DATA 1030 Final Report: Predicting Successful SUD Treatment Among Opioid-Dependent Individuals

Lauren To

Data Science, Brown University

November 2019

https://github.com/tolauren19/DATA1030-Final-Project

1 Introduction

The U.S. opioid epidemic of the late 1990's has dramatically worsened within the past decade, and a major challenge that remains for public health organizations is identifying not only the areas where they are most successful in treating opioid dependency, but also the areas in which they struggle the most. In an effort to identify salient influences, my project is a classification problem centered around predicting whether or not an individual with an opioid dependency is able to successfully complete a substance use disorder (SUD) treatment program.

I utilize a public data set from 2017 made available by the Substance Abuse and Mental Health Services Administration, titled the Treatment Episode Data Set - Discharges (TEDS-D-2017). The data set tracks admission and discharge cases for individuals ages 12 and older from SUD treatment centers across the nation. Among the features included are individual demographic information, as well as substance use characteristics. There have been previously published studies that address similar problems using the same data set(s).

2 EDA

TEDS-D-2017 has nearly 2 million data points; I restricted the data to opioid dependency outpatient cases, which returned about 188,000 observations. Of these, several rows had missing values due to the level of detail of some of the features. I deleted rows that had missing values for features that had less than 5% of values missing, or over 90% of values missing. I then ran an MCAR test and deleted features that returned a p-value higher than 5%. I finally used multivariate imputation for missing values on a random permutation of 100000 rows (for computational purposes).

All of the features in this data set were already encoded to some extent. All continuous features were collapsed into ordinal categories. There were some features, however, that were ordinally encoded in scenarios that were not very intuitive. I started by re-coding some of the labels so that some features would have fewer categories ('SUB1', 'SUB2', 'SUB3', 'SUB1_D', 'SUB2_D', 'SUB3_D'). Additionally, the default value for a "no" or "none" indicator was 2, which I changed to 0 ('ETHNIC', 'PREG', 'METHUSE', 'GENDER', 'VET', 'PSYPROB'). Finally, I one-hot encoded ordinal categories that should not have been ordinal ('REGION', 'PSOURCE', 'LIVARAG', 'LIVARAG_D', 'RACE', 'SUB1', 'SUB2', 'SUB3', 'SUB1_D', 'SUB2_D', 'SUB3_D', 'ALCDRUG', 'EMPLOY', 'ROUTE1', 'PRIMINC'). At the end of this process, I had a total of 115 features.

I re-coded my target variable 'REASON' to be 1 for successful completion and 0 for all other reasons. This gives a balance/baseline accuracy of 81.2%, with 0 being the most populous class.

I then used both f-score and mutual information classification to identify the most important features and created plots of the features that were returned by both lists. Here, I include an analysis of 4 features.

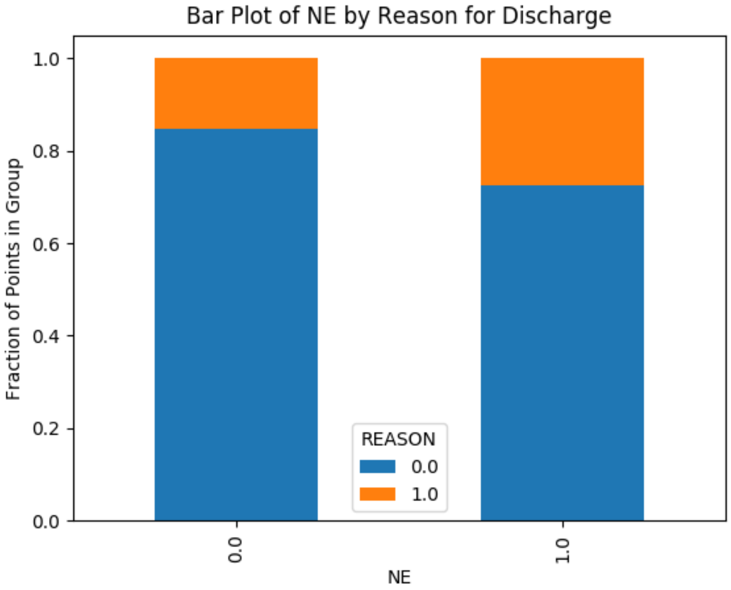

Figure 1: The NE variable is an indicator for the northeast region of the US. A higher proportion of the success group was from the NE area than the unsuccessful group, possibly signalling that those from NE have qualities or resources that facilitate successful treatment completion.

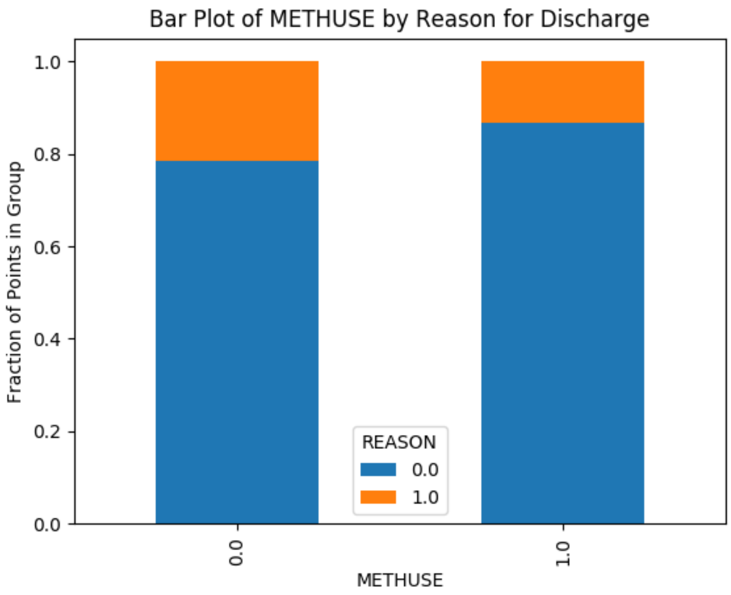

Figure 2: The METHUSE variable is an indicator of methadone use. Methadone is an opioid that is often used in treatment of narcotic drug addiction. One can see that the proportion of methadone users is higher among those who did not successfully complete treatment. This could be indicative of a confounding factor; those with more severe addictions could be more likely to use methadone and also struggle more with going through a treatment program. Alternatively, adhering to methadone treatment itself could be a barrier to program completion.

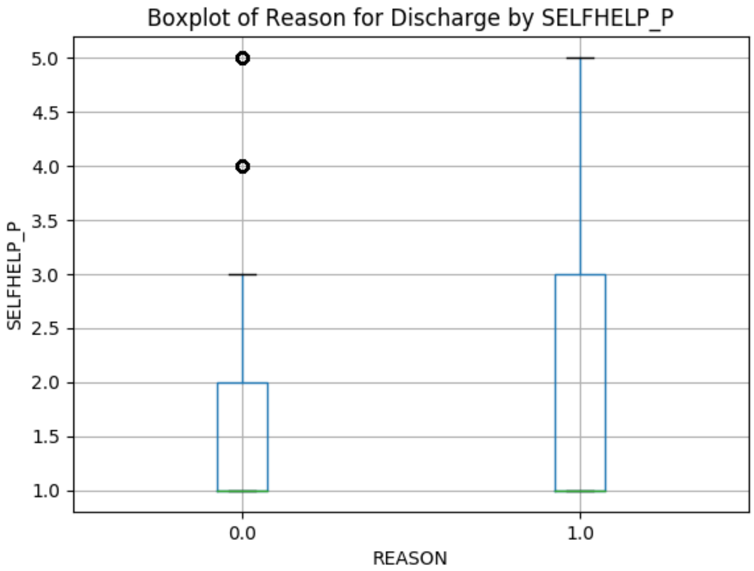

Figure 3: SELFHELP_P is a continuous variable representing the frequency with which an individual attended a self-help therapy group 30 days prior to admission (see codebook for categories). The means of both groups start at no attendance, although any attendance in general seemed to have already been more common among the successful group.

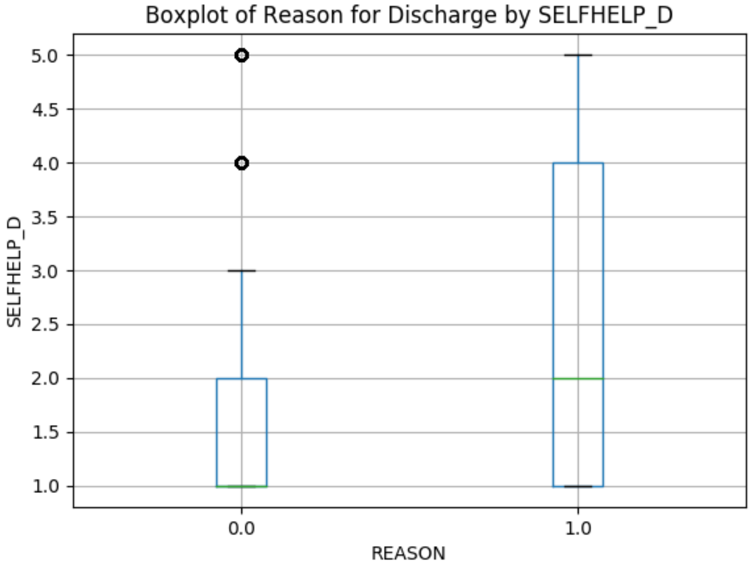

Figure 4: SELFHELP_D represents the frequency with which an individual attended a self-help therapy group 30 days prior to discharge. The mean attendance of the success group is much higher than that of the unsuccessful group; the latter's boxplot is virtually unchanged.

3 Methods

I used a stratified train-test split on the saved permutation of data (see `training_permutation.csv`), saving 20% for testing. The other 80% was used for stratified cross validation with 4 folds. I made sure to select distinct individuals from the beginning of the project to avoid data leakage. Since my data set is still rather large despite only using a sub-sample, I opted to not use an SVC, as those are not well-suited to large data sets. Instead, I used a random forest classifier and a logistic regression classifier. I used a consistent random state for all functions and methods that required it (e.g. `train_test_split`, `RandomForestClassifier`).

The permuted data were subjected to the same row/feature deletion and recoding that was done during pre-processing. For cross validation, I passed a SimpleImputer and OneHot encoder for categorical variables into the pipeline, and an IterativeImputer for the continuous/ordinal variables. I then added a Standard Scaler for all of the columns before the use of the actual estimator so that I could use feature importances and linear coefficients later. For the random forest classifier, I tuned the `max_depth` and `min_samples_split` parameters in ranges of 2-7 and 2-10, respectively. For the logistic classifier, I adjusted the `penalty` and `alpha` parameters, trying both l1 and l2 penalties, and using an alpha range from 1e-3 to 1e4.

Since this is a classification problem, I use accuracy score as my metric of performance.

I trained both models 5 times using 5 different random states to calculate mean test accuracy scores and standard deviations as a gauge of uncertainties that result from splitting.

4 Results

The results of both the logistic regression classifier and the random forest classifier are presented in the following table.

Both models do marginally better than the baseline score of 81.2%. The best test score for the logistic model is 83.7%, and 83.6% for the random forest. The best parameters for the logistic model

	Logistic Model	**Random Forest Model**
Baseline score	0.8123	0.8123
Best test score	0.837	0.836
Mean training score (5 runs)	0.8373	0.8322
Mean test score (5 runs)	0.8368	0.8328
Standard deviation (5 runs)	0.0006	0.0017
Best training parameters	Penalty = ‘l2’	Max depth = 7
	C = 0.01	Min samples split = 9

Table 1: Model Testing Results

are ridge penalization, and an alpha of 1e2. The best parameters for the random forest is a max depth of 7 (the deepest parameter I allow) and a minimum of 9 samples split. It should be noted that the standard deviation for the logistic model is extremely small. The standard deviation of the random forest classifier is larger, but still rather small; the mean test score falls about 15 standard deviations above the baseline. Overall, the logistic model performed consistently better than the random forest model in both training and testing, although the mean test accuracy score of the random forest surpasses its mean CV/training score.

To compare the models more specifically in terms of their performance, I plotted their confusion matrices.

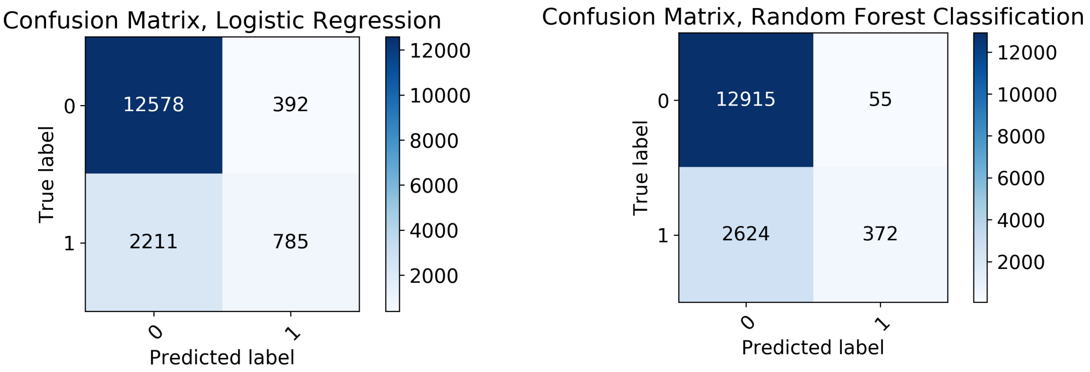

Figure 5: Overall, it appears that the logistic model was better at correctly identifying successful treatment, but classified more false positives. The random forest model was quite accurate in predicting 0’s, but it was also less sensitive in predicting 1’s.

I use a number of methods to measure global feature importance. Starting with feature permutations, I shuffled each feature one at a time using a reserved test set for both models to see how the test accuracy score would be affected. The permutation plots are shown below.

Shuffling incurred up to a 3% drop in accuracy for the logistic model, falling below the *baseline* score. Naturally, one would expect to see a strong correlation between substance use at time of discharge and reason for discharge, as SUB1_D is inherently indicative of whether or not an individual was considered to have successfully completed a treatment program or not.

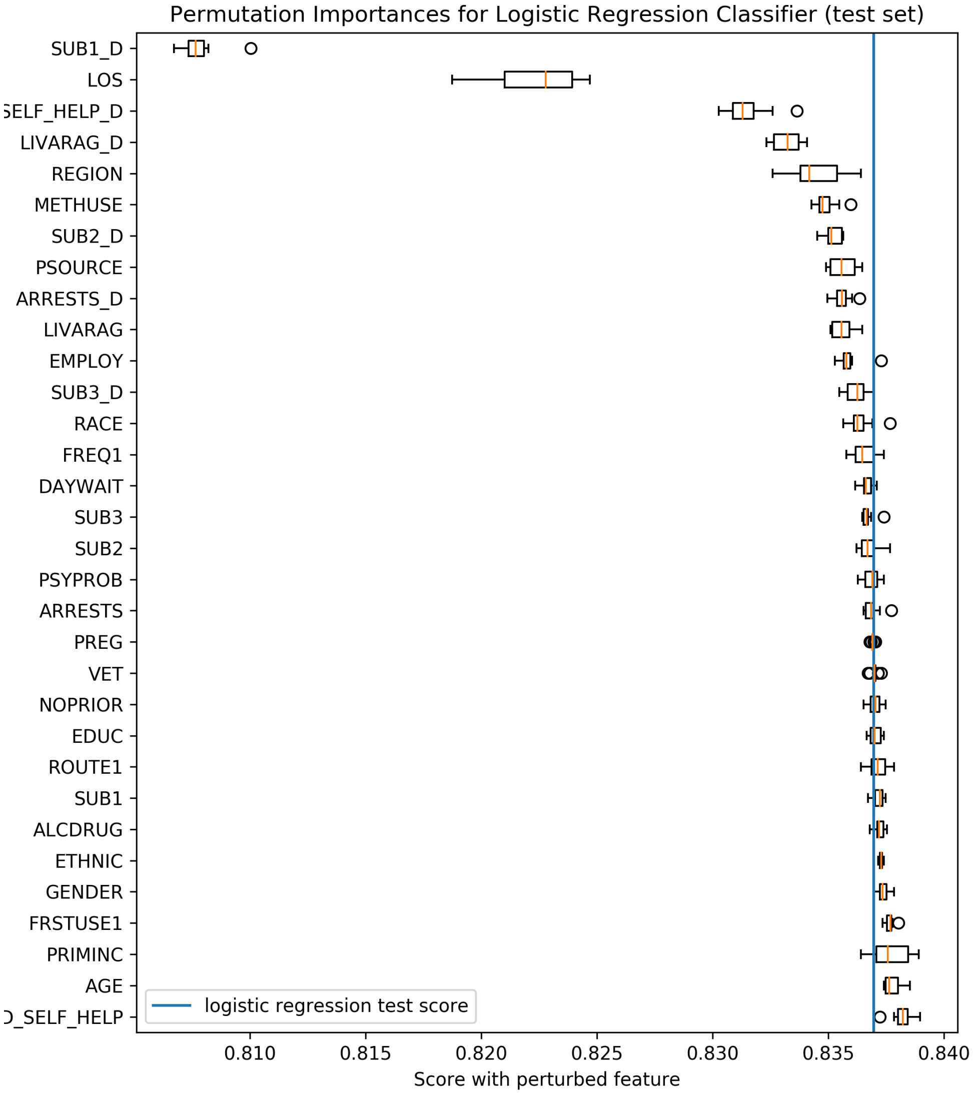

Figure 6: The three most influential features are SUB1_D (primary drug use at time of discharge), LOS (length of stay), and SELF_HELP_D.

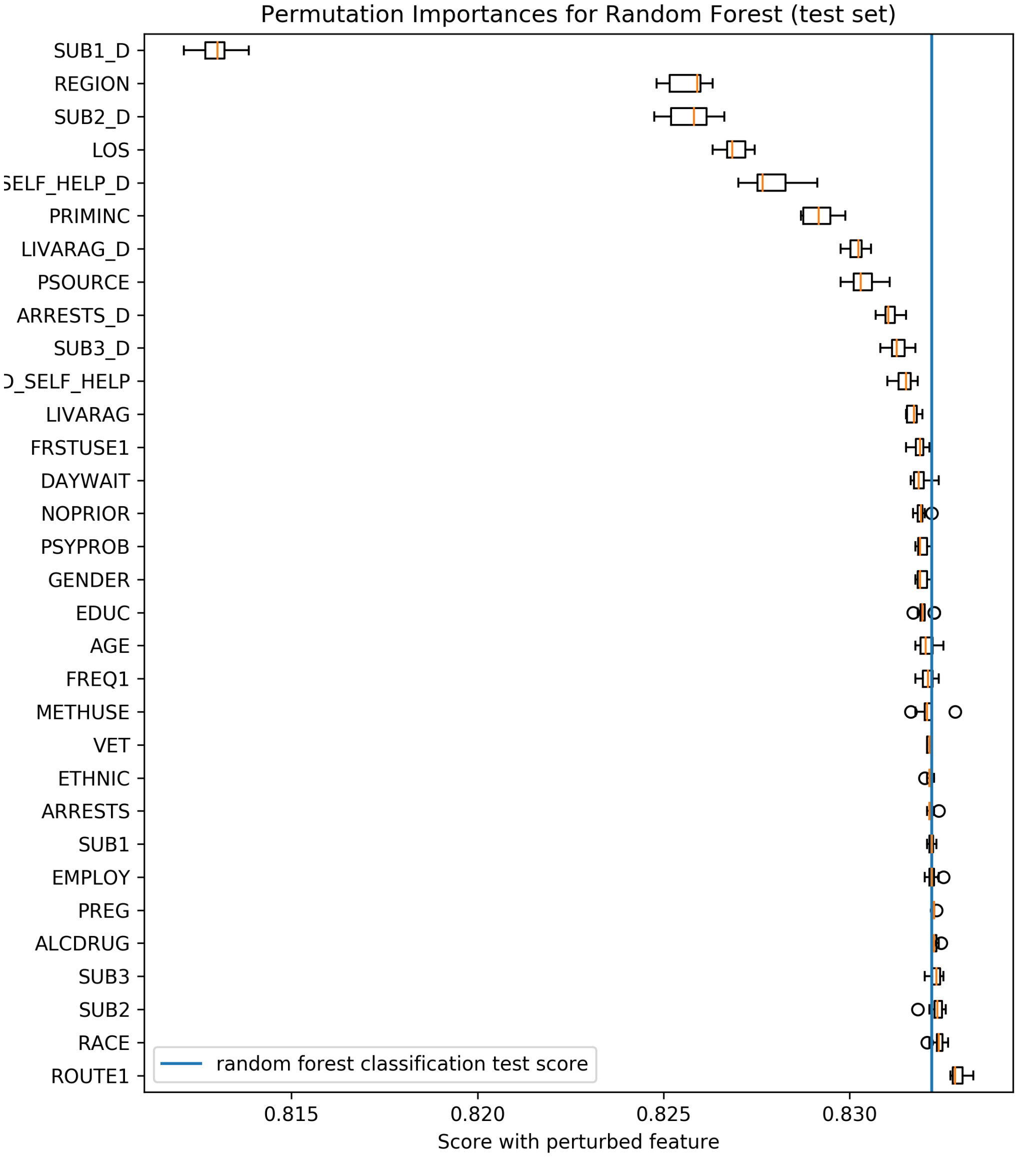

Figure 7: The permutation importances for the random forest are similar; the top three features are instead SUB1_D, REGION, and SUB2_D.

There are several features that, when shuffled, actually increase testing scores. This suggests little to no correlation between these features and our target variable. It could also suggest that the estimator is using other correlated features to calculate a prediction, actually resulting in higher accuracy due to reduced noise. These findings are interesting in that one might expect to see a stronger relationship between age and reason for discharge, for example (e.g. younger people who have been using opioids for a shorter amount of time might be see more success in treatment).

We see a ∼2.5% drop in testing accuracy for the random forest model linked to SUB1_D. There are much more modest changes for the remaining features. It is also interesting to note that the features that increased the random forest model's score are rather different from those that increased the logistic model's score. There is some overlap (SUB1, ALCDRUG, ROUTE1) but otherwise it looks like the estimators based their decisions using different weights for several features.

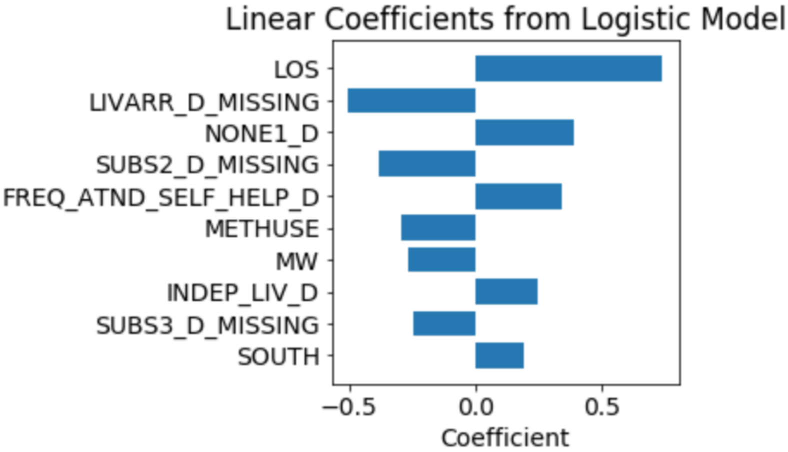

Figure 8: These are the 10 strongest linear coefficients returned by the logistic model. LOS has a very strong positive correlation with successful treatment. It is odd that missing values for living arrangement at discharge would have such a strong negative correlation.

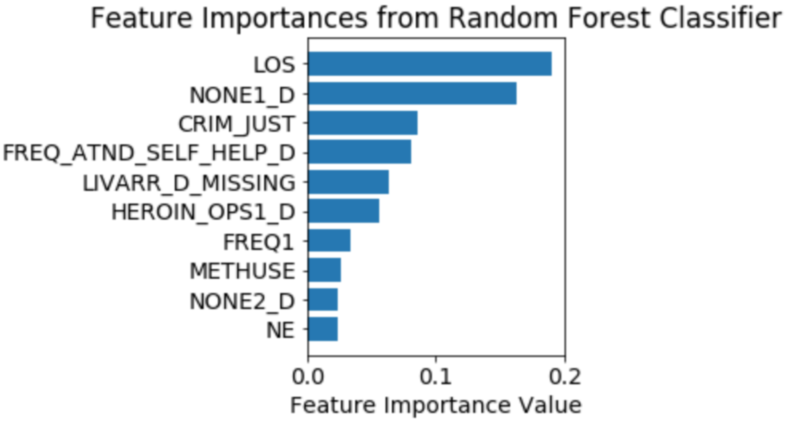

Figure 9: These are the 10 strongest feature importances returned by the random forest model. LOS again has the highest importance value.

Because I have so many features, a detailed dissection of local feature importance would not fit within the scope of this project. For brevity, I will discuss the random forest's collective force plot of the two categorical features mentioned in the **EDA** section, generated from the SHAP TreeExplainer. The plots use the first 1000 observations from the data set.

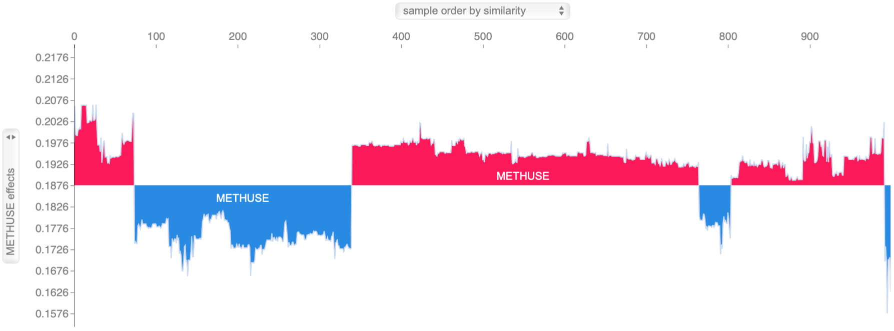

Figure 10: The blue regions are where METHUSE = 1, and the red, METHUSE = 0. Positive methadone use pushed the prediction downwards, while no methadone use pushed it upwards. Because this relationship is consistent across (what seems to be) all METHUSE observations, the model is not suggestive of any feature interactions with METHUSE.

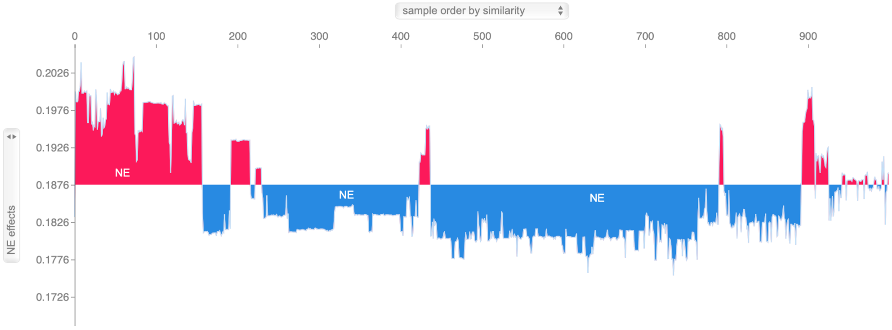

Figure 11: The red regions are where NE = 1, and 0 in the blue. There is again a consistent relationship between the feature and the direction of the weight, although the strengths of the weights are lower than those of METHUSE.

These results do suggest that both models work within reason; length of stay and substance use at time of discharge naturally have positive correlations with our target. Other less obvious features can be framed in a real-world context. Different region indicators, for example, seem to have strong predictive effects in differing directions. One could compare quality of care metrics or organizational structures across regions to see how they influence individuals' experiences, as some regions might excel in ways that others do not. The results we see with the METHUSE indicator could draw attention to tailoring programs to more closely accommodate those who are also taking methadone to treat opioid addiction. Centers can also look into encouraging more self-help therapy, as it seems to be helpful. Finally, centers could evaluate the mechanisms that involve someone's housing situation (LIVARAG, LIVARAG_D) and adapt their programs accordingly.

5 Outlook

I chose to use logistic and random forest models because they have high interpretability. Since there are so many missing values in the data, however, this project might benefit from estimators like XGBoost that are able to accept missing values. Given more time, I would also love to do more local feature importance analysis; since I am using human data, it would be helpful to pinpoint exactly what feature interactions influence an individual's ability to complete a treatment program. These analyses are especially important for creating tailored public health programs to specific communities or subpopulations. Another weak spot of my approach is how many features there are. I could perform PCA or another dimension reduction technique to cut down the number of features, improving computational runtime and interpretability. Lastly, the TEDS data set is published annually. To potentially improve the balance of my data set, I could attempt to combine more data from multiple years, although time-series data are not necessarily *iid*.

6 References

United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Office of Applied Statistics. Treatment Episode Data Set—Discharges (TEDS-D), 2017. https://www.datafiles.samhsa.gov/study-dataset/teds-d-2017-ds0001-teds-d-2017-ds0001-nid18480

United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Office of Applied Statistics. Treatment Episode Data Set—Discharges (TEDS-D) Codebook, 2017. 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

Acion L, Kelmansky D, van der Laan M, Sahker E, Jones D, Arndt S (2017) Use of a machine learning framework to predict substance use disorder treatment success. PLoS ONE 12(4): e0175383. https://doi.org/10.1371/journal.pone.0175383