Predict Substance Abuse Treatment Completion: Apply and Compare the Decision Tree, Random Forest, Naive Bayes, and CatBoost Classifiers

ChangYi Liu, Bela Bairy Ganapayya, Chia-Hsuan (Sharon) Lee, Kathy Wang

Problem Definition

This project will build a binary classifier to predict if an individual admitted for a substance abuse treatment program will complete treatment or dropout. The project will use individuals' demographics, substance use, and treatment information as initial features. We will identify the top relevant features to treatment completion status and use these selected features to build a binary classifier.

Background

People of all ages and from all areas of life are impacted by mental health and substance use disorders. The estimated lifetime prevalence rates in the US population were 6.1% for other drug dependency and abuse, 13.5% for alcohol dependence and abuse, and 22.5% for any non-substance abuse mental disease (Darrel A. Regier, 1990). These illnesses can influence how we interact with others and make decisions. Although these conditions are frequent, recurring, and frequently serious, there are substance abuse treatment programs available, and many people do recover. However, not all participants admitted to a program complete the full course and dropout without recovering.

Studies have shown that organizational factors and human resources in Public units influence client treatment success (Peter D.Friedmannmd, 1999).The dataset used for this project analyzes changes from treatment program admission to discharge among records that have been linked for admission and discharge. Multiple tables repeat the characteristics of primary substance of abuse, frequency of use, and employment position at admission. This is done to ensure the inclusion of crucial information that could otherwise be missed and to allow for an adequate comparison of the features at admission and discharge.

As substance abuse is a major problem in the U.S., this topic is of great importance for determining the factors that may affect treatment dropout, being able to predict dropout, and offering effective countermeasures. For instance, altering the treatment, providing extra care at the individual level, or reexamining the treatments for these groups at a higher level.

Dataset

Treatment Episode Data Set (TEDS)

https://www.samhsa.gov/data/data-we-collect/teds-treatment-episode-data-set

The source of this dataset is the U.S. Government's Substance Abuse and Mental Health Services Administration (SAMHSA). The dataset, TEDS-D (2019), collected from treatment facilities, are discharge records for substance use treatment by state in 2019. It has 1,722,503 observations and 74 features with categorical values. Missing values are indicated by -9, and there are 17,751,219 missing values in total. We will use features concerning demographics, treatments, substance use, and reasons for discharge of participants in the dataset to find the relationships between features and reason for discharge and conduct a classification analysis.

Methods

Data Exploration and Visualization

We explored the structure of our dataset and features, missing values frequency, univariate analysis of feature patterns, and bivariate analysis of feature relationships. We visualized the dataset using bar charts to see the relative frequency of missing values per feature, the number of unique values per feature, and value counts for each feature. We plotted count plots and box plots for each feature separated by the response variable "Reason" (treatment completion or dropout), to see if we can identify any patterns in each feature relative to treatment completion status. We plotted a correlation heatmap between any two attributes to visually identify the presence of highly correlated variables. We also divided all attributes into three families: demographics, treatment, and substance use, and then explored them separately.

Data Preprocessing

We performed four tasks in the data preprocessing phase: (1) variables preprocessing, which includes missing values removal and variables modification, (2) removing correlated features based on Spearman correlation coefficient of 0.8 or higher, (3) selecting key features using Chi-Squared test, and (4) one-hot encoding, to transform our nominal categorical features into an alternate numerical form to better apply classifier packages in python. We used Spearman correlation method and Chi-Squared test for feature selection because all the variables in our data set are categorical, and the analysis to be made is a nonparametric analysis.

Classification Models

To classify whether a patient discontinues treatment or not, we adopt four different classifier methods: decision tree, random forest, Naïve Bayes, and CatBoost classifier. We train a classifier using each method and compare the performance to choose the best-performing classifier. We construct three decision tree models. One is a fully grown tree with over 20 layers; the other two are pre-pruning trees having 7 and 10 layers respectively. The purpose of constructing pre-pruning trees is to avoid the over-fitting problem and to make it clearer the interpretation of the model. We use random forest with 1000 decision trees as well to reduce the overfitting problem of decision trees. We adopt Naïve Bayes Classifier because it handles large datasets more properly than the decision tree or random forest. We have about 42437 data instances for data training, which qualifies as a large dataset; and to rely only on the decision tree could result in overfitting and over-complicated decision rules. We use CatBoost Classifier to perform gradient boosting on decision trees with categorical variables.

In model construction, 70% of the preprocessed dataset will be used to train the model and the remaining 30% will be used for testing the model. We evaluate each model's performance by calculating accuracy and F1 scores on the test set. In addition, for the decision tree and Naïve Bayes models, k-fold cross-validation is implemented to further estimate and compare the generalization ability of each model, and thus increase the robustness of the model.

Experiment: Experiment Setup and Analysis Results

Data Exploration and Visualization

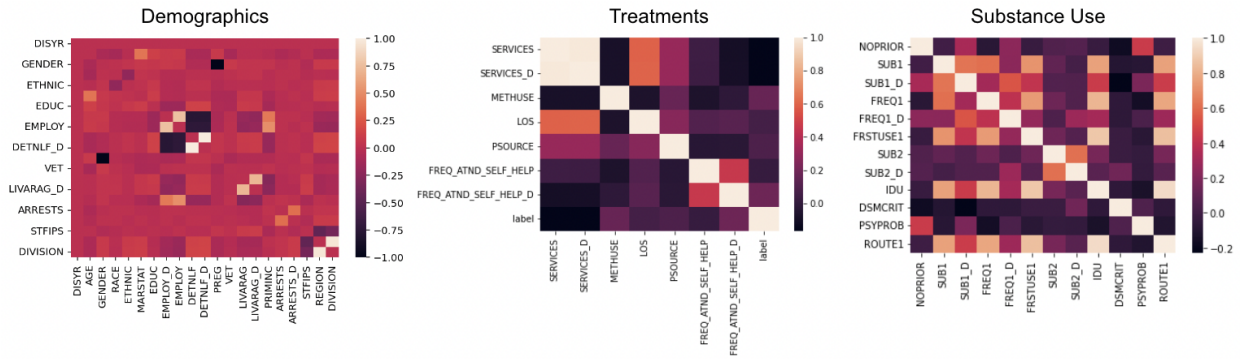
From the exploration of the whole dataset and referencing the codebook (which describes features and encoded values), we found that all features have coded values and are categorical, except for 2 ID variables with no predictive value, which were dropped. After filtering the dataset for response variable "Reason" as treatment completion and dropout only (1,158,539 total instances), we found that the dataset is imbalanced, with 725,909 instances of completion and 432,610 instances of dropout. There were 1,156,127 rows with at least one missing value, so we could not simply delete all rows with missing values. We calculated and plotted the relative frequency of missing values in each column, and identified 11 variables containing at least 50% missing values.

From the plots, we saw that the number of unique values per feature ranged from 2 to over 100. Almost all variables' numerical encodings were not ordinal. From the correlation heatmaps, we found many variables were correlated and redundant. From just observing the plots color-separated by the response variable, we could not identify any features with notable differences between value count ratios for completion versus dropout.

For splitting the features into 3 families, we subjectively categorized each feature. Demographic variables relate to basic patient information, such as age, gender, and work status. Treatment family variables include treatment types, duration, and treatment resources. The family of substance use otherwise contains measures of the type of addictive substances patients use and the frequency of use. For each family, we plotted a correlation heatmap to better visualize correlations between features (Figure 1), and identified redundant variables.

Figure 1

Correlation heatmaps of Demographics, Treatments, and Substance Use families



Data Preprocessing

Variables Preprocessing: Since we are only interested in instances of treatment completion or dropout, we filtered our response variable, discharge "Reason", accordingly. In our dataset, some variables have a value of -9, indicating missing/unknown/not collected/invalid values. We identified 3 variables for which some instances of -9 should be encoded as a different value instead of a missing value. For instance, the variable "Pregnant" had all males labeled as -9, so we relabeled males to the new label 3. Many variables contained missing values (-9), so we removed 11 variables that contained 50% or more missing values and then removed any rows containing missing values. The resulting dataset had 60625 rows and 63 columns.

Removing Correlated Features: To remove redundant features that may affect our model, we calculated the Spearman correlation coefficient between each pair of variables. We removed 8

variables with a correlation coefficient of 0.8 or higher with another variable.

Feature Selection: Before constructing the classification models, we used the Chi-Square feature selection method to filter out features that are not strongly correlated with the dependent variable— 'drop'. Methodologically, we do the hypothesis testing for each attribute under the conditions that:

$$\begin{cases} Null\ hypothesis\ H_0: an\ attribute\ is\ independt\ of\ dependent\ variable\ 'drop' \\ Alternative\ hypothesis\ H_1: an\ attribute\ is\ not\ independt\ of\ dependent\ variable\ 'drop' \end{cases}$$

Any attribute whose null hypothesis is rejected (under confidence level 0.005) is considered to be correlated with 'drop'. In our case, we hope to find ten of these cases, therefore the ten features with the lowest p-value under the Chi-Square distribution become our main features that will be used for modeling, as shown in Table 1.

Table 1.

Main features for analysis

Features	Description
ROUTE1	Route of administration
SUB1	Substance use at admission
PSOURCE	The person referring the client to treatment
HERFLG	If a record of heroin is reported at admission
ALCFLG	If a record of alcohol is reported at admission
FREQ_ATND_SELF_HELP_D	Attendance at self-use group in past 30 days prior to discharge
DIVISION	Census division
SUB3	Substance use at admission
DSMCRIT	Client's diagnosis of substance use
IDU	Current IV drug use reported at admission

One-Hot Encoding: We performed one-hot encoding on our 10 selected categorical variables to define a numerical meaning on the coordinate system for each variable. One-Hot Encoding turns all classes of features into a 0 & 1 vector; the length of the vector is equal to the number of classes under one feature (Pau Rodríguez, 2018). For example, the feature 'ROUTE1' in our model has five different classes, so the One-Hot Encoding vector of 'ROUTE1' is a length=5 vector with each class occupying one place in the vector. Class '1' under 'ROUTE1' becomes a vector [1,0,0,0,0]; class '2' becomes [0,1,0,0,0]; and class '4' is [0,0,0,1,0]…etc. Below is a comparison of data points prior to and after One-Hot Encoding (Figure 2). We then transform the dataset with the one-hot encoded vectors to have each column correspond to a unique value per

feature. The final preprocessed dataset that will be used for modeling has 60625 rows and 89 columns.

Figure 2

Sample of data before One-Hot Encoding (left) and sample of data after (right)

	HERFLG	ROUTE1	SUB1
0	0	1	2
1	0	1	2
2	1	4	5
3	1	2	5

	HERFLG	ROUTE1	SUB1
0	(1.0, 0.0)	(1.0, 0.0, 0.0, 0.0, 0.0)	(1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
1	(1.0, 0.0)	(1.0, 0.0, 0.0, 0.0, 0.0)	(1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
2	(0.0, 1.0)	(0.0, 0.0, 0.0, 1.0, 0.0)	(0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...
3	(0.0, 1.0)	(0.0, 1.0, 0.0, 0.0, 0.0)	(0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...

Classification Models

After the data encoding is complete, the classification features are now all possible categories for each feature. We have 89 features for classifying data points with each occupying a place in vector space. Then we use the machine learning packages in the scikit-learn library to build the decision tree, the random forest and the Naïve Bayes classifiers. And for CatBoost classifiers, the model is offered by a pre-set package of 'catboost'. Our strategy is to operate these models separately and to compare the differences of validity among them. Additionally, we run the 10-fold cross validation for the decision tree and the Naïve Bayes model. We believe that the average accuracy of 10 different validation trainings is more significant than the accuracy rate of a single model. However, cross-validation is not applied to the random forest and CatBoost model as they already form the model through many iterations.

In the selection of the model environment, the decision tree models follow the entropy-gain strategy; the random forest follows the same strategy from 1000 trees. The random forest model has a random state=20 to control the model shuffling. The Multinomial Naïve Bayes is implemented for its best fit for categorical variables after we tried Gaussian Naïve Bayes model but it disappointed us. Lastly, the CatBoost algorithm adjusts its classifiers by iterating 1000 times and uses the f-1 score to evaluate the performance per iteration.

Model Evaluation

We compared each classifier's performance by comparing the accuracy and F1 scores on the test set (Table 2). The CatBoost classifier has both the highest accuracy and F1-score on the testing set, 0.751 and 0.837 respectively.

Table 2

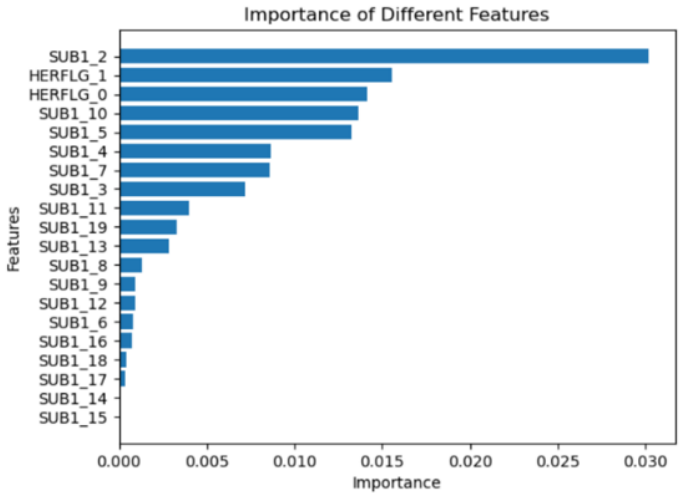
Accuracy and F-1 Score Comparison

Model	Accuracy Rate	F-1 Score
Full Decision Tree	0.728	0.813
10-layer Decision Tree	0.742	0.832
7-layer Decision Tree	0.739	0.833
Random Forest	0.739	0.825
Naïve Bayes	0.676	0.762
CatBoost	0.751	0.837

We also generated and plotted the top 20 feature importance scores from the random forest classifier (Figure 3). The features SUB1 (primary substance use at admission) and HERFLG (heroin use) appear to be the most important features for the model's classification.

Figure 3

Random forest classifier feature importance scores



Observation and Conclusion

The problem of substance abuse occurs at all levels of society. Sufferers of mental illness tend to abuse substances, which further deteriorates their physical and mental health, but also negatively impacts our society as a whole. Although treatment institutions can alleviate the

situation, their patients often fail to complete the entire treatment sessions. The purpose of this study is to predict whether a client will complete addiction treatment and study the features useful for the prediction, ensuring clients receive more effective treatments.

We select the 10 most statistically significant features associated with treatment completion, which generally reflect the following considerations: whether the main types of abused substance are alcohol and heroin, substance route of administration, frequency of self-help groups participation during treatment, reason for treatment, and whether the client is self-referred.

Then, we calculate the accuracy of the model with different classifiers and choose the optimal classifier with the highest accuracy. Due to having imbalanced dataset, we consider F1-score as a better metric to evaluate the model. According to the result shown in table 2, the CatBoost classifier has both the highest accuracy and F1-score on the testing set, 0.751 and 0.837 respectively, indicating a high predictive power.

In summary, one can predict if a client can complete the treatment by using CatBoost classifier and can improve treatment services and quality based on the selected features to increase treatment completion, such as encouraging participation in self-help groups, and strengthening treatment for specific substance addiction.

References

1. Regier, D. A., Farmer, M. E., Rae, D. S., Locke, B. Z., Keith, S. J., Judd, L. L., & Goodwin, F. K. (1990). Comorbidity of Mental Disorders with Alcohol and Other Drug Abuse. Results from the Epidemiologic Catchment Area (ECA) Study. *JAMA*, *264*. 2511-2518
2. Friedmann, P. D., Alexander, J. A., & D'aunno, T. A. (1999). Organizational Correlates of Access to Primary Care and Mental Health Services in Drug Abuse Treatment Units. JSAT, 16(1), 71－80.
3. Pau Rodríguez, Miguel A. Bautista, Jordi Gonzàlez, Sergio Escalera, Beyond one-hot encoding: Lower dimensional target embedding, Image and Vision Computing, Volume 75,2018,Pages 21-31, ISSN 0262-8856, https://doi.org/10.1016/j.imavis.2018.04.004.