CSCI547 Machine Learning
Substance Abuse Risk based on Personality Evaluation

Zachary Falkner
Department of Computer Science
University of Montana

May 12, 2018

1 Abstract

2 Introduction

This paper attempts to expand on work in done in the paper ”The Five Factor Model of personality and evaluation of drug consumption risk.” The original work did extensive searches over a variety of basic machine learning algorithms to a dataset consisting of demographic information, reporting of substance use, and evaluation of personality traits based on standardized surveys. The original study produced quality results . The plan is to extend this work to a Deep Neural Network Regressor leveraging the Tensorflow library in hopes that its sophistication can reveal deeper subtleties than more basic machine learning algorithms could yield.

2.1 Background

The NEO Five Factor personality model is a quantifiable way of assessing personality trains in individuals It is well established and widely accepted amongst the psychological community. It asses the traits of Neuroticism, Extraversion, Openness to experience, Agreeableness, and Conscientiousness. [2] Similarly the Barratt Impulsivity test quantifies how impulsive an individual is and the Sensation seeking test quantifies how likely someone is to seek new or pleasurable sensation. The hypothesis at hand is that in particular combinations, these traits may put an individual at higher risk for substance abuse. If a link can be identified at risk individuals could be targeted for preventative services.

2.2 Dataset

The dataset was collected in 2012 by Elaine Featherman via an anonymous on-line survey. The survey included demographic information of Age, Gender, Education, Country of Residence, and Ethnicity. In addition it provided an NEO Five Factor Revised questionnaire in addition to a Barratt Impulsivity Questionnaire and a Impulsive Sensation Seeking Questionnaire. This categorical data was quantized and normalized. Finally, the questionnaire asked about their use of 19 substances ranging from Caffine, Chocolate, and Nicotine, to Heroine, Amphetamines and everything in between. The questionnaire also include a fake drug ”Semer” as a control to remove bad data point resulting from false/over reporting. [1]

3 Methods

Preprocessing for this dataset was rather limited. Because the original research produced the best results with decision tree’s were used as a baseline. It was then attempted to improve upon these results with an SVM, linear and polynomial kernels produced similar result. All tests were run with a 2/3 1/3 train test split.

3.1 preprocessing

Much of the preprocessing was done in advanced on this dataset which is both nice and unfortunate. Participants who had reported Semer were pruned from the dataset as their responses are considered unreliable. However it would have been nice to be able to explore correlations of personality traits with false reporting. The data was otherwise quantified from categorical responses to numerical and normalized centering around 0 and a standard deviation of 1. The targets were left as categorical data CL0-CL6 corresponding with Never Used to Used within the last day. These were simply converted to a numeric 1-6 so they are feed able as targets for models. I also experimented with quantizing these by time period with CL0 and CL1 being given a score of 0 (as they are non drug users) and all others being given a score of $\frac{-1}{\log(\frac{1}{dt})}$ where dt was the time period in days. CL6 is 1 day giving a log of 0 so it is constrained to the max float value as to avoid infinite values. This results in similarly strong correlations between features and classes.

3.2 Decision Tree

Decision tree’s were run and it was found that sample splits of 200 and more yielded best results. This was used as teh baseline model as it gave the best results in the cited publication.

3.3 SVM

SVM's with linear and polynomial kernels were run on the model. They tended to yield poor results and the polynomial kernel performed no better than the linear.

3.4 Random Forest

Because results with decision tree's were good, Random Forests were applied. These are randomly grown ensembles of Decision tree's. From testing best results came from restricting tree's to be shallow, with a max depth of 2.

3.5 DNN Regressor

Theoretically a deep neural network should be able to accurately model any function. For this model a network consiting of 3 layers of sizes 11, 22, and 11 respectively were constructed with the middle layer being fully connected. a learning rate of 0.001 and a dropout of 0.2 were set. The training metric is set to minimize loss (Mean Squared Error) and AOC (Area under curve) is reported as the accuracy.

4 Results

Results were interesting. Forst many substances, these modesl at best produced a guess. however for Caffine, Ketamine, Methadone, VSAs, Heroine and crack all models performed quite well. Of the personality traits, N-Scores (Neurotocism), O-Score(Openness to new experience), and SS (Sensation Seeking) where always present as the prime contributors to the models. Neural networks tened to perform as a coin flip with the exception of the control Semer. I attribute this to lack of Random Parameter Search.

5 Discussion

There seems to be some correlation to neurotocism, openeness to expeience, and sensation seeking and abuse potential for more serious drugs such as

Table 1: Results for various models

Substance	SVM Train	SVM Test	RF Train	RF Test	DT Train	DT Test	NN AUC
Alcohol	0.41	0.39	0.41	0.39	0.43	0.37	0.499
Amphetamine	0.52	0.52	0.52	0.52	0.52	0.5	0.5
Amyl	0.68	0.7	0.68	0.7	0.69	0.67	0.6
Benzo	0.54	0.52	0.53	0.52	0.54	0.52	0.5
Caffine	0.75	0.7	0.75	0.7	0.75	0.7	0.499
Cannabis	0.41	0.4	0.39	0.4	0.41	0.38	0.5
Chocolate	0.44	0.44	0.45	0.4	0.49	0.33	0.499
Cocaine	0.56	0.52	0.56	0.52	0.58	0.53	0.5
Crack	0.87	0.85	0.86	0.88	0.87	0.86	0.55
Ecstasy	0.56	0.55	0.54	0.55	0.56	0.55	0.5
Heroin	0.86	0.83	0.86	0.83	0.86	0.83	0.52
Ketamine	0.79	0.78	0.79	0.78	0.79	0.79	0.6
LegalH	0.65	0.62	0.62	0.6	0.64	0.6	0.5
LSD	0.59	0.62	0.56	0.59	0.6	0.62	0.5
Methadone	0.75	0.76	0.75	0.76	0.75	0.76	0.63
Mushrooms	0.55	0.54	0.52	0.52	0.57	0.52	0.5
Nicotine	0.4	0.4	0.4	0.4	0.4	0.39	0.5
Semer	0.99	0.99	0.99	0.99	0.99	0.99	0.96
VSA	0.79	0.74	0.78	0.74	0.79	0.74	0.53

Heroine, Crack, and Ketamine. Certainly more data could shine better light on this. More data from a larger demographic could certainly attribute to better models as the sample size was quite small and the distribution of reported users was quite small. Furthermore, time did not allow for large scale feature/parameter subset searches. Applying these searches, particularly to the neural network classifier could produce better results.

References

[1] E.M. Mirkes V. Egan A.N. Gorban E. Fehrman, A.K. Muhammad. The five factor model of personality and evaluation of drug consumption risk, 2017.

[2] Robert R. McCrae, John E. Kurtz, Shinji Yamagata, and Antonio Terracciano. Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, 15(1):28–50, 2011. PMID: 20435807.