Lauren To
DATA1030

DATA1030 Final Proposal

1. What is the target variable?

The target variable is the successful completion of a drug treatment program of those with opioid dependence using program discharge data from the Treatment Episode Data Set 2017, funded by SAMHSA. Several reason codes are offered in the dataset, so I will be coding all observations with the reason "Treatment completed" as 1, and all other reasons as 0.

2. Is the problem regression or classification?

The problem is mainly regressive in nature, as I am identifying what individual characteristics or behaviors are predictive of successful program completion.

3. Why is this interesting/important?

It is well known that the opioid epidemic has become a major public health issue in the past 5 years, and this project could be important in identifying influential factors that determine whether or not an individual enters substance use disorder (SUD) treatment, and whether or not they finish it satisfactorily.

4. Number of data points and number of features.

The TEDS is collected and distributed every year, with 1-2 million data points each year comfortably. However, since I will be focusing on those with opioid dependence, my dataset is around 300000 points. If necessary, I can append data points from previous years. Before pre-processing, there are 22 features that include demographic information, nature of primary/secondary/tertiary drug use, source of referral treatment, number of prior treatment episodes, reason for discharge, length of stay, type of service provided, and several flags for more specific drug usage.

5. If dataset is from Kaggle/UCI/already described, write a short description about 2-3 public projects where the data has been used, and how the features were used.

This dataset is rather well-documented. The documentation for each variable and for methodologies can be found here: http://samhda.s3-us-gov-west-1.amazonaws.com/s3fs-public/field-uploads-protected/studies/TEDS-D-2017/TEDS-D-2017-datasets/TEDS-D-2017-DS0001/TEDS-D-2017-DS0001-info/TEDS-D-2017-DS0001-info-codebook.pdf

TEDS-D 2006-2011 were used in a 2017 paper by Acion et al. that uses machine learning to identify factors predictive of treatment completion specifically for Hispanic populations. The author of this paper used more stringent exclusion criteria than I am planning to (she dropped all observations that had any missing values, whereas I plan to code for a worst case scenario), and was limited to a much smaller dataset due to the ethnicity specification. She used 17 different models of machine learning, dividing her data (n ~ 20000) in 80% training and 20% testing, and compared their performances. Interestingly enough, not very much is discussed

regarding what the predictive factors themselves are, just the strengths and weaknesses of the models used. My project also differs in that I will be using more recent datasets (they are coded differently after 2011), which makes the pre-processing steps quite different.

Another paper by Mennis and Stahler, 2016 uses the TEDS-D to measure racial and ethnic disparities among patients based on drug type and treatment service provided. They use fixed effect logistic regression to test associations between race/ethnicity and treatment completion while controlling for different socioeconomic and environmental factors. There are also other papers that evaluate racial disparities regarding referral source and criminal justice mandates and psychiatric comorbidities.

6. Preprocess the dataset.

The dataset was not downloadable as a CSV, so I read it in as a Stata file with pandas. I started with a condition statement selecting only the data points where the DSM diagnosis was opioid dependence. I then dropped many variables that I was less interested in including in my models, as well as ones that had a significant number of values missing, and in which it did not make sense to recode with an average value. It should be noted that the variables with many missing values did not seem as essential to analysis, and they were generally bits of information that were more specific, such as veteran status or reason for not being in the labor force. I do not have many concerns, therefore, about these values severely affecting analysis. After these variables were removed, I then dropped rows that had missing values because I did not want to drop variables that I hypothesize to be important to my analysis. I recoded all unknown education values to be the average of the dataset. I also changed some binary variables coded as (1, 2) to (0, 1). I had to hard code the target variable to reflect (0, 1) labels, as multiple values needed to be changed to 0 while only one needed to be 1. I imputed education and employment using an IterativeImputer.

All of the variables in this dataset were already ordinally or label encoded. I left those that are ordinally encoded as is; these include: education, length of stay, frequency of primary substance usage, arrests, and age of first primary drug use. These variables represent measurements of time, or easily ordered metrics, so it made sense for them to remain ordinal. The exact codes of each label can be found in the codebook given above. I recoded most of the label encoded variables to be one-hot encoded. This includes: services, referral sources, gender, race, ethnicity (Hispanic marker), living arrangement, primary/secondary/tertiary drugs, route of primary drug usage, census region, marriage status, employment status, and an alcohol/drug combination indicator. It was not intuitive to leave numerical labels, as the assignments were rather arbitrary, so I made new columns for each category of these variables. I did, however, collapse the drugs into 8 major categories instead of the original 19. There are indicators for each category for all three levels of drug usage. The variables psychiatric problems, methadone usage, and pregnancy status were label encoded, as they are binary in nature and can be labelled simply with a 0 or 1.

I had no continuous variables in my dataset, so no scaling was necessary. After pre-processing, I have 78 columns of indicators and variables, with 125918 data points.