2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)

Predicting Opioid Use Disorder (OUD) Using A Random Forest

Adway S. Wadekar
Saint John's High Shrewsbury, MA 01545
adwayw4@gmail.com

Abstract – **Opioid Use Disorder (OUD), defined as physical or psychological reliance on opioids, is quickly becoming a public health epidemic. This research demonstrates how supervised machine learning can be used to predict adults at risk for OUD by considering interactions between various demographic, socioeconomic, physical, and psychological features in an integrated manner. A labeled data set was built from the responses to the 2016 edition of the National Survey on Drug Use and Health (NSDUH). This labeled data set was used to train and test a random forest classifier while accounting for class imbalance. The classifier can predict adults at risk for OUD accurately (sensitivity = 0.81, specificity = 0.76, AUC = 0.86), although the prevalence of OUD is only about 1%. Early initiation of marijuana (prior to 18 years of age) emerges as the dominant predictor for developing OUD in adult life. This is surprising because it ranks higher than both mental illness and disability; two conditions that are often comorbid with substance use disorders. Thus, curbing early initiation of marijuana may be the best prevention strategy. This highlights the crucial role that educators, counselors, and parents can play in alleviating the United States' opioid overdose crisis.**

Index Terms – Opioid Use Disorder, Machine Learning, Random Forest, Early Initiation of Marijuana

INTRODUCTION & MOTIVATION

Opioid Use Disorder is defined as a physical or psychological reliance on opioids, a substance found in many prescription drugs, and illegal drugs such as heroin [1]. The United States is in the midst of an opioid overdose epidemic, and it has now become a public health crisis [2]. Currently, an average of 130 US citizens dies of an opioid overdose epidemic daily [3].

Several factors may be attributed to increasing the likelihood of addiction, including mental illnesses, issues with personal and social relationships, proximity to other drug users, and past traumatic events [5]. Thus, contemporary studies have associated some demographic, socioeconomic, physical and psychological features with substance use disorders. Most of these studies, however, attempt to explain substance use disorders by considering each feature in isolation, or one at a time. For example, the Department of Health and Human Services (DHHS) estimates that living with a disability increases the risk for substance abuse [6]. Similarly, the National Bureau of Economic Research reports a connection between mental illness and substance abuse [7]. It is necessary to move beyond explanation to prediction, where intricate associations between these features can be considered to determine who is likely to develop OUD. This can help public health officials in planning and offering effective intervention and support strategies. This research explores the use of supervised machine learning to build a classifier to predict individuals that are at risk for OUD based on an integrated consideration of demographic, socioeconomic, physical, and psychological features.

ANALYSIS APPROACH

Supervised machine learning requires a labeled data set where each individual is identified as having opioid use disorder or not. Because a labeled data set was not readily available, it was built using the responses from the 2016 edition of the National Survey on Drug Use and Health (NSDUH) conducted by the Substance Abuse and Mental Health Services Administration [4]. The 2016 NSDUH survey includes questions about opioid use for the first time.

The 11 independent features and their levels included in the labeled data set were: (i) gender (two levels), (ii) age (four levels), (iii) race (seven levels), (iv) income (four levels), (v) employment (four levels), (vi) education (four levels), (vii) one or more of cognitive, visual, dressing, walking, hearing and running errands disability (two levels), (viii) any mental illness including psychological distress and suicidal thoughts (two levels), (ix) first use of alcohol before 18 years (two levels), (x) first use of marijuana before 18 years (two levels), and (xi) overall health (four levels). Each observation was tagged as "OUD" or "No-OUD" depending on whether or not there was opioid dependence or abuse.

The labeled data set consisted of 42,324 observations, of which only 459 or approximately 1% are labeled as OUD. Thus, the OUD class is very rare. This imbalance between the OUD and No-OUD groups makes the classification problem very challenging. To address class imbalance, the labeled data set was split 80-20 using stratified sampling. The partition containing 80% of the data was downsampled [11] to create a balanced training set. The partition with 20% data was used as the test set. This downsampling procedure creates a balanced training set, but the test set remains imbalanced.

978-1-7281-2607-4/19/$31.00 ©2019 IEEE
DOI 10.1109/COMPSAC.2019.00161

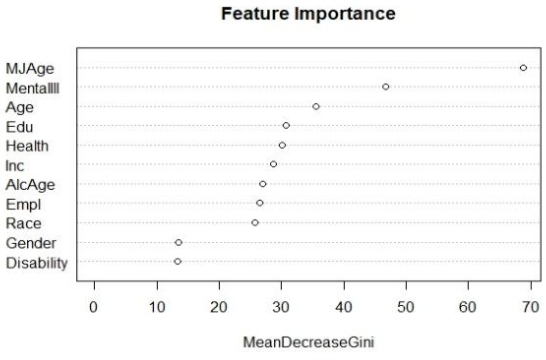
A random forest classifier [11] was trained on the balanced training set, and evaluated on the test set. This train-test process was repeated 50 times. For each run, the sensitivity, specificity, and the AUC [11] of the classifier were computed from the confusion matrix. The average metrics were computed across all runs. The analysis was done in R, using packages: randomForest [13], pROC [15], and caret [12].

RESULTS & DISCUSSION

The random forest classifier can predict adults likely to develop OUD accurately (avg. sensitivity = 0.81, avg. specificity = 0.76, avg. AUC = 0.86). For clinical psychology applications such as this one, AUC values of greater than 0.7 indicate good performance [14]. Thus, the forest classifier registers good performance although the prevalence of the OUD class is only about 1% in the imbalanced test set.

The average sensitivity of the classifier is higher than average specificity. That means, that it correctly identifies many individuals with OUD, but at the cost of flagging some individuals incorrectly. Had it been the other way, that is, if the specificity was higher than sensitivity, the classifier would miss a higher percentage of individuals with OUD, which is not a very desirable scenario.

The classifier also ranks the 11 independent features in order of their importance for predicting OUD in its importance plot shown in Figure 1. Early initiation or "First use of marijuana before 18 years" emerges as the strongest predictor of OUD. This is unobvious and surprising because it ranks higher than mental illness and disability, two conditions often comorbid with OUD. Therefore, curbing early initiation of marijuana may be the best preventive strategy. This highlights the crucial role that parents, educators, and counselors can play in alleviating the opioid crisis.



RELATED RESEARCH

Using machine learning to predict various dimensions of substance use disorders is considered novel and promising [10]. Acion et al. compare several machine learning techniques to predict success in substance use treatment [8]. Young et. al. [9] identify behavioral biomarkers for opiate and stimulant dependence. While both these works consider certain aspects, to the best of my knowledge, this is the first study to show the potential of machine learning to predict OUD by considering the interplay between various features.

CONCLUSIONS & FUTURE RESEARCH

This research demonstrates how machine learning can be used to predict adults at risk for OUD. Machine learning is also promising in identifying the interactions among features that increase this risk. It also shows that public domain data sets such as NSDUH hold vast information that can be mined to improve our understanding of substance use disorders and mental illnesses. These data sets comprise a large number of observations, a variety of characteristics are collected about each observation, the data is cleaned and recoded, and the surveys are conducted in a statistically sound manner.

Future research involves trying different machine learning techniques such as support vector machines and gradient boosting on the NSDUH data. Investigating and mining other public domain data sources, which may have other informative features, is also a topic of future research.

REFERENCES

1. Center for Disease Control, Module 5: Assessing and Addressing Opioid Use Disorder (OUD). (n.d.). Retrieved Feb, 2019, https://www.cdc.gov/drugoverdose/training/oud/accessible/index.html.
2. Center for Disease Control, CDC Injury Center, Understanding the Epidemic. Retrieved Feb. 2019. https://www.cdc.gov/drugoverdose/epidemic/index.html,
3. Center for Disease Control, CDC Injury Center, Drug Overdose Deaths. https://www.cdc.gov/drugoverdose/data/statedeaths.html, Retrieved Feb. 2019.
4. [NSDUH] Substance Abuse and Mental Health Services Administration. (2016). 2016 National Survey for Drug Use and Health [Data file and code book]. Available from Substance Abuse and Mental Health Data Archive: https://www.datafiles.samhsa.gov/, Retrieved Feb. 2019.
5. National Institute on Drug Abuse. The Science of Drug Use: Discussion Points. February 2017. https://www.drugabuse.gov/related-topics/criminal-justice/science-drug-use-discussion-points, Retrieved Feb. 2019.
6. National Rehabilitation Information Center, Substance Abuse and Individuals with Disabilities, vol. 6, no. 11, January 2011, https://naric.com/?q=en/publications/volume-6-number-1-january-2011-substance-abuse-individuals-disabilities, Retrieved Feb. 2019.
7. National Bureau of Economic Research, Mental Illness and Substance Abuse, https://www.nber.org/digest/apr02/w8699.html, Retrieved Feb. 2019.
8. L. Acion, D. Kelmansky, M. van der Laan, E. Sahker, et. al., Use of a machine learning framework to predict substance use disorder treatment success, April 2017, https://doi.org/10.1371/journal.pone.0175383.
9. W. Y. Ahn, and J. Vassileva, Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence. Drug and Alcohol Dependence. 161. 10.1016/j.drugalcdep.2016.02.008, 2016.
10. J. Bresnick, For Opioids and Substance Abuse, Big Data Analytics Is Just the Beginning, Health IT Analytics, https://healthitanalytics.com/features/for-opioids-and-substance-abuse-big-data-analytics-is-just-the-beginning, Retrieved Nov. 2018.
11. M. Kubat, An introduction to machine learning (2nd ed.). Cham, Switzerland: Springer, 2017.
12. M. Kuhn, Classification and Regression Training. November 2018, Retrieved from The Comprehensive R Archive Network database.
13. A. Liaw, Breiman and Cutler's Random Forests for Classification and Regression. March 2015, Retrieved from The Comprehensive R Archive Network database.
14. M. Rice & G. Harris. Comparing Effect Sizes in Follow-up Studies: ROC Area, Cohen's d, and r. Law and Human Behavior, 29(5), 615-620, 2005. Retrieved from http://www.jstor.org/stable/4499443.
15. X. Robin, N. Turck, A. Hainard, . . . S. Siegert, September 2018, Package 'pROC'. Retrieved from The Comprehensive R Archive Network database