Final Project Proposal

Rethinking Substance Abuse Relapses

Christopher Carbonaro

Department of Psychology, University of Notre Dame

MDSC 40122-01: Machine Learning for Social and Behavioral Research

7 November 2019

Introduction

In the United States, more people use mind altering substances than those who abstain. The Substance Abuse and Mental Health Services Administration (SAMHSA, 2019) estimates approximately 60.2 percent of Americans have consumed either alcohol, tobacco, or an illicit drug within the past month (p. 6). Of these 164.8 million past-month users, 31.9 million used an illicit drug within that timeframe and 20 percent are estimated to have used an illicit drug within the past year (SAMHSA, 2019, p.1). Unfortunately, the SAMHSA (2019) also estimates that approximately "21.2 million people aged 12 or older needed substance use treatment" in 2018 (p. 3). However, while this translates to roughly 7.8 percent of the population being in need of treatment, only 1.4 percent "of people aged 12 or older received any substance use treatment within the past year;" of those 3.7 million individuals who received treatment, only 2.4 million received treatment at a special facility (SAMHSA, 2019, p. 3). This strongly suggests there are more individuals in need of treatment than individuals receiving it.

Yet the SAMHSA (2019) does not spend substantial time profiling the individuals who receive treatment, nor does it discuss what factors may drive a user to seek aid. Surprisingly, this topic has received relatively little attention from the academic community. Several studies have been conducted to assess the efficacy of various treatment options but the question of what initially engenders treatment has been widely ignored; worse, the handful of articles which do address this question have reached conflicting conclusions (Boyle, Polinsky, & Hser, 2000; Taylor, Caudy, Blasko, & Taxman, 2017; Battjes, Gordon, O'Grady, Kinlock, & Carswell, 2003). To fill this lacuna, our paper aims to build on the few pre-existing articles which address this topic and provide clarification regarding what characteristics prompt substance abuse treatment. While some neuroscientists have studied drug abuse treatment by focusing on the brain, such as by identifying neurological pathways associated with addiction (Venniro et al., 2017), this paper adopts a different tack. Rather than identifying physiological characteristics which cause substance abuse, this paper focuses on sociological characteristics which prompt addiction. These characteristics are assessed by examining individual users; for example, how does a substance abuser's level of income impact the likelihood of their reception of treatment? This is made possible through use of the SAMHSA's 2018 National Survey on Drug Use and Health (NSDUH, 2019), a dataset containing the records of 55,160 participants and over 3,000 predictor variables.

This proposal proceeds as follows: first, the current literature on what prompts substance abuse treatment will be discussed. Following this review, we will detail the algorithms we propose using and the metrics which will allow us to evaluate the results. Admittedly, addiction is a tortuously complex phenomenon with a myriad of potential causes. To tackle this problem, this paper employs several statistical learning algorithms, commonly referred to as "machine learning" algorithms, to help resolve the research complications introduced by grappling with numerous predictors (in this case, numbering over several thousand). These algorithms help nullify the researcher's preconceptions, permitting the data to speak for itself. They will also help conduct variable selection, as will be detailed shortly. Namely, we will use a form of regularized linear regression referred to as the "elastic net" and an additive decision tree algorithm colloquially referred to as "boosting."

The Research Gap

There are a handful of pre-existing papers which address the sociological characteristics which may motivate substance abusers to seek help. Our article was predominantly motivated by the work started in Boyle et al.'s article, *Resistance to Drug Abuse Treatment* (2000). In this article, the researchers studied 283 drug users who were offered a treatment referral assessment by using descriptive statistics and multiple regression analysis (p. 555, 562). The article aims to "assess the correlates of accepting (vs. declining) an offer of a referral to drug treatment" by assessing variables in three separate domains: demographic characteristics, socioenvironmental factors, and drug-related characteristics (p. 560). The researchers ultimately found "there were no statistically significant differences between the acceptors and the decliners in terms of gender, ethnicity, age, or education" (p. 562). However, individuals who had used illegal drugs in the past month or failed a drug test were found to be "significantly more likely to have used illegal drugs (other than marijuana)" (p. 562). In addition, Boyle et al. found that "users of 'harder' drugs such as heroin and cocaine were more likely to accept assessment than users of a 'softer' drug such as marijuana" (p. 566).

Yet several years later, Battjes et al. (2003) conducted a study comparable to Boyle et al's (2000) and produced conflicting results. Battjes et al. (2003) investigated 196 youths "admitted to an adolescent outpatient substance abuse treatment program" and analyzed the results using multiple regression (p. 221). In their results, Battjes et al. (2003) remark that an individual's "severity of substance use" was not indicative of their willingness to undergo treatment (p. 221). Battjes et al. (2003) attribute this to the increased likelihood of experiencing negative consequences from "harder" drug usage (p. 228). However, while this finding differs from those of Boyle et al.'s (2000) article, they share Boyle et al.'s finding of no significant differences between the motivation of different ages, genders, or race/ethnicities (p. 228).

Finally, Luongo et al. (2016) conducted a recent noteworthy study among Canadians who use illicit drugs and their propensity towards addiction treatment. This article examines the interactions between social marginalization, low income individuals, alternative income generation, and addiction treatment readiness (p. 160). Their hypothesis is that "socioeconomically marginalized people who use illicit drugs often engage in alternative income generating activities to meet their basic needs," and given the health/social risks common to these illegal activities, these individual may "consider addiction treatment to reduce their drug use or drug-related expenses" (p. 159). Not only do Luongo et al. (2016) find support for their hypothesis, they find that illegal sex work serves as "an independent predictor of self-reported need for addiction treatment" (p. 162). Thus, Luongo et al.'s article suggests that there may be interactions between gender, lower levels of income, and illicit activities which can predict an individual's willingness for treatment.

Methodology

This paper aims to leverage the size of the NSDUH (2018) dataset to evaluate what sociological characteristics effectively predict an individual's likelihood of being treated. We will also attempt to use the data to discern whether there is a difference between what variables predict an individual who undergoes treatment and individuals who self-identify as needing treatment. The large number of cases in the dataset will help ameliorate the possibility of producing chance findings which may have been incurred by smaller surveys e.g. Boyle et al.'s survey (2000), which contained only 58 subjects who accepted treatment; as Yarkoni and Westfall (2017) note, "more data beats better algorithms" (p.1108). Additionally, the plethora of predictor variables in the NSDUH (2018) dataset allows for the possibility of revealing interactions which might have remained undetected in smaller surveys. Finally, our use of statistical learning algorithms will allow for the potential detection of non-linear effects in the data. It is worth noting that we are not the first researchers who have employed machine learning tools to study substance abuse treatment (see Cohen, Ilumoka, & Salehi, 2015); however, to the best of our knowledge, we are the first to have used it to address this research question.

Handling Overfitting with Cross-Validation

Before discussing the algorithms, we must discuss our methods of selecting the most appropriate algorithm and assessing the model's performance. This is imperative if we wish to accurately describe the generalizability of our models; otherwise, we will "mistakenly fit sample-specific noise as if it were signal," a phenomenon commonly referred to as "overfitting" (Yarkoni & Westfall, 2017, p. 1102). Unfortunately, as Yarkoni and Westfall (2017) succinctly remark: "To fit is to overfit … because the relationship between variables in any sample is always influenced in part by sampling or measurement error … a fitted model will almost invariably produce overly optimistic results" (p. 1102). Thus, we need some mechanism of accurately assessing how well our model generalizes to predict data which was not used to train the model. The ideal method of accomplishing this would be to collect a second set of data and test the trained model on the new sample which was never used in the training process. However, this is frequently impractical or impossible.

But there is a solution. The researcher can "randomly split the original dataset into two sets—a *training* dataset and a *test* dataset. The training half is used to fit the model, and the test half is subsequently used to quantify the test error of the trained model" (Yarkoni & Westfall, 2017, p. 1111). By doing this, the researcher can leave out data from the training process, providing themselves with a collection of data which was not used to produce the model. As Yarkoni and Westfall (2017) remark, this solution carries with it its own problem: reduced statistical power (p. 1111). This can lead to underfitting, or the failure to pick up on the signal in the data due to a lack of information. Thankfully, this can be solved by iterating the process multiple times. As Yarkoni and Westfall (2017) note, this approach is more generally "termed *K-fold cross-validation*, where K, the number of 'folds,' can be any number between 2 and the number of observations in the full dataset" (p. 1111).

Another solution which solves the same problem is commonly known as "bootstrapping" (James, Witten, Hastie, & Tibshirani, 2013, p. 187). As James et al. (2013) describe, "rather than repeatedly obtaining independent data sets from the population, we instead obtain distinct data sets by repeatedly sampling observations *from the original data set*" (p. 189). The sampling is done with replacement, meaning "the same observation can occur more than once in the bootstrap data set" (James et al., 2013, p. 189). This contrasts with k-folds CV; with k-folds, each observation is included only once. After including a pre-specified number of percentage of observations in the bootstrap sample, the remaining observations are used as the test set.

One final remark on cross-validation is the need for separating the model selection process from model assessment. To avoid biasing the model selection process, the researcher must use cross validation to check the performance of various algorithms against each other. However, to avoid incurring bias in the model selection process, the researcher must use nested cross-validation. This is done by beginning with a bootstrap or k-folds training set. Then, both models are assessed by performing k-folds cv or bootstrapping cv on the training set. Once the more apt algorithm is selected by comparing the results of the nested cross validation, the model can be trained on the original training set and assessed on the test set. This is done for each sub-sample produced by k-folds or bootstrapping. For our research purposes, we will be using bootstrapping for both model selection and model assessment. This is in response to the comparatively small number of individuals in the dataset who have received substance abuse treatment.

Algorithm 1: Regularized Regression and the Elastic Net

The first algorithm we will be testing is a variant of the tool used by most of the other researchers who have previously studied the sociological predictors of substance abuse treatment: linear regression. The best-known variant of a linear regression can be written as:

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta_j}$$

for *p* predictors and coefficients (Hastie, Tibshirani, & Friedman, 2008, p. 12). However, this version of the model always rewards more complex models. More variables will improve explanatory power. Unfortunately, this is not always desirable. There are two well known variants of ordinary least-squares regression (OLS) which can help us mitigate this problem: the ridge regression and the lasso. Both introduce a penalty term which discourages overly complex models. This penalty term is multiplied by a shrinkage parameter, λ, which allows the researcher to experiment and find the optimal level of penalty.

The ridge regression takes the form (Hastie et al., 2008, p. 63):

$$\hat{\beta}^{\text{ridge}} = \text{argmin} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\, \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

while the lasso regression takes the form (Hastie et al., 2008, p. 68):

$$\hat{\beta}^{\text{lasso}} = \text{argmin} \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\, \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

The key difference between the two forms is the way the coefficients are incorporated into the penalty term. The ridge regression squares the coefficients before summing them. This produces shrunken coefficients but does not perform variable selection (Hastie et al., 2008, p. 64). However, it can calculate a regression even if several of the variables are highly colinear. In comparison, the lasso takes the absolute values of the coefficients before summing them into the penalty term. The consequence of this change is the constrained region produced by the penalty will be defined as a parallelepiped in j dimensional space rather than a j dimensional ovaloid, increasing the likelihood of the least squares error function intersecting the constraint function when one of the variables equals zero (Hastie et al., 2008, p. 71). As Hastie et al. (2008) remark, “thus the lasso does a kind of continuous subset selection” by reducing some variables to zero (p. 69). Unfortunately, the lasso cannot easily handle colinear variables (Hastie et al., 2008, p. 72).

The failings of both the ridge and the lasso produced the variant of OLS which we will use in this paper: the elastic net. This version of OLS incorporates a penalty term which is partially defined by the ridge and partially defined by the lasso. The equation takes the following form (Hastie et al., 2008, p. 72):

$$\hat{\beta}^{\text{lasso}} = \text{argmin}\,\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\,\beta_j)^2 + \lambda\sum_{j=1}^{p}(\alpha\beta_j^2 + (1-\alpha)|\beta_j|)$$

This allows for both variable selection and handling of collinearity. Hastie et al. (2008) refer to this as “a different compromise between ridge and lasso” (p. 73). Thus, this model will allow us to automatically select significant variables in the dataset while handling handle colinear variables. This model will allow us to search for linear relationships within the data.

Algorithm 2: Boosted Regression Trees

However, it is possible that not all the relationships in the data are linear. Consider the impact of an individual’s income on the likelihood of their propensity towards substance abuse treatment. Individuals with an abundance of disposable income may be more able to take time off from work to visit a special facility than a blue-collar worker; similarly, an unemployed individual receiving insurance through Medicare may not need to worry about the costs of treatment. However, the median American earner may have more difficulty affording substance abuse treatment. This would signify a non-linear trend in the data. Thankfully, boosting serves as an excellent tool for detecting such trends.

As Hastie et al. (2008) describe, “boosting was a procedure that combines the outputs of many ‘weak’ classifiers to produce a powerful ‘committee’” (p. 338). Boosting can be thought of as a means of using gradient descent to minimize a loss function (p. 342). The general process involves fitting a model, adjusting the residuals, and then fitting a new model to the adjusted residuals. This process is iterated until the loss function is minimized. In this paper, given that we want to classify individuals as recipients of substance abuse care, we use the AdaBoost algorithm to search the predictor space and produce a series of additive classifiers. The formal process is outlined on in Hastie et al. (2008, p. 339): first, produce observation weights for each observation. The initial weights will simply be $w_i = \frac{1}{N}$ for all *i* where *N* is the number of observations. Then, fit a classifier to the training data. Next, compute the error by summing the weights of the incorrectly classified observations and dividing by the total sum of the weights. This is described by the following equation:

$$\mathrm{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}$$

where $G_m(x_i)$ is the classifier. Next, scale the weights “by a factor e^{α_m}, increasing their relative influence for inducing the next classifier $G_{m+1}(x)$ in the sequence” (Hastie et al., 2008, 339). This value is multiplied by the original weight if classified incorrectly; otherwise, the weight remains unchanged. Finally, after *m* models, a new variable is predicted by feeding the predictors to each of the classifiers and scaling the results by α_m for each classifier. These values are added and the sign of the final value determines whether the prediction should be classified as 1 or -1, i.e. by majority vote.

Resources

Battjes, R. J., Gordon, M. S., Ogrady, K. E., Kinlock, T. W., & Carswell, M. A. (2003). Factors that Predict Adolescent Motivation for Substance Abuse Treatment. *Journal of Substance Abuse Treatment, 24*(3), 221–232. doi: 10.1016/s0740-5472(03)00022-9

Boyle, K., Polinsky, M. L., & Hser, Y.-I. (2000). Resistance to Drug Abuse Treatment: A Comparison of Drug Users Who Accept or Decline Treatment Referral Assessment. *Journal of Drug Issues*, *30*(3), 555–574. doi: 10.1177/002204260003000304

Cohen, J., Ilumoka, A., & Salehi, I. (2015). Neural Network-Based Drug Abuse Treatment Optimization. *Procedia Computer Science, 61*, 454–459. doi: 10.1016/j.procs.2015.09.186

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition* (2nd ed.). New York, NY: Springer New York.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An Introduction to Statistical Learning: With Applications in R*. New York: Springer.

The Substance Abuse and Mental Health Services Administration. (2019). The Key Substance Use and Mental Health Indicators in the United States: Results from the 2018 National Survey on Drug Use and Health. Retrieved from https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHNationalFindingsReport2018/NSDUHNationalFindingsReport2018.pdf

Luongo, N. M., Dong, H., Kerr, T. H., Milloy, M. J. S., Hayashi, K., & Richardson, L. A. (2017). Income Generation and Attitudes Towards Addiction Treatment Among People who use Illicit Drugs in a Canadian Setting. *Addictive Behaviors, 64*, 159–164. doi: 10.1016/j.addbeh.2016.08.041

Venniro, M., Caprioli, D., Zhang, M., Whitaker, L. R., Zhang, S., Warren, B. L., … Shaham, Y. (2017). The Anterior Insular Cortex→Central Amygdala Glutamatergic Pathway Is Critical to Relapse after Contingency Management. *Neuron, 96*(2). doi: 10.1016/j.neuron.2017.09.024

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. doi: 10.1177/1745691617693393