American Journal of Public Health Research, 2020, Vol. 8, No. 5, 140-146
Available online at http://pubs.sciepub.com/ajphr/8/5/2
Published by Science and Education Publishing
DOI:10.12691/ajphr-8-5-2

A Machine Learning Approach to Identify Socio-Economic Factors Responsible for Patients Dropping out of Substance Abuse Treatment

Prateek Gautam[1], Pradeep Singh[2,*]

[1]Marquette High School, Chesterfield, MO, USA
[2]Department of Mathematics, Southeast Missouri State University, Cape Girardeau, MO, USA
*Corresponding author: psingh@semo.edu

Received July 05, 2020; Revised August 07, 2020; Accepted August 16, 2020

Abstract In recent years, the subject of substance abuse has drawn considerable attention from researchers and policymakers alike. Researchers have been utilizing the wealth of patient level data available from various agencies to develop prediction models for the relationship between socio-economic factors and substance abuse issues. According to the Substance Abuse and Mental Health Services Administration (SAMHSA), in 2017, 26% of patients admitted to treatment facilities drop out prematurely, which is significant when considering that roughly 1.5 million people are admitted to these treatment facilities every year, thereby revealing the need for an analysis to identify variables associated with such a large number of people not completing treatment. This study applies Multiple Logistic Regression (MLRM) as well as Random Forest Classification (RF) model to determine significant socio-economic factors responsible for patients prematurely dropping out of substance abuse treatment for opioid misuse. A MLRM has its limitations when the dataset has a large number of categorical variables; machine learning methods such as RF have proven more effective and accurate when dealing with such data. Patient level data from the Treatment Episode Dataset - Discharge (TEDS-D 2017) was analyzed and the models were compared using the Area Under the Curve (AUC) operating characteristic. The MLRM was found to have an AUC of .68 while the RF model had an AUC of .89, thereby demonstrating the advantage of machine learning methods. The factors deemed significant from the RF model can provide healthcare professionals as well as administrative officials with the necessary information to help address the issue of patients prematurely dropping out of opioid misuse treatment.

Keywords: *opioid abuse, random forest, machine learning, TEDS-D, treatment dropout*

Cite This Article: Prateek Gautam, and Pradeep Singh, "A Machine Learning Approach to Identify Socio-Economic Factors Responsible for Patients Dropping out of Substance Abuse Treatment." *American Journal of Public Health Research*, vol. 8, no. 5 (2020): 140-146. doi: 10.12691/ajphr-8-5-2.

1. Introduction

According to the National Survey on Drug Use and Health (NSDUH), in 2018, about 164.8 million Americans aged 12 or older were past month substance users, including tobacco, alcohol, and illicit drugs [1]. A large proportion of these illicit drugs are opioids, a class of drugs that has contributed to a national emergency due to their overly addictive nature.

States have employed several tactics, such as investing in Prescription Drug Monitoring Programs (PDMP) to curb opioid related deaths, and improvements can be seen. Patrick et. al [2] finds that a state's implementation of a PDMP was associated with an average reduction of 1.12 opioid-related overdose deaths per 100,000 population in the year after implementation.

Additionally, databases like the Substance Abuse and Mental Health Services Administration (SAMHSA) has provided the means for researchers to begin to employ data driven tactics [3] to control physician prescribing patterns, especially because physicians in the past have established that analyzing demographics plays a major role in their opioid prescribing habits. Wright et. al [4] demonstrates the benefits of a data-driven approach by effectively utilizing the "socio-ecological context" in their study. Using community level data they determined that the structure of a local healthcare system is a "major determinant in community-level access to opioids," highlighting the necessity to monitor the prescribing patterns of physicians.

In the following years, this approach was expanded upon to include the analysis of economic factors in communities. Hollingsworth et. al [5] examined macroeconomic conditions at the state level and determined that increased unemployment rates, as well as general macroeconomic shocks, lead to an increase in overall drug rates, which is "driven by rising opioid deaths". Brown and Wehby [6] sought to utilize both economic and demographic characteristics at the state level to identify predictors of increased opioid-related death rates. They found that economic

downturns that substantially reduce house prices can increase opioid related deaths, indicating the need for heightened access control during these periods.

Still, while much of this research has the potential to control one aspect of opioid misuse (prescription patterns and access), another equally important aspect of opioid misuse has not been addressed on a large scale: opioid misuse *treatment*. According to Treatment Episode Data Set - Discharge (TEDS-D) 2017, roughly 26% of patients at substance abuse treatment facilities across the nation drop out of treatment prematurely.

Acion et. al [7] utilized Super Learning (SL) methods and logistic regression to analyze TEDS-D data spanning from 2006-2011 to find common socio-economic factors among Hispanics in adulthood who prematurely drop out of substance abuse treatment for alcohol, cocaine/crack, marijuana/hashish, prescription opioids/synthetics, and methamphetamines. The scope of research conducted by Acion et al [7], however, was limited to simply comparing the accuracy of various prediction models rather than identifying socio-economic factors that truly contribute to patients prematurely dropping out of substance abuse treatment.

Stahler et. al [8] used TEDS-D data from 2013 to identify racial/ethnic disparities in substance use treatment completion. While they were able to determine that Black and Hispanic patients were less likely to complete treatment than white patients, their study was limited to only 42 of the largest US metropolitan areas and only included race/ethnicity as a predicting factor instead of including other socio-economic variables. Godinet et. al [9] adopted a similar approach using TEDS-D data from 2016. Despite including other demographic characteristics as predictors, their focus was on the Asian American and native Hawaiian and Other Pacific Islander populations. Both of these studies, while having the capacity to help address issues regarding substance abuse treatment completion, lack a comprehensive perspective.

There has not yet been a comprehensive study that examines people of all demographics on a large scale to identify socio-economic factors that contribute to patients dropping out of treatment. This paper looks at over 500,000 patients from TEDS-D 2017 with two goals: First, it compares the relative strengths of Random Forest and Multiple Logistic Regression prediction models. Secondly, it identifies specific socio-economic factors that contribute to patients dropping out of opioid misuse treatment.

This research will have an immense impact on efforts to combat the opioid epidemic by providing socio-economic factors that can be used by public health leaders to develop more comprehensive interventions that ensure complete treatments for all patients, and help healthcare providers administer more targeted and equitable treatments that meet the individual needs of all patients.

2. Materials and Methods

2.1. Data Collection

There are several agencies that collect information on patients and make the data available for public use. We first examined the National Survey on Drug Use and Health (**NSDUH**). This set includes roughly 67,500 patient responses from both publicly funded and private treatment facilities, collected through face to face interviews. However, since the number of responses in comparison to the total number of substance abuse treatment patients across the nation was low, and the survey was voluntary, this data set likely does not properly represent all patients being treated for opioid misuse. The next dataset examined was the National Survey of Substance Abuse Treatment Services **(N-SSATS).** In contrast to the NSDUH set, N-SSATS includes almost 1.1 million data points collected from private and public treatment facilities. Still, the voluntary nature of the survey, coupled with the allowance of multiple responses for certain variables, led us to conclude that this dataset may be skewed as well. The final set we looked at and eventually used in our analyses was the Treatment Episode Dataset: Discharge **(2017 TEDS-D).** TEDS-D includes data on over 1.5 million patients, collected not through voluntary surveys, but rather from state administrative tracking systems. This meant that along with not being biased, we could be sure that TEDS-D included almost all government backed treatment facilities and the results from our analyses could be used by administrations to implement changes.

2.2. Choosing Variables

The TEDS-D dataset records 77 different socioeconomic variables related to all admitted patients, including age, gender, employment status, etc. However, some of the variables overlapped in scope, which meant that including all variables in our analyses could result in redundant information being analyzed. This exclusion criterion eliminated 48 variables. Table 1 provides a list of the variables used.

Table 1. List of Variables Used

Census State FIPS Code	Level of Education	Marital Status	Type of Service Provided at Admission
Length of Stay	Treatment Referral Source	Number of Previous Substance Use Treatment Episodes	Number of Arrests 30 Days Prior to Admission
Employment Status at Admission	Co-Occurring Mental and Substance Use Disorders	Pregnant at Admission	Gender
Veteran Status	Living Arrangements at Admission	Type of Service Provided at Discharge	Employment Status at Discharge
Number of Arrests 30 Days prior to Discharge	Age	Race	Primary Source of Income
Primary Substance Used at Admission	Frequency of Substance Used at Admission	Age of First Use of Primary Drug	Health Insurance at Admission
Primary Source of Payment for Treatment	Frequency of Attending Self Help Sessions	Frequency of Attending Self Help Sessions at Discharge	Census Geographic Division
Alcohol Use at Admission			

TEDS-D includes patients admitted for any type of substance abuse treatment; we had to identify only those data points that represented patients that had opioid abuse as their primary reason for admission to a treatment center. We were then left with 546,945 points out of the original 1.5 million data points, representing roughly ⅓ of the total set. In the original TEDS-D dataset, the response variable, "Reason for Ending Treatment," is a polychotomous variable (7 responses: completed treatment, dropped out of treatment, terminated by facility, transferred to another facility, incarcerated, death, other). We were interested in a dichotomous variable, where a patient either did or did not drop out of their treatment program prematurely; this would allow us to conduct a Multiple Logistic Regression analysis. A "1" was assigned to all data points where patients dropped out of treatment, and a "0" was given to all other points (treatment ended successfully, transfer to another facility, etc.). About 28% of patients (≈ 150,000) were found to have not completed their treatment.

2.3. Statistical Analysis

A Multiple Logistic Regression Model (MLRM**)** is widely used when two or more independent variables (IV) are being used to predict a dichotomous variable (DV), whose outcomes are usually denoted with 0's and 1's. The goal of a multiple logistic regression is to find an equation that best predicts the probability of a value of the response variable as a function of the independent variables. In this study, the probability of dropping out of opioid abuse treatment is being predicted by socio-economic variables. While a MLRM provides easier interpretation of the relation between independent and response variables, there is often a compromise in its accuracy and specificity when the model includes a large number of categorical and independent variables [10]. In recent years, machine learning methods have become increasingly popular analyzing tools for large datasets. One such method is Random Forest Classification (RF).

A Random Forest consists of many individual decision trees. Each decision tree consists of a class prediction created by splitting the data at nodes, points where the data is separated on the basis of some characteristic so that each resulting group is as different as possible, and members of each subgroup are as similar as possible. For each time a split in a tree is considered, a random sample of 'm' predictors is chosen as split candidates from the full set of 'p' predictors. The split is allowed to use only one of these 'm' predictors. A fresh sample of 'm' predictors is taken at each split, and typically we choose 'm' to be approximately the square root of the total number of predictors. Forcing each split to consider only a subset of the predictors increases the chance of each predictor to be part of the tree. In this study, p=29 and m=5.

Bootstrapping was used to reduce the variance by taking repeated samples from the same dataset. Every single decision tree uses a new, random, bootstrapped data sample to ensure the entirety of the dataset is being properly represented. RF Classification can predict outcomes with a much higher level of accuracy and reliability. However, interpretation of the relation between variables is much more difficult.

We used the open source statistical analysis program 'R' to carry out our analyses. For the MLRM, we split the data into training (70%) and test (30%) sets. The model was obtained using the training data and validated using the test data.

For Random Forest Classification, 200 decision trees were generated, and out-of-bag data was used for the validation of these trees.

3. Results and Discussion

Table 2 provides an analysis of the maximum likelihood estimates of all 29 independent variables used in the study. Any variable with a p-value less than .05 is deemed significant, and 22 out of the 29 independent variables were found to be significant predictors of a patient dropping out of treatment.

Generally, a person's health insurance is a factor in the duration of hospitalization and quality of care they receive. However, Health Insurance (HLTHINS) does not contribute to the probability of a patient dropping out. Similarly, having a co-occurring mental and substance use disorder would often be expected to play a role in whether or not someone drops out of treatment. Yet, the variable (PSYCHPROB) was not found to be helpful in predicting dropout rates.

Figure 1 shows the Receiver Operating Characteristic (ROC) curve for the model, a graph that represents how successful the model is at correctly classifying patients as dropouts or non-dropouts. The Area Under the Curve (AUC) is .68, indicating the model accurately classifies patients as dropouts or not-dropouts 68% of the time.

Figure 2 is a graph of the Mean Decrease Gini Value of each independent variable from the Random Forest analysis. For a Random Forest Model, the Gini Index is used to measure the purity of a node, the extent to which the node contains only one class. By adding up the total amount the Gini Index is decreased across all the splits for a predictor, and then averaging it over the number of trees we can determine the importance of a variable. The higher the Mean Decrease Gini, the more significant a variable is, and the graph shows variables in decreasing order of importance.

The "Length of Stay", "State in which a patient lives", "Census Region in which a patient lives", "Employment Status at the time of dropping out", and the patient's "Age", were the five most important variables in predicting a patient prematurely dropping out of treatment. A patient's "Age of First Use for opioids", "Number of arrests in the 30 days prior to discharge", and their "Frequency of attendance at substance use self-help groups in the 30 days prior to discharge" were also relatively important factors that contribute to dropout.

Figure 3 shows the ROC curve for the Random Forest analysis. The AUC is .89; the model accurately classifies patients as dropouts or not-dropouts 89% of the time. The Random Forest model does a much better job of predicting patient dropout than the MLRM.

Table 2. Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard error	Wald Chi-square	Pr > ChiSq
Intercept	1	0.3856	0.0541	50.784	<.0001
STFIPS	1	-0.0140	0.000400	1222.6	<.0001
EDUC	1	-0.0325	0.00211	237.13	<.0001
MARSTAT	1	-0.0309	0.00156	393.54	<.0001
SERVICES	1	-0.0687	0.0152	20.495	<.0001
LOS	1	0.000969	0.000331	8.5628	0.0034
PSOURCE	1	-0.0455	0.00143	1007.8	<.0001
NOPRIOR	1	0.00450	0.00237	3.6018	0.0577
ARRESTS	1	0.2100	0.00286	5395.5	<.0001
EMPLOY	1	-0.0371	0.00225	272.14	<.0001
PSYPROB	1	0.00175	0.00162	1.1624	0.2810
PREG	1	0.0247	0.00174	201.45	<.0001
GENDER	1	-0.2153	0.0176	150.16	<.0001
VET	1	-0.0206	0.00137	225.30	<.0001
LIVARAG	1	0.0836	0.00205	1665.0	<.0001
SERVICES_D	1	0.1669	0.0153	119.15	<.0001
EMPLOY_D	1	-0.00216	0.00133	2.6477	0.1037
ARRESTS_D	1	-0.1990	0.00187	11352	<.0001
AGE	1	-0.0150	0.00192	61.014	<.0001
RACE	1	0.0227	0.00228	99.361	<.0001
PRIMINC	1	0.0491	0.000920	2844.5	<.0001
SUB1	1	-0.1274	0.00487	685.34	<.0001
FREQ1	1	-0.00649	0.00151	18.408	<.0001
FRSTUSE1	1	0.00268	0.00190	1.9807	0.1593
HLTHINS	1	-0.00124	0.00110	1.2745	0.2589
PRIMPAY	1	-0.00940	0.000894	110.47	<.0001
FREQ_ATND_SELF_HELP	1	0.00809	0.00141	32.741	<.0001
FREQ_ATND_SELF_HELP_D	1	-0.0150	0.00117	164.19	<.0001
DIVISION	1	-0.0923	0.00209	1950.3	<.0001
ALCDRUG	1	-0.0221	0.0109	4.1484	0.0417

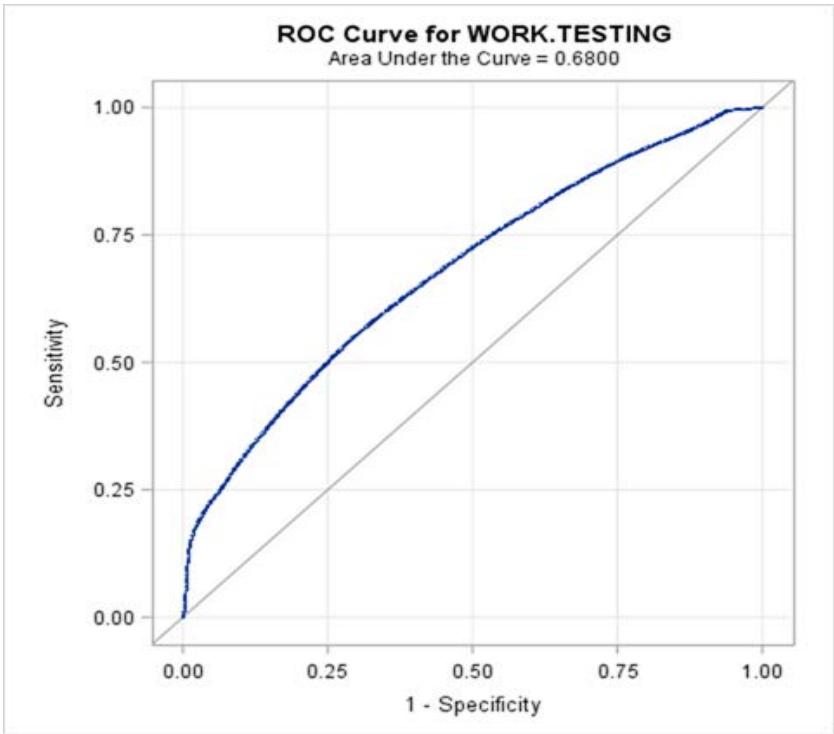

Figure 1. MLRM ROC Curve

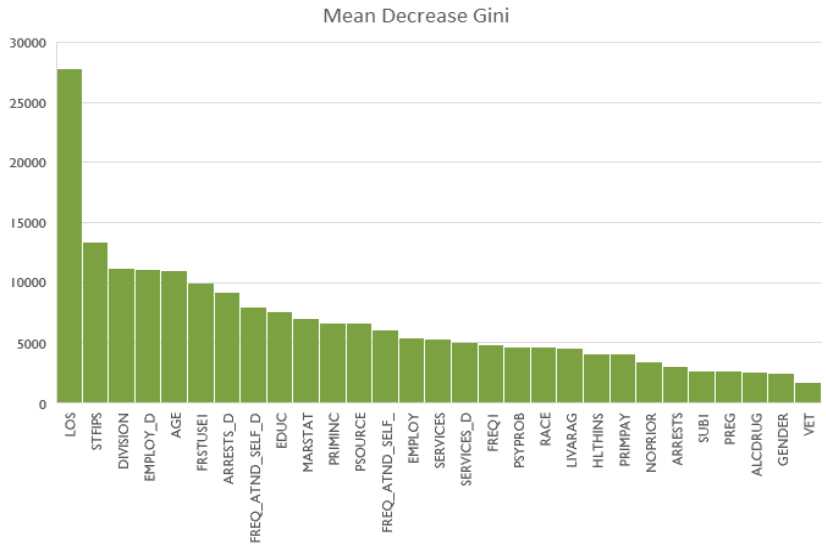

Figure 2. Mean Decrease Gini of Random Forest

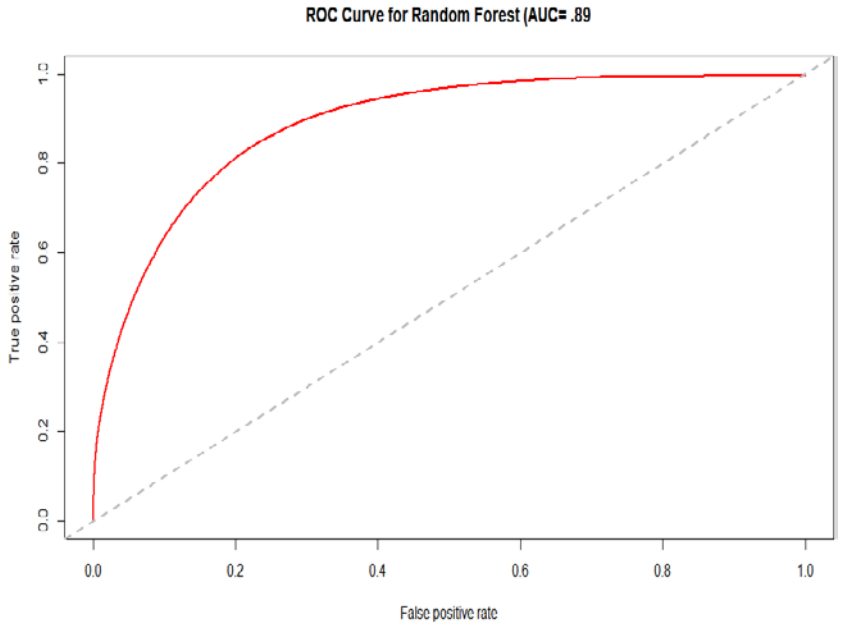

Figure 3. Random Forest ROC Curve

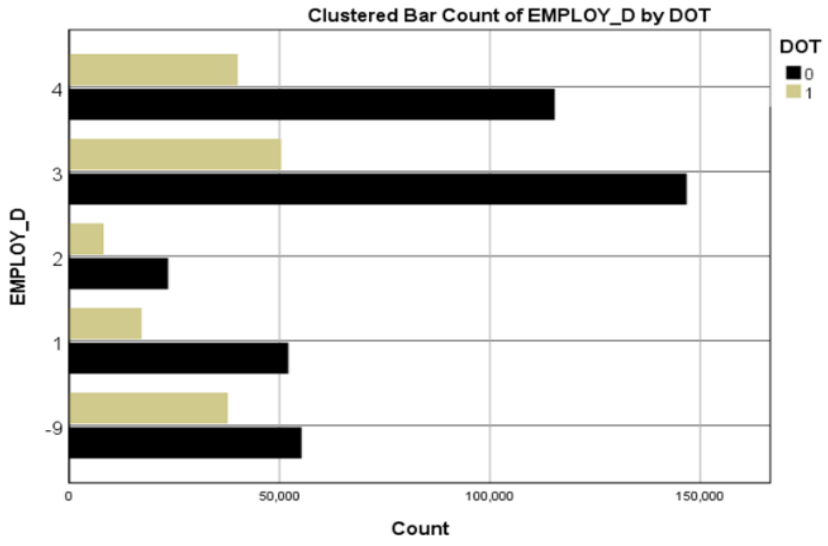

Figure 4. Distribution of Dropout Patients by Employment Status

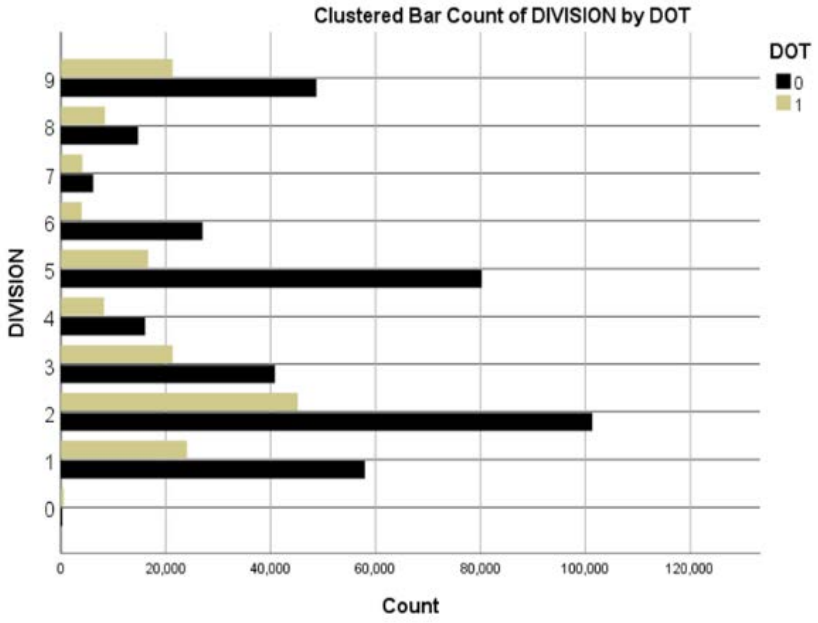

Figure 5. Distribution of Dropout Patients by Census Region

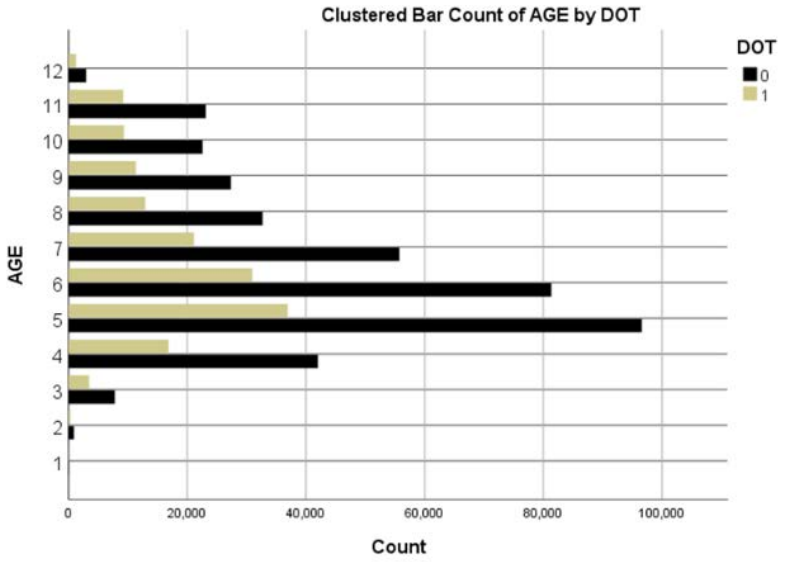

Figure 6. Distribution of Dropout Patients by Age

Graphs depicting the stratification of top variables were created to indicate which response for those variables were associated with higher rates of treatment dropouts. Figure 4 shows the distribution of responses for the independent variable "Employment Status at Time of Discharge," which was determined to be the fourth best predictor of premature dropout. The graph indicates that the most common employment status of patients who did not complete their treatment was "Unemployed."

Figure 5 shows the distribution of responses for the independent variable "Division," which collected responses on the census division region a patient lived in. The Middle Atlantic (New Jersey, New York, Pennsylvania) was found to be the region most likely to have patients drop out of opioid abuse treatment.

Figure 6 shows the distribution of responses for the independent variable "Age." The graph indicates that patients aged 25-29 were the most likely to drop out of treatment, followed by ages 30-34.

4. Conclusion

A machine learning approach such as Random Forest Classification is a much better way to classify patients than traditional methods like Multiple Logistic Regression. It provides a more accurate and focused range of variables, allowing for treatment facilities to carry out more targeted solutions to cater towards every individual patient.

The Length of Stay of a patient at a treatment facility was the most important variable in predicting whether or not the patient would drop out of treatment prematurely; patients were most likely to drop out within the first two days of treatment or after the 180th day of treatment. Based on this information, treatment facilities should give extra care and attention to patients upon first arrival and make extra efforts to retain patients as their length of stay increases.

The "state" and "region" of treatment were both the next most important factors in predicting patient dropout.

Since the Middle Atlantic region was identified as being the most prone to patient dropouts, healthcare professionals and administrators in that region can attempt to address the issue by comparing facility structures and policies to other regions with lower rates of patient dropouts.

A patient's "age" and "employment status" were the next most important variables in determining whether a patient drops out of treatment. Treatment facilities should develop specialized solutions that would help decrease the dropout rates for patients that are categorized as ages 25-29 or unemployed.

Although the MLRM was not as accurate as the Random Forest the Analysis of Maximum Likelihood Estimates gives us the ability to generate an equation using the independent variables that can predict the likelihood of a patient dropping out of opioid abuse treatment. Administrative officials must determine which model best suits their needs and ability to implement policies to help curb the opioid epidemic. The outcomes of this analysis can give healthcare providers insight on how to identify patients more at risk of dropping out to better administer more targeted treatment plans to avoid that risk. At the same time, this analysis can be used by administrators to ensure that opioid abuse treatment is being provided equitably and that discrimination based on any socio-economic factor is minimized.

The implications of this research are twofold: first, machine learning methods have the potential to accurately identify factors responsible for patients dropping out of treatment for not only opioid use, but any substance. Second, the work in this study can be expanded upon by analyzing TEDS-D data spanning several years to determine whether changes in policies or practices at certain facilities affected the significance of any of the independent variables.

References

[1] "Key Substance Use and Mental Health Indicators in the United States: Results from the 2018 National Survey on Drug Use and Health." Substance Abuse and Mental Health Services Administration, SAMHSA, 2018, www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHNationalFindingsReport2018/NSDUHNationalFindingsReport2018.pdf.

[2] Patrick, Stephen W., et al. "Implementation of Prescription Drug Monitoring Programs Associated With Reductions In Opioid-Related Death Rates." Health Affairs, vol. 35, no. 7, July 2016, pp. 1324-1332.

[3] "SAMHSA Strategic Plan – FY2019-FY2023." Substance Abuse and Mental Health Services Administration, SAMHSA, 2018, www.samhsa.gov/sites/default/files/samhsa_strategic_plan_fy19-fy23_final-508.pdf.

[4] Wright ER, Kooreman HE, Greene MS, Chambers RA, Banerjee A, Wilson J. The iatrogenic epidemic of prescription drug abuse: county-level determinants of opioid availability and abuse. Drug Alcohol Depend. 2014; 138: 209-215.

[5] Alex Hollingsworth & Christopher J. Ruhm & Kosali Simon, 2017. "Macroeconomic conditions and opioid abuse," Journal of Health Economics.

[6] Brown E, Wehby GL. Economic Conditions and Drug and Opioid Overdose Deaths. Med Care Res Rev. 2019; 76(4): 462-477.

[7] Acion, Laura & Kelmansky, Diana & Laan, Mark & Sahker, Ethan & Jones, Deshauna & Arndt, Stephan. (2017). Use of a machine learning framework to predict substance use disorder treatment success. PLOS ONE. 12. e0175383.

[8] Stahler, Gerald J., and Jeremy Mennis. "Treatment Outcome Disparities for Opioid Users: Are There Racial and Ethnic Differences in Treatment Completion across Large US Metropolitan Areas?" Drug and Alcohol Dependence, vol. 190, Sept. 2018, pp. 170-178.

[9] Godinet, Meripa & McGlinn, Lindsey & Nelson, Dawna & Vakalahi, Halaevalu. (2019). Factors Contributing to Substance Misuse Treatment Completion among Native Hawaiians, Other Pacific Islanders, and Asian Americans. Substance Use & Misuse. 55. 1-14.

[10] Sharma D. Improving the art, craft and science of economic credit risk scorecards using random forests: Why credit scorers and economists should use random forests. Academy of Banking Studies Journal, 11(1): 93-116, 2012.