Foundations of Data Science
Project Proposal

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

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 in prescription data, overdose deaths data with respect to different regions.

We will be using “*Causes of Death by Opioid Consumption”* dataset to answer the following problem by considering educational, demographics and geographical factors:

1. Identifying the areas affected by Opioid Consumption?
2. To predict what chemicals when used in significant quantities creates an opioid drug?
3. To use predictive analysis to predict the likelihood that a given doctor is a significant prescriber of opiates. The motive behind this is that a systematic way of identifying sources may reveal trends in particular practices, fields, or regions of the country that can be used to combat this problem effectively.
4. To predict instances of drug overdoses and prevent them from occurring?

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.

Publications:

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

Thus we will continue this research into articles such as these and get important datasets from relevant websites such as data.gov The Offices of the NYC Chief Medical Officer and NYU Langone Hospital may provide us more insight into drug overdoses and the prescription procedures so that we may better understand the process of prescription to overdose.

Adding more on this, we have a contact of *“The United States Pharmacopeia”* which is an industry in the US working on standardizing medical products all over the world. They have a compendium of drug information which will be helpful for us for fetching interesting insights from the dataset.

Data Source Loading

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

1. **Information on drug prescriber:**
   (https://www.cms.gov/ :*"Medicare Provider Utilization and Payment Data: Part D Prescriber")*

   This information contains key insights into different drug prescribers in order to create a predictor to predict the likelihood of a given doctor being a significant prescriber of opiates.

   This data has the following features:

Attribute	**Description**	**Data Type**	**Attribute Type**
NPI	Unique National Provider Identifier Number	Numerical	Categorical
Gender	Gender of the Subject (M/F)	Char	Categorical
State	U.S. State by abbreviation	Char	Categorical
Credentials	Set of initials indicative of medical degree	Char	Categorical
Specialty	Description of type of medicinal practice	Char	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. **Data on Overdose Deaths:**
   (https://www.data.gov/ - *"Accidental and Drug Related Deaths"*)

   This dataset contains a listing of each accidental death associated with the drug overdose in Connecticut from 2012 to June 2017. Data are derived from an investigation by the Office of the Chief Medical Examiner which includes the toxicity report, death certificate, as well as a scene investigation.

The features in this data include:

Attribute	Description	Data Type	Attribute Type
CaseNumber	Accidental Death Unique Case Number	Numerical	Categorical
Date of Death	Registered date of death	Date	Categorical
Sex	Gender of the subject	Char	Categorical
Race	Race of the subject	Char	Categorical
Age	Age of the subject	Char	Categorical
Residence (City, State, County)	Residential Details of the subject	Char	Categorical
Death Location (City, State, County)	Location of caused death of the subject	Char	Categorical
Immediate Cause	Cause of Death of the subject	Char	Categorical
List of Drugs/Chemicals caused the death	Indicator of drugs in cause of death (For example if the Feature Cocaine was found in the drug, then it will be indicated by a “1” value.)	Binary	Categorical
MannerOfDeath	Indicator whether death was accidental or not	Char	Categorical

Overall, the kind of data we are targeting is data that tells us about the kind of Opiates/drugs that are prescribed by doctors to patients and the kind that are detected at the time of accidental overdose.

Additionally, we are also targeting historical data of Overdose deaths to try and determine trends over time.

Proposed Methodology

1. **Data Transformation - Merging the data**
    - Before working on the dataset, the collected data should be in a single format.
    - Joining tables.

2. **Data Cleaning**
    - Eliminate unnecessary, irrelevant and duplicate data records.
    - Clean the data by adjusting the missing values (Techniques like Gaussian Distribution, Taking Mean if data is numeric can be used).
    - Find errors (potential spelling errors) in the data and fix them.

3. **Data Exploration**
    - Data Summary and Exploratory analysis for finding main characteristics of the opioid dataset.
    - Plot graphs to explore the data to help us understand the data better and draw additional inferences.
    - Find correlations between different features and across datasets if possible.

4. **Feature Engineering**
    - Creating new features from the existing data source for modeling the data.
    - We list the feature importance, and find correlations between the most important features.
    - We can also list feature importance based on specific data points such as the Speciality of the doctor as seen in the first dataset. For example, if the doctor is a surgeon, then it is possible that the features (drugs) that are prescribed prominently (more important) are the ones used for pain relief. This can help us understand the prescribers better.

5. **Cross Fold Validation and Data Modelling**
    - Build a classifier (Decision tree or Gradient boosted tree) to fit the data and predict the problem statements. This is because there may be features that are correlated with each other, and a tree model like the above can easily handle that, while regression cannot do it as well. Gradient Boosted trees can work with multi-collinear features without being disrupted.
    - Getting the importance matrix with top predicted feature.
    - Performing 5-fold cross-validation to estimate the skill of the model on new data.

6. **Performance Evaluation, and Model Selection**
    - A feature importance matrix and accuracy score will play a major role in performance evaluation and model selection.
    - Performance Evaluation can be done by comparing the training dataset with the test data and calculate the accuracy.

We will evaluate our approach by using the tables such as:

	Technique1	**Technique 2**	**Technique3**
Parameters: x, y min_samples,min_leaves	Prediction Accuracy	Prediction Accuracy	Prediction Accuracy
Parameters: x1, y1	Prediction Accuracy	Prediction Accuracy	Prediction Accuracy

And we go on adding various rows and columns to this.

Assumptions

The following are the assumptions made while defining the problem:

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 a significant amount and this assumption will later support the theory with evidence. So if our assumption is false, then we will get to know about it in the feature importance table, and then we will be better equipped to decide whether to consider it for predictions or not.

We are also assuming that 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. So if our assumption is false, we will get to know about it in the feature importance table. If our assumption is true, we will see that in the feature importance table as well. That will surely benefit us as we want to find trends with respect to the region anyway.