Final Report: Predicting Risk of Homelessness Amongst Substance Abuse Patients

Ayush Lahiri

PPOL 565- 01

5th May 2023

1. Research Question and Background

The aim of this study is to predict the risk of homelessness amongst those suffering from substance abuse, utilizing individual level demographic characteristics, via supervised learning methods. Among the characteristics of unsheltered persons, the aim of the prediction is focused on, homelessness which is comorbid with substance abuse disorders and/or mental illnesses. For the data available at the case level (discussed in subsequent sections), we aim to answer the following question:

Given, demographic, employment, substance abuse disorder, mental health disorder information (among others) of a patient during admission to substance abuse disorder treatment facility, can we predict their risk of homelessness by the time their treatment ends.

The issue of homelessness is a nationally persisting problem. In 2022, 582,462 people were experiencing homelessness (de Sousa et al., 2022) and 4.6 million adults reported to be likely to lose their homes (Bhattarai & Siegel, 2022). Of key importance is the strong co-occurrence of substance abuse, mental health and homelessness. The California Policy Lab found in its 2019 survey, that 78% and 75% unsheltered homelessness persons reported having mental health conditions and suffering from substance abuse conditions respectively and 50% and 51% reported that their mental health conditions and substance abuse problem respectively, contributed to their loss of housing (Rountree et al., 2019).

Current tools used to assess risk for housing needs, such as the VI-SPDAT (OrgCode Consulting, 2015) have been shown to be biased towards disadvantaged groups, especially Black women (Cohen, 2022), inconsistently applied across programs and rely on personal judgement of case managers leading to delays or inaccurate rejection for housing (Cohen, 2022; Kertesz et al., 2017). This study therefore looks to contribute towards creating objective assessments of risk of homelessness, taking into account community-based characteristics to aid in the Housing First policy and contribute towards quicker, targeted interventions and easier inter-agency communication.

2. Data

The study uses the most recent Treatment Episode Data Set, Discharges (TEDS-D) from 2020 (Center for Behavioral Health Statistics and Quality, 2020), a public dataset published by the Substance Abuse and Mental Health Services administration, which contains approximately 1.4 million unique cases. Data is recorded for all states barring Idaho, Maryland, New Mexico, Oregon, Utah and West Virginia. 76 different characteristics of each patient are recorded regarding their socioeconomic conditions, mental health, substance disorder and living arrangement of substance abuse patients through their treatment journey, in publicly funded treatment facilities. Data is collected both at admission and at discharge.

Based on variable importance in literature, significance of correlations between our output variable and predictor variables and presenting available demographic characteristics we limit our discussion to 6 variables presented in Table 1. for brevity.

While the data set contains approximately 1.4 million cases, our output variable, i.e. Living Arrangement at Discharge is missing for about 22% of the dataset. As an output variable, we are unable to impute these values and hence these cases are not used in development of our prediction model. Therefore, our analytical sample and hence the descriptive statistics presented below are based on cases for which the output variable is not missing.

Total N = 1,083,555				
Variable		**n**	**% of non-missing values**	**Missing values**
Race				17598 (1.6%)
	White	747699	70.14	
	Black	193813	18.18	
	Alaska Native \ American Indian	25446	2.38	
	Asian	5723	0.53	
	2 or more races	93276	8.75	
Gender				470 (0.043%)
	Male	704832	65.07	
	Female	378253	34.92	
Primary Substance of Addiction				5883 (0.54%)
	Alcohol	357813	33.2	
	Heroin	253555	23.52	
	Methamphetamine \speed	143977	13.36	
	Opiates \Synthetic substances	142845	13.25	
	Marijuana \Hashish	114652	10.64	
	Cocaine \Crack	64830	6	
Age when first used primary substance				36319 (3.4%)
	11 years and under	56934	5.43	
	12-17 years	436652	41.7	
	18-29 years	422122	40.3	
	30 years and older	131528	12.56	
Previous treatment episodes in any substance use treatment program				50056 (4.62%)
	Yes	651696	63.06	
	No	381803	36.94	
Comorbidity with Mental Disorder				100780 (9.3%)
	Yes	509003	51.8	
	No	473772	48.2	
Living Arrangement at Discharge				0
	Independant Living	731638	67.52	
	Dependant Living	206577	19.06	
	Homeless	145340	13.41	

Table 1: Descriptive Frequencies of select variables

We see a few important implications for our predictive models, from the descriptive statistics presented in Table 1. Our sample has a slightly lower representation for white individuals, slightly higher representation for Black and Alaskan Native/American Indian individuals, and much lower representation for females than the national average (U.S. Census Bureau, 2022).

For our output label, i.e. living arrangement at discharge, our analytical sample is imbalanced with only 13% of the cases being homeless at discharge. Therefore, we incorporatere techniques such as synthetic minority oversampling, random under sampling and undertake stratified randomized allotment to training and testing splits. We also find, that all predictor variables presented (and others not presented) have some missing values. Since missingness is not co-occurring between all the variables, dropping missing values for each predictor would significantly reduce the size of the final data used. Hence, I use mode-based imputation, as all variables in our dataset are categorical. While

nearest-neighbors based imputations have the potential to impute with better accuracy, it is computationally expensive (discussed in subsequent sections).

As discussed previously, the large proportion of missing values for our output label is concerning and requires further investigation to ascertain whether any systemic problems exist with dropping missing values for living arrangement at discharge causing our data to be biased.

State	% of missing values for living arrangement at Discharge within state	% of missing values for living arrangement at discharge
Arizona	100	46
New York	26	16.5
California	48.3	16.2

Table 3: Top 3 states contributing to missing valus for living arrangement at discharge

Reason for end of treatment	% of missing values for living arrangement at discharge
Treatment completed	61
Dropped out of treatment	30
Terminated by Facility	3

Table 4: Top 3 Reasons contributing to missing values for living arrangement at discharge

Since the source data for TEDS-D is data reported by various facilities in different states, Table 3 looks at whether state-based laws/practices contribute to this problem. Column 2 of Table 3, shows the percentage of missing values for the output label within each state, for the top 3 states with missing values: Arizona, New York, and California. Column 3 indicates the proportion of missing values for the output label contributed by the three states. These states account for almost 80% of the missing values for the output label. Notably, in Arizona, no cases have recorded values for living arrangement at discharge. Therefore, it may be necessary to include Arizona in the list of states with missing values in the TEDS-D dataset. However, New York and California have a large amount of data missing as well, but also have cases that do record living arrangement at discharge.

Table 4 shows, that the reason for discharge, strongly accounts for missing values. Specifically, when the reason for end of treatment episode is due to treatment completion and when patients self-dropout of treatment, 91% of the missing values for the output label are accounted for. This implies, that for most of the missing values, facilities may either simply choose not to record living arrangement at discharge or are unable to record so due to loss of contact with the patient. For New York, we find that 73% of the missing values of the output label, are for cases where treatment was completed and 24 % was for cases who dropped out of treatment. In California, 97% of the missing values in the output label was for cases who dropped out of the treatment.

Therefore, dropping all cases where values are missing may pose a limitation of our final analytical sample, since a large proportion of missing values can be attributed to patient dropout and factors causing treatment dropout may be correlated with living arrangement at discharge.

3. Descriptives Analysis

We next look at some patterns gained from exploratory data analysis on the direction of correlation between select predictor variables and our outcome variable.

From figures 1 through 4, we see that those who are homeless at discharge have a disproportionately higher proportion of racial minorities, specifically Black and American Indian/Alaskan Native populations, higher proportion of males, higher proportion of patients who have mental disorders co-occurring with a substance abuse disorder and a higher proportion of patients who have 1 or more times seeked substance abuse disorder treatment in the past as compared to those who have dependent or independent living arrangements available at discharge.

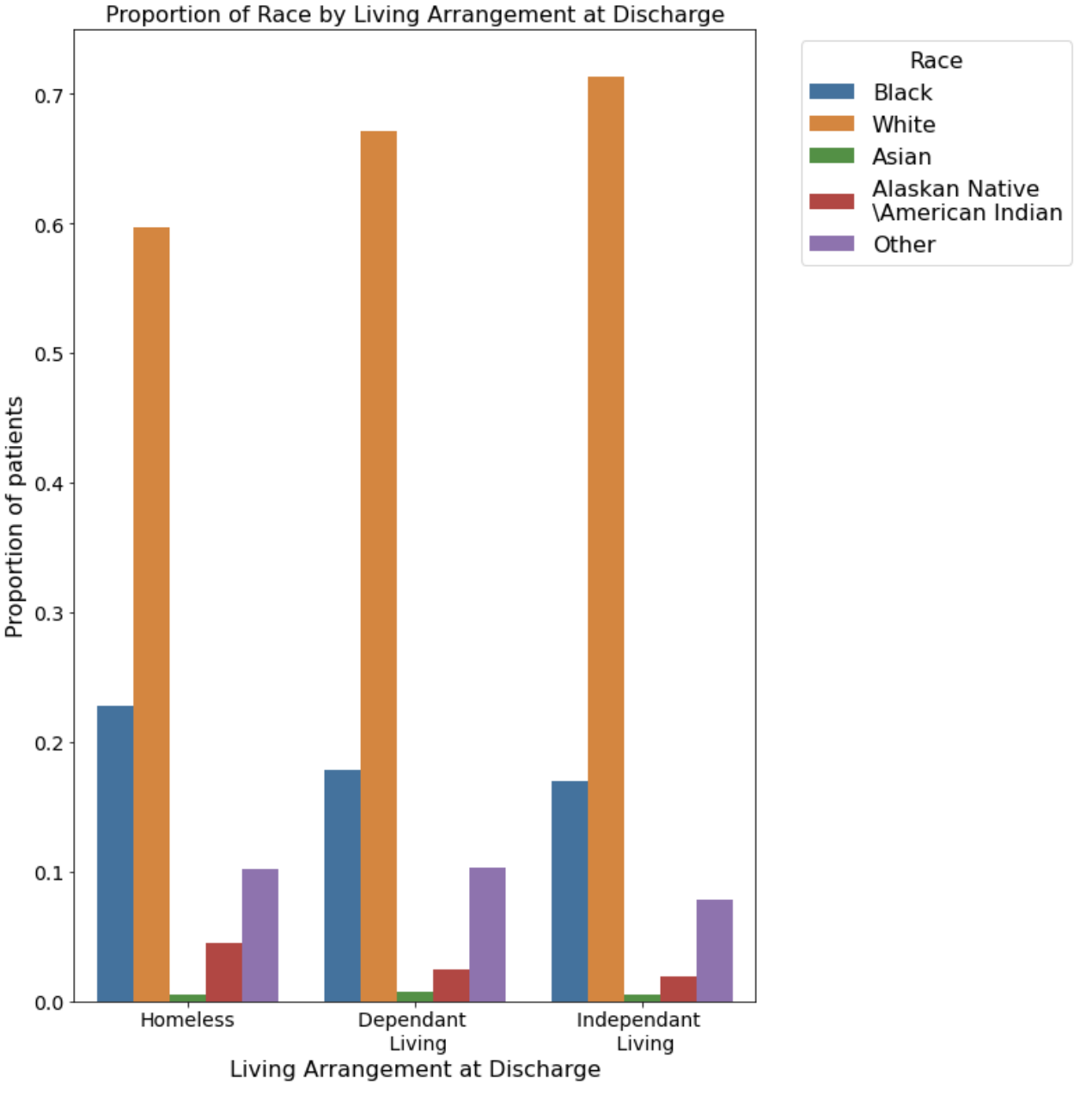

Figure 1.

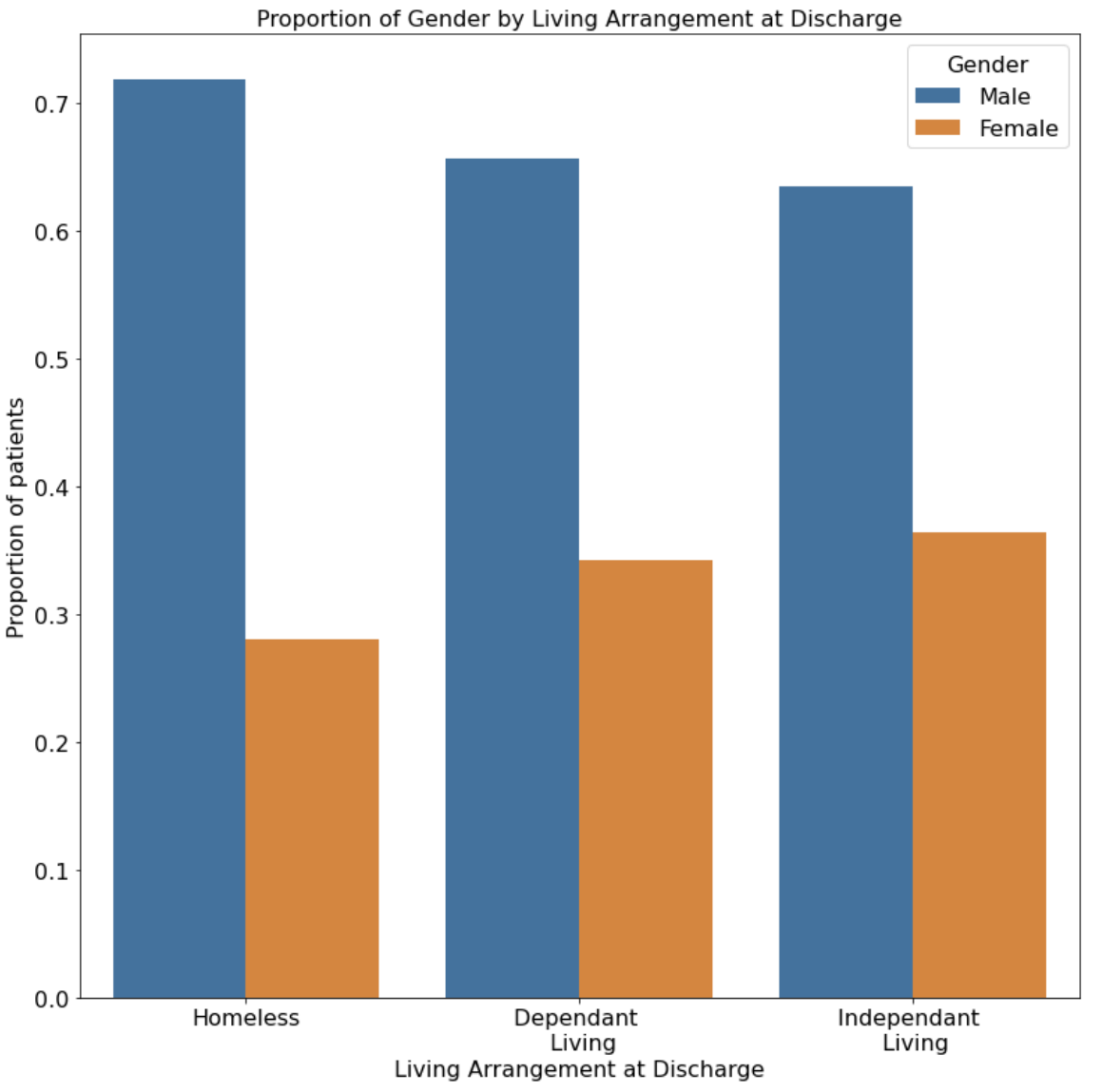
Proportion of Gender by Living Arrangement at Discharge
Gender
Male
Female
Proportion of patients
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Homeless
Dependant Living
Independant Living
Living Arrangement at Discharge

Figure 2.

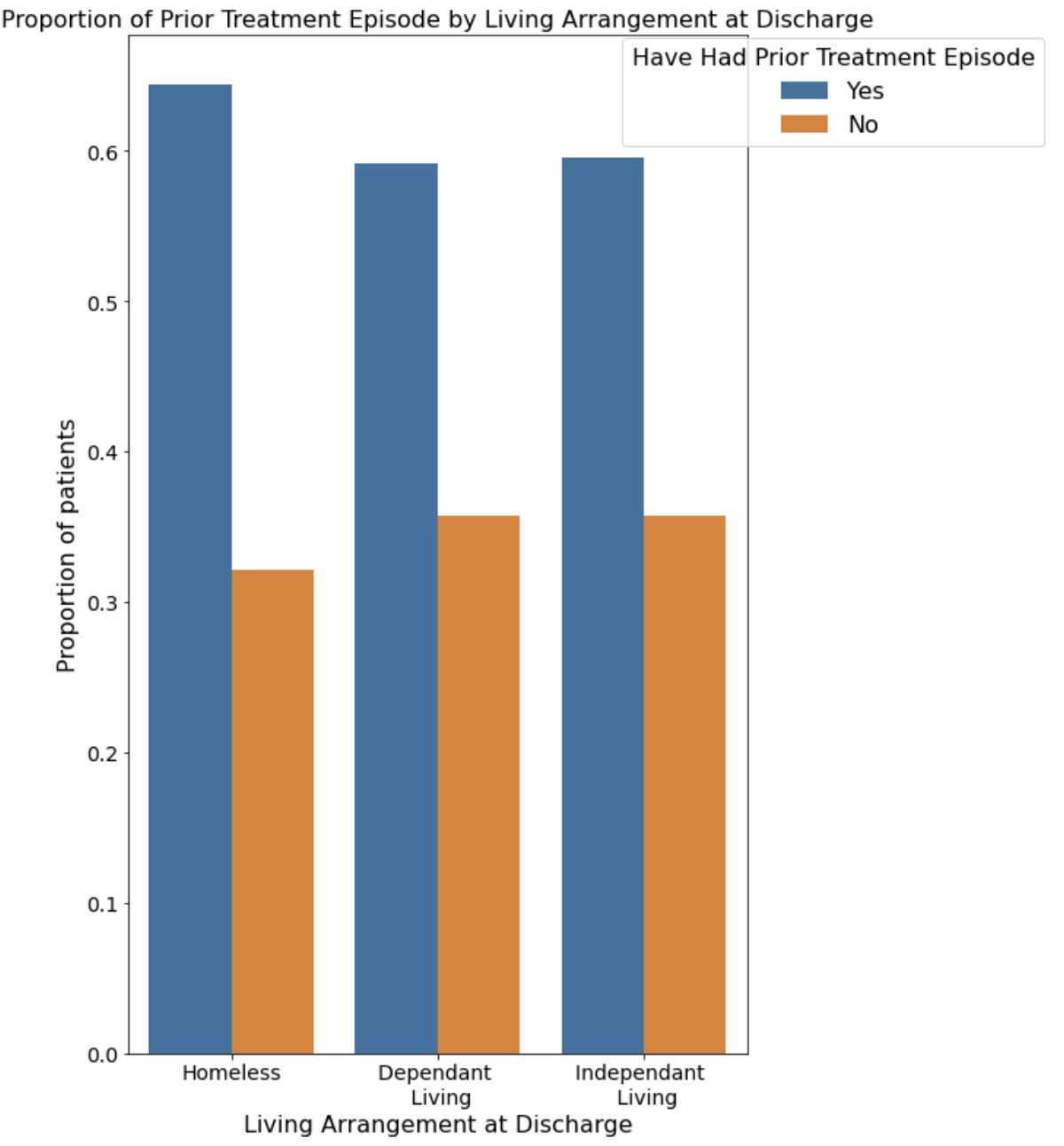
Proportion of Prior Treatment Episode by Living Arrangement at Discharge
Have Had Prior Treatment Episode
Yes
No
Proportion of patients
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Homeless
Dependant Living
Independant Living
Living Arrangement at Discharge

Figure 3.

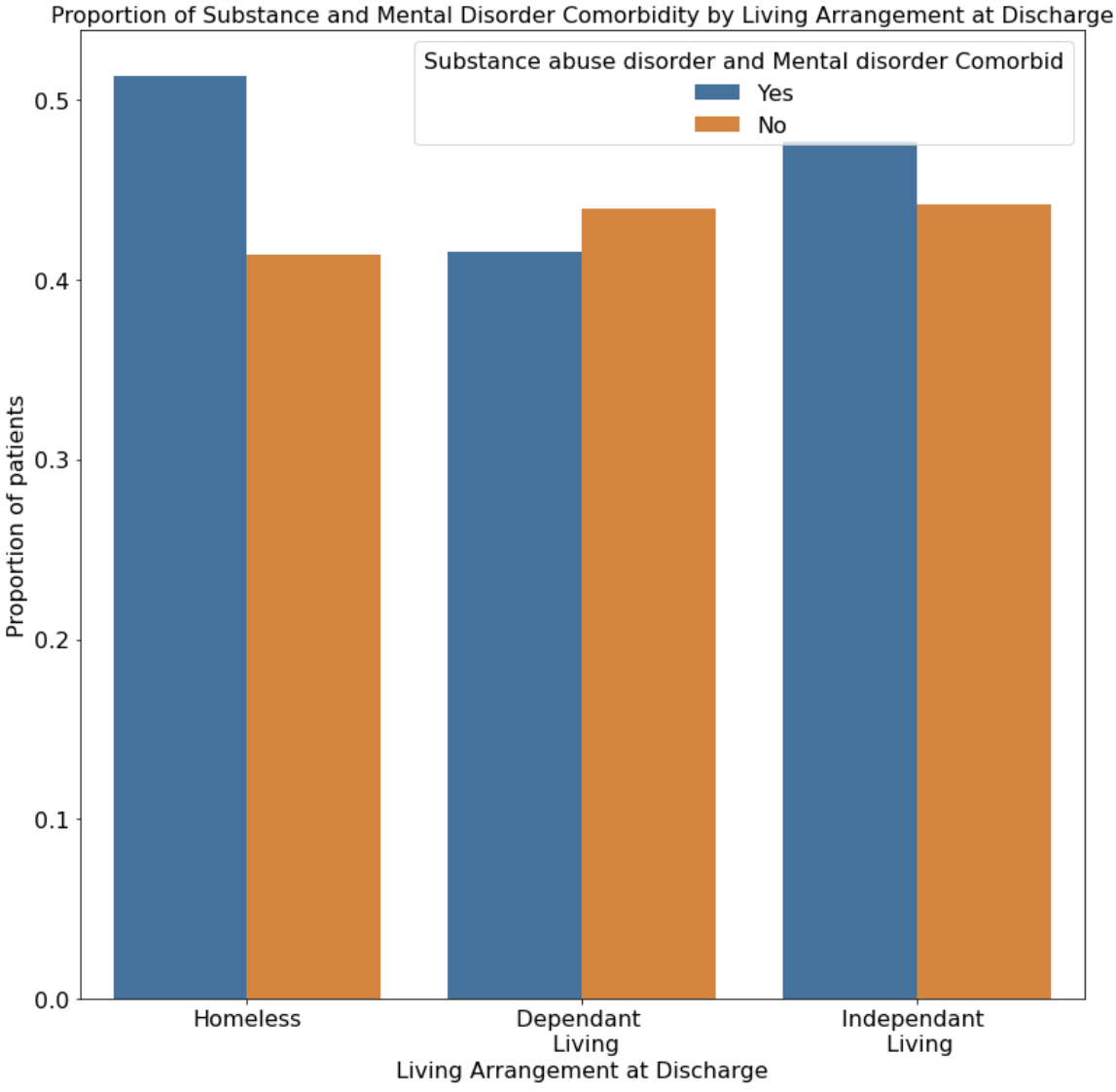

Figure 4.

As found by Moxley et al. (2020) and observable in our sample, seen in figure 5, it is evident that individuals tend to maintain their initial living arrangement throughout their stay, regardless of the type of living arrangement. This suggests a strong association between the living arrangement at admission and the living arrangement at discharge.

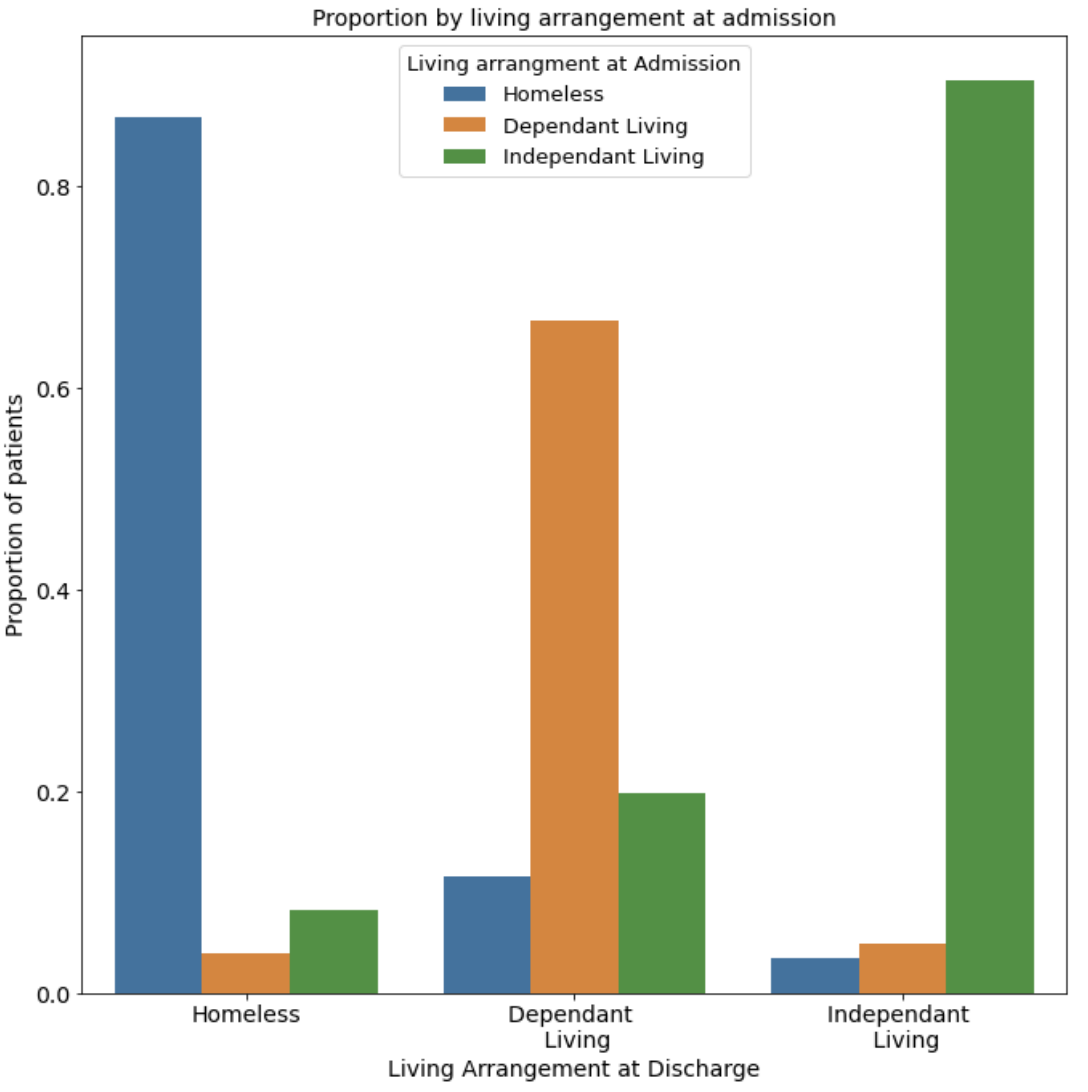

Figure 5.

Therefore, the adversarial case for our prediction model, is for those who are homeless/not homeless at admission and not homeless/homeless at discharge (5 % of analytical sample). In fact we can see that for homeless and independent living, almost 80% of the times living arrangement at admission is a perfect predictor for living arrangement at discharge. Firstly, our goal is to create a generalized model which aims to predict risk of homelessness based a health and socio-economic conditions of an individual. Having homelessness status at admission trivializes this problem, since if we could observe the housing status for the underlying population, then there is no requirement for predictions, as can be seen through a data. In our methodology, we therefore do not include homelessness status at admission as a feature.

Next, we investigate how the previously discussed set of selected variables are distributed for groups where homelessness status changes through treatment journey and cases where patients are not homeless at either points of record in the treatment journey, in figures 6 through 9.

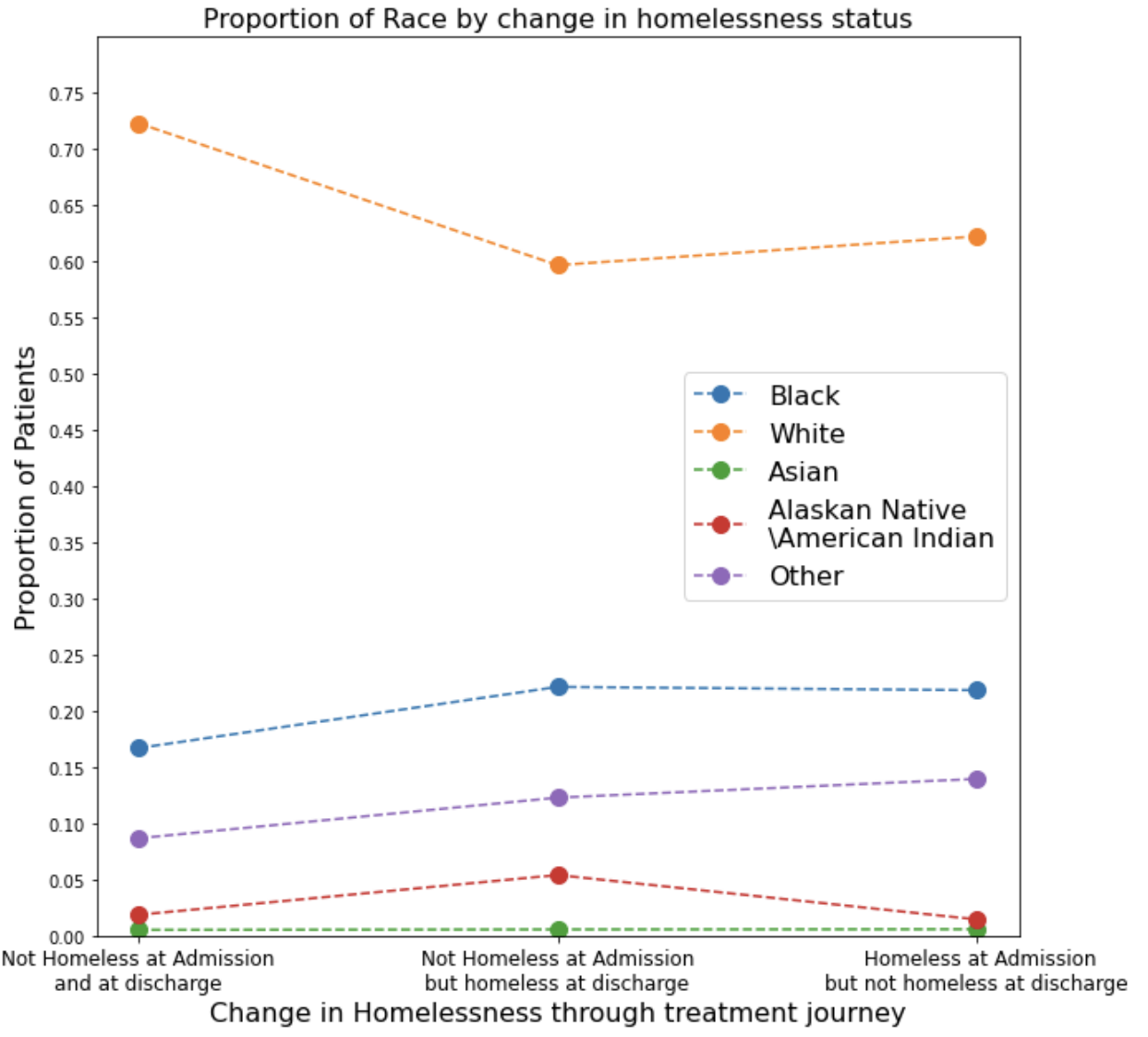
Proportion of Race by change in homelessness status
Proportion of Patients
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Black
White
Asian
Alaskan Native
\American Indian
Other
Not Homeless at Admission
and at discharge
Not Homeless at Admission
but homeless at discharge
Homeless at Admission
but not homeless at discharge
Change in Homelessness through treatment journey

Figure 6.

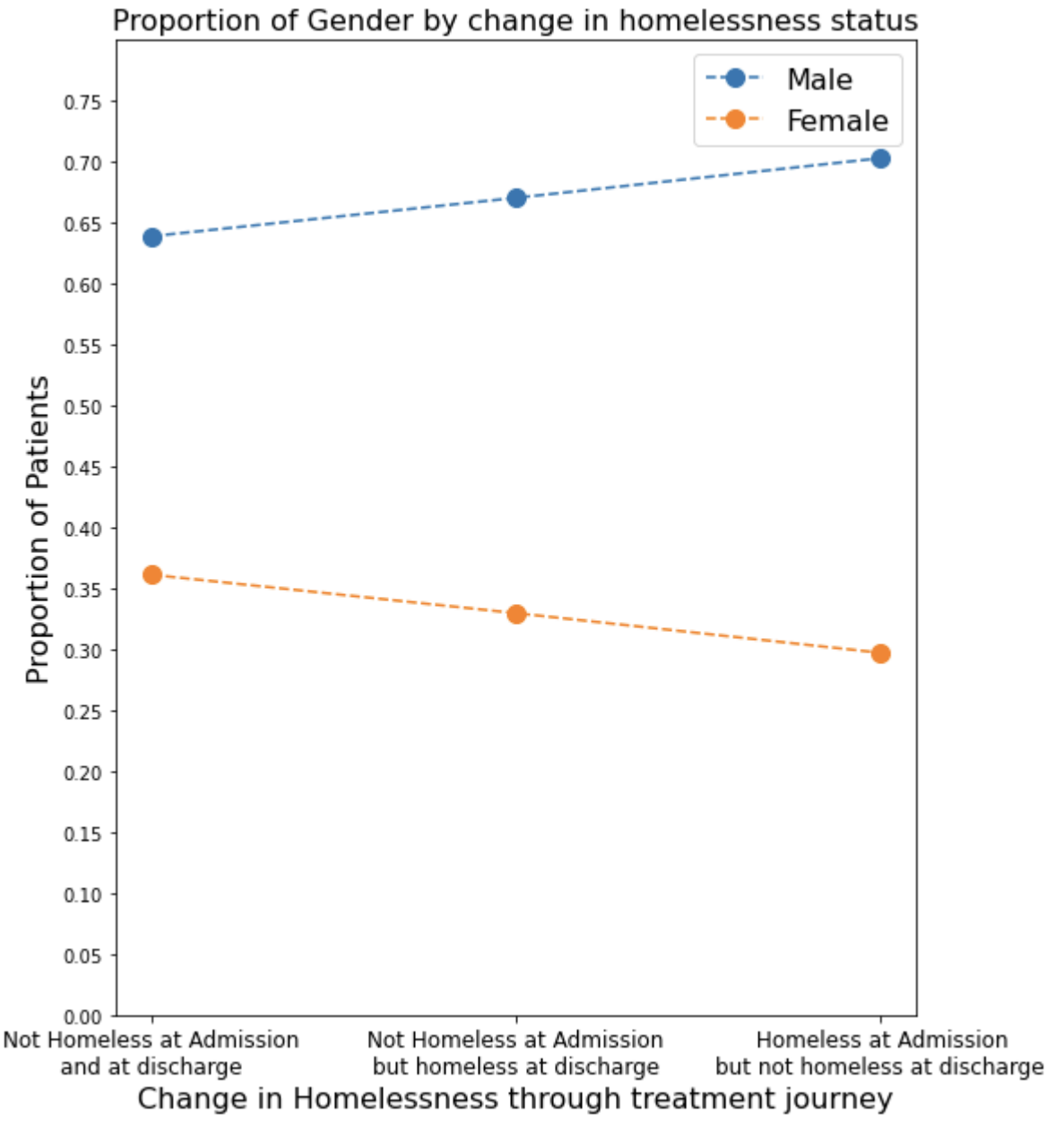
Proportion of Gender by change in homelessness status
Male
Female
Proportion of Patients
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Not Homeless at Admission and at discharge
Not Homeless at Admission but homeless at discharge
Homeless at Admission but not homeless at discharge
Change in Homelessness through treatment journey

Figure 7.

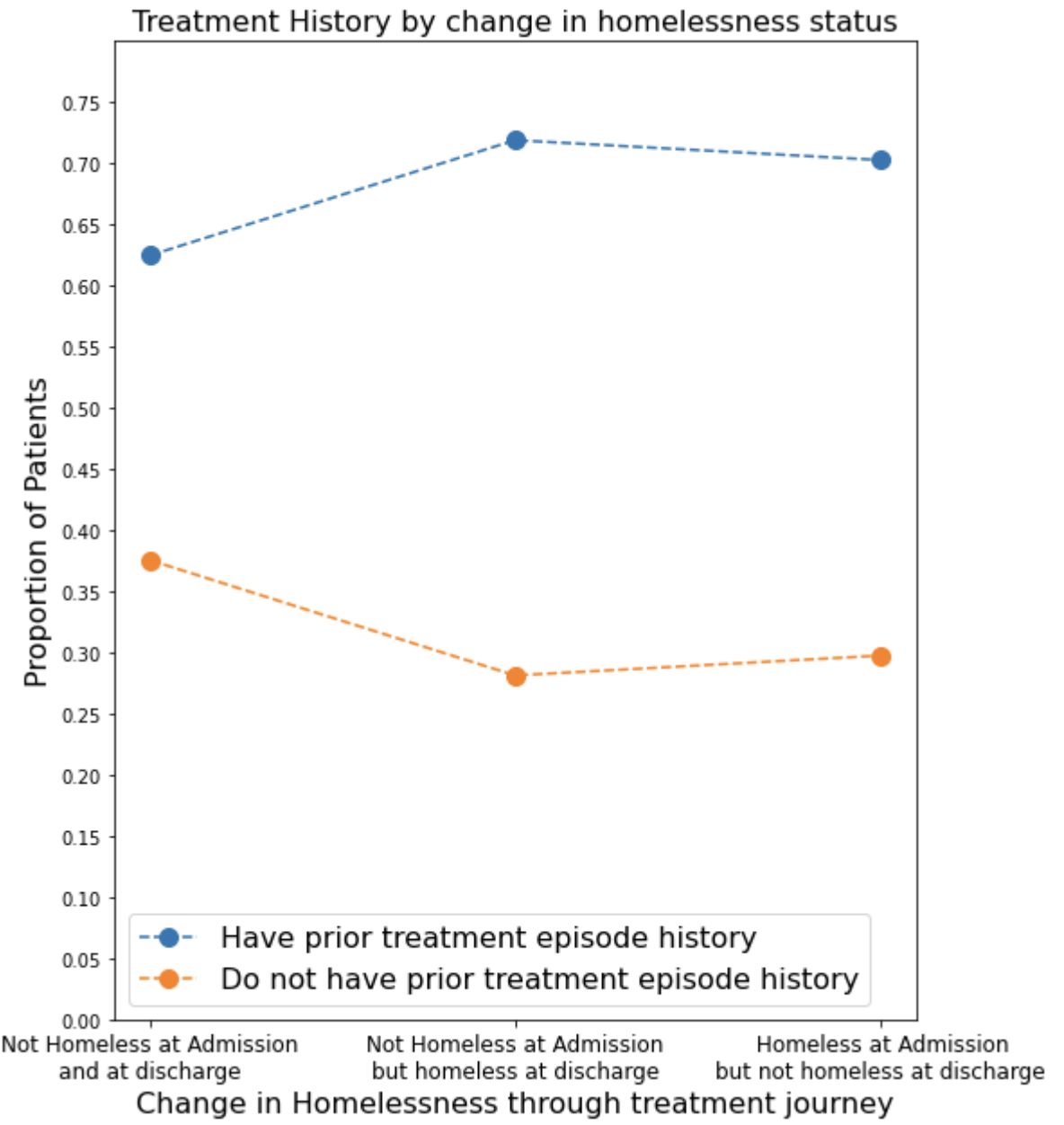
Treatment History by change in homelessness status
Proportion of Patients
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Have prior treatment episode history
Do not have prior treatment episode history
Not Homeless at Admission and at discharge
Not Homeless at Admission but homeless at discharge
Homeless at Admission but not homeless at discharge
Change in Homelessness through treatment journey

Figure 8.

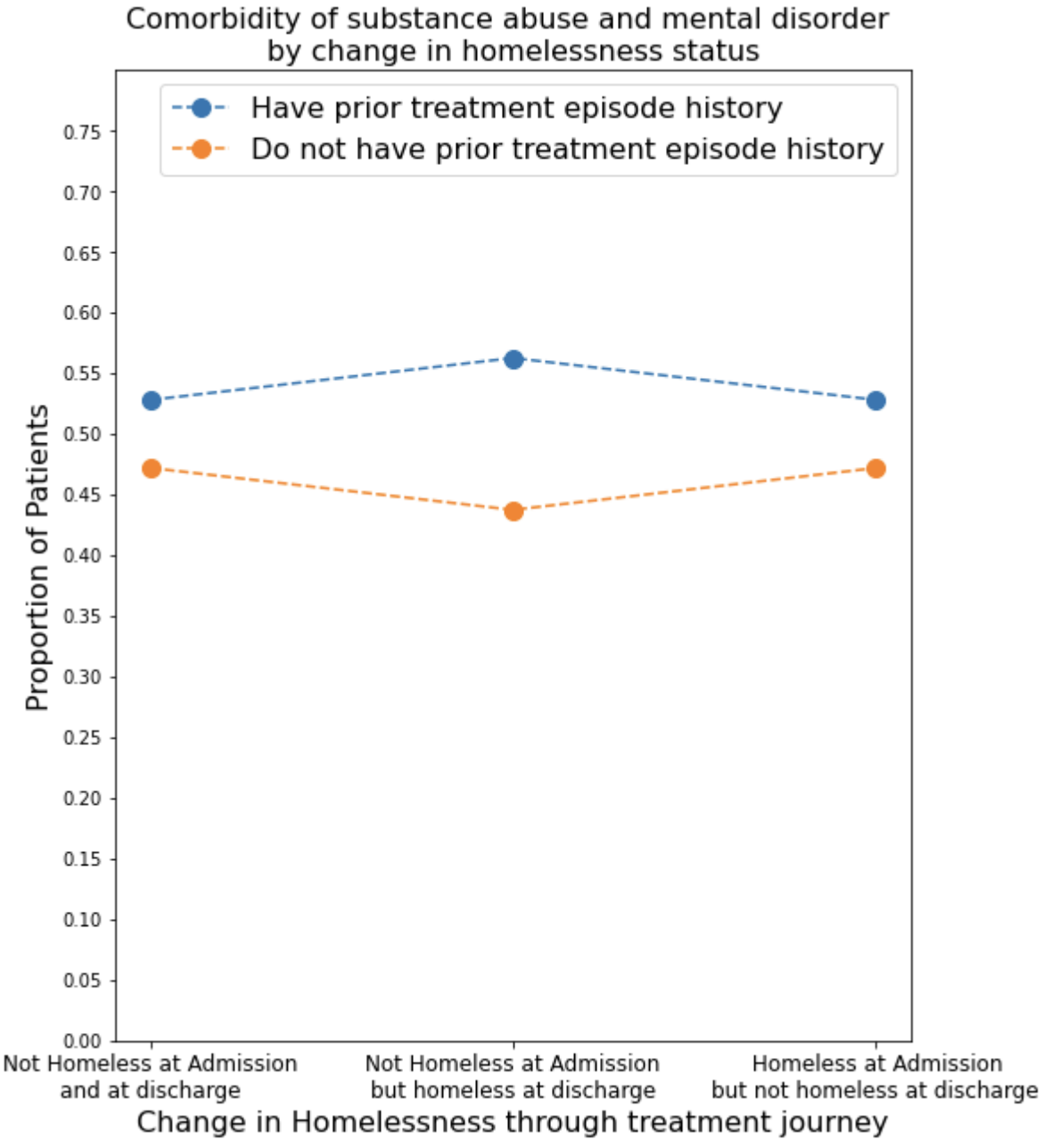

Figure 9.

In comparing those who become homeless by discharge and those were not homeless at admission and discharge we see that that the former group has a higher proportion of Black and Alaskan Native /American Indian populations, males, those with multiple past treatment episodes and those with co-occurring mental disorders. However, comparing those who become homeless by discharge (and not homeless at admission) with those who are no longer homeless by discharge (and homeless at admission) we see for the select variables there is little difference in the

proportions of these groups. Our prediction error, can therefore stem from our model being unable to differentiate between these groups.

The change in homeless status analysis, provides us valuable insights on how model errors may be distributed and inform how our model accuracy may need evaluation beyond traditional type 1 and type 2 errors. We also may benefit from stacking multiple models, if our base models are able to learn different types of information about the data.

4. Methodology

Since the task at hand is predicting a binary variable, i.e. homelessness status during discharge from treatment, our methodology incorporates model experimentation with parametric and multiple non-parametric methods and a combination of the two.

(A) Parametric Methods

Logistic Regression

Logistic regression (LR) is a widely used technique for classification and is employed to estimate the likelihood of an event occurring, such as "homeless" or "not homeless". Since the outcome is a probability, the dependent variable is limited between 0 and 1, by applying a *logit transformation* on the odds, i.e. probability of success divided by the probability of failure and can be represented as:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_n X_n \qquad \text{(James et al., 2021, p. 132).}$$

Here we estimate the beta parameter or coefficient using maximum likelihood estimation and then predict the probability of outcome based on values and coefficients of predictor variables. For binary classification, a probability greater than 0.5 is usually interpreted as predicting 1, while a probability less than 0.5 is interpreted as predicting 0. However, the thresholds depend on our final use case and domain knowledge & requirements.

Logistic regressions primary advantage lies in their interpretability as a parametric method. Each coefficient has a magnitude and direction, which can provide feature specific insights over predictions. They are also computationally efficient and handle small data well. However, LR assumes a linear relationship between the features and the log-odds of the target variable and such a simplified relationship may not represent the true nature of how the target variable is related to the features. Secondly, LR is sensitive to outliers and class imbalance and can have low predictive accuracy for such types of data.

(B) Non-Parametric Methods

B.1. Random Forests

Random forests (RFs) are based on the algorithm of decision trees. A decision tree creates multiple nodes representing features of the data, and holds decision rules at each node which attempt to optimally split the data. At each node of the tree, the algorithm chooses the feature that best splits the data on the basis of maximizing the information gain or minimizing the impurity in the target variable. The leaves or final nodes of the tree represent the predicted outcomes or classifications. The goal is to maximize the homogeneity of the subsets created by the split, such that the variance of the target variable is minimized within each subset.

RFs are an ensemble of decision trees, where each tree is trained on a random subset of the data with replacement and a random subset of the features. This subset generation method is known as *bagging* (Breiman, 1996). Each tree in the forest makes a prediction, and the final prediction is determined by combining the predictions of all the trees in the forest, through simple/weighted averaging.

The random subsets of features used to train each tree in the forest introduce additional randomness into the model, which helps to reduce overfitting and improve generalization. RFs also allow for finding feature importance in making

predictions adding to the model's interpretability. However, RFs may underperform when modelled on data for small or low-dimensional data, as the randomness becomes greatly reduced

B.2. Gradient Boosting

Gradient Boosting, utilizes an ensemble of weak learners. The algorithm begins with initializing, a weak learner such as a decision tree, linear/logistic regression etc. and is used to make predictions. The errors then made by this model are evaluated and the next iteration of this models uses the previous stages error to train a new model that focuses on reducing those errors. This process is repeated multiple times, with each new model building upon the errors of the previous model (Friedman, 2001). The key innovation in gradient boosting is the use of a gradient descent optimization algorithm by descending the gradient through iterative introduction of new models, as compared to gradient descent in artificial neural networks where 'descend' in the gradient is brought by updating network parameters.

Gradient boosted trees are a special case where the base/weak learner is a decision tree and the (E)Xtreme Gradient Boosting or XGboost (XGB) is one of the fastest implementations of gradient boosted trees.

It does this through parallelization and tackling one of the major inefficiencies of gradient boosted trees: considering the potential loss for all possible splits to create a new branch, by looking at the distribution of features across all data points in a leaf and using this information to reduce the search space of possible feature splits. Further it has multiple regularization methods to tackle overfitting.

However, despite its advantages and measures taken in the development of XGboost, gradient boosting is prone to overfitting especially if the data is noisy or there are too many weak learners. Further its computational complexity increases steeply, as the size of the dataset/number of features increases.

(C) Stacked Models

Model Stacking is a way to improve model predictions by combining the outputs of multiple models and running them through another predictive model called a "meta-learner". Essentially a stacked model works by running the output of multiple models through the meta-learner (usually a linear regressor/classifier, but can be other models like decision trees). The meta-learner attempts to minimize the weakness and maximize the strengths of every individual model. Stacked models rely on the fact, that different models are learning different aspects of the output and its associated information in the feature space.

In our modelling attempt, we look to combine results from our three base models i.e. LR, RF AND XGB to train our stacked classifier with our meta classifier as a random forest.

Stacked models in general have shown high performance due to their ability to combine the results and extract what each of type of model is learning about the data. However, they are computationally very demanding and hyper-parameter tuning for stacked models can be complicated both due to computational expense and the multiple layers of hyperparameters for each base model and the meta-model.

5. Literature Review

Nusinovici et al. (2020) in their study on predicting risk of diabetes found that LR was the best performing model for predicting one of their 4 possible outcome variables, with the highest area under the receiver operating characteristic curve of all the models tested. Further, the differences in performance between other machine learning models (Support Vector Machines and Neural Networks) and LR were not statistically significant. With regard to imbalanced and rare events data, and/or small samples, results from LR can be inconsistent unless techniques such as prior correction and weighting can be applied. King & Zeng (2001) applied these corrections to their LR model, and showed that they can make a difference when the population probability of interest is low.

Breiman (1996; 2001) demonstrated that substantial gains in classification and regression accuracy can be achieved by using ensembles of trees, where each tree in the ensemble is grown in accordance with a random parameter and

final predictions are obtained by aggregating over the ensemble. RFs are fast and easy to implement, can produce highly accurate predictions and can handle a very large number of input variables without overfitting (Biau, 2012). Using data on 5767 European listed companies, Ballings et al. (2015) compare the stock price direction predictive performance of RFs, with multiple models include LR. Using company fundamentals as features, they find that that ensemble methods like RFs have greater prediction accuracy over a one-year prediction period. In predicting clean energy stock prices Sadorsky (2021) also finds that decision tree bagging and random forests predictions of stock price direction are more accurate than those obtained from LR models.

Chen and Guestrin (2016) demonstrated through their original XGboost algorithm that tree boosting with sparsity aware algorithms can provide highly accurate results. With parallelization and distributed computing and exploiting out-of-core computation, the algorithm is computationally efficient for large datasets. Their study also highlights XGboost's performance in popular Kaggle competitions. Out of the 29 winning solutions published on Kaggle's blog in 2015, 17 solutions utilized XGBoost, while the second most popular method, deep neural nets, was used in only 11 solutions. Li et al. (2020) further show that in their application for predicting quality of personal credit loans, XGboost far outperforms other classification models including random forests and linear regression,

Thorne at al. (2017) demonstrate the efficacy of using stacked models in fake news detections. They show compared to each of the base classifiers individually which include neural networks, logistic regressions and gradient boosting, the stacked classifier combining the results of all the base models shows approximately 10% higher testing accuracy. Further Alves (2017) also demonstrated that stacked classifiers provide high accuracy in the context of identifying a new neutral Higgs boson. The found that while deep neural networks outperformed the stacking classifier it was also accompanied with high computational complexity. For a marginal decrease in accuracy, and with significantly lower computational complexity, stacked classifiers provided the most efficient solution.

Shifting the focus to previous research on homelessness, it is important to note that most of these studies have been primarily inferential in nature. The US Department of Housing and Urban Affairs' research focuses on forecasting homelessness, at an aggregated level to predict demand for Continuum of Care programs, using macroeconomic and aggregated demographic information (Nisar et al., 2019). While valuable from a funding and broader policy perspective, the unit of analysis does not allow for insights at the individual level and is unable to take into account a homeless individual's situation.

Studies utilizing TEDS most often focus on gaining insights on topics other than homelessness, such as treatment success (Acion et al., 2017) and disparities in addiction treatment received (Saloner & Cook, 2013). To the best of my knowledge, a study by Moxley et al. (2020) is the only research which uses the TEDS dataset to infer relationships between homelessness, substance abuse and racial disparities, but does not focus on prediction. Further they do so only for the state of Utah with 1862 cases, which as the authors also note, is not accurately representative of national demographics, especially for racial minorities. The study further creates multiple binary variables which reduce insights from their model such as classifying case records as having 'mental illness or not'. That certain types mental illnesses and similarly addiction to certain types of drugs may be relatively more strongly correlated with risk of homelessness is lost in the analytical framework of binaries.

This project is primarily focused on prediction at the individual level, with more nationally representative data and utilize more interpretable and informative variables developing on previous studies.

6. Model Development and Performance

The final feature list used is presented in Appendix A. Importantly, living arrangement at admission was not used as a predictor variable, since that leads to the prediction task being trivial. Since the goal of this research, is to contribute towards prediction of risk of homelessness of the general population based on demographic and mental health criteria, not using living arrangement at discharge further contributes towards this goal of generality. Below we discuss in further detail, the steps take in model development and our testing results and performance.

(A) Feature Selection

Feature selection was firstly based on data availability. All variables for which missing values were >40% were dropped. Such features may not be recorded due to state-based practices or privacy laws and hence cannot be classified as missing at random in order to be imputed. Further since all our features, barring one, are categorical variables, they are imputed with mode-based frequency. For the variables we do not use due to missingness, mode-based imputation with a relatively high percentage of missing values can introduce high bias, and create unrepresentative distributions of the categories.

Secondly multiple features were binary representations for specific classes, captured in other features. For eg. A separate feature captures "heroin in system at admission", while the variables primary/secondary/tertiary substances in system also captures the same as nominal variables. Such binary features were dropped, since creating dummy variables would lead to one of the dummy variables having perfect multicollinearity, with its respective binary feature, originally included in the data.

(B) Hyper-Parameter Tuning

Hyper-parameter tuning was undertaken using the Optuna Library in Python. Optuna allows for hyper parameter tuning for an efficient and exhaustive search in the hyperparameter space by utilizing Bayesian optimization methods to search for global optimization by iteratively building a probabilistic model of the function mapping from hyperparameter values to the objective function.

In our case, hyper parameter tuning was done through 5-fold cross validation, with our optimization parameter as balanced accuracy instead of accuracy. Due to the size of the dataset, while 10-fold cross validation may have yielded relatively more representative parameter values for higher performance, it also requires higher computational capabilities which were unavailable. Appendix C, outlines our intermediate feature engineering and data processing in detail.

(C) Modelling Data

- **Sampling and Imputation:** We firstly undertake mode based imputation in order to deal with missing values. Other more accurate methods of categorical imputation such as KNN and associated computational limitations are discussed in Appendix B**.** Due to the highly imbalanced nature of the dataset with only 13% of the data representing those who are homeless at discharge, we employ a combination of over and under sampling.
  Firstly, we keep all cases where homelessness status changed between admission and discharge. These cases are 5% of the entire dataset. It is possible there is something systemically different about this sub-group and present an important use case for our model. We therefore look to retain all possible information about such cases. Further due to their relatively low frequency they do not add significantly to computational complexity. We then under-sample the majority class, i.e. who are not homeless at discharge (and at admission) to 250,000 cases. Through experimentation, we found that such a magnitude of cases was computationally efficient and also predictions for this class label was accurate at an acceptable level. Finally, we use SMOTE-NC to over sample, the cases who were homeless at discharge (and at admission) to be equal to 250,000. This was the maximum number we could over-sample to without facing computational time-out.

- **Cross Validation**: We present our cross-validation accuracy, in order to understand each models generalized performance over different subsets of the data. We do this through Stratified 10-fold cross validation, where validation and test sets are split to ensure a comparably proportional number of cases for each output label. Due to possible differences in distribution of features between the training and testing dataset, based on the randomized state, cross-validation accuracy can provide a better sense of how our models are performing over the entire dataset.

- **Training and Testing**: We employ 70% of the data for training and employ 30% as testing data while undertaking stratified shuffling in order to ensure that the training and testing sets have comparable proportions of each output class label.

(D) Results

D.1 Performance

We present the performance results of our models in Table 5.

Model	CV Accuracy %	Testing Accuracy %	Balanced Accuracy %	Recall %
Logistic Regression	68	65.1	65.3	66
Random Forest	**70**	**67.5**	**67.4**	**67.2**
XGBoost	67	64.5	65.6	69
Stacked Classifier (Meta classifier: Random Forest)	71	68.4	67.5	65

Table 5: Modelling Experimentation Results

D.1.1. Discussion: Performance

Firstly, we see that our stratified cross validation accuracy, is comparable to our testing accuracy, which implies that our training set and testing set had comparable distribution of features.

Importantly, we find that Random Forests and Stacked classifiers have the highest performance for our given use case. Since, false negatives, i.e. predicting those who are homeless as not homeless, is a more impactful error we also give higher consideration to our recall score, i.e. number of correctly predicted positive cases as percentage of all true positive cases. For a small reduction in accuracy as compared to our stacked classifier, we gain a higher recall in our random forest model. While XGboost achieves a higher recall than either of the models, we can see that it has lower balanced accuracy. This implies that it is generating higher false positives, thereby inflating the recall scores.

We can further observe tradeoffs between false and true positives each of our models make, from the area under the receiver operating characteristics curve in Figure 10.

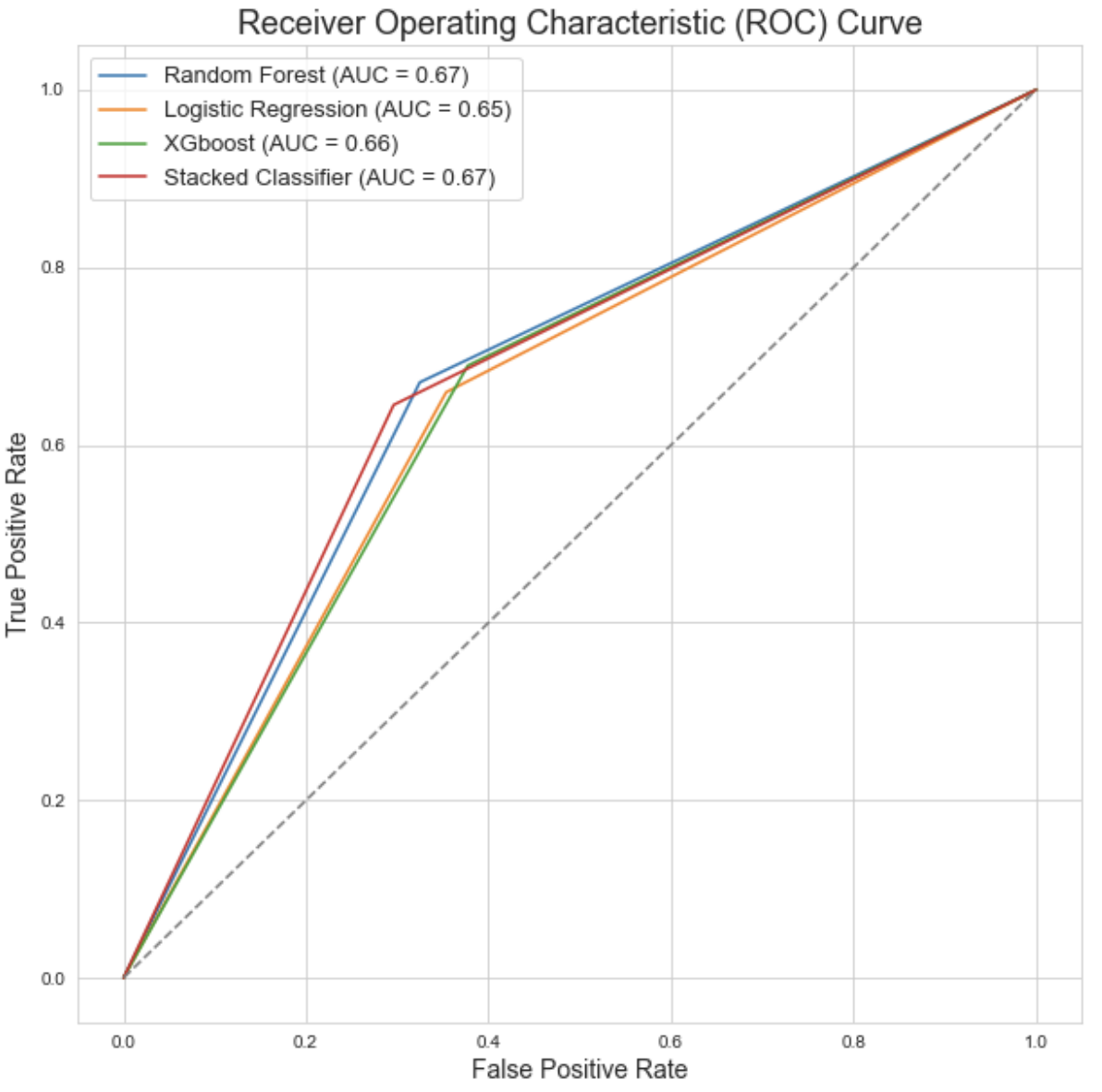

Figure 10.

D.1.2. Discussion: Receiver Operating Characteristics Curve

We see that amongst all models, our stacked classifier is able to achieve the highest true positive rate for the comparable false positive rate. However, all three models perform roughly the same overall. This informs us that the models are not significantly different in terms of what they are learning about each of the output classes. Random forests tend to marginally overidentify cases as homeless at discharge compared to the stacked classifier, at the cost of the stacked classifier under identifying true positive cases. As we discussed XGB's high recall previously, we can observe from XGBs ROC, that for the threshold where true positive rates are highest and roughly equivalent true positive rate to RF and SC, the false positive rate for XGB is much higher. Overall, ROC confirms our models less than satisfying performance, with our highest AUC being 0.68, compared to the base case when the false positive rate is equal to the false negative rate represented by the dashed line at 45 degrees, where area under the curve is 0.5 and correct prediction of cases is equivalent to the flip of a coin.

D.2. Errors

We next explore the nature of the errors our model is making.

From our confusion matrices presented in Appendix C, we find that type 1 error, is the majority driver of the errors in our model. That is all of our models are primarily misclassifying cases who are not homeless at discharge as

homeless at discharge. Such errors are unexpected from the model, since we initially saw in our explanatory analysis, that the differences in distributions across important variables were indeed different. We therefore, asses our errors, by further breaking them down by homelessness status at admission in Table 6.

Living Status at Admission	Living Status at Discharge	% Misclassified
Not Homeless	Homeless	50
Homeless	Homeless	30
Homeless	Not Homeless	60
Not Homeless	Not Homeless	27

Table 6: Distribution of Errors

D.2.1. Discussion: Errors

The errors stem from those whose homelessness status changes during the course of the treatment. Misclassification for either of the cases facing change in homelessness indicates, that features might be more closely related for homelessness at admission for such cases. That is, those who are not homeless at admission but are so at discharge, are more similar to those who are not homeless at either points during treatment and vice versa. This indicates, that there may be some systemic difference not captured in our variables or alternatively, people recorded in such cases face an exogenous shock during the period of treatment, which leads to their change in homelessness status. This is a key finding from our model, as it helps us find that risk of homelessness appears in two forms: those who are systemically at risk of homelessness and those who are not yet, but due to a sudden change in conditions become homeless.

7. Feature Importance

In order to understand, what features are important predictors of homeless, we look at feature importance from our random forest model, presented in figure 11.

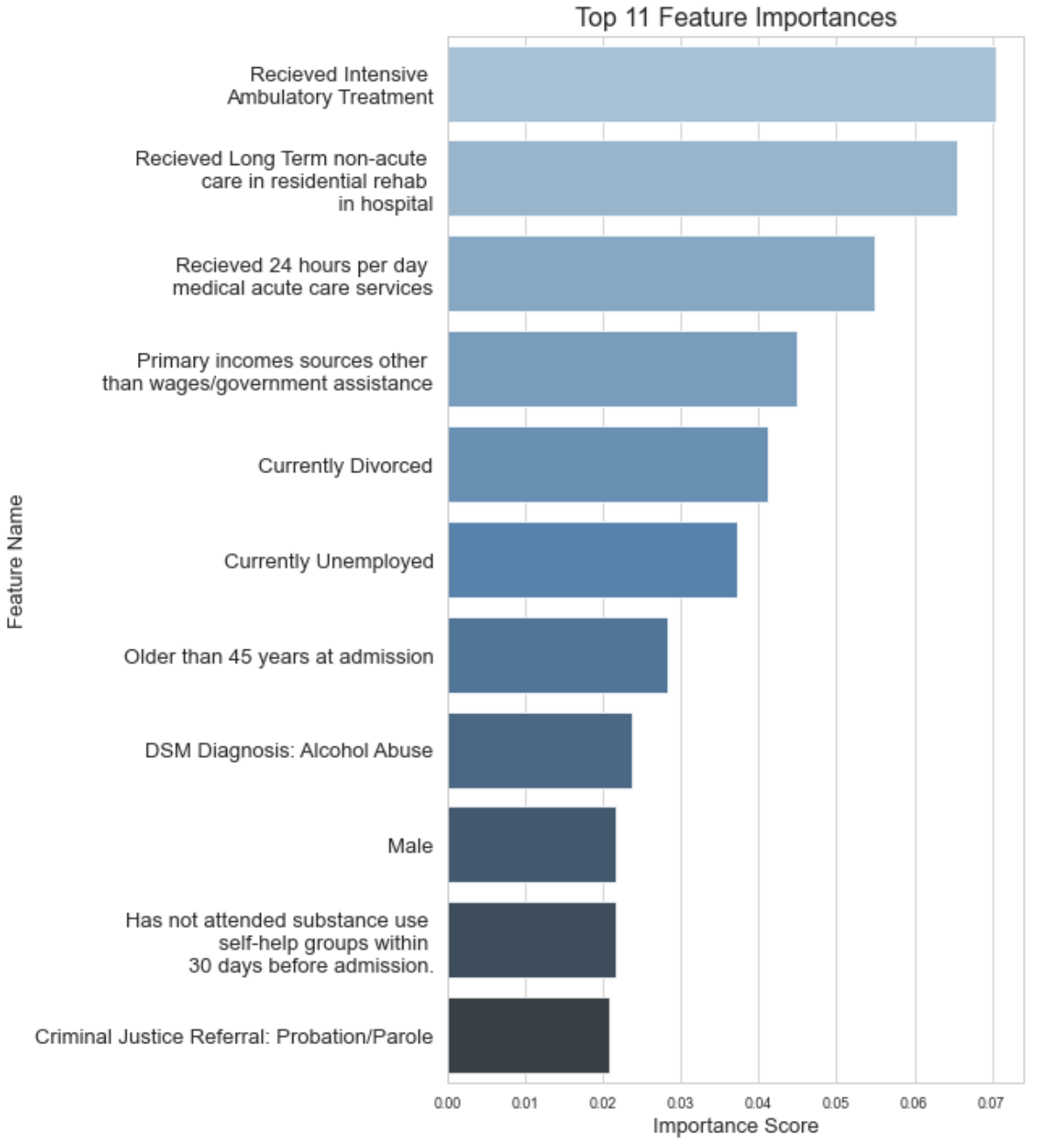

Figure 11.

7.1. Discussion: Feature Importance

We see that types of treatment received at admission are the top three variables in deciding the output label. In the TEDS-D dataset, types of treatment are captured as a detailed ordinal variable and can be interpreted as a proxy of how serious/acute one's substance abuse problem is, what type of social support structures exist for such patients, how socially secure are patients (for example, in their ability to undertake residential rehab possibly at the cost of foregone wages) etc. We also see, that a DSM diagnosis of alcohol abuse is an important predictor, highlighting that mental disorders restricting one's ability to judge the impact of alcohol in their lives, has a particularly significant effect as compared to other possible mental disorders/ substance abuse mental disorders. We thus show that treating mental disorders as a "having or not" binary or popular narratives of focus on illegal synthetic substances such as opiates may not fully capture the underlying issues in homelessness and substance abuse. Among other important features, we interestingly find that attendance in self-help groups and treatments undertaken due to legal directives of their probationary/parole status are important predictors of homelessness status at discharge. The importance of these variables identifies the role that support systems and criminal justice systems might play for those who are at risk of being homeless. Given that such variables can be influenced through policy and public

encouragement programs, we also gain important insights in terms of what factors can be exogenously influenced to understand our outcomes.

8. Policy Implications and Conclusion

From our above modelling attempt, we do find evidence that a person's characteristics today provide predictive information regarding a person's homelessness status in the future. Importantly, we show the same, without using any information about a person's current homelessness status and find that acuteness of the substance abuse problem and its associated treatment, income sources, mental disorders surrounding alcohol abuse, attendance of self-help groups and criminal justice referrals are important predictors of homelessness. Despite, less than expected accuracy, our distribution of errors also provides important policy insights. In particular, there are a small number of cases, where we find that one's homelessness status changes over time but whose features are difficult to distinguish from cases where the homelessness status does not change. This difference between those who have faced long term homelessness versus those whose housing status changes in short period of time, presents an important line of inquiry towards factors distinguishing them and understanding factors that explain homelessness. If there is a possibility of 'shock' events changing one's homelessness status suddenly then this requires important consideration for robust policy design. For example, if a person who is otherwise not at risk of homelessness, suddenly faces the loss of the primary earner on whom they depend and do not have capabilities to sustain themselves, then interventions for such shock events require separate mechanisms for identification and improving outcomes. Such level of identification can be very resource intensive, given that these events cannot be captured easily through 'data' in the context of machine readability and requires in-depth interaction through surveys and interviews. For our model and the data available to us, these cases present a unique challenge, in their relative rarity of occurrence - possibly not providing enough information about such cases and/or being affected by factors not captured in the features. Further research, into how change in homelessness can be explained and data more generalizable to the overall population at risk of homelessness would provide significant insights towards this question.

9. Limitations

A. <u>Sample</u>

One of the primary limitations of the data, is that it is at the case level, i.e., if the same person is admitted to the program twice, two separate cases are logged. Due to each case ID being unique such cases cannot be grouped together. Therefore, some data points do not contribute unique information to the model, which may affect accuracy and bias of the model. Secondly, the data is not fully representative of how substance abuse and other conditions predict homelessness for primarily three reasons. Data for 6 states have not been recorded and additionally Arizona is not being considered in our sample as none of the outcome variables for the state have been recorded. Secondly TEDS only contains information about cases where someone suffering from substance abuse 1) agrees to seek treatment 2) agrees to seek treatment at a publicly funded facility. If community-based characteristics are correlated with seeking/completing such treatment, then the model trained on this data may be biased. We also could not use our total available sample and more data points could have contributed to higher accuracy.

<u>Methodology</u>

We also identify some additional limitations in our methodology. Some features were dropped due to a high number of missing values. If such features captured important differences in information about our output labels, this would explain the bias in our prediction results. Further, we undertake mode-based imputation which may change the distribution of our features and further bias our models. For categorical features using K nearest neighbors with a 'most-frequent' decision making strategy may have yielded a more representative imputed dataset. However due to computational complexity, we could not undertake KNN based imputation. We also drop all cases for which the outcome variable is missing and as discussed previously, these cases show correlation for those who are unable to complete the treatment and drop out mid-way. If factors affecting dropout are correlated with homelessness status, then our model may be further biased.

Bibliography

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

Alves, A. (2017). Stacking machine learning classifiers to identify Higgs bosons at the LHC. Journal of Instrumentation, 12(05). https://doi.org/10.1088/1748-0221/12/05/t05005

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 42(14), 7046-7056.

Bhattarai, A., & Siegel, R. (2022). Inflation is making homelessness worse. *Washington Post*. https://www.washingtonpost.com/business/2022/07/03/inflation-homeless-rent-housing/

Biau, G. (2012). Analysis of a Random Forests Model. Journal of Machine Learning Research, 13(38), 1063-1095

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32.

Center for Behavioral Health Statistics and Quality. (2020). Treatment Episode Data Set (TEDS), Discharges, 2020 [Data set]. Substance Abuse and Mental Health Services Administration (SAMSHA). https://www.samhsa.gov/data/data-we-collect/teds-treatment-episode-data-set

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 785–794. https://doi.org/10.1145/2939672.2939785

Cohen, E. (2022). The effect of Housing First programs on future homelessness and socioeconomic outcomes. (Federal Reserve Bank of Kansas City, Research Working Paper no). https://doi.org/10.2139/ssrn.4071014

de Sousa, T., Andrichik, A., Cuellar, M., Marson, J., Prestera, E., & Rush, K. (2022). The 2022 Annual Homelessness Assessment Report (AHAR) to Congress. *The U.S. Department of Housing and Urban Development*. https://www.huduser.gov/portal/sites/default/files/pdf/2022-AHAR-Part-1.pdf

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189–1232. http://www.jstor.org/stable/2699986

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R (2nd ed.). Springer

Kertesz, S. G., Austin, E. L., Holmes, S. K., DeRussy, A. J., Van Deusen Lukas, C., & Pollio, D. E. (2017). Housing first on a large scale: Fidelity strengths and challenges in the VA's HUD-VASH program. *Psychological services*, 14(2), 118–128. https://doi.org/10.1037/ser0000123

King, G., & Zeng, L. (2001). Logistic regression in rare events data. Political Analysis, 9(2), 137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Li, H., Cao, Y., Li, S., Zhao, J., & Sun, Y. (2020). XGBoost Model and Its Application to Personal Credit Evaluation. IEEE Intelligent Systems, 35(3), 52–61. https://doi.org/10.1109/mis.2020.2972533

Moxley, V. B. A., Hoj, T. H., & Novilla, M. L. B. (2020). Predicting homelessness among individuals diagnosed with substance use disorders using local treatment records. *Addictive Behaviors*, 102, 106160. https://doi.org/10.1016/j.addbeh.2019.106160

Nisar, H., Vachon, M., Horseman, C., & Murdoch, J. (2019). Market Predictors of Homelessness: How Housing and Community Factors Shape Homelessness Rates Within Continuums of Care. *The U.S. Department of Housing and Urban Development*. https://www.huduser.gov/portal/sites/default/files/pdf/Market-Predictors-of-Homelessness.pdf

Nusinovici, S., Tham, Y. C., Chak Yan, M. Y., Wei Ting, D. S., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C.-Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. Journal of Clinical Epidemiology, 122, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

OrgCode Consulting. (2015). Vulnerability Index—Service Prioritization Decision Assistance Tool (VI-SPDAT): Prescreen Triage Tool for Single Adults. http://everyonehome.org/wp-content/uploads/2016/02/VI-SPDAT-2.0-Single-Adults.pdf

Rountree, J., Hess, N., & Lyke, A. (2019). Health Conditions Among Unsheltered Adults in the U.S. *California Policy Lab*. https://www.capolicylab.org/wp-content/uploads/2023/02/Health-Conditions-Among-Unsheltered-Adults-in-the-U.S..pdf

Saloner, B., & Cook, B. L. (2013). Blacks And Hispanics Are Less Likely Than Whites To Complete Addiction Treatment, Largely Due To Socioeconomic Factors. *Health Affairs*, 32(1), 135–145. https://doi.org/10.1377/hlthaff.2011.0983

Sadorsky, P. (2021). A random forests approach to predicting clean energy stock prices. Journal of Risk and Financial Management, 14(2), 48. https://doi.org/10.3390/jrfm14020048

Thorne, J., Chen, M., Myrianthous, G., Pu, J., Wang, X., & Vlachos, A. (2017, September 1). Fake news stance detection using stacked ensemble of classifiers. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/W17-4214

U.S. Census Bureau. (2022). QuickFacts: United States. Retrieved April 9, 2023, from https://www.census.gov/quickfacts/fact/table/US/PST045221

Appendix A. Features and Target Variables

- ***LIVARAG_*D: Living arrangement at time of discharge (*Target Variable*)**
- *EDUC*: Highest level of education completed
- *MARSTAT*: Marital status at time of admission
- *SERVICES*: Type of service setting where treatment was received
- *DETCRIM*: Legal status at admission based
- *NOPRIOR*: Number of prior treatment episodes
- *ARRESTS*: Number of times arrested in the past year
- *EMPLOY*: Employment status at admission
- *PSYPROB*: Presence of emotional/psychiatric problems at admission
- *PREG*: Indicates if the client is pregnant
- *GENDER*: Gender
- *VET*: Veteran status
- *DSMCRIT*: Primary substance abuse and/or dependence diagnosis according to DSM-IV criteria
- *AGE*: Age at admission
- *RACE*: Race
- *ETHNIC*: Ethnicity
- *PRIMINC*: Primary source of income
- *SUB1*: Primary substance of abuse
- *n_sub*: Number of substances in system (*Feature Engineered as SUB1+ SUB2+ SUB3*[1]*)*
- *ROUTE1*: Primary route of administration for the primary substance of abuse
- *FREQ1*: Frequency of use for the primary substance of abuse
- *min_firstuse*: Minimum of age at first use of primary substance, secondary substance, tertiary substance (Feature Engineered)
- *FREQ_ATND_SELF_HELP*: Frequency of attendance at self-help groups in the past year
- *IDU*: Indicates if the primary substance of abuse is injected
- *ALCDRUG*: Indicates if the primary substance of abuse is alcohol or drugs

Appendix B. Feature Engineering and Over-Sampling Process

- Four our target variable, we create a binary variable as homeless o not homeless. Those having status as dependent/independent living, belong to the latter category.
- For our RACE variable, we combined Alaska Natives and Native Americans as one category of Native Americans. Further different categories for multiple races were combined into a single category denoting two or more races
- For primary/secondary/tertiary substances we group all substances with low occurrences (between 0% and 1% of total samples) as the 'other' substances category.
- For DSM criteria, we group DSM diagnosis by substance i.e for example, alcohol disorder and alcohol abuse are grouped in the category alcohol.
- We generate 6 age groups as : 12-14, 15-18, 19-24,25-34, 35-44 and 45 years of age and above.
- We also generate two variables: *n_sub* and *min_firstuse*. The former, takes values 1,2,3 highlighting if they had only one, two or three substances in their system at admission respectively. The latter, is the youngest age at which patient notes first use of any of the primary, secondary and tertiary substance they report.

[1] SUB2 AND SUB3, refer to the secondary and tertiary substances in system at time of admission.

- We then undertake mode-based imputation to deal with missing data. KNN based multiple imputation using 'most-frequent' was attempted but could not be undertaken due to computational complexity.
- As a final step we generate dummy variables for all features barring *n_sub*, since it is not a categorical feature and drop the first class of each variable as the reference category, to avoid multi-collinearity. We use SMOTE-NC methodology for scikit learn available in python 3.0, since we are dealing with categorical variables in our case.

Appendix C. <u>Confusion Matrices</u>

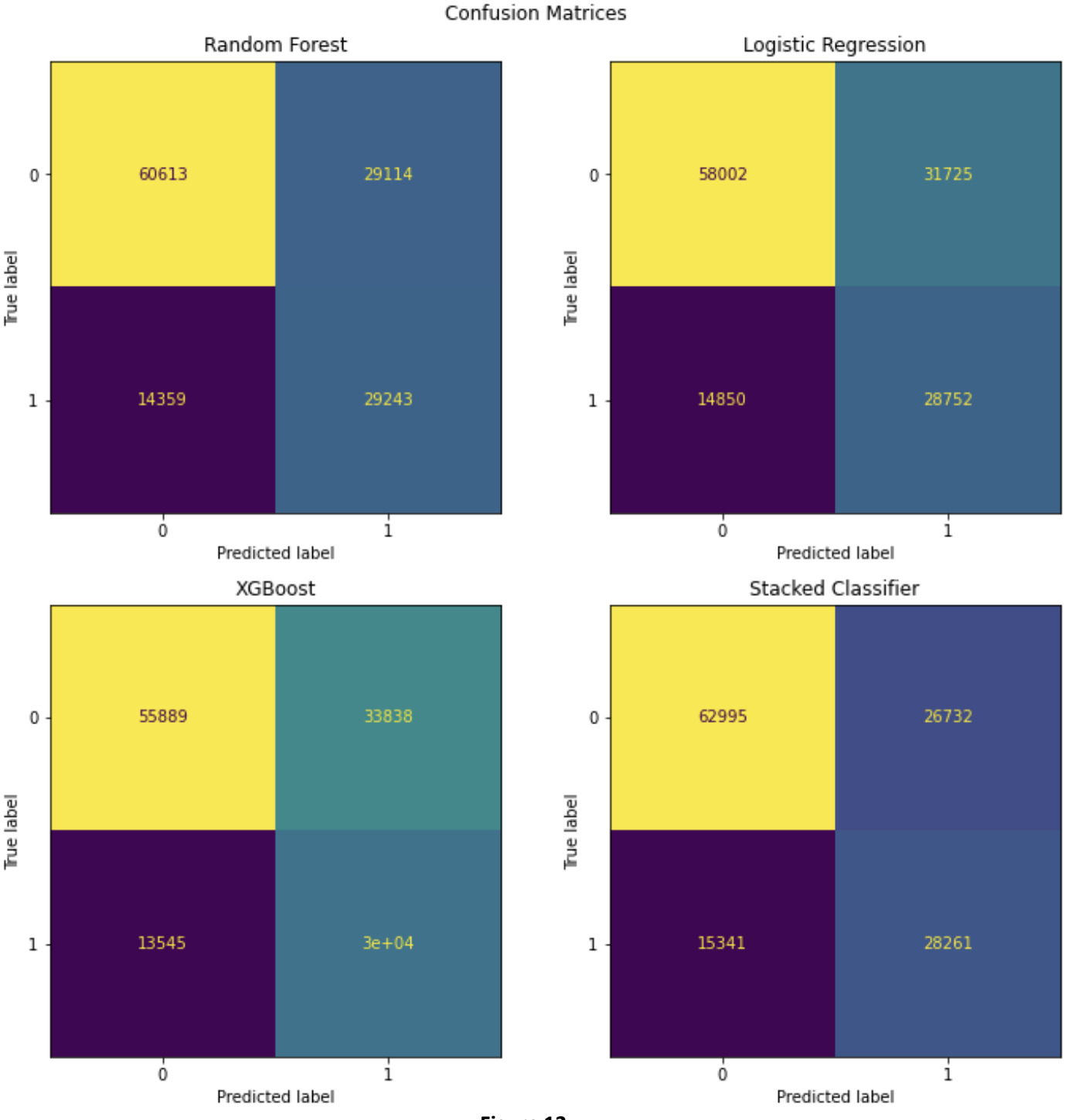

Figure 12.