What Factors Predict Dropout Rates for Inpatient Substance Treatment Programs?

Marjete Vucinaj

Data Science, CUNY School of Professional Studies

DATA 698: Master's Research Project

Dr. Arthur O. Connor

November 25th, 2024

Abstract:

Substance use disorders (SUDs) affect over 46 million individuals in the U.S., and treatment dropouts pose a significant challenge to recovery. This study used the 2021 Treatment Episode Dataset: Discharge (TEDS-D) to examine factors contributing to inpatient treatment dropouts by applying machine learning models, including XGBoost, Random Forest, Neural Networks, and Lasso Logistic Regression. While all model results were comparable, XGBoost proved the most effective model, achieving the highest AUC (0.70) and F1 score (73.62).

Key predictors of treatment outcomes were identified, such as frequency of self-help program attendance, employment status, geographic region, age, and frequency of substance use. Cross-validation confirmed consistent model performance with minimal overfitting, while external validation using 2020 data demonstrated the stability and generalizability of the results. This study emphasizes the impact of the frequency of attendance in self-help programs on inpatient treatment outcomes. Future research should further explore the field of addiction and substance treatment in inpatient settings. This research aims to implement approaches, such as offering more self-help programs, to improve completion rates and promote sustained recovery outcomes.

Introduction:

Addiction is a public health crisis. In 2022, approximately 46.3 million individuals in the U.S., or 16.5% of people (aged 12 and older), were reported to have a substance use disorder (SUD), which includes alcohol and drug use disorders (SAMHSA, 2022). The *Diagnostic and Statistical Manual of Mental Disorders Fifth Edition* (DSM -5) has a range of behaviors to fit the SUD diagnosis, such as impaired control over use. Even if an individual has a 'mild' SUD, addiction is known to progress and worsen, especially as one becomes more physically dependent (Volkow, 2021).

In 2022, the CDC reported the highest growth of overdose deaths, of over 107,000 people; this equates to one overdose death every five minutes. Fentanyl and, more recently, xylazine (tranquilizer) are responsible for over two-thirds of overdoses (Kariisa et al., 2023). In the past few decades, treatment options such as harm reduction programs and methadone maintenance therapy have become more popular to prevent overdose and reduce use and use more safely. However, methadone can still provide a 'high' and can be misused by individuals who are not prescribed. Harm reduction programs encourage using new needles, testing strips for fentanyl, and having Naloxone (Narcan) available when using. However, individuals who have impaired control over use may not consistently practice these suggestions (Crapanzano, 2018). While harm reduction techniques can and have been used to reduce overdose deaths in this past year (2023), the best option to prevent overdose death is recovery (Centers for Disease Control and Prevention, 2024).

Throughout society, individuals with SUD are blamed for having an addiction, which subsequently results in stereotypes. The societal stigma often leads to self-stigma, such as self-blame, especially about why the individual cannot stop using, making it more challenging to break the cycle of use or seek recovery. Relapsing several times is also part of the process in folks with addiction, which further causes stigma. Notably, research suggests that terms like ‘addict’ and ‘alcoholic’ are often associated with social deviance, further complicating access to treatment and recovery efforts. However, some research suggests that if one is willing to accept these terms, it can have a positive effect on treatment (Crapanzano, 2018).

In this project, we will examine the risk factors that predict inpatient substance treatment dropout rates using R. If individuals drop out of inpatient treatment, they are more likely to continue in their addiction. While completion of inpatient substance treatment does not guarantee recovery or sobriety, those who do are one step closer to living a life free from addiction. The variables that will be examined are social determinants of health, comorbidity, substance use severity, age of first use, participation in self-help programs, and referral source; current literature and studies focus on one or a few of these factors.

Literature Review:

Several studies have used observational data, questionnaires, and electronic health data to measure and model the factors responsible for dropping out of Substance Abuse Treatment. Gautam and Singh (2020) reviewed several datasets and selected the Treatment Episode Dataset: Discharge (2017 TEDS-D) from the Substance Abuse and Mental Health Services Administration. The researchers also indicate that opioid misuse treatment has not been thoroughly investigated and point out that the treatment dropout rate is about 26%. They specifically selected data where admissions reflected opioid use only and did not distinguish any treatment setting (inpatient vs outpatient).

Gautam and Singh (2020) analyzed TEDS-D 2017 using MLRM and RF to model dropout rates and focus on socioeconomic factors that impact the dropout rates of substance use treatment. They assigned transfer to another facility as treatment completion. The top 5 variables that impacted dropout rates were the “Length of Stay,” “State in which a patient lives,” “Census Region in which a patient lives,” “Employment Status at the time of dropping out,” and the patient’s “Age.” They did not find health insurance or co-occurring mental and substance use disorders to be significant variables. The researcher specifically pointed out that patients drop out after the first few days of treatment, or if they have a more extended treatment stay, they tend to drop out after 180 days; they suggested that staff pay more attention during these times to prevent dropouts. Results also indicated that the RF model (AUC of .89) performed better than the MLRM (AUC of .68).

Acion et al. (2017) also focused on the Treatment Episode Dataset: Discharge using data from 2006-2011. To learn more about disparities in outpatient treatment settings, they investigated a subset of the population in the data, Hispanic patients. Since they were using data spanning 5 years, they only selected data where the individual had no prior treatment for a drug or alcohol. They used 17 algorithms to model the data, in addition to Super Learning (SL), a targeted learning method, which takes the weighted average of all the models. Using the AUC metric, Acion found that SL performed significantly better than some models; while SL performed the best with RF following, these two were similar. They identified the top 10 variables of importance: “length of stay, age, principal source of referral, primary problematic substance, its age of first use, and its frequency of use, employment level, SUD type (i.e., only alcohol, only drugs, or both alcohol and drugs), education level, and patient’s specific Hispanic origin.”

Baird, Cheng, and Xia (2022) also use the TEDS dataset for 2017-2019 to examine disparities in substance use treatment completion rates. They assigned ‘transfer to another facility’ as drop out of treatment, and then remodeled and only kept drop out and completion as is with no other variables; they found minor differences in this change. They concluded that a two-stage virtual twins model (random forest + decision tree) performed better than XGBoost, neural network, and logistic regression. Their key findings were that those with no co-occurring mental health diagnosis were more likely to complete treatment, and those with employment were more likely to complete treatment.

While previous studies offered insight on a national scale, Gottlieb (2022) used two local datasets from the Houston Emergency Opioid Engagement System (HEROES) to examine retention rates for folks with opioid use disorder. The datasets contained 715 patients for a 3-month program and 691 patients for a 4-month program. They found varying dropout rates based on the time of stay, with a 15% dropout rate for 90 days and a 24% for the 120-day program. They evaluated the rate of dropout every 30 days using several models: Logistic Regression, Radial Basis Support Vector Machines AdaBoost, Gentle Boost, Logit Boost, Robust Boost, Total Boost, and Random Forest; the RF model performed the best. They also found that a history of overdose was one of the top contributing factors to drop-out rates.

Previous studies used U.S.-centered data and machine learning, while Andersson et al. (2018) analyzed data from Norway and St. Olavs University Hospital in Trondheim with Cox proportional hazards regression. The dataset included a total of 454 patients from 5 inpatient substance use treatment centers. Besides completion rate information, the data also included self-reported questionnaire data, which measured patients' motivation and mental distress. This study’s treatment length was 2 months to 12 months, depending on the target group. The authors used a Cox Regression and found that mental distress was a critical factor in the dropout rate, and motivation was a key factor in completion. They also found that patients with ADHD were more likely to drop out of treatment.

Research Question

Do the following factors contribute to dropout rates of inpatient treatment for addiction?

(a) *Social determinants of health include* age, gender, race, ethnicity, marital status, education, employment status, living arrangements, past arrests, region, co-occurring disorder, and DSM diagnosis.

(b) *Substance use severity:* frequency of use, medication-assisted opioid therapy, primary substance, current IV drug use, age of first use, Previous substance use treatment episodes, Route of administration, Substance use type (drugs only, alcohol only, or both).

(c) *Choice in treatment:* referral source, length of stay, participation in substance use self-help programs.

Data and Variables:

For this project, the dataset selected is from Substance Abuse and Mental Health Services Administration, Treatment Episode Dataset: Discharge, 2021 (TEDS-D). TEDS has been funded and managed by the Center for Behavioral Health Statistics and Quality (CBHSQ) since 1992. Data is collected as administrative records from all 50 U.S. states, the District of Columbia, and U.S. territories and standardized for consistency regardless of location. However, there was insufficient data for the following states: Delaware, Idaho, and Oregon, and they were excluded from the 2021 TEDS (Substance Abuse and Mental Health Services Administration [SAMHSA], 2023).

TEDS includes reporting for admissions in treatments that are publicly funded and does not generally include any information from private treatment due to difficulty in accessing that data. However, many states vary due to "highly diverse state data collection systems." Public funding and state-to-state differences impact the data, as some states do not have medication-assisted treatment; based on the practices of the local criminal justice, "coercion plays a role in referral to treatment (Substance Abuse and Mental Health Services Administration [SAMHSA], 2023)."

Like other datasets with Private Health Information (PHI), various methods are used to protect confidentiality—data swapping de-identifies records by exchanging certain variables, such as swapping between a census region. More unique variables, such as age, are recorded to prevent identifying older and younger folks, who are a smaller subset of the data. These methods preserve the dataset's accuracy, reliability, and usability (Substance Abuse and Mental Health Services Administration [SAMHSA], 2023).

Data cleaning and feature engineering:

The dataset had all unknown data coded as -9; this would harm predictive modeling as it would encode it as a unique category. Therefore, these values were replaced with NA, and all NA observations were dropped. After all data cleansing, there were 32,449 observations; 21,058 represented completed treatment, and 11,391 dropped out. Most variables in the dataset are categorical with numeric encoding, so outliers were not of concern.

The treatment outcome variable at discharge is this project's target variable; this variable was recoded into a binary outcome. Dropped out, Terminated, Incarcerated, or Death was coded as drop-out; Transferred was removed as it is not a clear indication of dropping out or completing the program. Completion was encoded as 0 and dropout as 1, so the model will treat the dropout as the positive class, and variable importance will reflect dropout rather than completion.

Length of stay (LOS) was moderately correlated (-0.54) with the target variable (before binning). It initially had 37 values, where values of 1-30 represented each corresponding day; their values were binned into ranges, 0–7, 8–14, etc.; 31- 37, which already represented a range, remained as is and saved as new numeric values. Note that this variable was still slightly correlated after binning as well (-0.37) with the target variable) Moreover, it was later excluded from the second set of modeling to compare the models with and without this variable.

DSM diagnosis (DSMCRIT) had 19 values in the initial dataset; this many categories, in addition to few observations in some, would lead to noise in the model. Therefore, this variable was also binned into five values: mild substance, moderate substance, severe substance, mental health disorder, and other disorders. Note that this variable is a primary diagnosis where only one value can be present, so while it is not indicative of co-occurring substance use and mental health, it does provide information on the severity of substance use.

Data Preprocessing:

Data Splitting: using the createDataPartition function in the caret package, the train_data was split into 70% of the data for model training, which is used to create the model, and test_data was divided into 30% of the data for model testing, used to qualify performance.

Using the recipes package to create a preprocessing pipeline and the themis package to add further functionality (in the order presented below):

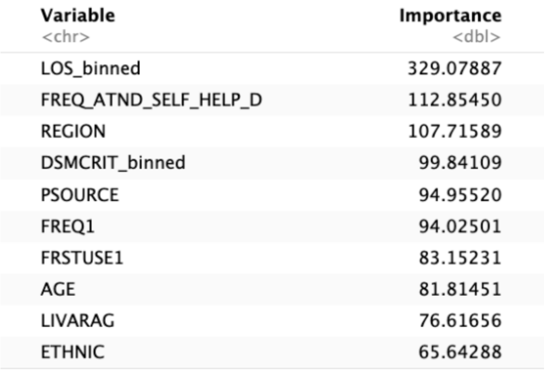
> Data Balancing: The distribution of the outcome variable was not balanced, with 21,058 admissions where treatment was completed and 11,391 dropouts. To avoid bias toward the majority class, use step_smote()only on the training set to generate synthetic examples of the minority class. This method also avoids overfitting.
>
> Scale and Center data: Use the step_normalize function to scale and center the predictors needed for models like logistic regression and neural networks to ensure standard mean and standard deviation.

Variables

The inpatient service setting was selected due to the importance of having support while individuals would be withdrawing- without support, withdrawals can be deadly (while the service setting for 24-hour detox is also an inpatient category, it was excluded as it would not account for withdrawal symptoms lasting up to 7-10 days). Inpatient service settings include Rehab/residential, short-term (30 days or fewer), and Rehab/residential, long-term (more than 30 days).

Further efforts were made to reduce the number of variables to simplify the model. The near-zero variance function was applied to the dataset but did not locate variables with near-zero variance. Additionally, a correlation plot was used to assess multicollinearity, but it did not show any variables with concerningly high correlations. However, it did show that the variables with the highest positive correlation were Employment and substance use; the highest negative correlations were frequency in substance use self-help programs, age, and education.

The Random Forest model, which is robust to noise and irrelevant features, was used to determine variable importance. The results below show that length of stay is the most important variable, followed by the frequency of self-help programs. The top nine variables were selected for the first round of modeling. Ethnicity was excluded due to the 11-point difference between the previous variable of importance and to simplify the model.



Variable	Importance
<chr>	<dbl>
LOS_binned	329.07887
FREQ_ATND_SELF_HELP_D	112.85450
REGION	107.71589
DSMCRIT_binned	99.84109
PSOURCE	94.95520
FREQ1	94.02501
FRSTUSE1	83.15231
AGE	81.81451
LIVARAG	76.61656
ETHNIC	65.64288

Figure 1 Top Variables

Inclusion criteria were based on the selecting variables related to the hypothesis and dimension reduction for round 1 of modeling:

(a) *Social determinants of health:* age, living arrangements, region, and DSM diagnosis.
(b) *Substance use severity:* frequency of use, age of first use
(c) *Choice in treatment:* referral source, length of stay, participation in substance use self-help programs.

The boxplot below shows the relationship between length of service and treatment outcome. The median and IQR are different between treatment completion (0) and treatment dropout (1), showing that those with a lower length of service are more likely to drop out of treatment. This suggests a potential predictive nature of the length of stay variable.

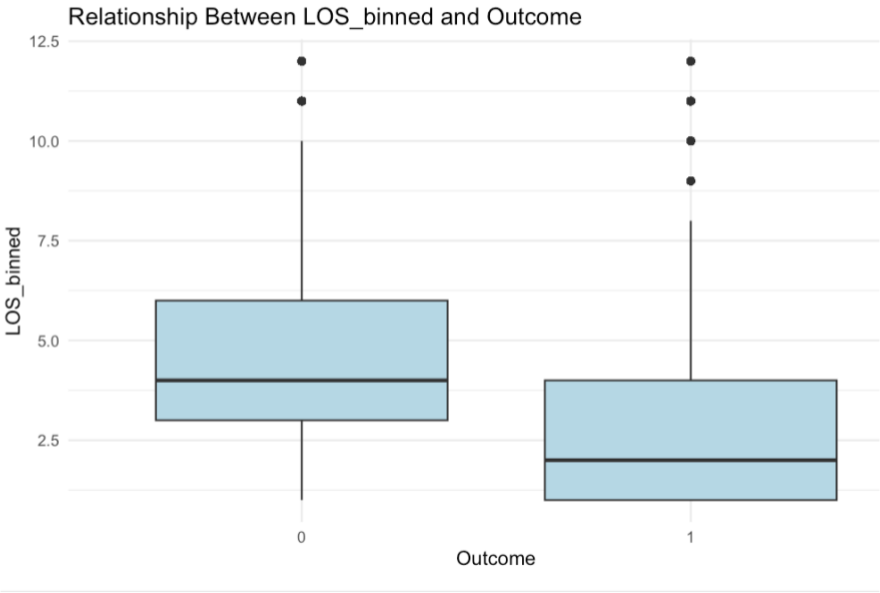
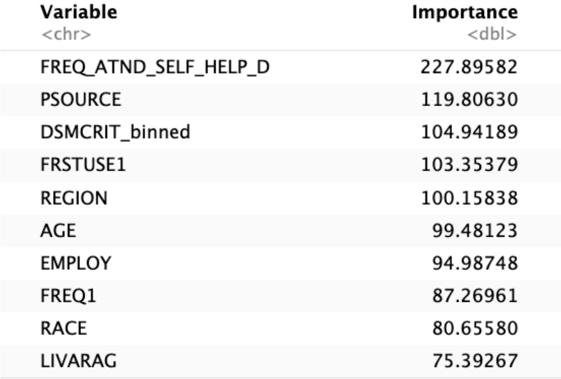


Figure 2: Boxplot of LOS variable

The length of stay variable was excluded for the second round of modeling. Random Forest feature importance was used to confirm any changes to the top variables for inclusion. Other features were now identified as necessary: employment and race. Note that the values for importance have also increased for the top 4 variables listed below.



Variable <chr>	Importance <dbl>
FREQ_ATND_SELF_HELP_D	227.89582
PSOURCE	119.80630
DSMCRIT_binned	104.94189
FRSTUSE1	103.35379
REGION	100.15838
AGE	99.48123
EMPLOY	94.98748
FREQ1	87.26961
RACE	80.65580
LIVARAG	75.39267

Figure 3 Top variable importance with LOS removed

The inclusion criteria were based on selecting variables related to the hypothesis and dimension reduction for round 2 modeling.

(a) *Social determinants of health:* age, living arrangements, region, employment, race, and DSM diagnosis.
(b) *Substance use severity:* frequency of use, age of first use
(c) *Choice in treatment:* referral source, participation in substance use self-help programs.

The frequency of attendance in substance use self-help programs at discharge was the variable used (a similar admission metric was also included in the dataset). This was selected because some individuals may not have been exposed to self-help programs outside of rehab programs, or the frequency might have increased during their inpatient stay, influencing treatment outcomes.

Statistical Methods:

Parallel Computing is a technique for reducing computation time. It divides tasks into smaller sub-stacks, executed simultaneously and aggregated to produce final results. This method is ideal for large-scale data, model training, and cross-validation.

Random Forest Classification (RF):

An ensemble learning technique where multiple decision trees are trained using bagging (bootstrap aggregating) and each tree is trained on a random subset of each tree. This method improves accuracy by aggregating the prediction using the class with the most votes to produce the final prediction. RF also reduces overfitting, even when the number of trees is large. The mean decrease in Gini is a metric used to measure the importance of this technique. RF can handle complex datasets and is ideal for nonlinear relationships. Some drawbacks are that this method might be less interpretable than a simpler one and computationally expensive.

To create this model, the target variable Outcome was specified, predictors from the balanced training data were used to train the model, and feature importance calculation was enabled. Then, the trained RF model was used to make predictions of the processed test data. To prevent overfitting, the number of trees was set to 1000, the number of features per split was reduced to 3, the leaf size or node was limited to 5, and the depth of trees was limited to 50.

Extreme Gradient Boosting (XGBoost)

This model adjusts the gradient boosting algorithm to be more scalable and efficient, giving it the XGBoost name. It uses lasso (L1) and Ridge (L2) regularization to prevent overfitting; it also improves generalizability by randomly sampling a subset of features for each tree. It optimizes decision trees with max-depth, also known as the maximum delta step; this pruning method eliminates ineffective splits. The model also used parallelization to train the model faster. XGBoost increases accuracy with first and second-order gradients to optimize the loss function. The model is highly efficient for large datasets and highly customizable; however, it is computationally expensive and less interpretable than a simpler model.

To model XGBoost, the data was first converted with the xgb.DMatrix (the required matrix format for this model) will be used to create the training and test matrices. Next, the cross-validation parameters were defined. To optimize the model and reduce overfitting, the maximum depth of each tree was set to 3, and the learning rate (eta) was decreased to 0.05; while this slows down the model, it generalizes better. The row sampling (subsample) specifies the training data of 70%; feature sampling was also set to 0.7, improving generalization by preventing the model from relying too heavily on any feature. L2 regularization, or lambda, was set to 2, and ridge regularization was used to stabilize the model and reduce overfitting. L1 regularization or alpha was increased to 1, which uses lasso regularization to force coefficients to zero and improves feature selection. The minimum sum of instance weights needed in a child node was set to 5 to reduce insignificant subsets of data. The minimum loss reduction of Gamma was set to 1 to reduce overfitting. Additionally, the model has 500 boosting rounds and stops training early if there are no improvements for 20 consecutive rounds.

Lasso Logistic Regression (LR):

LR is a linear model used for binary classification. It predicts the outcome using a linear combination of features. This method uses the sigmoid, or logistic function, to convert the log odds into probabilities between 0 and 1. LR is an interpretable and straightforward model. Logistic regression is sensitive to multicollinearity and outliers and struggles with complex and nonlinear relationships. Lasso regression was applied to improve the model and reduce overfitting.

The glmnet() function is used for this model, and the family is set to binomial, ensuring that logistic transformation is applied. The alpha parameter corresponding to lasso regression is set to 1 to apply L1 regularization. This shrinks the less relevant coefficient to zero, improving variable selection and simplifying the model. Cross-validation was used to generalize the model on unseen data to identify the optimal lambda, which determines the strength of the penalty for all coefficients.

Neural Networks (NN):

Inspired by how the brain works, NN uses interconnected neurons to analyze and covert data to make predictions. The networks consist of various node layers - input, hidden, and output. This model uses activation functions such as nonlinearity and optimizes weights to identify complex patterns. This model is ideal for large amounts of data and, if not adequately regularized, can be prone to overfitting. It is also computationally expensive and needs to be more transparent in decision-making. However, NN is flexible, captures nonlinear relationships well, and handles structured and unstructured data (Kuhn & Johnson, 2018).

To model NN, use nnet package, to train the neural network; define the nu on the balance training dataset. The number of hidden node layers was set to 5, making the model more straightforward. L2 regularization was applied by setting the decay value to 0.2, which penalizes larger weights, reduces the chance that the model will memorize patterns, and prevents overfitting. Maxit, the maximum number of iterations, was limited to 150 to stop training before introducing noise. Also, the lineout was set to false to indicate classification.

Performance Metrics:

After using the test set to predict all the models, performance was evaluated using the confusion matrix in the caret package. This matrix provides the metrics of True Positives (TP): correctly predicted class 1; True Negatives (TN): correctly predicted class 0; False Positives (FP): incorrectly predicted class 1; and False Negatives (FN): incorrectly predicted class 0.

The matrix also provides metrics of:

Accuracy - proportion of total correct,
Sensitivity – the percentage of correct, actual positive cases (also known as recall)
Specificity – the proportion of actual negative cases accurately identified
Precision – the proportion of correct predictive positive outcomes
F1 score - mean of precision and recall

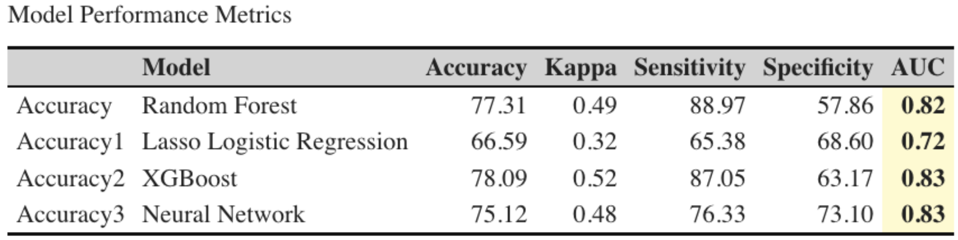
Use the pROC package to plot the ROC curve. This curve shows the tradeoff between correctly identifying the actual positives (sensitivity) and incorrectly identifying them (specificity). It can also show changes when the classification threshold is changed.

AUC measures the area under the ROC curve and the model's overall performance. This metric is also used to compare models; a higher value indicates that the model is better at differentiating between positive and negative classes.

Discussion of Results:

Round 1 of modeling: with length of stay variable:

This first round of modeling examined how the LOS variable affects the models. The models, including service length, perform well on the test set. XGBoost and Neural Networks performed the best, with an AUC of 0.83, and Random Forest followed with an AUC of 0.82. XGBoost had the highest accuracy and high sensitivity, capturing most of the true positives.



Model Performance Metrics

	Model	Accuracy	Kappa	Sensitivity	Specificity	AUC
Accuracy	Random Forest	77.31	0.49	88.97	57.86	**0.82**
Accuracy1	Lasso Logistic Regression	66.59	0.32	65.38	68.60	**0.72**
Accuracy2	XGBoost	78.09	0.52	87.05	63.17	**0.83**
Accuracy3	Neural Network	75.12	0.48	76.33	73.10	**0.83**

Figure 4: Model performance

Further analysis was done to observe the high-importance variables, specifically in the XGBoost model. Note that the NN, LR, and RF models also reflected this pattern of LOS's importance.

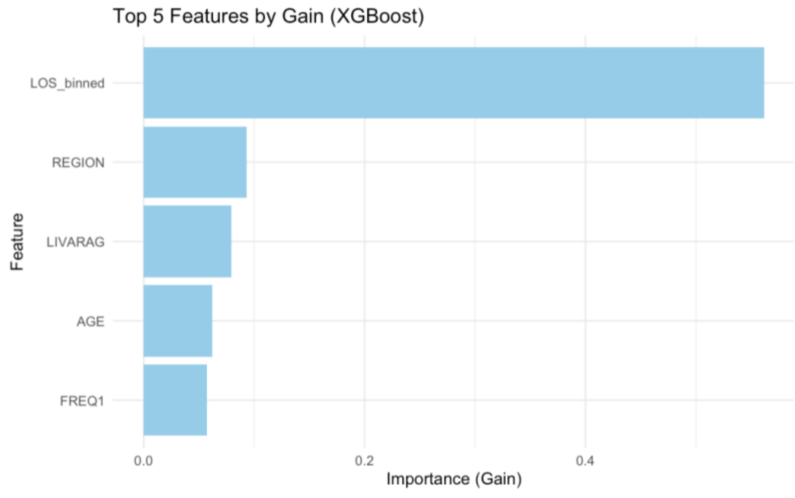


Figure 5 Top Features from XGBoost

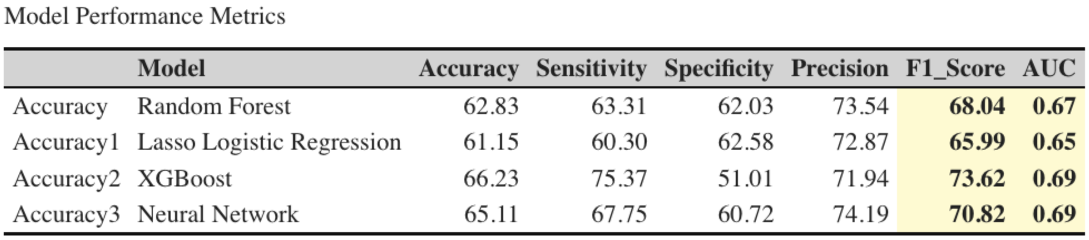
The Length of service variable seems more influential than other features. The dominance of this feature might reduce the influence of different variables in the model. Specifically, the length of service between days 7 to 30 drove the importance. The LOS variable likely predicted the outcome (dropout or completion). This seems likely as some rehab programs might vary from 7 days to 14 days to 28 days. Additionally, since this variable is calculated simultaneously as the treatment outcome, excluding it from the next modeling round is best.

In their predictive modeling, Gautam and Singh (2020) kept the length of service variable and found it to be the most crucial variable for predicting dropout. Additionally, their paper includes a visual of the mean decrease gini (top features for RF), showing that the service length was twice as significant as other variables. Although they used this variable, they selected all the service settings in their analysis, which differs slightly from this project.

Round 2 of modeling:

Ten predictive variables were used to develop these models, identifying factors predicting drop-out rates in inpatient substance treatment programs. This round serves as the key result.

The metrics below show the model performance on the test dataset. The area under the curve, which measures how well a model distinguishes between positive and negative outcomes across thresholds, shows that XGBoost and NN perform best. At the same time, RF and LR have comparable results. XGBoost has the highest F1 score of 73.62, highlighting its ability to balance predicting true positives and minimizing false negatives.



Model Performance Metrics

	Model	Accuracy	Sensitivity	Specificity	Precision	F1_Score	AUC
Accuracy	Random Forest	62.83	63.31	62.03	73.54	**68.04**	**0.67**
Accuracy1	Lasso Logistic Regression	61.15	60.30	62.58	72.87	**65.99**	**0.65**
Accuracy2	XGBoost	66.23	75.37	51.01	71.94	**73.62**	**0.69**
Accuracy3	Neural Network	65.11	67.75	60.72	74.19	**70.82**	**0.69**

Figure 6 Model Performance on test data

The ROC curve below visualizes the tradeoff between sensitivity and specificity. The straight line through the middle represents a random guess model. All models have relatively close curves, suggesting again that the results do not drastically differ from the model. The closer the curve is to the top left corner, the better the model performs; XGBoost performs the best (the exact metrics for AUC), and NN overlaps, making it hard to see the purple curve.

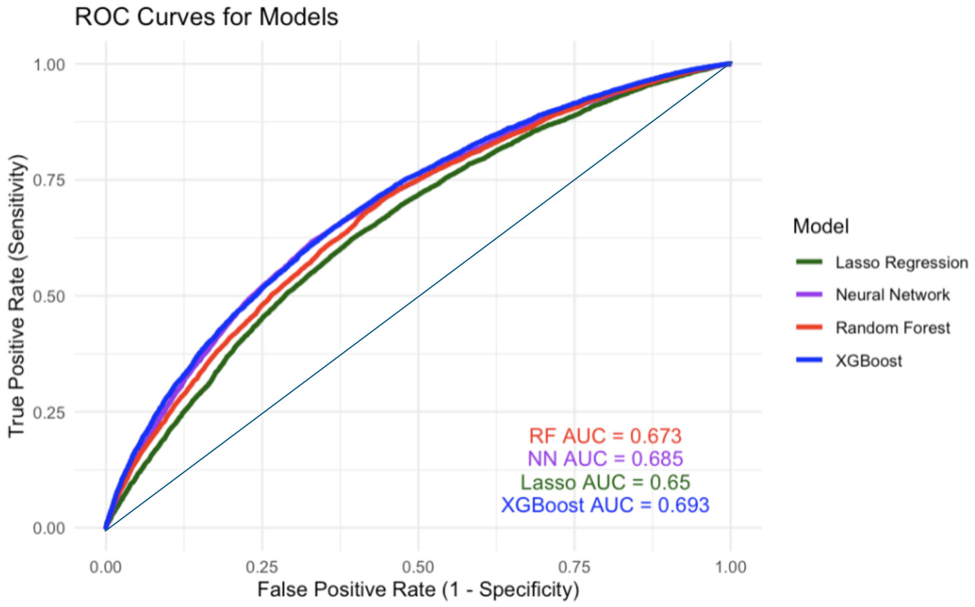
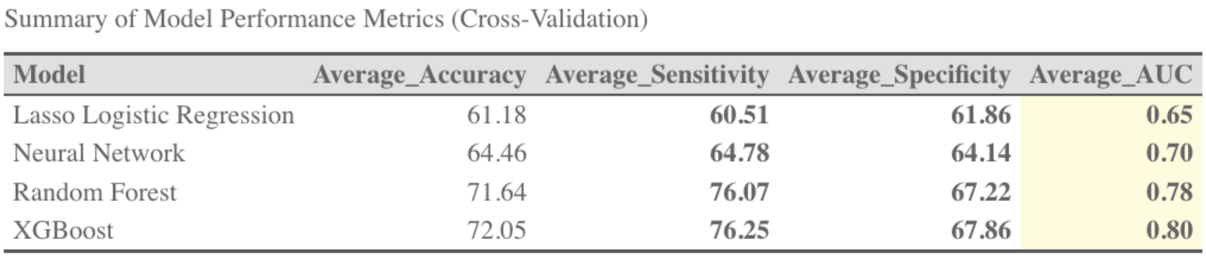


Figure 7 ROC Curve for all models

Cross-validation also evaluated metrics to confirm the models' stability and generalization. The results show that XGBoost and RF have similar metrics for sensitivity and specificity. However, XGBoost's accuracy (72.05) and AUC (.80) suggest outperforming the other models across multiple validation folds.

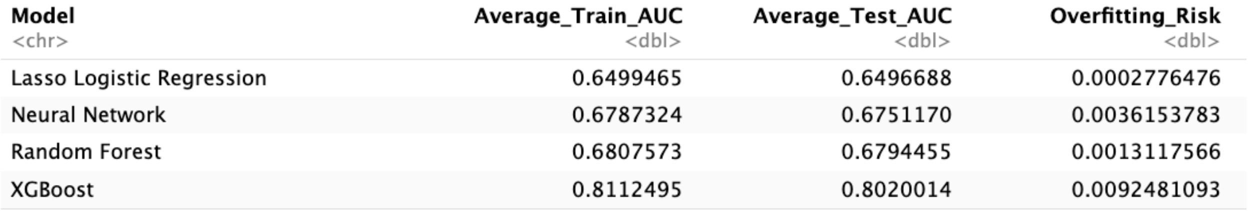


Summary of Model Performance Metrics (Cross-Validation)

Model	Average_Accuracy	Average_Sensitivity	Average_Specificity	Average_AUC
Lasso Logistic Regression	61.18	**60.51**	**61.86**	**0.65**
Neural Network	64.46	**64.78**	**64.14**	**0.70**
Random Forest	71.64	**76.07**	**67.22**	**0.78**
XGBoost	72.05	**76.25**	**67.86**	**0.80**

Figure 8 Model performance on cross-validation

Cross-validation was also used to confirm internal validity. The table below demonstrates that the difference between the average training AUC and the average test AUC, which represents the overfitting risk, is less than 0.05 for all models, indicating minimal overfitting.



Model <chr>	Average_Train_AUC <dbl>	Average_Test_AUC <dbl>	Overfitting_Risk <dbl>
Lasso Logistic Regression	0.6499465	0.6496688	0.0002776476
Neural Network	0.6787324	0.6751170	0.0036153783
Random Forest	0.6807573	0.6794455	0.0013117566
XGBoost	0.8112495	0.8020014	0.0092481093

Figure 9 Train and Test AUC with Cross Validation

The bar chart below shows the top five variables in the XGBoost model, using a gain metric that represents how much a feature contributes to the model's performance. The performance metrics on the x-axis are scaled in importance, and the chart shows that the frequency of attendance at self-help programs is the most important, followed by employment, age, region, and frequency of use.

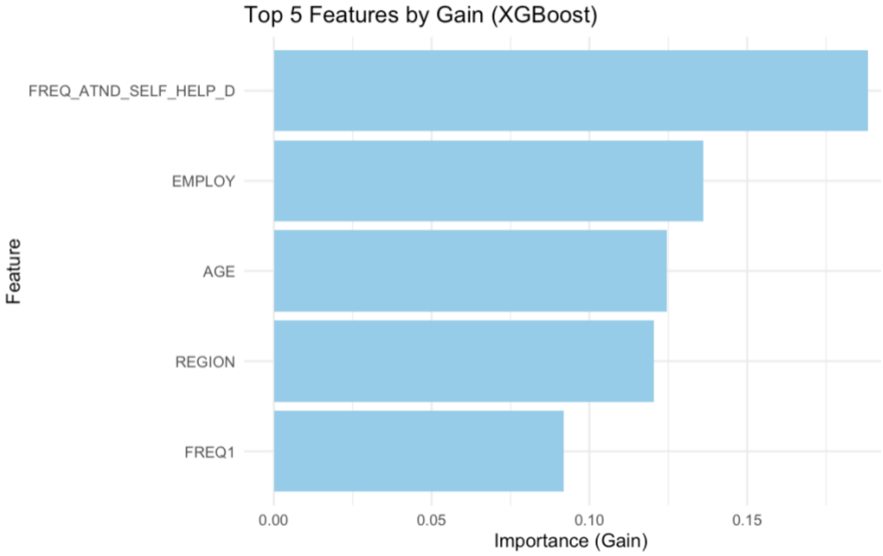


Figure 10 Top features from XGBoost

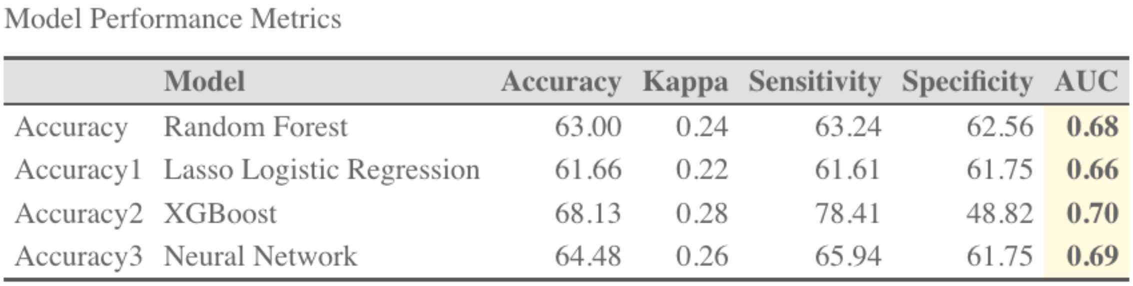
Given the comparable performance across models, analyzing the top predictive variables identified by each might offer insights. LM and RF also found that the frequency of attendance at self-help programs was of the highest importance (identified as third highest for NN). All models found employment and age to be among the top 5 features. The region was recognized as the top predictor in all models besides LM. Frequency of use was not a top 5 feature for the other models. Instead, several models identified living arrangements as the top feature, NN and RF. Additionally, NN and RF identified DSM diagnosis as an essential variable. The referral source was a top variable in LR and RF models.

Removing the Length of Service variable (from round 1 of modeling) allows the models to rely on other predictive variables. It provides a distinction in model performance, with AUC decreasing over 10% on the second round of modeling, meaning the models relied too heavily on that LOS variable.

Round 3 of modeling 2020 data:

This round aims to prove external validity. The same ten predictive variables were used to develop these models. TEDS 2020 assessed how generalizable the models in round 2 are to new data from the previous year.

The models performed similarly on the new data. Once again, all models were comparable, with XGBoost slightly outperforming the others with an AUC of 0.70. Based on the test data, NN was the next best model.

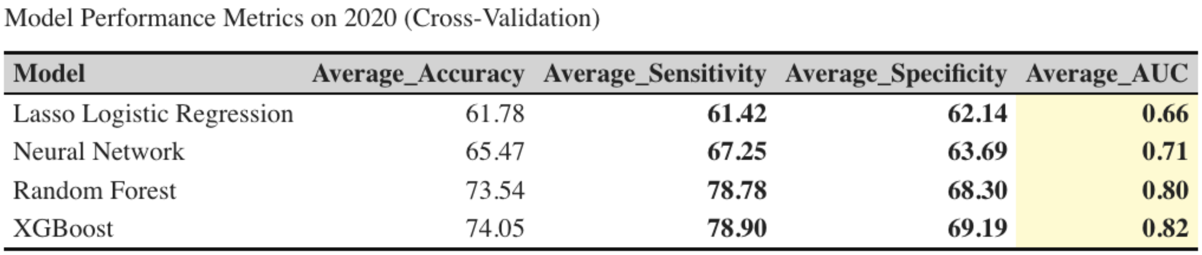


Model Performance Metrics

	Model	Accuracy	Kappa	Sensitivity	Specificity	AUC
Accuracy	Random Forest	63.00	0.24	63.24	62.56	**0.68**
Accuracy1	Lasso Logistic Regression	61.66	0.22	61.61	61.75	**0.66**
Accuracy2	XGBoost	68.13	0.28	78.41	48.82	**0.70**
Accuracy3	Neural Network	64.48	0.26	65.94	61.75	**0.69**

Figure 11 Model performance on test set

The cross-validation metrics also confirmed that the XGBoost model's performance was consistent across many folds. Similarly to the previous CV metrics, RF was the next best model.



Model Performance Metrics on 2020 (Cross-Validation)

Model	Average_Accuracy	Average_Sensitivity	Average_Specificity	Average_AUC
Lasso Logistic Regression	61.78	**61.42**	**62.14**	**0.66**
Neural Network	65.47	**67.25**	**63.69**	**0.71**
Random Forest	73.54	**78.78**	**68.30**	**0.80**
XGBoost	74.05	**78.90**	**69.19**	**0.82**

Figure 12 Model performance on cross-validation

The top variables found for XGBoost in 2020 were similar to those in the 2021 data. Frequency of attendance at self-help programs was the most important variable, followed by employment, age, and region. Unlike the 2021 results, the referral source was also among the top five features of the XGBoost model. Referral source was also a top feature in the RF model of the 2020 data.

Note that several more admissions for inpatient treatment were made in 2020 (compared to 2021), and slightly fewer dropouts occurred; this did not impact the model's performance, but it was a factor considering it occurred during COVID-19.

The goal of round 3 of modeling was to assess external validity. This project focused on developing models to identify factors that predict drop-out rates in inpatient substance treatment programs using 2021 data (most recent dataset). The XGBoost performed the best in this project, and the results from the 2021 data were validated using the 2020 data (which also found XGBoost to be a top model). The results from the 2020 data also showed that referral source was a crucial feature in the modeling; Acion et al. (2017) also found this to be the case.

Very few researchers in the literature modeled their predictions with XGBoost. The only research where XGBoost was mentioned was by Baird, Cheng, and Xia (2022), where they found that the virtual twins model (random forest + decision tree) performed better. Acion et al.(2017), with their 17 algorithm configurations, might have included the XGBoost model, but it needed to be noted in their research. Most past studies have found that Random Forest performed the best compared to other models. This is not of significant concern as RF performed very well in this Project.

Additionally, none of the previous research examined inpatient rehab treatment status specifically, which could be why they did not find the frequency of attending self-help programs as a top predictor. If researchers used this variable, they selected the frequency of attendance at admission, which caused this variable to be less of an important feature. Based on the findings of this project, patients who had a higher frequency of attendance at substance use self-help programs would be more likely to complete treatment than those who did not attend self-help.

Like the results of this project, Gautam and Singh (2020), Acion et al. (2017), and Baird, Cheng, and Xia (2022) found that employment status was a top predictor of treatment outcomes. Those unemployed or not in the workforce likely had higher dropout rates than those with part-time or full-time employment and a higher incentive to complete treatment and return to work.

Additionally, Gautam and Singh (2020) found that the region where a patient lives is a top predictor of treatment outcome. The results of this project support their findings, and rural areas need more resources and funding for programs impacting treatment outcomes. Similar to the findings of this project, Gautam and Singh (2020) and Baird, Cheng, and Xia (2022) found age as an essential feature in treatment outcomes. Acion et al. (2017) also found frequency of use to be a significant predictor of treatment outcome.

Conclusion:

This study also focuses on comparing machine learning models for predictive performance. The following models were evaluated: Random Forest, Neural Networks, Logistic Regression, and XGBoost. While existing literature found that Random Forest performed better, this study found that XGBoost proved to be the most effective, achieving the highest F1 and test AUC (0.70) and outperforming the other models in balancing sensitivity and specificity. Cross-validation found that the average AUC of XGBoost was 0.80. Models were also assessed for overfitting, and the results were minimal. Modeling was also computed based on data from the previous year, and the result confirmed external validity.

The goal of this project was to assess if social determinants of health, substance use severity, and choice in treatment contributed to dropout rates of inpatient treatment for addiction. Four variables in the social determinants of health category were found to be critical features: employment, region, and age. Frequency of use was found to be an essential variable in the substance use severity category. Participation in self-help programs was a crucial feature in the choice of treatment category. These results were from the XGBoost model, which performed the best, and suggest that some factors within these three categories: social determinants of health, substance use severity, and choice in treatment do contribute to dropout rates of inpatient treatment for addiction.

These findings partially align with existing literature, specifically identifying employment, region, age, and frequency of use as significant predictors. The ideal application of this study was to identify factors that led to dropout rates of inpatient substance treatment, to address them, or to focus on these groups so that more individuals would complete treatment and not be in active addiction. For example, more funding could be provided for regions with higher dropout rates. Further research could be done on what age-appropriate interventions and age-specific needs are needed to improve treatment outcomes. Individuals with a higher frequency of use could be enrolled in more structured care or added support such as frequency of counseling.

Unlike other research, this project identifies self-help program attendance as the most significant predictor for inpatient settings. Self-help programs like Narcotics Anonymous and Alcoholics Anonymous are of no cost; therefore, more significant efforts to arrange more meetings at rehab treatments and encourage participation would have an impact on treatment outcomes. Further research is needed to examine how often substance use self-help programs are offered in rehabs across the U.S. and if attendance at these programs is optional or required at rehab centers.

There is minimal existing literature on this topic in general and specifically on using TEDS data, which is focused only on inpatient rehab service settings. Some researchers mentioned evaluating models for all service settings, while others concentrated on outpatient settings. Further research is needed to improve treatment outcomes in this setting. The TEDS dataset also has limitations in collecting data from public sources nationwide and does not include data on private treatment settings, which could enhance generalizability. Another limitation is that each observation is not a person but an admission episode, meaning that one person could have been admitted and discharged several times in a year, resulting in several observations for one person. (Substance Abuse and Mental Health Services Administration [SAMHSA], 2023).

Another limitation of this study is that even if more people completed treatment, there is no guarantee that completion of inpatient programs or any service setting of treatment would result in success. There often is no post-treatment data regarding whether patients who completed treatment continued to abstain from substances.

Addiction is a public health issue, as millions of people suffer from substance use disorders. More individuals could complete treatment if targeted funding and improvements were implemented in treatment, particularly in areas identified as the top predictors. By prioritizing these changes, not only would individuals in treatment benefit, but this would also result in societal benefits, including reduced healthcare and legal costs.

References

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

Andersson, H. W., Steinsbekk, A., Walderhaug, E., Otterholt, E., & Nordfjærn, T. (2018). Predictors of dropout from inpatient substance use treatment: A prospective cohort study. *Substance Abuse: Research and Treatment, 12,* 1–10. https://doi.org/10.1177/1177872718786108

Baird, A., Cheng, Y., & Xia, Y. (2022). Use of machine learning to examine disparities in completion of substance use disorder treatment. *PLoS ONE, 17*(9), e0275054. https://doi.org/10.1371/journal.pone.0275054

Centers for Disease Control and Prevention. (2024, August 12). *Opioid use disorder: Treatment and prevention of opioid overdose*. https://www.cdc.gov/overdose-prevention/treatment/opioid-use-disorder.html

Crapanzano, K. A., Hammarlund, R., Ahmad, B., Hunsinger, N., & Kullar, R. (2018). The association between perceived stigma and substance use disorder treatment outcomes: A review. *Substance Abuse: Research and Treatment, 10,* 1-12. https://doi.org/10.2147/SAR.S183252

Gautam, P., & Singh, P. (2020). A machine learning approach to identify socio-economic factors responsible for patients dropping out of substance abuse treatment. *American Journal of Public Health Research, 8*(5), 140-146. https://doi.org/10.12691/ajphr-8-5-2

Gottlieb, A., Yatsco, A., Bakos-Block, C., Langabeer, J. R., & Champagne-Langabeer, T. (2022). Machine learning for predicting risk of early dropout in a recovery program for opioid use disorder. *Healthcare, 10*(2), 223. https://doi.org/10.3390/healthcare10020223

Kariisa, M., O'Donnell, J., Kumar, S., Mattson, C. L., & Goldberger, B. A. (2023). Illicitly manufactured fentanyl–involved overdose deaths with detected xylazine — United States, January 2019–June 2022. *MMWR Morbidity and Mortality Weekly Report*, *72*(26), 721–727. https://doi.org/10.15585/mmwr.mm7226a4

Substance Abuse and Mental Health Services Administration. (2022). *2022 National Survey on Drug Use and Health (NSDUH)*. U.S. Department of Health and Human Services. https://www.samhsa.gov/data/sites/default/files/reports/rpt42731/2022-nsduh-nnr.pdf

Substance Abuse and Mental Health Services Administration. (2023). *Treatment episode data set discharges (TEDS-D) 2021: Public use file (PUF) codebook*. Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. https://www.samhsa.gov/data/

Volkow, N. D., & Blanco, C. (2021). The changing opioid crisis: Development, challenges and opportunities. *Molecular Psychiatry, 26*(1), 218-233. https://doi.org/10.1038/s41380-020-0661-4