A Predictive Modeling and Analysis on Causes of Death by Opioid Prescription

Prof. Rumi Chunara
CS6053 - Foundations of Data Science (Fall 2018)

Presentation Compiled by: Akshay Tambe, Aditya Bhatt
Net ID: apt321, apb462

Data Science for Social Good

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 higher rate. Overdose of Opioids leads to death.

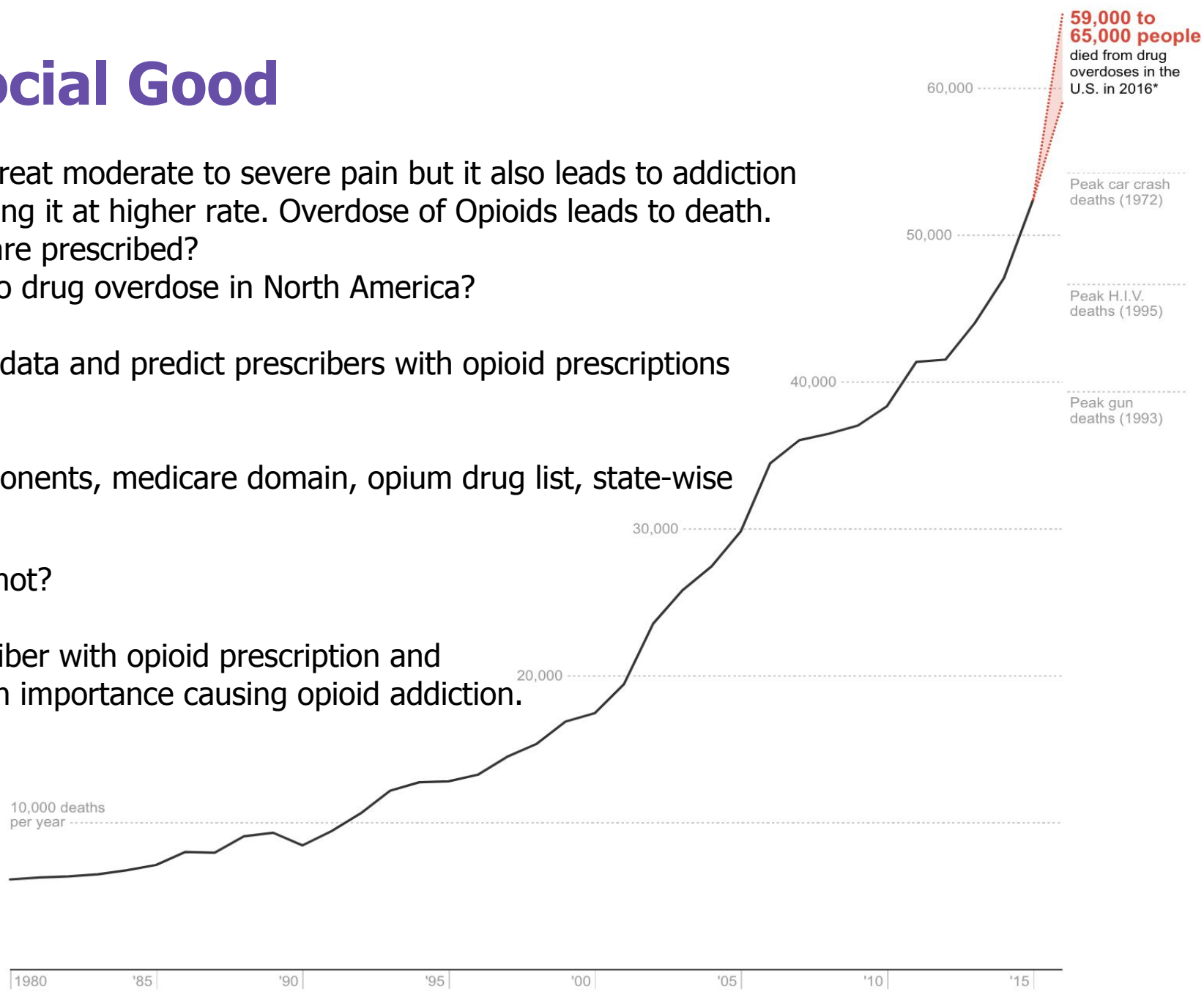
1. How to improve ways the opioids are prescribed?
2. How to reduce the death toll due to drug overdose in North America?

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?

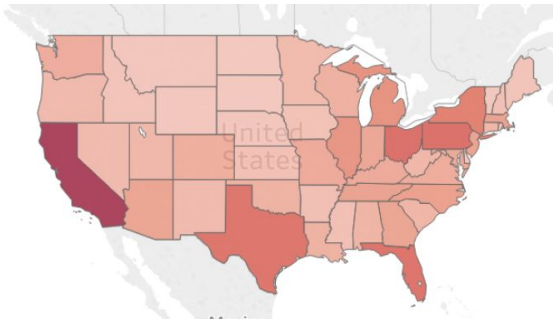
Impact: Model is able to predict prescriber with opioid prescription and individual chemicals and features of high importance causing opioid addiction.

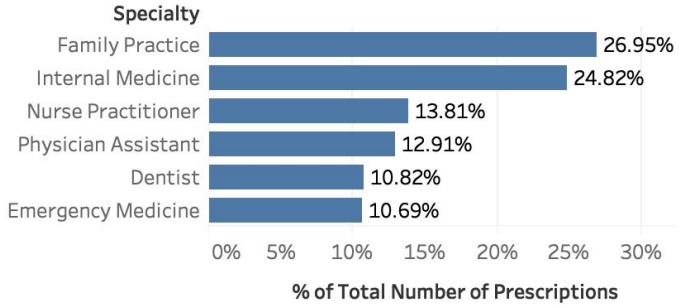


Approach So Far ...

Data Collection: Initialized analysis on 3 Datasets.

Data Preprocessing: Replacing Missing Values with String Value, Clean up states, Removing Commas in Numerical Values, Removing Spacing and Special Characters.

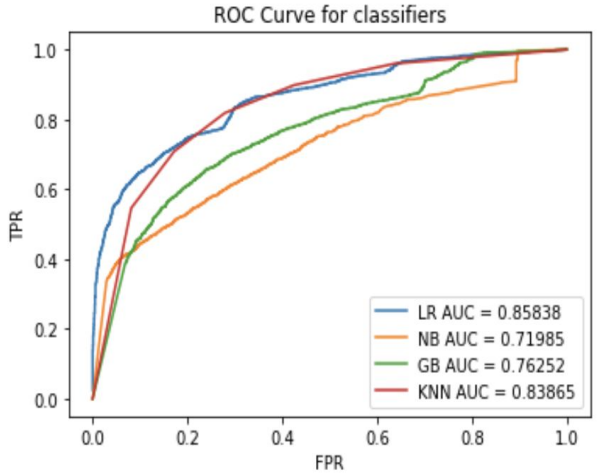




Data Exploration:

- Most no. of deaths due to opioid overdose is in California and Ohio.
- There are 11 opioid drugs out of the 250 drugs mentioned in the prescribers data.
- Prescribed Opioids is higher in the case of Male.
- Use of opioids is higher in specialties which involve the use of Painkillers/Inhibitors, based on the top specialities.
- States like CA, NY, FL, TX have higher opioid prescribers which corresponds with high death rates reported due to opioid overdose.

Feature Engineering: Identifying Features for opioid detection and eliminating rest, Factorizing: converting categorical columns to numerical.



Model Selection with Cross Fold Validation:

Using 10-Fold Cross Validation, KNN model does well with 0.82 mean AUC Score.

Model Evaluation and Results

Avg. Precision-Recall Score: 0.77
Accuracy from Confusion Matrix: 0.7780

We can conclude that, so far KNN produces better results in predicting target variable. There is no significant jump in accuracy after Cross Validation.

Work in Progress

- Performing data modeling using Decision Tree and Random Forest.
- Performing Dimensionality Reduction (PCA), finding feature importance matrix to detect trends, correlations between overdose deaths, opium components and speciality of medicare domain.

Data Modeling:

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

The LR and KNN model is better at ranking the test set than the other models by comparing their AUC scores.

Model	Mean AUC	Max AUC
KNN	0.8264	0.8331
Gaussian NB	0.7480	0.7568
Bernoulli NB	0.7059	0.7158
LR	0.6879	0.6933

T/F	T	F
T	1481	570
F	529	2372