Group 6 Final Project Planning/Proposal ~First Meeting~

“Drug Courting”

* Drug court participants, predicting what characteristics make for a candidate likely to graduate from a drug court program. What makes them likely to complete the program?
  + Graduated--1 (cleaned themselves up, avoided jail)
  + Terminated--0 (drug test failure, other arrests, etc)
* Multiple data points on each participants
* Matthew already has the data set from work in a SQL database
* Data is for the whole state of Georgia (some courts are specific to family/people with children, veterans, mental health, etc)
* Look at graduates within the last year, decide on elements we want to include in our analysis
* Binary analysis-- participants either graduate (1) or they don’t (0)

Statistical Analysis:

Descriptive statistics will be computed for all variables of interest. We will examine the distribution of all continuous variables to determine whether they were normally distributed and examine frequency counts of all categorical variables. Depending on the distribution of the data, techniques such as binning or combining categories will be discussed.

The data will be cleaned by examining relevant data points as well as items that will move forward in the final analysis. Variables with greater than 25% of missing data will be removed from the analysis, whereas mean, median, and mode imputation will be used for those variables with less than 25% missing data will. Will will examine models with and without imputation to determine whether imputation had an impact on our final estimates.

Data will be pre-processed by created dummy variables for categorical variables (.get\_dummies) and continuous variables will be normalized with StandardScaler.

We will examine several machine learning techniques to determine what method and model provides the best fit. Initially, we will use random forest for feature selection procedures. Our random forest models will be computed in an iterative process. We will examine the accuracy of difference models with differing numbers of features. We will determine a relevant stopping based on change in multiple indices (i.e. accuracy, F1, etc). In particular, a significant drop in accuracy and other scores will denote a stopping point.

Next….

Defalt values from some features

Flask app.

Machine learning library?

Other stuff we want to include?

Problem statement:

Application will be used by the court system to predict whether arrestee’s are good candidates for drug court. How determinate is the risk profile of the candidate as to whether or not a candidate will graduate from drug court? How determinate is the time frame between referral and acceptance to drug court in a candidate’s graduation status?

Which learning model produces the greatest accuracy?

Which dates are important? Amount of time between the arrest/referral date could be especially important.

Next steps:

Decide what we want to include in the model? Random forest?

Write proposal/submit

Pull/clean up data

**Data points:**

Acceptance date

Arrest date,

Referral date

Risk level

Acceptance type

Exit date

Exit status

Referral source

Demographic info: DOB, Education level(could change), Employment status(at entry), Gender, Income level, Employment stability, Military service, Race

Program type (is encoded in the ID)

Clinical Diagnosis and Level

Diagnosis Reason

Number of drug tests

Number of positive or negative drug tests

Employment assistance referral

Count weekly judicial status meetings

Primary drug of choice

Secondary drug of choice

Time between phases of the program (maybe?) number of days in each phase

Count number of sanctions maybe

Number of treatment sessions