TEDS Executive Summary

Peter Li

6/3/2019

https://peternu2020.github.io/datascience3/

The Treatment Episode Data Set (TEDS) consists of approximately 2.9 million substance abuse treatment records. The dataset is collected and administered by the Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration (SAMHSA) and also consists of data collected by states in the United States. The dataset consists of both treatment admissions and discharge records collected from 2015-2016.

The primary goal was to develop models to predict treatment completion, length of stay, and substance use frequency at treament discharge from the demographics of individuals and other factors such as the type of substance abuse and frequency of use before treatment, history of previous treatment, etc. Another constraint was that the models developed had to be fairly intrepretable. This constraint was self-imposed and arbitrary, but it allows for specific predictor variables to be identified and further evaluated. However, this constraint limits the use of black-box methods such as neural networks that may fit more accurate models at the cost of intrepretability.

All of my modeling methods used a training set for fitting and a test set for assessment of model accuracy. Furthermore, each model had parameters optimized through 10-fold cross validation on the training set with selection based on the lowest train error (missclassification) rate. All of the dependent variables being predicted were multi-class and not binary, thus, methods such as SVMs and binary logistic regression were not used.

The first methods used for predictive modeling were lasso and ridge logistic regression. The next methods used were ensemble tree methods. Random forests were used with the CV tuned number of randomly sampled features at each split. All of these numbers were less than the total number of predictors, thus, bootstrap aggregation was excluded. All boosting models used the CV tuned parameters of a 0.200 learning rate, depth of 2, and rounds of 100. Overall, random forests outperformed the other two modeling methods for all three dependent variables being predicted.

The dependent variable with the lowest test error rates across all three modeling methods was the primary substance usage frequency at treatment discharge. This variable had the least amount of possible classes out of all the dependent variables. The variable arguably provides as much information, if not more, on treatment efficiency as the treatment completion variable. A treatment being completed also does not guarantee an individual has reduced or stopped using substance(s). However, a caveat is that if there is no difference between the individual's primary substance use frequency before and after the treatment then the treatment efficiency should not be considered significant.

Predicting stay duration:

```
## # A tibble: 3 x 2
##   model                          test_error_rate
##   <chr>                                    <dbl>
## 1 lasso/ridge logistic regression          0.775
## 2 boosting                                 0.717
## 3 random_forest, mtry = 4                  0.69
```

Predicting treatment completion:

```
## # A tibble: 3 x 2
##   model                          test_error_rate
##   <chr>                                    <dbl>
## 1 lasso/ridge logistic regression           0.492
## 2 boosting                                  0.469
## 3 random_forest, mtry = 8                   0.41
```

Predicting primary substance usage frequency at treatment discharge:

```
## # A tibble: 3 x 2
##   model                           test_error_rate
##   <chr>                                     <dbl>
## 1 lasso/ridge logistic regression           0.353
## 2 boosting                                  0.288
## 3 random_forest, mtry = 8                   0.221
```

https://wwwdasis.samhsa.gov/dasis2/teds.htm https://wwwdasis.samhsa.gov/dasis2/teds_pubs/TEDS/Discharges/TED_D_2015/teds_d_2015_codebook.pdf https://wwwdasis.samhsa.gov/dasis2/teds_pubs/TEDS/Discharges/TEDS_D_2016/2016_teds_d_codebook.pdf

Substance Abuse and Mental Health Services Administration, Treatment Episode Data Set (TEDS): 2015. Rockville, MD: Substance Abuse and Mental Health Services Administration, 2018. Substance Abuse and Mental Health Services Administration, Treatment Episode Data Set (TEDS): 2016. Rockville, MD: Substance Abuse and Mental Health Services Administration, 2018.