Merck Datathon 2023: Investigating Dropout Rate in Substance Abuse Treatments

Team 5

May 21, 2023

1 Non-Technical Executive Summary

1.1 Introduction

Substance abuse is a common issue nowadays which may lead to many social, mental, and physical problems. Various treatment services are available to help resolve or reduce these problems. In the surveyed treatment data set, we observe that nearly one in four of the admissions end with the patient dropping out of the treatment. Dropping out of treatment is a significant problem since it directly relates to the likelihood of relapse or resuming substance use. It is crucial to comprehend the patterns that drive the rates of treatment drop-outs in order to implement the necessary enhancements and improve the outcome of treatment programs. In this study, we aim to determine the average rates of treatment dropouts and identify factors that predict dropouts from a data perspective, with the objective of reducing these rates in the future.

We make use of the treatments and facilities data sets provided by the challenge. We first examine the data set from a broad perspective and analyze the treatment dropout rates by different demographic and social factors. Given the high dimensionality of the data sets, we focus the analysis on factors that have a significant relationship with the dropout target. Factors that are highly correlated with or strong predictors of patients dropping out of treatment are analyzed by correlational methods and causal graphs to determine causality. To predict future potential dropout events, we build a decision tree classification model to predict whether dropouts will happen and identify factors that are strong predictors of treatment dropout. Furthermore, we perform the survival analysis to predict the dropout rate for different treatment services.

1.2 Primary Findings

In the SAMHSA TEDS-D data sets, around 1 million cases end up dropping out of treatment among a total of 4 million treatments from 2017-2020. We explore the data sets to reveal dropout trends from four perspectives: patient demographics, geographic locations, substance abuses pattern, and the treatment service angle.

Firstly, at the patient level, we find people in the following condition have a higher probability to drop out from treatments: aged 18-20 years old, female, single, other single race (Hispanic), relying on public assistance as the primary income, with education less than 8th-grade level, with Medicaid insurance, currently not in the labor force, and have independent living arrangements and no arrests in the past month before admission. Pregnant women, non-veteran, patients without co-occurring mental and substance use disorders, and patients receiving treatment services for free are at higher risks to drop out of treatments than their counterparts.

Secondly, at the state level, patients in Georgia, Louisiana, Wisconsin, and Michigan have higher dropout rates. At Core-Based Statistical Area (CBSA) level, patients in Flint (MI), Fresno (CA), and Los Angeles-Long Beach-Anaheim (CA) have over 50% dropout treatments and these three CBSAs are the top 3 areas with the highest dropout rate among CBSAs with over 5,000 treatment cases between 2017-2020.

Thirdly, we find that alcohol, heroin, and marijuana/hashish are the top 3 substances with the most dropout events. Marijuana/hashish, heroin, and cocaine/crack are the top 3 substances with the highest probability to drop out of treatment. The route of substance administration also plays a role

in the dropout rate with smoking leading the dropout rate, followed by injection and inhalation routes of administration. Interestingly, people who use two drugs have a higher probability than those who use three drugs.

Lastly, from the treatment service perspective, patients who receive ambulatory non-intensive services, those who have zero attendance at substance use self-help groups in the past 30 days prior to discharge, and those who have at least one prior treatment have a higher dropout rate than other treatment services. Patients who have 8 to 14 days waiting to enter substance use treatments have the highest dropout rate, interestingly, and patients who wait 0 days have the lowest dropout rate. Patients referred by schools have the highest probability of dropout, while referrals from the court/criminal justice have the lowest probability of dropping out of treatments.

Causal inference is conducted to reveal causal relationships between various factors and treatment dropout. In general, length of stay, time spent waiting for treatment, and location are likely factors in dropout outcomes. For rehab, living arrangements, location, and being arrested have causal relationships with dropping out of treatment. Encouraging more frequent attendance of self-help groups in those who already attend may help prevent dropping out. Next, the classification model reveals that the location of the patient, treatment service type, and waiting time are strong predictors of treatment dropout. In addition, we find different treatment services have different survival rates for dropout outcomes.

Based on our findings, we recommend the following:

1. Since young patients of ages 12-29 have the highest dropout rate among all ages, we recommend treatment facilities offer specialized treatments for these young patients, especially treatments targeted at adolescents.

2. Building needed facilities to provide accessible service to reduce waiting time at early states.

3. Reduce homeless and create more self-help groups for the population to decrease the dropout.

These policies and solutions will influence treatment success positively and prevent substance abuse relapse.

In conclusion, our studies help identify patients at risk of dropping out and explain the risk factors that led to its prediction, and suggest policy changes to address patient dropouts.

2 Technical Exposition

2.1 Preliminary Investigation

In the data set, we see many reported values are missing or not applicable. Rows with missing values cannot simply be dropped outright, because while missing values make up a small portion of each column, the vast majority (99.97%) of rows have at least one missing value. In the treatment data, these values appear as -9. Many library functions can drop missing values that are relevant to one column, so a function was written to replace the -9 values with 'None'. In all the bar plots, we denote -9 by missing.

We investigate the characteristics of the patients who have higher dropout rates through four lenses: patient demographics, geographic locations, substance abuses pattern, and treatment services.

2.1.1 Demographics

Understanding which group of patients is more likely to drop out of treatment is essential to prevent dropout. We identify the following demographic information of patients with higher rates: age, race, ethnicity, education, gender, marital status, income source, and employment.

In Figure B.1, we show that young adults from ages 18-20 have the highest dropout rate. Overall, patients of ages 12-29 have a higher dropout rate. In terms of education, the figure demonstrates that as the education level increases, the dropout rate decreases.

Figure B.2 shows that among common racial groups, the "other single" group has the highest dropout rate. Among the ethnicity of the patients, Mexicans have the highest dropout rate.

Figure B.3 shows that among each group of marital status, missing value and the never-married groups have the highest dropout rate. Similarly, the missing value group has the highest dropout rate

compared to females and males, while females are more likely to drop out than males. As expected, Figure B.4 shows being pregnant corresponds to a higher dropout rate. One caveat is that male patients with lower dropout rates belong to the non-pregnant group. We also examine whether veteran status affects the dropout rate. As shown in the figure, veteran status has a weak impact on the dropout rate.

Figure B.5 shows that patients who rely on public assistance as a primary income source have the highest dropout rate among all groups. This is consistent with the dropout rate distribution among employment groups. Patients who are not in the labor force have the highest dropout rate. Note that here unemployed group is those who were laid off or looking for jobs.

Living styles and criminal history in the past 30 days prior to admission might also have an impact on the dropout rate as one would imagine. Figure B.6 shows that the effects of these two factors are not significant. Patients who live independently and do not have criminal history in the past 30 days have the highest dropout rate.

2.1.2 Geographic Locations

The dropout rates of substance abuse treatment can vary across different geographic areas. Factors such as availability and accessibility of treatment programs, funding and resources allocated to treatment services, cultural attitudes towards addiction and treatment, and variations in healthcare systems can contribute to the differences in treatment dropout rates. We have assessed the treatment dropout events and dropout rates at a state level and CBSA level to reveal areas that need more intervention.

We first explore the substance abuse trend in each state and find that New Jersey, New York, California, and North Carolina are leading in the total number of treatment events. However, patients in Georgia, Louisiana, Wisconsin, and Michigan have higher dropout rates. Meanwhile, the evaluation of the treatment facilities data sets from 2016-2020 reveals that California, New York, and Florida have the most treatment facilities and also have provided the greatest quantity of treatment services including assessment, testing, transition, ancillary, and other services.

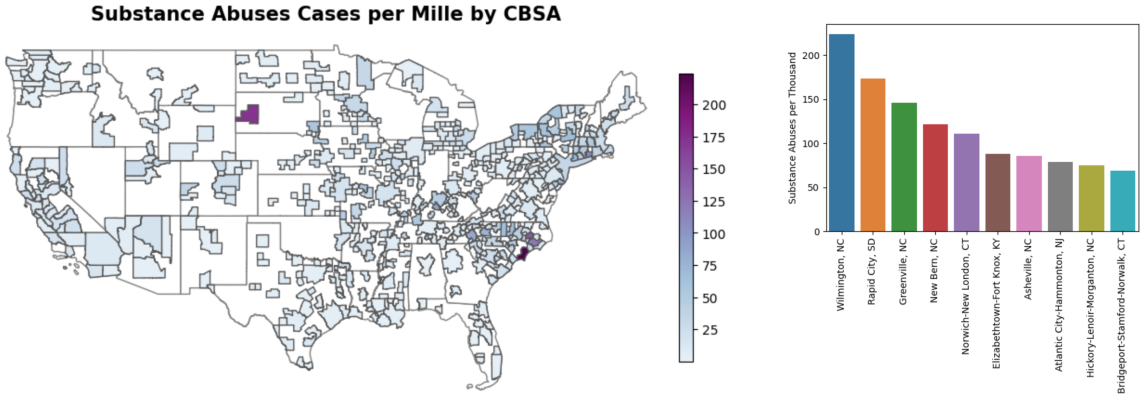

Figure 1: Substance Abuse Cases per Mille at CBSA level (left: map, right: top 10 CBSAs).

Using the SAMHSA TEDS-D data sets, we then evaluate the treatment dropouts at Core-based Statistical Area (CBSA) level and find that Wilmington, NC, Rapid City, SD, and Greensville, NC have the highest number of substance abuse cases per thousand people (Figure 1). Some CBSAs have less than 1000 treatment cases from 2017-2020 and have shown a dropout rate of above 80%. The low number of cases may be caused by state agency policies and data collection bias, therefore are excluded from the analysis. We focus on CBSA with over 5000 treatment cases and identify that patients in Flint (MI), Fresno (CA), and Los Angeles-Long Beach-Anaheim (CA) have over 50% dropout treatments and these three CBSAs are the top 3 areas with the highest dropout rate among CBSAs with over 5,000 treatment cases between 2017-2020 (Figure 2).

2.1.3 Primary Substance Use

The dropout rate in substance abuse treatment can vary depending on the specific substance being treated. Factors such as the addictive nature of the substance, the severity of the addiction, the

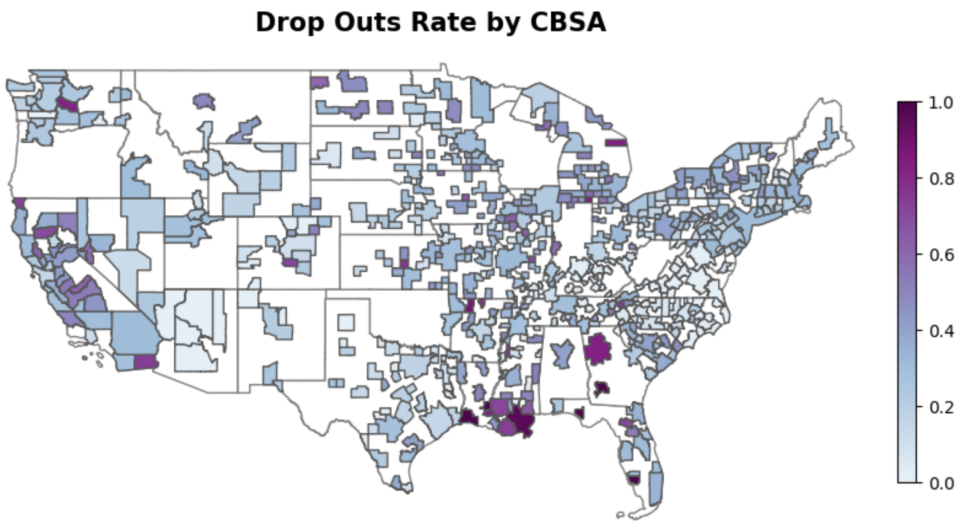

Figure 2: Treatment Dropout Rates at CBSA level.

available treatment options, and individual characteristics can influence the likelihood of dropout from treatment. By understanding the dropout rate for different substances, efforts can be made to improve treatment outcomes and provide more effective support for individuals seeking recovery.

There are 18 different types of substances recorded in the SAMHSA TEDS-D data sets, we find the most popular substances are alcohol, heroin, and marijuana/hashish. Interestingly, these 3 substances are the top 3 with the highest dropout events. The large quantity of dropout cases may be due to the fact that these 3 substances have a large base audience. When assessing the dropout rates for different substances, we identify that marijuana/hashish, heroin, and cocaine/crack are the top 3 with the highest probability to drop out of treatment (Figure B.7). The route of substance administration also plays a role in the dropout rate with smoking leading the dropout rate, followed by injection and inhalation routes of administration (Figure B.8). Interestingly, people who use two drugs have a higher probability than those who use three drugs (Figure B.9).

2.1.4 Treatment Service

Previously, we have found that people who depend on public assistance as a primary income source have the highest dropout rate. Figure B.10 provided additional support. Patients with Medicaid or without insurance are more likely to drop out. Interestingly, the figure also shows that when the treatment is free, the dropout rate is highest.

Attending a self-help group is a common and useful way to address substance abuse. Figure B.11 shows that although missing value and unknown frequency groups have higher dropout rates among all groups, we can see clearly that the dropout rate decreases as the frequency of attending self-help groups increases. The type of treatment services at admission can also have an impact on the dropout rate. The data shows that ambulatory, non-intensive treatment and ambulatory, intensive outpatient, and long-term rehab groups are most likely to drop out.

Patients referred by schools have the highest probability of dropout, while referrals from the court/criminal justice have the lowest probability of dropping out of treatments (Figure B.12). Among referrals by the court justice, patients on "Probation/parole" have the highest dropout, while those on "DUI/DWI" have the lowest dropout (Figure B.13).

2.2 Correlation Matrix

Two metrics of correlation were used to evaluate the relation between the variables and a positive result of dropout. The first metric was Cramer's V (Figure 3), which is a value between 0 and 1 based on the chi square metric. A larger Cramer value indicates association between the attributes, based on the number of possible values the attribute can take.

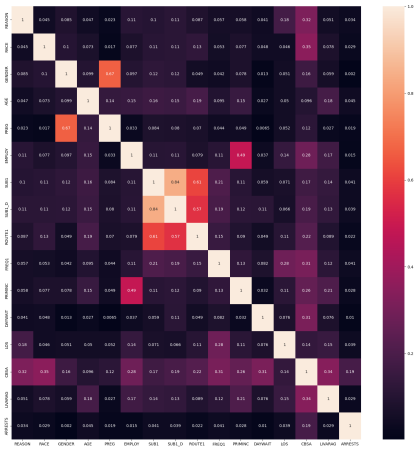

Figure 3: Cramer's V metric of correlation between categorical variables for treatment data.

All	V	Detox	V	Rehab	V	Ambulatory	V
CBSA	0.320	CBSA	0.324	CBSA	0.327	CBSA	0.341
STFIPS	0.303	STFIPS	0.298	STFIPS	0.302	STFIPS	0.323
REGION	0.213	REGION	0.239	FREQ1_D	0.201	FREQ1_D	0.251
DIVISION	0.206	DIVISION	0.231	LOS	0.197	ARRESTS_D	0.231
ARRESTS_D	0.195	DETCRIM	0.177	FREQ3_D	0.184	LOS	0.230
FREQ1_D	0.192	HERFLG	0.169	DIVISION	0.182	REGION	0.230
LOS	0.180	MTHAMFLG	0.164	REGION	0.179	DIVISION	0.222
FREQ3_D	0.175	LOS	0.159	NOPRIOR	0.149	DETNLF_D	0.203

Table 1: Cramer's V metric associated with each variable for different treatments, top 8 sorted descending.

In many instances, such as for the classifier, one-hot encoding was applied to the categorical attributes, and the drop out reason was treated as a binary variable. The second metric used Pearson's correlation coefficient. This is useful as it allows signed identification of which particular values for an attribute explain dropouts. It further identified missing values that have a large effect on dropouts and needed to be adjusted for.

2.3 Causal Inference

Causal links in the data set were first examined using the PC (Peter-Clark) algorithm. The PC algorithm identifies causal relationships by starting with the graph of all undirected connections between variables. It then uses statistical independence tests on the data set and conditional independence rules to reduce the graph to only its causal relationships. In this case, due to the majority of variables being nominal, the chi squared test was used to evaluate independence.

The PC algorithm is limited by its time complexity. As the number of attributes increases, the number of possible graphs is $3^{\frac{1}{2}(n-1)n}$, which grows faster than exponentially. For this data set, tradeoffs must be made between number of rows and number of variables included. Because a small sample of the data is used, random stratified sampling is used to preserve the frequency of key population variables. This includes age, gender, race. For the general analysis, it was also stratified on services (detox, rehab, ambulatory).

The run time of the algorithm can be further reduced by making use of prior knowledge. In the treatments set, some attributes are measured at certain intervals. These attributes are sorted into categories based on chronological order. Attributes in later categories can be affected by those in earlier categories, but not vice versa.

- The first category includes factors such as the substances, substance routes, veteran status, primary income source, and others occur before treatment and as such cannot be causally affected by attributes that occur afterward.

Attribute (all)	Pearson Correlation	Attribute (detox)	Pearson Correlation
cbsatitle_Anaheim, CA	0.094	HERFLG	0.158
DAYWAIT	0.092	SUB1_5	0.152
REGION_1	0.089	IDU_1	0.139
DIVISION_2	0.084	ROUTE1_4	0.135
statename_California	0.080	FREQ1_3	0.128
DIVISION_9	0.072	SUB1_D_5	0.127
METHUSE_1	0.071	cbsatitle_B-C-N, MA-NH	0.127
LOS	0.065	statename_NH, MA	0.127

Table 2: Top 8 variables of greatest positive Pearson correlation with REASON_2, for all services and detoxification. Subscript denotes the value of the variable. Correlation with missing values (-9) were excluded. REGION_1 refers to Northeast, DIVISION_2, 9 refers to Middle Atlantic, Pacific respectively, METHUSE_1 refers to a Yes for client uses opioid therapy. SUB1_5 refers to Heroin, IDU_1 refers to no IV drug use, ROUTE1_4 refers to Injection, FREQ1_3 refers to Daily use.

Attribute (rehab)	P Correlation	Attribute (ambulatory)	P Correlation
cbsatitle_Anaheim,CA	0.126	DIVISION_2	0.152
statename_Louisiana	0.078	statename_California	0.141
DAYWAIT	0.076	REGION_1	0.137
cbsatitle_NewOr-Metairie,LA	0.075	DAYWAIT	0.135
FREQ1_D_2	0.069	LOS	0.133
statename_Michigan	0.067	DIVISION_9	0.126
cbsatitle_DT_W_Dearborn,MI	0.059	cbsatitle_Anaheim,CA	0.111
LIVARAG_1	0.055	statename_New York	0.109

Table 3: Top 8 variables of greatest positive Pearson correlation with REASON_2, for rehab and ambulatory treatments. Subscript denotes the value of the variable. Correlation with missing values (-9) were excluded.

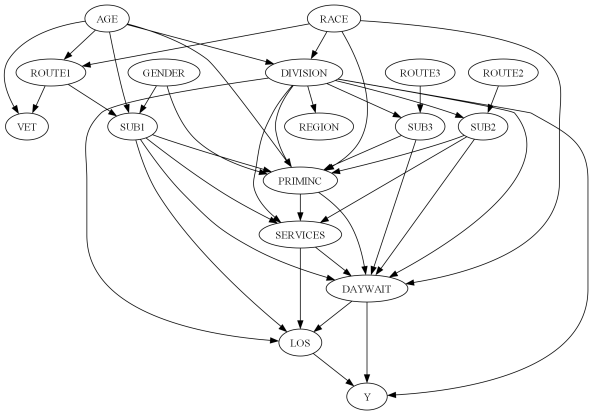

Figure 4: Causal graph for variables correlated with dropout in the general set

- The second category are the attributes that are decided at the beginning of treatment, such as the wait time.
- The third category are variables that occur during treatment, such as the length of stay or whether medication-assisted opioid therapy is applied. This includes all the variables in the data set that have the subscript 'D', which can be obtained up to a year later than the earlier data.
- Additionally, a few attributes are inherent characteristics that are not causally affected by any others. This includes age, gender, and race.

Having a map of causality is useful because it permits asking counterfactual questions: "What would happen if instead..?". In this case, we consider the effect of scenarios where attributes are fixed to certain values. This average causal effect in all cases was calculated with a drawing of 900,000 samples after the causal model was fit. It represents the change in expected value of the dropout variable, given an intervention on another variable. A positive/negative value represents a greater/lesser likelihood of dropping out respectively.

2.3.1 General Dataset

In the general model produced, dropout is directly affected by length of stay, days waiting, and division(figure 4). Dropout is not affected by veteran status or region. Correlation with region appears to be instead explained by division. It was expected that switching a patient from welfare to wages/salary might improve financial situation and lower dropout rate. However, this was not shown: The average treatment affect was 0.00115, which represents a very small increase in dropout rate. A possible interpretation of this is that the time commitment of full time work might make it difficult to spend time in treatment.

Reducing the wait time to receive treatment, in the beginning, is a plausible way to reduce general treatment dropouts (Figure 5). Reduction of wait time appears most effective in the beginning, and plateaus to non-effect as length increases.

2.3.2 Rehabilitation Treatments

In particular, it is important to understand the process in which a patient drops out of rehab, because it takes place over a longer period of time. We produce a model using the most correlated features (Figure 6).

Patients getting arrested is a cause of dropouts. The average causal affect of being arrested once is 0.0148, and being arrested more than once is 0.037. Policies that disincentivize patients from committing crimes is one way to address this. To do so, one might target related variables such as health insurance, primary pay source, and alcohol consumption.

Policies that reduce homelessness could be potent for keeping patients in treatment. Altering the living arrangement at discharge variable such that a homeless person receives a supervised residence affects dropout by -0.09, or 9 in 100. For an independent living space, the ACE is -0.06.

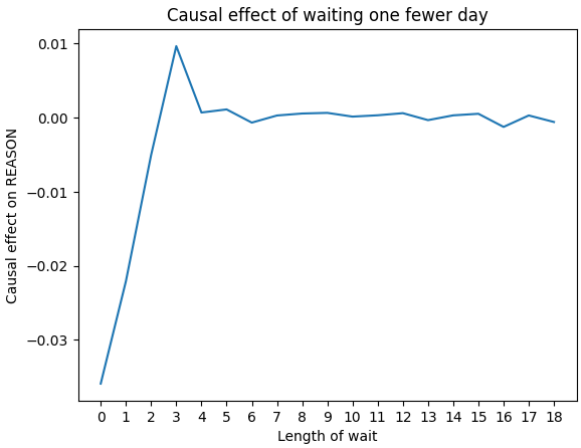

Figure 5: Causal effect of reducing wait time by one day, by overall wait time length.

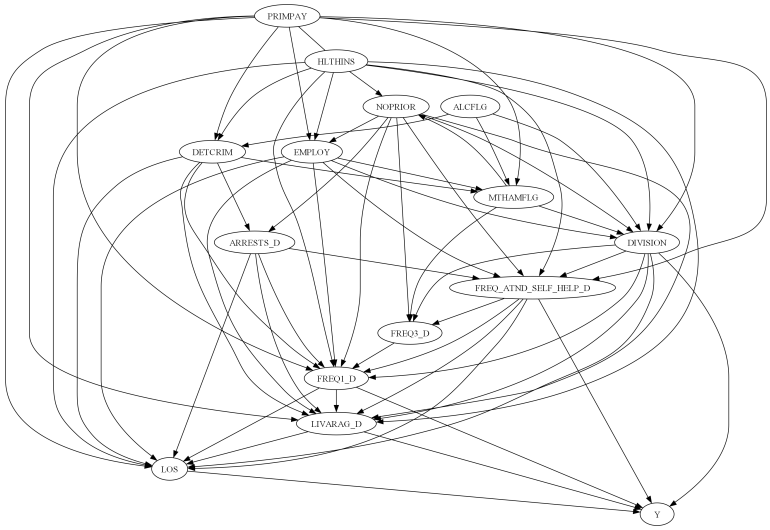

Figure 6: Causal graph for variables correlated with dropout of rehab treatments.

A further policy target is self-help groups. Small increases in attendance of those who already attend could diminish dropouts (Figure B.14). Attending '1-3' days instead of '0' does not appear to help, but if attendance is increased further from '0' to '4-7' days the ACE is -0.02.

2.4 Classification model to predict dropout event

Based on the studies on the factors that may affect the dropout cases in the previous sections, we here present a classification model in order to predict whether a patient may drop out of a treatment. Tree based model, namely "lightgbm", is used given its recognized good performance in classifications and interpretability.

The following variables are taken as input to the classifier: all substances used of the patient, route and frequency of the primary substance, frequency of self help activities, location of the patient, prior substance use, gender, financial status including employment, insurance and payment, treatment service type, service waiting time, and length of stay. Any variables about status at discharge are not included because they are not helpful in predicting whether a person may drop out of a treatment.

We conducted a few data prepossessing steps to keep as much useful data as possible for the modeling.

- Cases where discharge reason is unknown are discarded.
- For discharge reasons, create a new variable "Attrition" to mark dropout or non-dropout.

Importance rank	Attribute
1	LOS (Length of stay)
2	DIVISION_5 (located at South Atlantic)
3	SERVICES_2 (Detox free-standing)
4	DIVISION_9 (located at Pacific)
5	SERVICES_4 (Rehab, short term)
6	DAYWAIT_0 (0 days on waiting services)
7	DIVISION_1 (located at New England)
8	DIVISION_6 (located at East South Central)
9	SERVICES_7 (Ambulatory, non-intensive outpatient)
10	SERVICES_6 (Ambulatory, intensive outpatient)

Table 4: Dropout classifier top 10 importance factors.

- Added a count of the total number of substance use for a patient and make it as a new feature.
- For the variable which represents length of stay, we map it to actually days instead of using categorical values. The median value is taken for a given time range.
- For model training, categorical variables are converted to numerical items by splitting categories and use a binary value for each.

We noticed that about one fourth of the total cases are dropout cases. In the model training, the balance of dropout/non-dropout population was considered. We tested the training using the whole data set or balanced data set which means we sample a subset of the non-dropout cases. It turns out that using the whole data set gives better classification accuracy.

Training and test data are created in 70/30 of the whole data set. Cross validation is employed using the "StratifiedKFold" method. We finally get a classifier accuracy of 0.798.

The feature importance ranking can be found in Table 4. Overall, the location, length of stay, waiting time and the service type are most importance factors which may influence a dropout decision.

2.5 Survival Analysis

To understand the treatment dropout dynamics over time, survival analysis is performed to assess the survival probabilities and survival functions for the time-to-dropout outcome. We first evaluate the high-level survival for all treatment services together then later evaluated the survivals for different treatment service types including detox, rehab/residential, and ambulatory (Figure 7).

To prepare the data sets appropriately, we first transform the categorical variable 'REASON' into a binary variable to indicate whether dropout happens. The categorical variable 'length of stay' (LOS) has values from 1 - 37 with 1 - 30 represent the actual number of days which are left as is, and 31 - 36 represent different ranges of days which are transformed to the midpoint number of day by taking the average of lower limit and upper limit in the day ranges. For example, LOS value 31 represents 31 - 45 days which transformed to 38 days. A value of 37 in 'LOS' variable represents treatment length of more than a year which is excluded from the survival analysis with a purpose to investigate the time-to-dropout event within one year time frame. The treatment service variable 'SERVICES' is grouped into 3 subgroups: detox comprised of "Detox, 24-hour, hospital inpatient" and "Detox, 24-hour, free-standing residential", rehab/residential comprised of "Rehab/residential, hospital (non-detox)", "Rehab/residential, short term (30 days or fewer)", and "Rehab/residential, long term (more than 30 days)", ambulatory comprised of "Ambulatory, intensive outpatient", "Ambulatory, non-intensive outpatient" and "Ambulatory, detoxification".

We find the overall median survival time for time-to-dropout is 273 days (95% CI [273, 273]). The median survival time for detox, rehab, and ambulatory groups are 75.5 days (95% CI [75.5, 75.5]), more than a year, 273 days (95% CI [273, 273]) respectively. Ambulatory services have the highest survival rate at 6 months, followed rehab services and detox services. The survival rates between detox, rehab/residential, ambulatory service types are significantly different with each other groups by pairwise logrank test ($p < 0.0001$).

The survival analysis has limitations in terms of generalizability due to the lack of length of stay information at a more granular level. Our analysis focus on treatment services only, and there may be

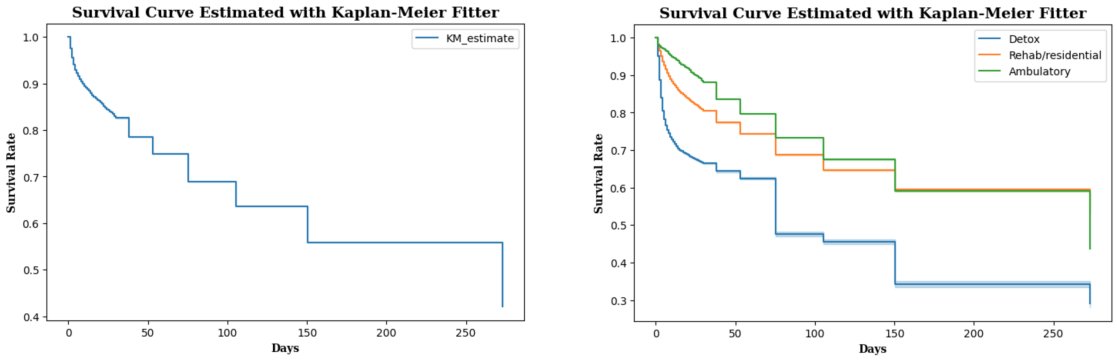

Figure 7: Overall Survival Curve (left) and Survival Curve by Services (right)

other relevant factors that were not included in the model which would benefit from a much lengthier investigation with more time. Therefore, caution should be exercised when applying these findings to broader populations or different settings.

3 Code

Please refer to the attached zip file for the scripts. The code is for demonstration of the analysis process.

References

A Limitations

The work done is limited by time and resource constraints. The causal framework is a cursory analysis which is assumed to be true for the sake of the exercise, but might require a much lengthier investigation with stronger computing resources in a professional setting. It should be noted the policy recommendations are contingent on the graph sufficiently modelling real-world effects.

Classification model is trained with a baseline model, and there is room for improvement with fine tuning on the model parameters.

B Figures

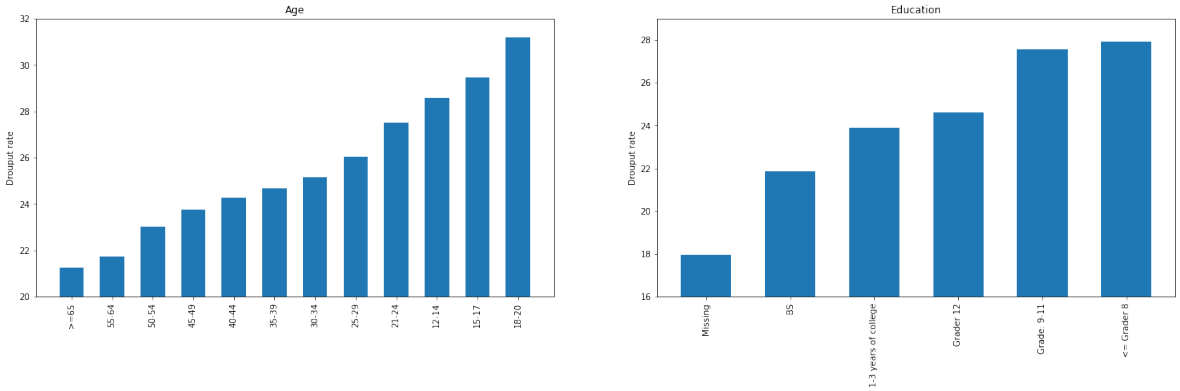

Figure B.1: Left: dropout rate of each age group. Right: dropout rate of each education group.

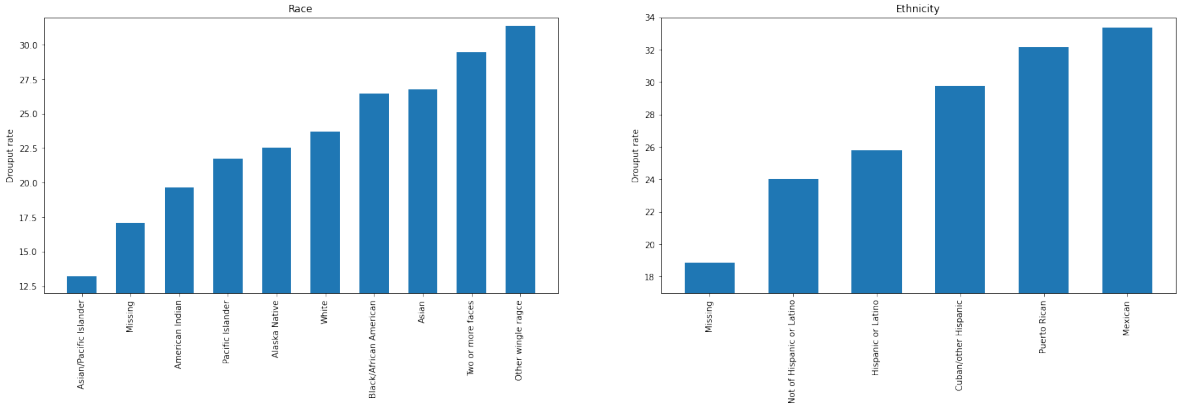

Figure B.2: Left: dropout rate of each racial group. Right: dropout rate of each ethnic group.

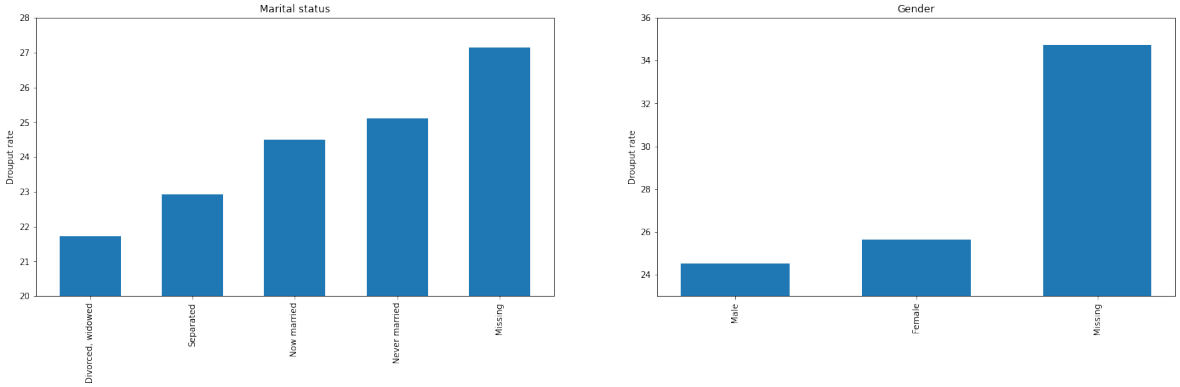

Figure B.3: Left: dropout rate by marital status. Right: dropout rate by gender.

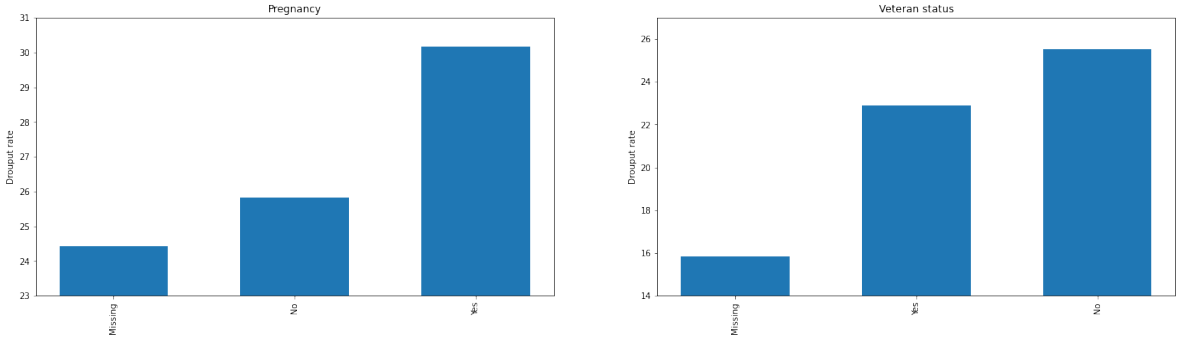

Figure B.4: Left: dropout rate by pregnancy status. Right: dropout rate by veteran status.

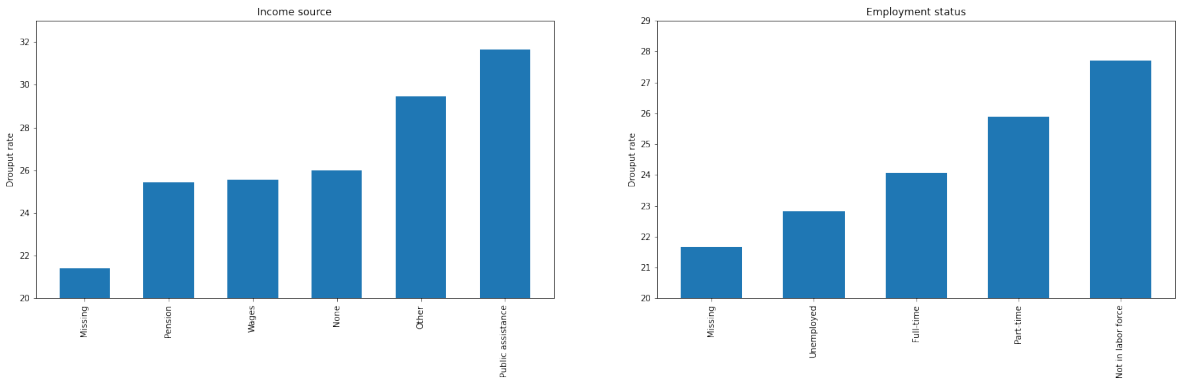

Figure B.5: Left: dropout rate by income source. Right: dropout rate by employment.

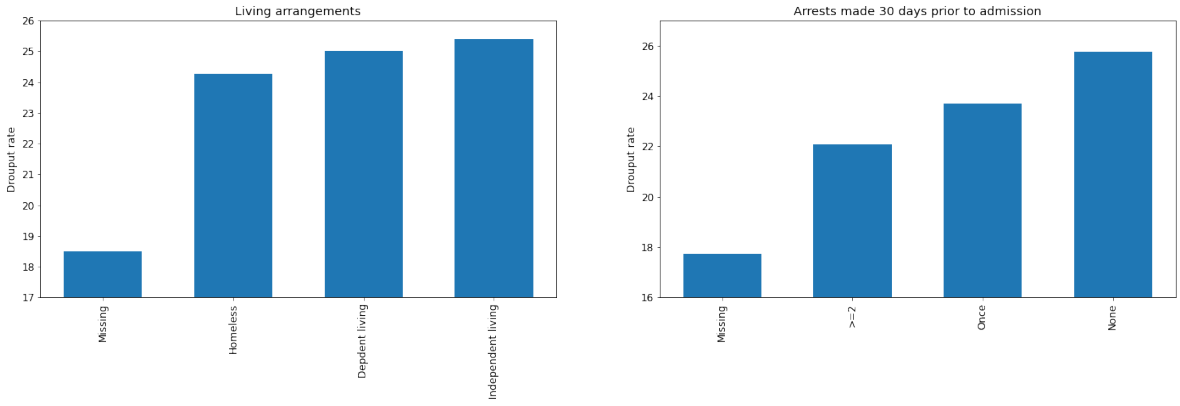

Figure B.6: Left: dropout rate by living arrangements. Right: dropout rate by arrest history.

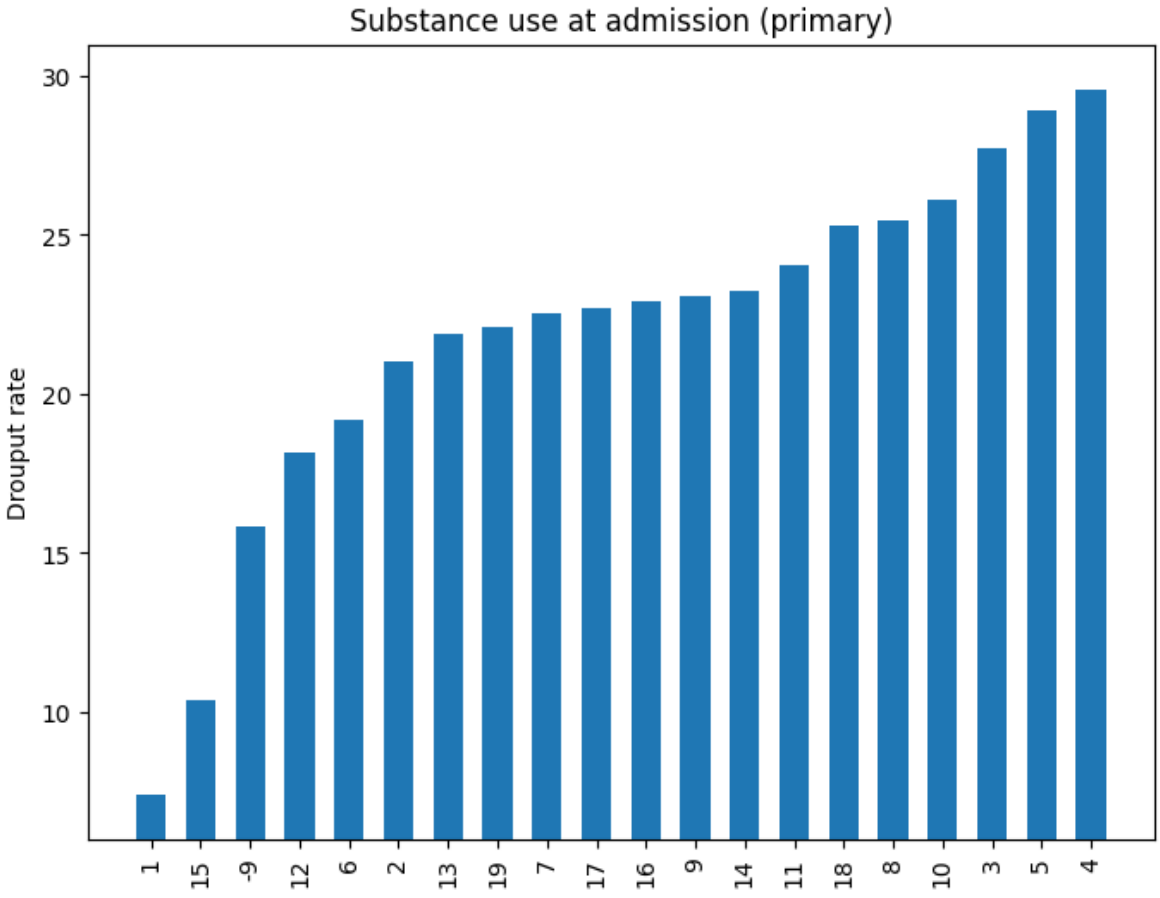

Figure B.7: Dropout Rate by Primary Substance use at admission

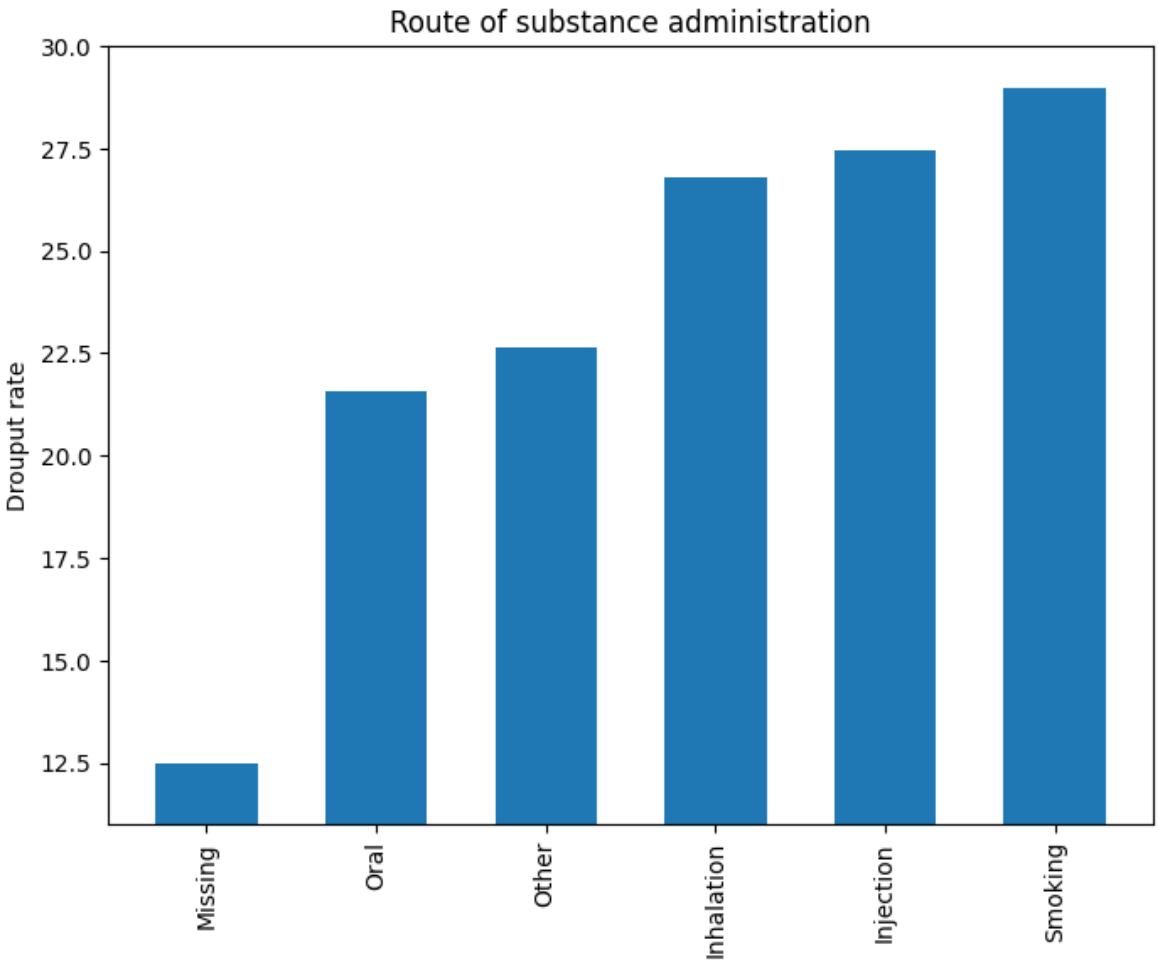

Figure B.8: Dropout Rate by Substance Route of Administration.

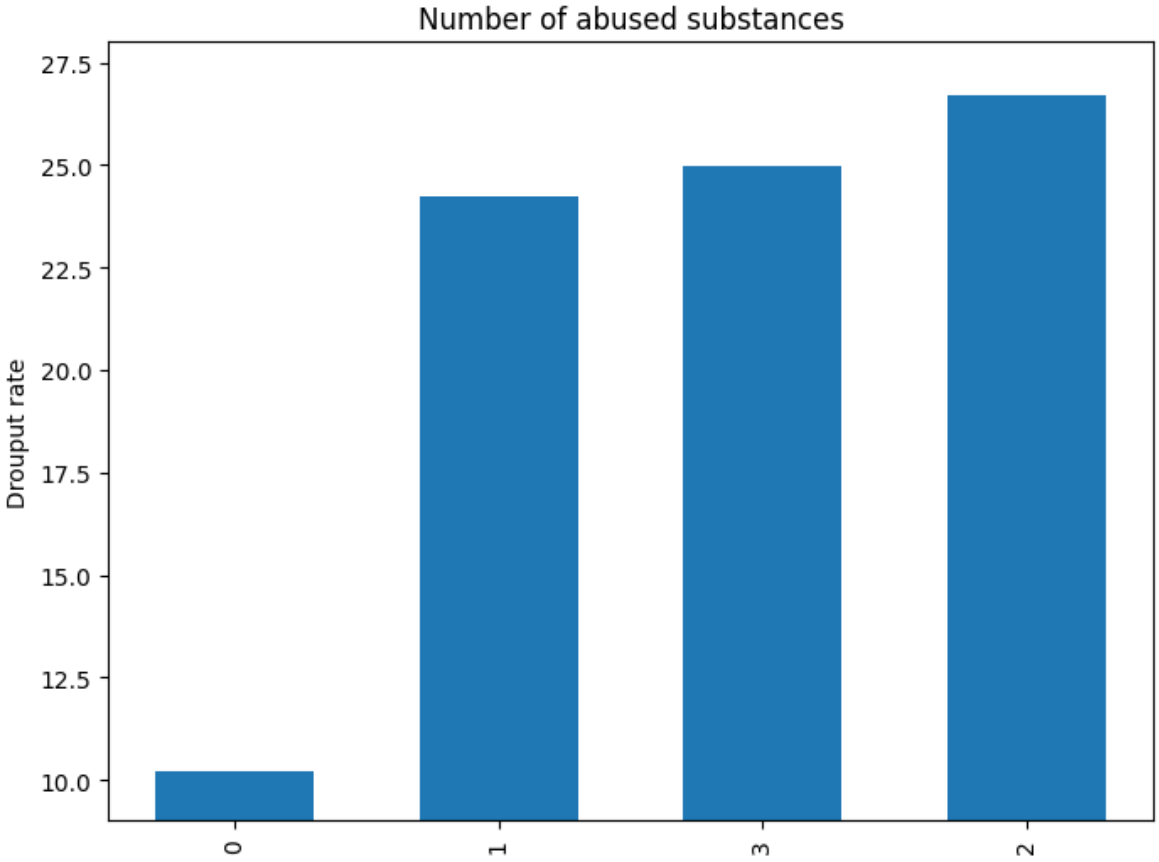

Figure B.9: Dropout Rate by No. of Substances Abused

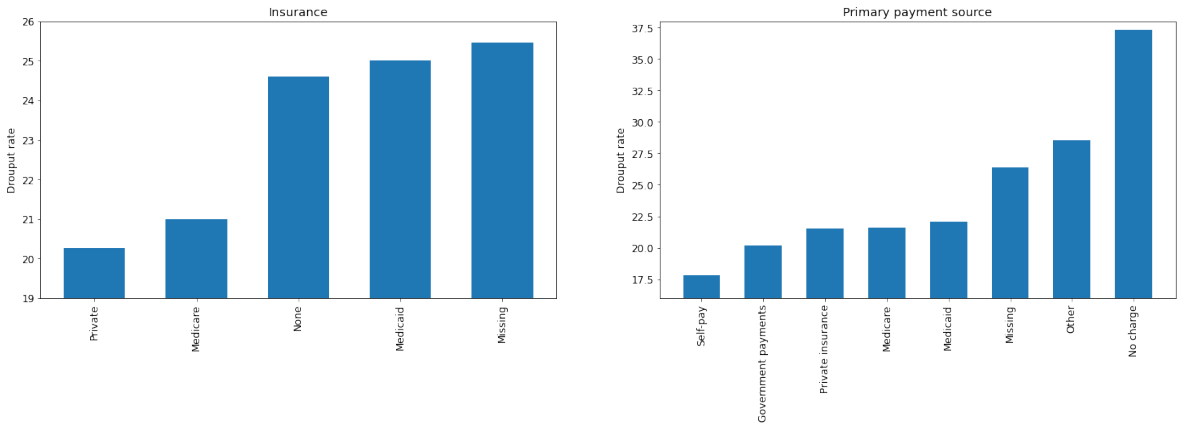

Figure B.10: Left: dropout rate by insurance type. Right: dropout rate by primary payment source.

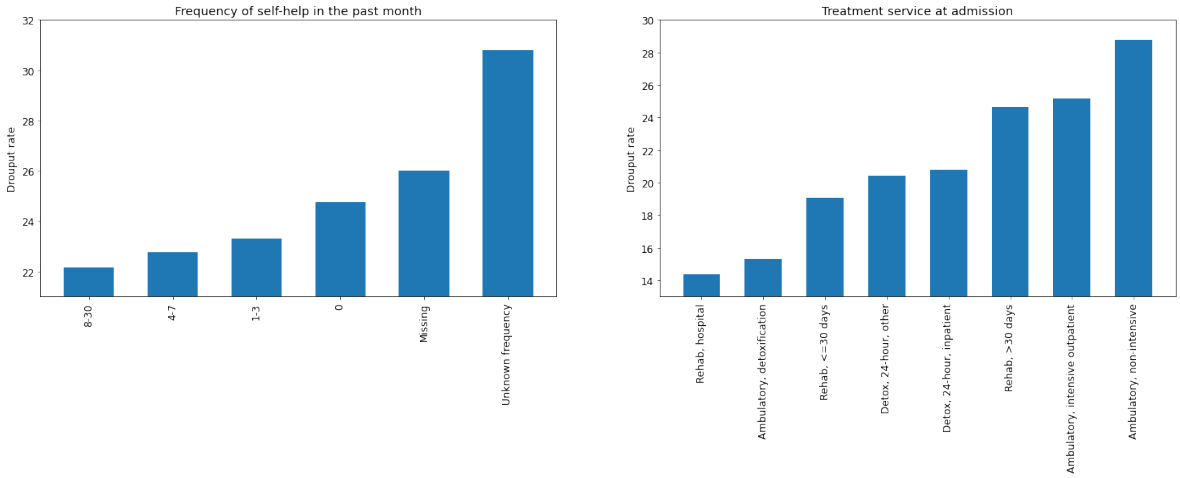

Figure B.11: Left: dropout rate by frequency of attending self-help group. Right: dropout rate by treatment service at admission.

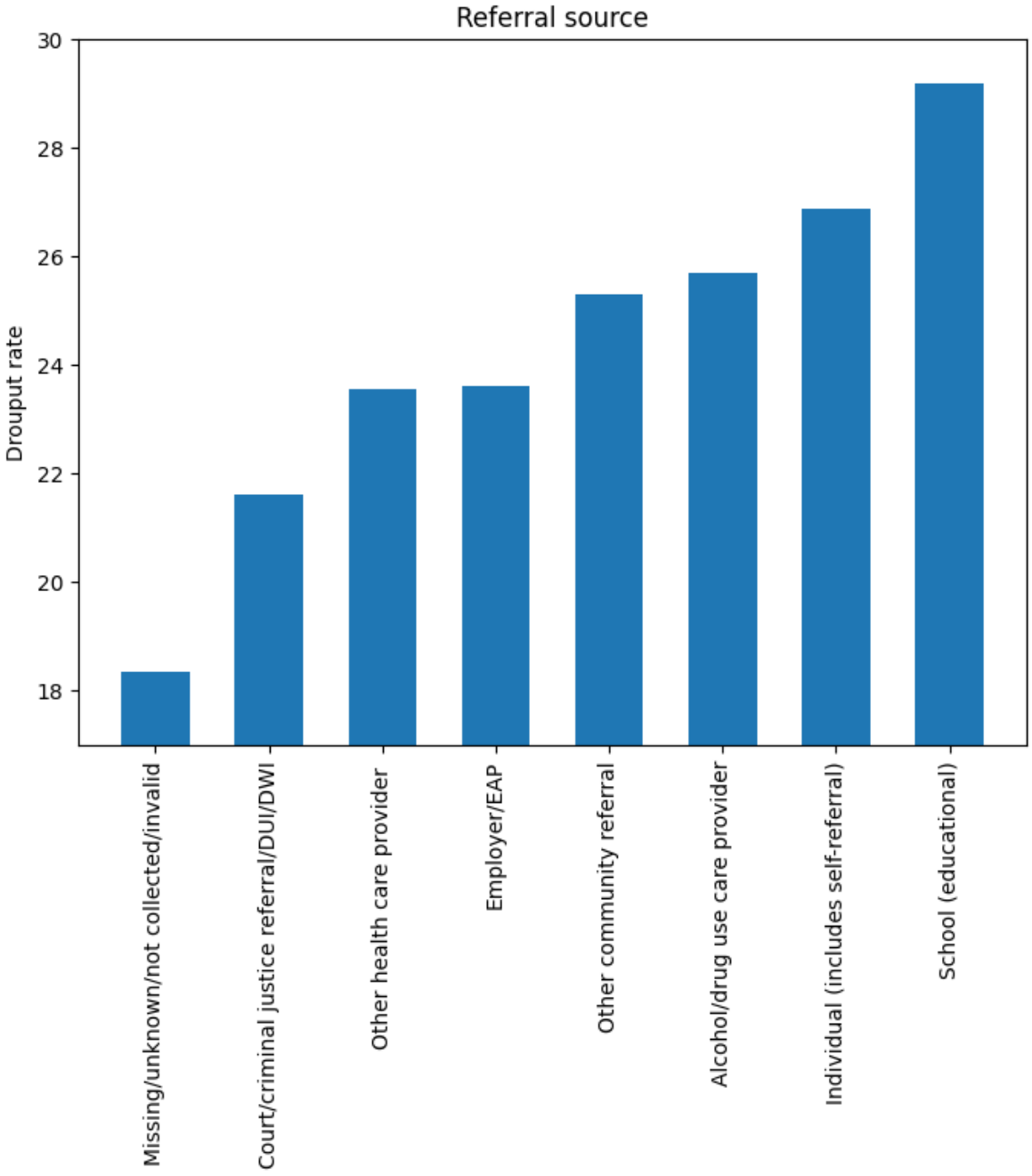

Figure B.12: Left: dropout rate by frequency of attending self-help group. Right: dropout rate by treatment service at admission.

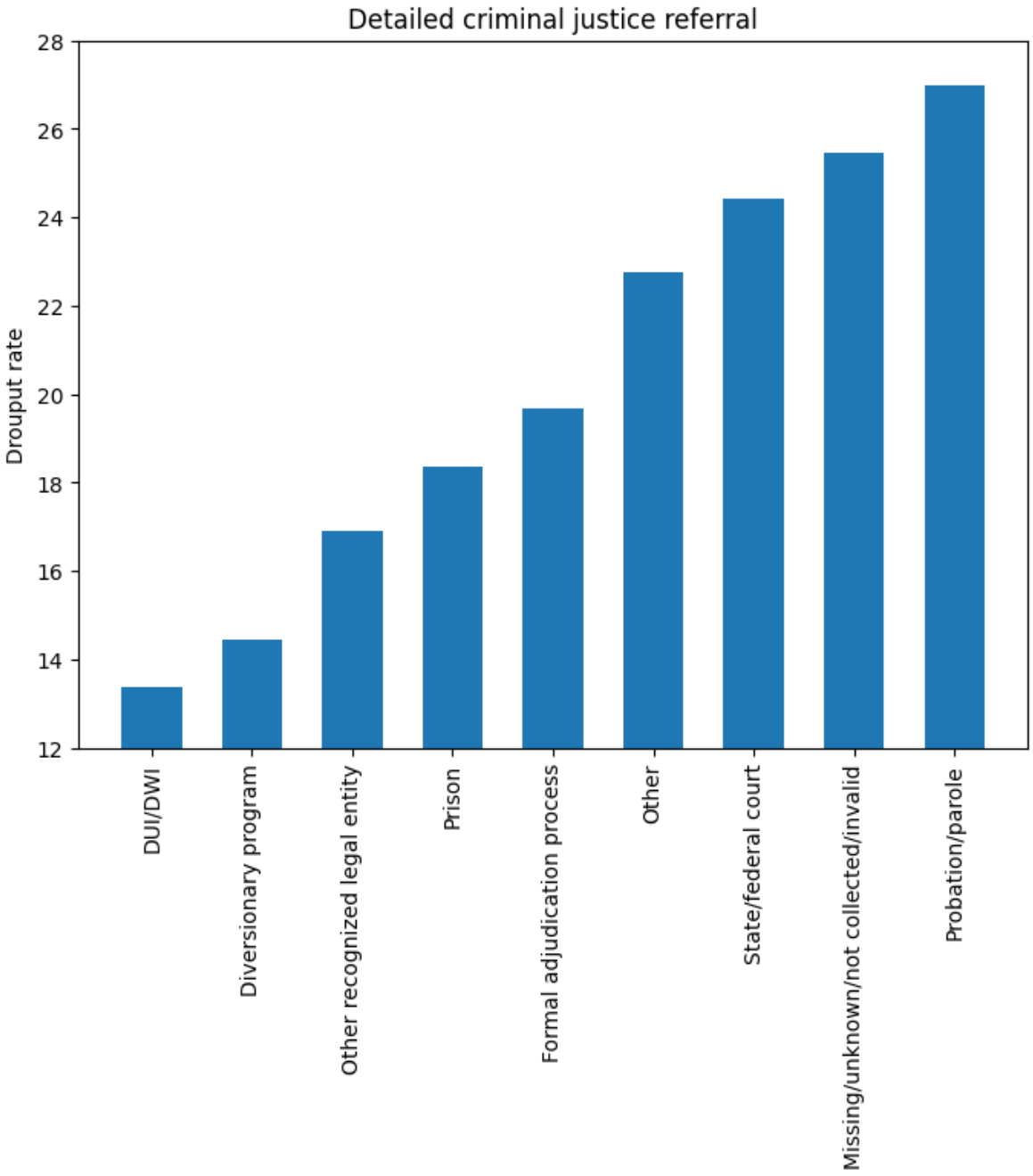

Figure B.13: Left: dropout rate by frequency of attending self-help group. Right: dropout rate by treatment service at admission.

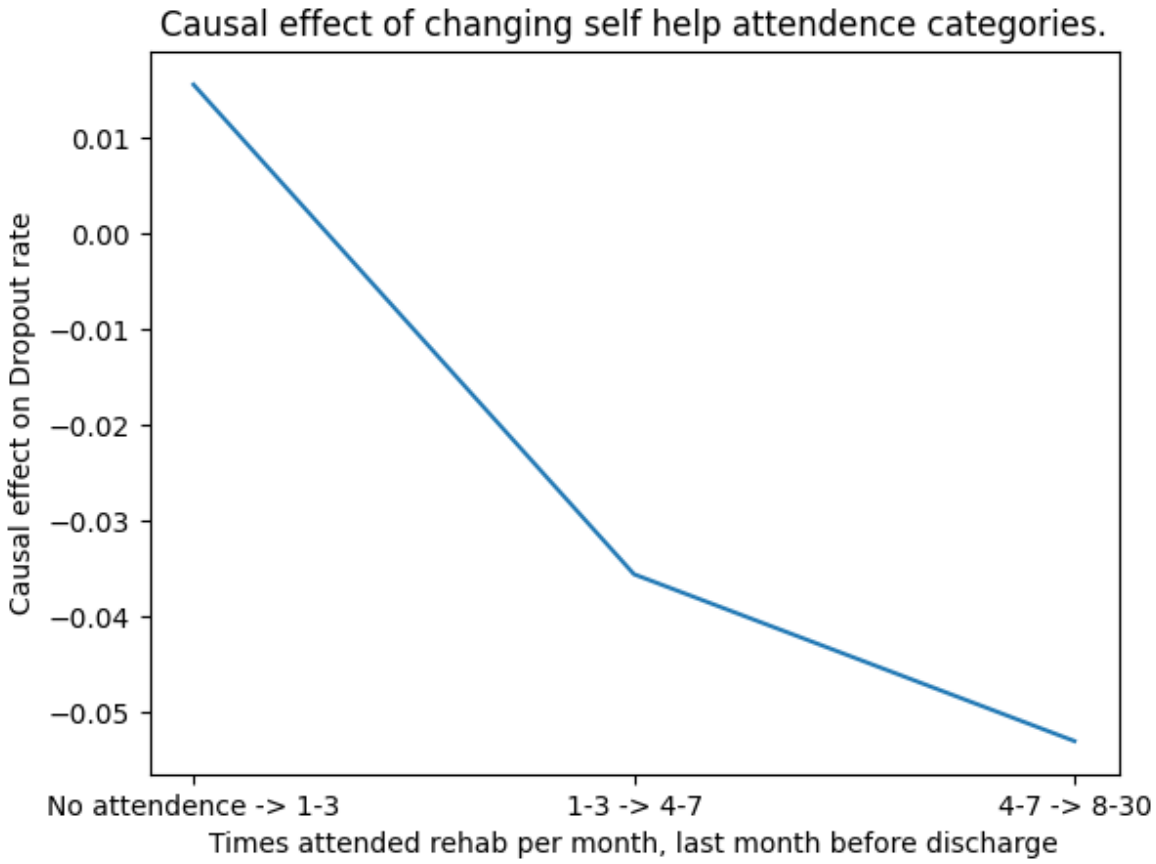

Figure B.14: Effect of changing self help groups on Dropout