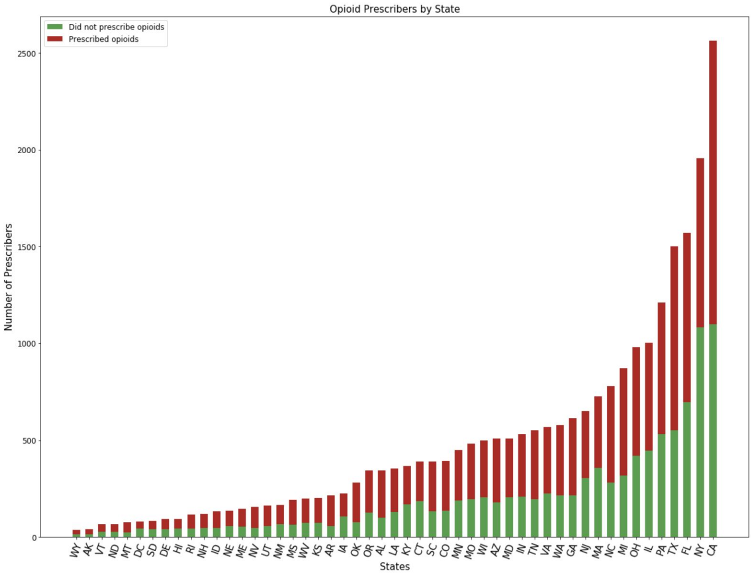


Insights:

- Most number of deaths due to opioid overdose are in California and Ohio.
- Use of opioids is higher in specialties which involve the use of Painkillers/Inhibitors, based on the top specialities.





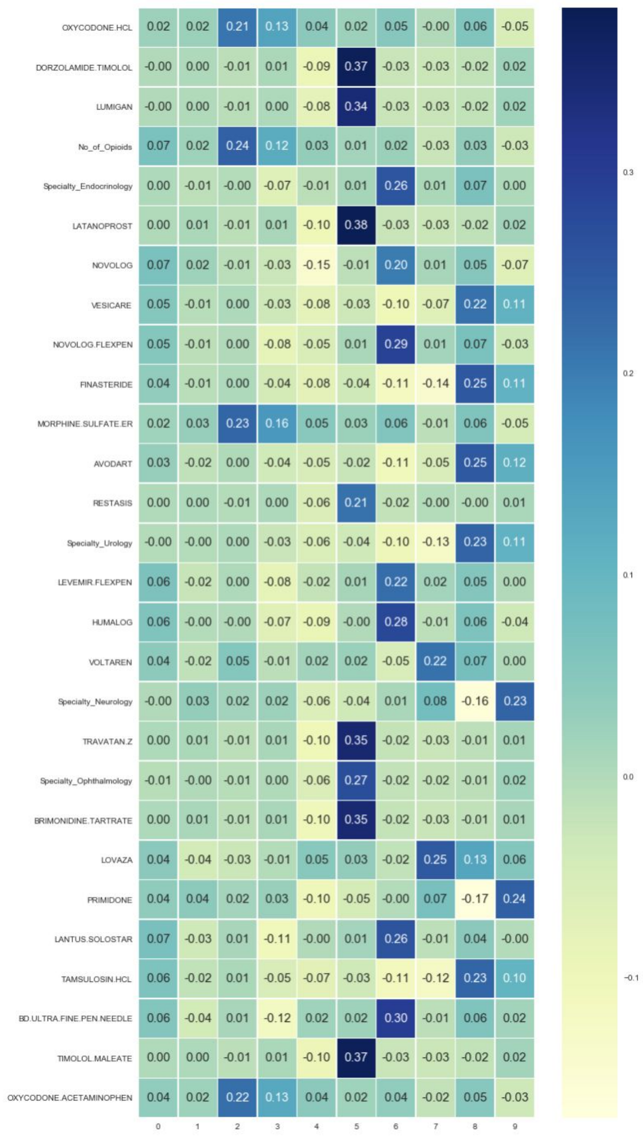


Insights:

- If a prescriber increase the number of prescriptions, it is likely to increase the opioid prescription by 10% (Box Plot Value).
- There are 11 opioid drugs out of the 250 drugs mentioned in the prescribers data. (Refer Notebook Code).
- 60% of the prescribers on this list are opioid prescribers.
- Prescribed Opioids is higher in the case of Male.
- States like CA, NY, FL, TX have higher opioid prescribers which corresponds with high death rates reported due to opioid overdose.

Feature Extraction with Principal Component Analysis

We performed PCA to identify latent factors and co-relation between opium components and speciality of medicare domain.



	0	1	2	3	4	5	6	7	8	9
OXYCODONE.HCL	0.02	0.02	0.21	0.13	0.04	0.02	0.05	-0.00	0.06	-0.05
DORZOLAMIDE.TIMOLOL	-0.00	0.00	-0.01	0.01	-0.09	0.37	-0.03	-0.03	-0.02	0.02
LUMIGAN	-0.00	0.00	-0.01	0.00	-0.08	0.34	-0.03	-0.03	-0.02	0.02
No_of_Opioids	0.07	0.02	0.24	0.12	0.03	0.01	0.02	-0.03	0.03	-0.03
Specialty_Endocrinology	0.00	-0.01	-0.00	-0.07	-0.01	0.01	0.26	0.01	0.07	0.00
LATANOPROST	0.00	0.01	-0.01	0.01	-0.10	0.38	-0.03	-0.03	-0.02	0.02
NOVOLOG	0.07	0.02	-0.01	-0.03	-0.15	-0.01	0.20	0.01	0.05	-0.07
VESICARE	0.05	-0.01	0.00	-0.03	-0.08	-0.03	-0.10	-0.07	0.22	0.11
NOVOLOG.FLEXPEN	0.05	-0.01	0.00	-0.08	-0.05	0.01	0.29	0.01	0.07	-0.03
FINASTERIDE	0.04	-0.01	0.00	-0.04	-0.08	-0.04	-0.11	-0.14	0.25	0.11
MORPHINE.SULFATE.ER	0.02	0.03	0.23	0.16	0.05	0.03	0.06	-0.01	0.06	-0.05
AVODART	0.03	-0.02	0.00	-0.04	-0.05	-0.02	-0.11	-0.05	0.25	0.12
RESTASIS	0.00	0.00	-0.01	0.00	-0.06	0.21	-0.02	-0.00	-0.00	0.01
Specialty_Urology	-0.00	-0.00	0.00	-0.03	-0.06	-0.04	-0.10	-0.13	0.23	0.11
LEVEMIR.FLEXPEN	0.06	-0.02	0.00	-0.08	-0.02	0.01	0.22	0.02	0.05	0.00
HUMALOG	0.06	-0.00	-0.00	-0.07	-0.09	-0.00	0.28	-0.01	0.06	-0.04
VOLTAREN	0.04	-0.02	0.05	-0.01	0.02	0.02	-0.05	0.22	0.07	0.00
Specialty_Neurology	-0.00	0.03	0.02	0.02	-0.06	-0.04	0.01	0.08	-0.16	0.23
TRAVATAN.Z	0.00	0.01	-0.01	0.01	-0.10	0.35	-0.02	-0.03	-0.01	0.01
Specialty_Ophthalmology	-0.01	-0.00	-0.01	0.00	-0.06	0.27	-0.02	-0.02	-0.01	0.02
BRIMONIDINE.TARTRATE	0.00	0.01	-0.01	0.01	-0.10	0.35	-0.02	-0.03	-0.01	0.01
LOVAZA	0.04	-0.04	-0.03	-0.01	0.05	0.03	-0.02	0.25	0.13	0.06
PRIMIDONE	0.04	0.04	0.02	0.03	-0.10	-0.05	-0.00	0.07	-0.17	0.24
LANTUS.SOLOSTAR	0.07	-0.03	0.01	-0.11	-0.00	0.01	0.26	-0.01	0.04	-0.00
TAMSULOSIN.HCL	0.06	-0.02	0.01	-0.05	-0.07	-0.03	-0.11	-0.12	0.23	0.10
BD.ULTRA.FINE.PEN.NEEDLE	0.06	-0.04	0.01	-0.12	0.02	0.02	0.30	-0.01	0.06	0.02
TIMOLOL.MALEATE	0.00	0.00	-0.01	0.01	-0.10	0.37	-0.03	-0.03	-0.02	0.02
OXYCODONE.ACETAMINOPHEN	0.04	0.02	0.22	0.13	0.04	0.02	0.04	-0.02	0.05	-0.03

Insights from Correlation Matrix [1]:

Factor 5: Related to eye diseases

- Latanoprost (0.38)
- Timolol.Maleate (0.37)
- Travatan.Z (0.35)
- Brimonidine.Tartrate (0.35)
- Lumigan (0.34)

Factor 2: Related to general pain

- Morphine.Sulphate (0.23)
- Oxycodone.Acetaminophen (0.22)
- Oxycodone.HCL (0.21)

Factor 6: Related to Type 1 and 2 Diabetes Patient

- BD.Ultra.Fine.Pen.Needle (0.30)
- Levemir.Flexpen (0.22)
- Lantus Solostar(0.26)

Negative Correlations (Non-Opiate): Related to Prostate Enlargement

- Finasteride (-0.14)
- Avodart (-0.11)

Factor 8 and 9: Related to Seizure and Epilepsy

- Primidone(0.24)

From the research of above factors (2, 5, 6), we found that people are undergoing treatment for Glaucoma, Diabetes and related diseases have a high risk of being prescribed treatment that contains Opioids. These patients are likely to prescribe Opioids which can lead to addiction and overdose.

While in the case of enlarged prostate treatment and epilepsy/seizures (Factors 8 & 9), and with other endocrine disorders, the Prescriber would avoid prescribing opioids.

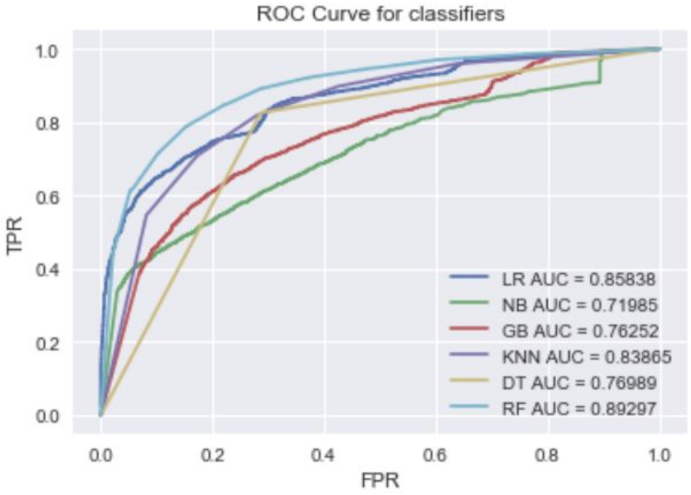
Feature Engineering

- Identified features for opioid detection and eliminated the rest.
- Removed Credentials and NPI Column in order to trim our features down.
- Performed Factorising with label encoding: Converted all categorical columns to numeric columns so that we can run classifiers like KNN and Logistic Regression which only works on numerical data.
- Split the dataset in 80:20 for data modeling.

Data Modeling

Fit the model to predict Opioid.Prescriber and perform basic initial evaluations:

- Logistic Regression( )
- Gaussian Naive Bayes( )
- Bernoulli Naive Bayes( )
- KNN Classifier( )
- Decision Tree( )
- Random Forest( )



Observations:

- The LR, KNN and Random Forest models are better at ranking the test set than the other models by comparing their AUC scores.

Model Selection (Cross-Validation & Hyperparameter Tuning)

- We attempted to improve our best 4 classifiers using cross-validation & hyperparameter tuning using GridSearchCV( ) and RandomSearchCV( ) to get the best parameters to train our model.
- For each iteration, train and test indices were generated, a model was trained on the generated train and test sets for every set of hyper-parameter and then the mean of all folds was recorded for each setting of the hyper-parameters.
- By doing this, we gained information about the effects of different hyper-parameters and also helped in generating generalized and unbiased information about the model.

Using High-Performance Cluster (HPC) at NYU:

We use the NYU HPC to re-run our models, we used 10 CPU cores simultaneously by enabling parallelism and used Grid/Random search with cross validation for 100 iterations which took about 3 hours to fully run (17 hours without parallelism, results of both were almost similar).

Best Parameter Settings:

Model	**Parameters**
Logistic Regression	C = 0.0001, solver = 'newton-cg', max_iter = 200
KNN Classifier	weights = 'distance', n_neighbors = 7, n_jobs = -1, leaf_size = 2, algorithm = 'ball_tree'
Decision Tree	criterion = 'entropy', max_depth = 90, max_features = 8, min_samples_leaf = 5, min_samples_split = 5
Random Forest	bootstrap = False, max_depth = 80, max_features = 'auto', min_samples_leaf = 1, min_samples_split = 10, n_estimators = 1000

Model Performance after hyperparameter tuning (Accuracy Metrics) using HPC:

Model	**Accuracy (Default Parameters)**	**Accuracy (Hyperparameter Tuning)**	**Results**
Logistic Regression	0.7608	0.7651	**Improved (~+0.43)**
KNN Classifier	0.7785	0.7815	**Improved (~+0.3)**
Decision Tree	0.7799	0.7789	**Not Improved**
Random Forest	0.8185	0.8375	**Improved (~+2)**

Model Performance with hyperparameter tuning and 10-fold Cross Validation (AUC):

Model	**Mean AUC**	**Max AUC**
Logistic Regression	0.8528	0.8575
KNN Classifier	0.8338	0.8409
Decision Tree	0.8464	0.8530
Random Forest	0.9112	0.9158

From hyperparameter tuning, we saw a significant improvement in Random Forest (+~2%). After applying the hyper-parameters to calculate AUC Scores of each model, based on accuracy and AUC Metrics, we found that Random Forest does well with 0.91 mean AUC Score and 0.84 Accuracy when considering without Gender column. Therefore, we choose Random Forest for Model Evaluation.

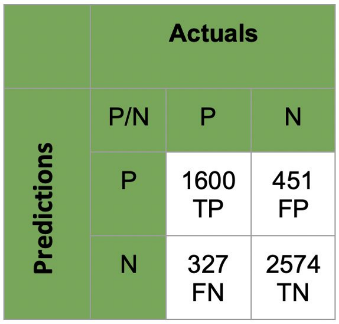
Model Evaluation

Model evaluation was done using Accuracy, Precision-Recall Curve Metrics.

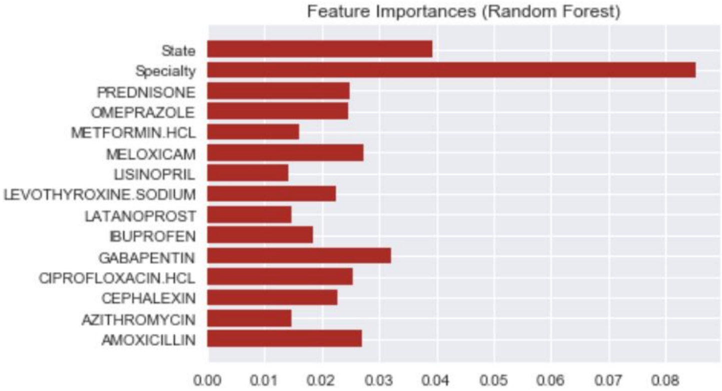
Out Sample Accuracy of Random Forest = 0.842891760904685

Average Precision-Recall Score = 0.82

Confusion Matrix and Feature Importance Matrix:



		Actuals	
Predictions	P/N	P	N
	P	1600 TP	451 FP
	N	327 FN	2574 TN



Results

- After experimenting with different classifier using hyperparameter tuning, we found that Random Forest is performing well. Therefore, we ran our predictions using Random Forest with significant good number of True Positives and True Negatives.
- We also plotted a Precision-Recall metric to evaluate classifier output quality. From the evaluation, we found a good precision score of 0.82.

Conclusion and Inferences

We believe that by improving drug prescription practices we can help reduce prescriptions that contain opiates. This in turn significantly reduces the chances of the drug overdose. In our analyses, we found that some of the drugs that are prescribed are without the mention to the patient that the drugs have opium components. So we realized there is no awareness in the community regarding opium components in their prescription.

In particular, when we performed Principal Component Analysis of our features, we took the features that were important weighted latent factors and did research on each of them to see where they are usually prescribed. After doing some research on strongly correlated components, we noticed that people undergoing treatment for Glaucoma, Diabetes and related diseases have a high risk of being prescribed treatment that contains Opioids. Such cases are likely to increase the probability of Prescriber to prescribe Opioids. This in turn can lead to addiction and overdose. While in the case of epilepsy/seizures, and enlarged prostate treatment with other endocrine disorders, the Prescriber would avoid prescribing opioids.

As there were only 11 opioids components out of 250 in the Prescriber data (from the descriptive analysis), our model justifies its accuracy by predicting a good number of True Negatives than True Positives. We noticed that since we predict more true negatives than true positives, most of our features in the top 15 feature importance were drug components that are non-opiates. There are important opiate components too and these correspond with our results from the Principal Components Analysis.

We would like to see this predictor and analysis to be used in the future in tools that will assist prescribers in prescribing different medications and better understand the probabilities of addiction in their prescriptions, the drug components correlated with opiates, the significant factors, etc. All this can go a long way in improving awareness, practices and increased vigilance during drug prescriptions.

Assumptions

- Prescribers of the same specialty in different states are under different local state laws that enforce prescription of certain drugs. This assumption is safe to make because ultimately we are predicting the target variable by considering all the features of the dataset anyway.
- There are trends within prescribers of the same specialty. The reason for this assumption is that doctors of a certain specialty are assumed to be prescribing similar drugs to patients. This assumption is okay because we are trying to predict if the doctor is more likely to have prescribed an opioid and this assumption supports the theory with evidence.
- The State Wise comparison in Overdose Data and Prescribers data is normalized and furthermore, we have compared the data considering percentage instead of numeric comparison.

Changes/Updates from Original Proposal

- In the initial proposal we discussed more problems such as predicting instances of drug overdoses, and we wished to get more data for this. However, our data did not have many essential features for this such the quantities of drugs that directly lead to death from overdose cases. This, coupled with the fact there is not enough descriptive and specific data on drug overdose deaths in general, made us focus on concrete topics as per instructor’s comments.
- We also initially wished to find important features from the prediction models, but later we decided that we would use Principal component analysis in addition to that to find the important latent factors, as we wanted to compare the feature importances from those two techniques.