MSD Datathon Report

Yuan Xue, Nandini Dadwal, Himanshu Jain, Piyush Jeena

21 May 2023

1 Overview

Relapse in substance abuse can be a disheartening and challenging experience. It often occurs when someone in recovery returns to using drugs or alcohol after a period of abstinence. It's essential to understand that relapse doesn't mean failure or lack of willpower; it's a common part of the recovery journey for many individuals. The path to recovery is seldom linear, and setbacks can serve as valuable learning opportunities. Relapse can be triggered by various factors, such as stress, emotional difficulties, social pressures, or encountering old environments associated with substance use. It's crucial to approach relapse with compassion and seek support from loved ones, professionals, or support groups. By learning from the experience, developing new coping mechanisms, and building a strong support network, individuals can regain their commitment to recovery and continue on the path toward a healthier, substance-free life.

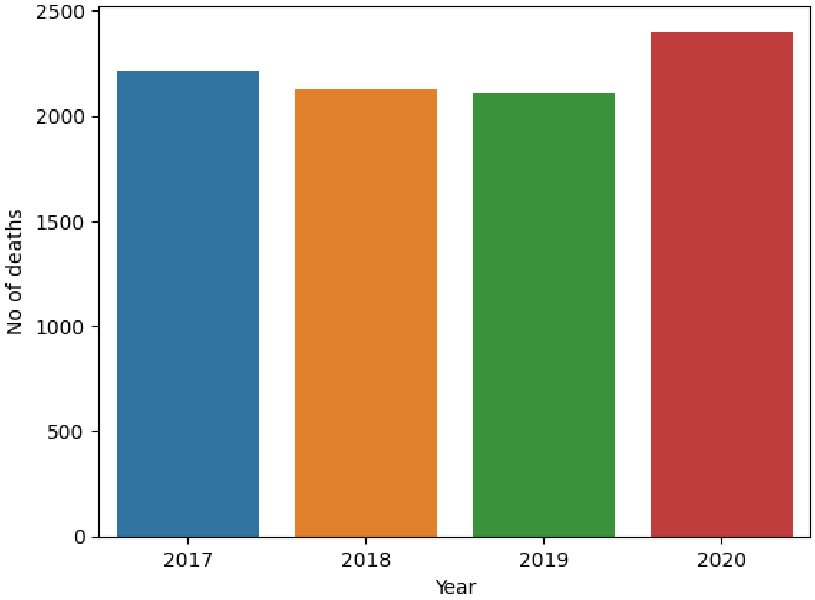


Figure 1

In Fig 1, we see that the number of deaths due to substance abuse increased by a slight margin in 2020. Interestingly, we observed that the proportion of clients dying is higher among recurring patients(0.28%) as compared to clients with no prior treatment episode(0.16%).

So our focus was **to investigate the different causes of relapse among clients which is resulting in their repeat admission**. To summarize we build a solution that covers the following issues:

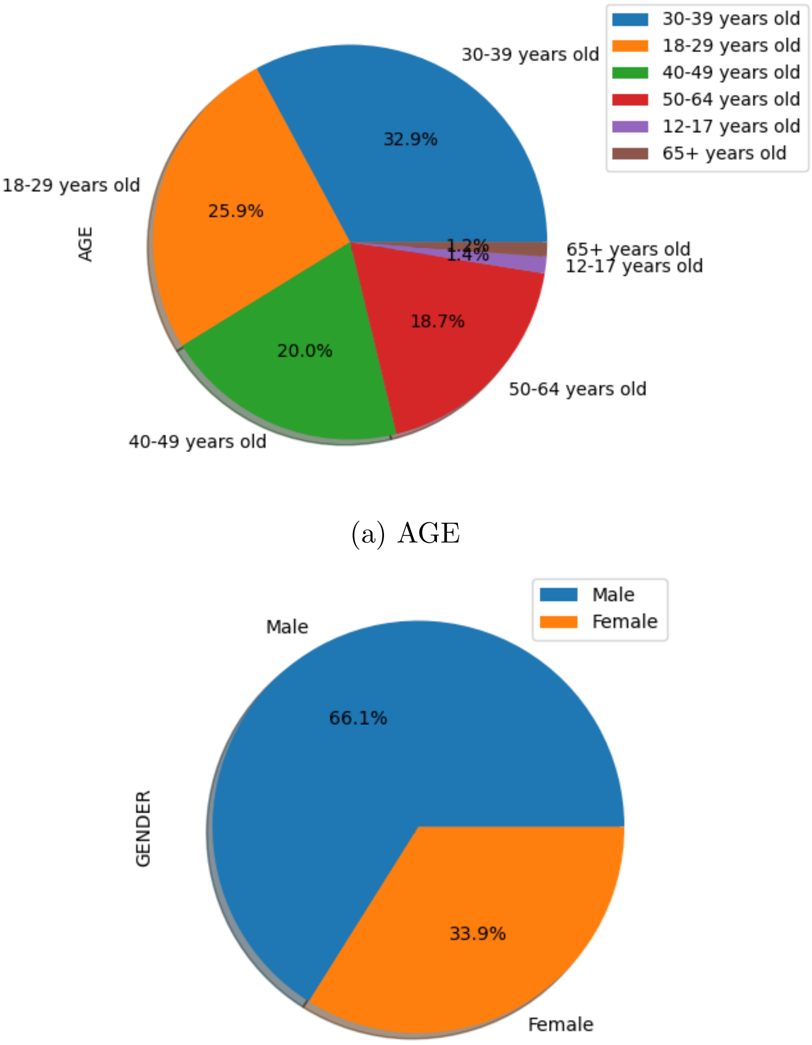
- Predict the patients who will return to substance abuse treatment after a relapse.
- Use explainable AI (shapely values) to identify client-level predictors contributing to relapse.

This can be extended to provide personalized treatment details which can be used by medical practitioners and reduce readmission.

2 Initial Insights

Firstly, we started with the data exploration of the (treatments 2016-2020) dataset which includes the state-level admission records.

2.1 Biographical attributes

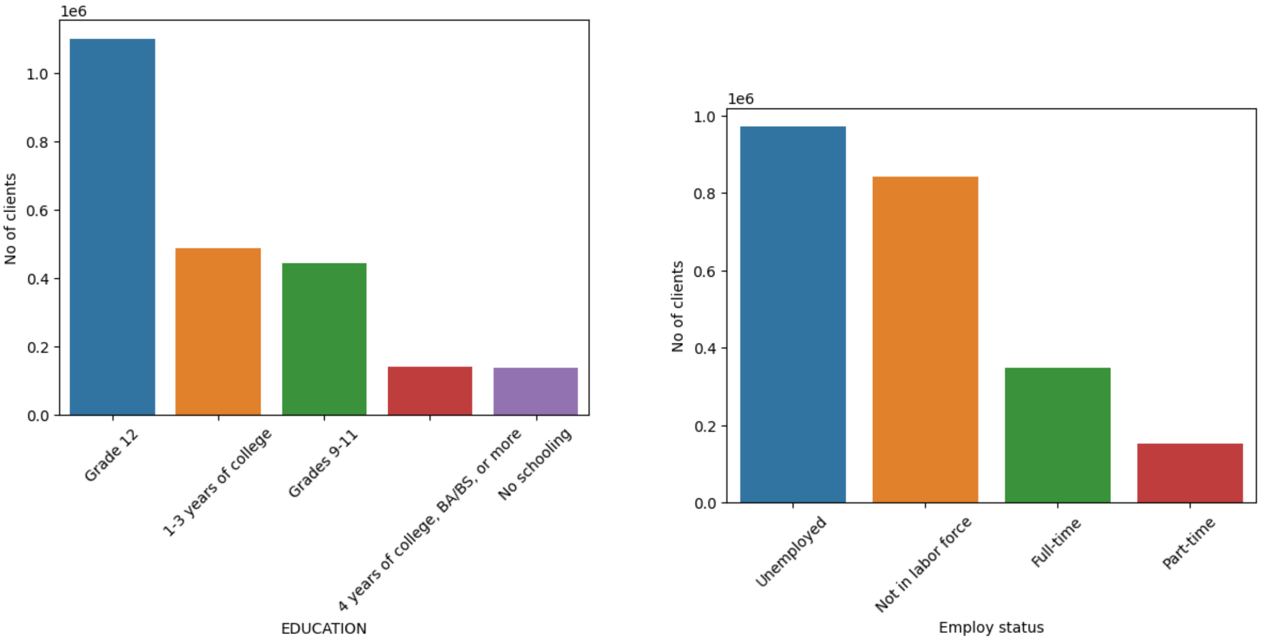


(a) AGE

(b) GENDER

Figure 2: Distribution of recurring clients based on (a) AGE (b) GENDER

From the above figures, we observe that clients between the age of 30-39 years old are more prone to relapse, and around 66% of repeat admissions are male.

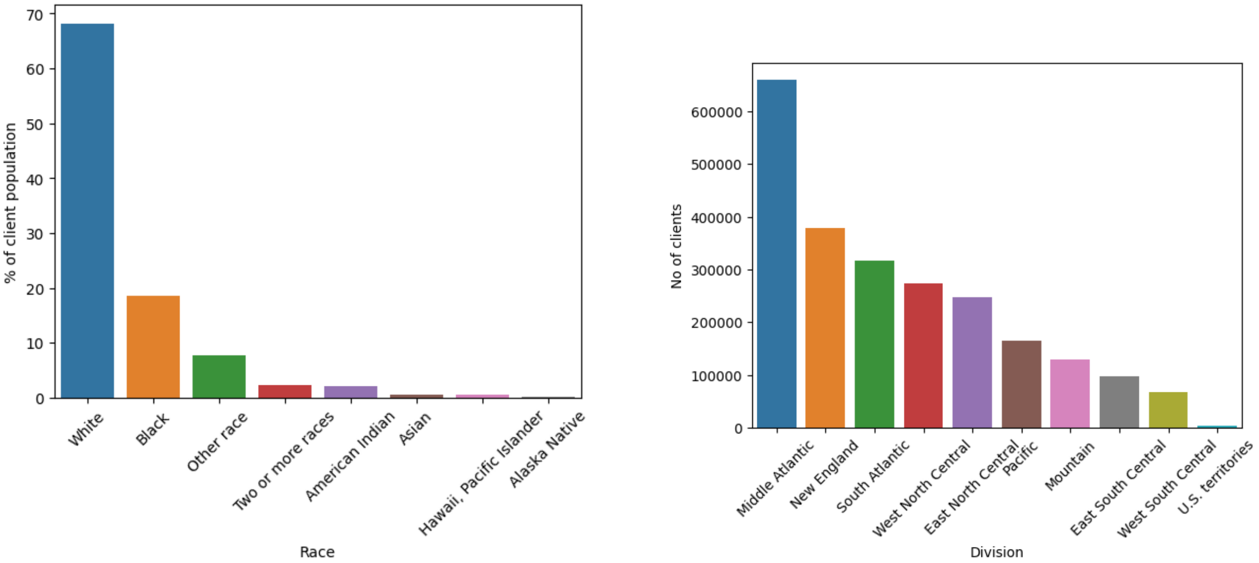


(a) EDUCATION

(b) EMPLOYMNET STATUS

Figure 3: Distribution of recurring clients based on (a) EDUCATION (b) EMPLOYMENT STATUS

People who are less educated and unemployed tend to relapse more which seems quite likely.



(a) RACE

(b) DIVISION

Figure 4: Distribution of recurring clients based on (a) RACE (b) DIVISION STATUS

From the above plots, we observe a recurring admission in substance abuse treatments in the Middle Atlantic division comprising big cities like New York, Philadelphia, and Washington. One interesting thing to note is the higher relapse rate among the white race which is quite contrasting since the American entertainment industry mostly portrays drugs etc within the black community.

2.2 Substance attributes

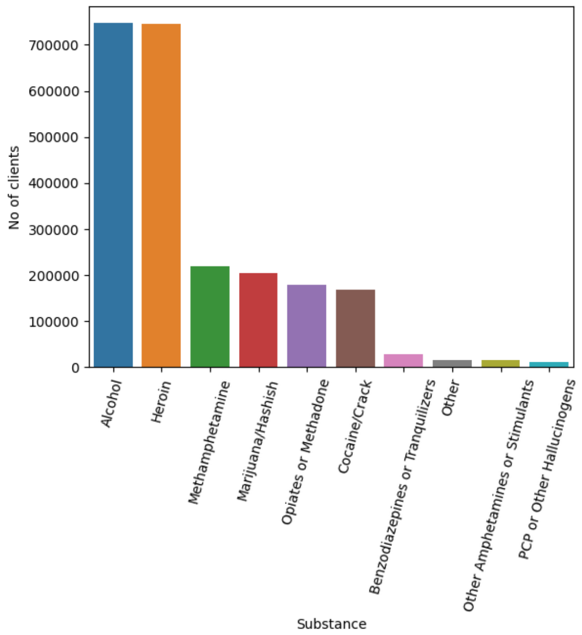


Figure 5: Distribution of patients based on first substance intake

We clearly see an expected dominance of alcohol and heroin among relapse patients which seems quite reasonable since both these substances are easily available as compared to others.

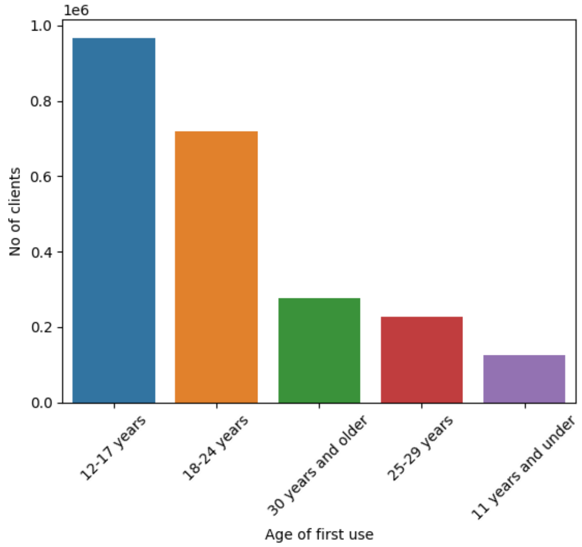


Figure 6: Distribution of patients based on age of first use of any substance

Teenagers tend to relapse more as can be seen from Fig 6. This seems feasible since teenagers are addicted to trying new things in almost every society. Even with simple exploratory analysis, we gained some string insights :

- Government should focus on making policies to register clients in employment agencies. More focus should be put in the cities based in Middle Atlantic.
- Substance abuse counseling programs should be provided to teenagers and their families to resolve problems and manage feelings and behaviors effectively without using drugs or alcohol.

3 Methodology

3.1 Data Preparation and Preprocessing

We stored the data in the parquet format to enable easy fetching of records. Only those records were considered where the SUB1(first substance) and NOPRIOR(number of previous treatment episodes) are known. This is done since we are trying to predict if a client will return or not so we need to know the NOPRIOR dependent variable. Secondly, we ignore patients who don't use any substance at the time of admission. Essentially, this ignores the clients who started using substances while being treated at the treatment facility.

Since all the categorical features are integers (with their respective descriptions in the data table), we mapped all of them with their respective names.

3.2 Feature engineering

- We created a new feature (NUM_SUBS) which tells us the no of drugs found within a client over a period of time. It's calculated by summing the flag variables for the primary, secondary, or tertiary drug usage at admission. It takes values : (0 subs, 1 subs, 2 subs, 3 subs).
- We performed ordinal encoding on AGE(age), EDUC(education level), ARRESTS(no of arrests prior to admission), FRSTUSE1(age of first use), NUM_SUBS since these features are ordinal in nature and we wanted to preserve that. For the remaining features, we use the regular One Hot encoding.
- We dropped features(health insurance) that had more than 50% missing values and imputed the relevant features using mode.

3.3 Model building

The problem statement is a binary classification task with the target variable, NOPRIOR being a binary variable(1: repeat patient, 0: No prior treatment).

- Split the data into train(70%) and test(30%) sets.
- Performed one hot encoding on the training data
- Used Xgboost(since it usually performs well on the tabular dataset) for modeling with basic tuning(using Randomized Search) on the learning rate.

3.4 Evaluation

We wanted to choose a metric that predicts high-risk patients better than low-risk patients since we want to deliver immediate help to high-risk patients. In fig 7 we show our model's confusion matrix on the test data. Our recall and precision turned out to be 0.85 and 0.77 respectively. In our case, we want to minimize False positives (Increase recall) since we want to give attention to the recurring patients. In other words, we want to correctly predict high-risk patients as compared to low-risk patients.

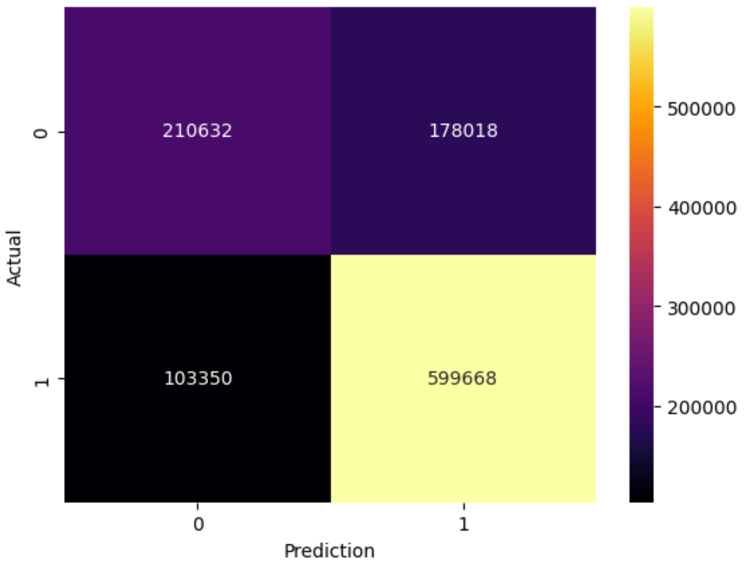


Figure 7: Confusion matrix; 1 : repeat admission, 0: first time admission

4 Model Interpretability

We used SHAP values to interpret our model. This analysis was performed on a subset of data(10000 data points) since the compilation time is quite high for shapely values. Let's first look at the individual level. In Fig 8, we see that :

- NUM_SUBS (no of drugs reported in a client) being 3 increases the repeat admission probability by 0.39%.
- Living in Mountain (Arizona, Colorado) decreases the repeat admission probability by 0.32
- Not attending self-help groups has increased the repeat admission probability by 0.29

Obviously this is not a full solution to the problem but these personalized insights along with the experience of medical practitioners can be used to diagnose clients who might be at high risk.

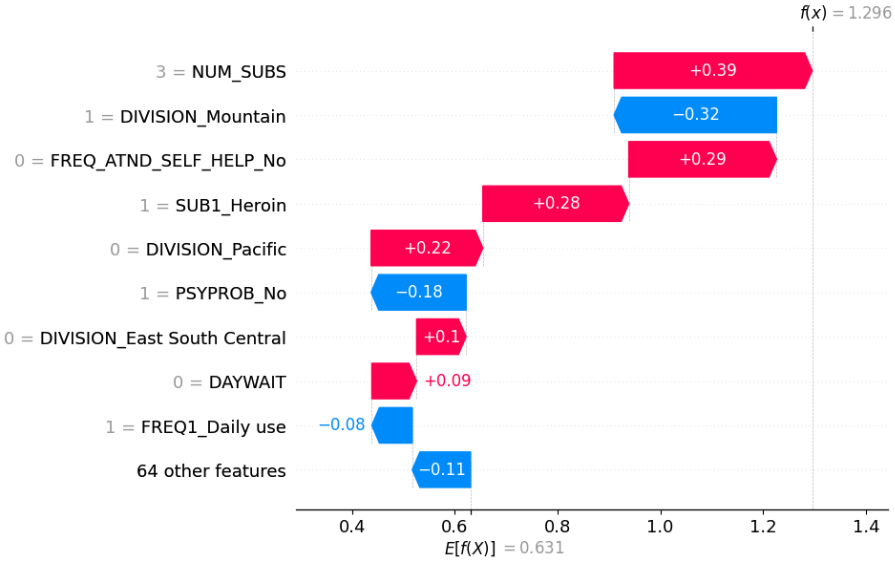


Figure 8: SHAP values at individual level

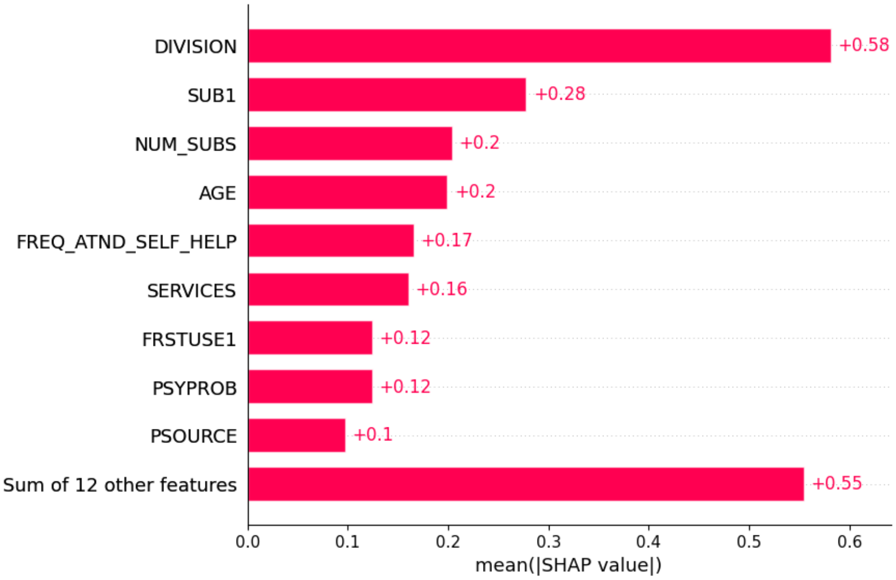


Figure 9: SHAP values at feature level

In Fig 9, we see the feature importance on our model. Features that have made large positive/negative contributions will have a large mean SHAP value. In other words, these are the features that have had a significant impact on the model's predictions. We observe that region, type of first substance consumed, total no of drugs reported, age, self-help groups play an important role in predicting readmission probability.

5 Future direction

- Improve the accuracy of correctly predicting the readmission patients.
- Using explainable AI to correctly identify the risk factors responsible for readmission at an individual level.
- Building a dashboard that can be used by medical practitioners for personalized treatment.

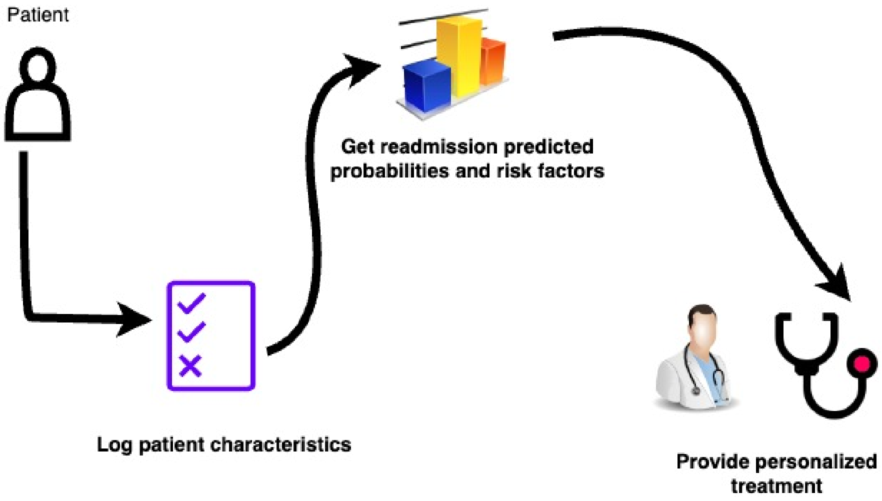

Patient
Get readmission predicted probabilities and risk factors
Log patient characteristics
Provide personalized treatment

Figure 10: Substance abuse treatment system