Project Report:
Prediction of Drug Abuse Probability

Robert Henzel, Ömer Sari

May 25, 2019

Abstract

The following project report covers the attempt to model drug abuse probability using a numerical score and consequently predict abuse probability relying only on non-sensitive and observable information. Concluding, a critical analysis of why the approach may have failed is provided.

1 Introduction

With the final project it is aimed to build a predictive model of hard drug abuse probability relying only on non-sensitive information and demographics using a small feature set. As the selected information set is going to be rather limited, high accuracy is not expected from the results, still the benefits of such modeling - should it proof feasible - are evident: The appeal of being able to target individuals with easy-to-ask questions and consequently derive the probability of drug abuse poses a non-intrusive instrument in handling drug abuse prevention and control for institutional actors as well as practitioners. The data used is based on a national drug use and health survey in the US conducted yearly.

In section 2 the structure and contents of the underlying data set are explained as well as applied pre-processing. Section 3 then covers modelling of a drug-abuse probability (score) which is not part of the data set and building of the predictive model. Consecutively section 4 offers examination and evaluation of training results.

2 Data Set

The following section gives detailed overview concerning the used data, it's structure and how it was pre-processed for the project's examinations.

2.1 Data Source

The examined data set is the public use data file of the 2015 National Survey Drug Use and Health (NSDUH) . The main purpose of NSDUH is to measure the correlates of drug is in the United States. The attendees are 12 years and older U.S citizens that are the part of noninstitutionalized population. The survey series contain information regarding the use of tobacco, alcohol and illicit drugs. There are also modules of questions about the mental health issues.

The data file contains 57,146 rows representing the number of observations and 2,666 columns that contain variables. Due to the large number of variables it is not viable to describe each variable since that would make hundreds of pages. In this report, a general descriptions of the main categories of the variables will be included. The complete documentation about the 2,666 variable containing variable names and descriptions can be found at samhsa.gov[1].

The main categories of the variables in the dataset are: Identification, Self-Administered Substance Use Sections, Imputed Substance Use, Other Self-Admind. Sections, Interview Information, Demographics, FI Debriefing Questions, Geographic, Sample Weighting and Estimation Vars. Following, a description of used categories is given:

Self-Administered Substance Use Sections These sections contains the questions regarding the use of both legal and illegal drugs. The legal drugs are tobacco types, alcohol and prescribed drugs. There are subsections for more than twenty illegal drugs and misuse of prescribed medication such as pain relievers, stimulants etc. There are fourteen identical questions for each drug section. These questions aim to assess the respondents substance use habits and their degree of dependency.

Risk Availability Questions in this section aim to find out the respondents' risk/availability. The respondents are asked about their opinion on other people's drug use. An example is: "How much do people risk harming themselves physically and in other ways when they smoke marijuana once a week?". The same question

[1]https://datafiles.samhsa.gov/study-dataset/national-survey-drug-use-and-health-2015-nsduh-2015-ds0001-nid16894

is asked changing the frequency term; e.g, once a month. The same format is used for cigarettes, alcohol, cocaine, heroin and LSD. Then availability questions for these drugs are asked in the following format: "How difficult or easy would it be for you to get some cocaine, if you wanted some?". The ending questions of this section target to get a sense of risk taking behaviour of the individual. The respondent is asked whether they enjoy doing things that are "a little dangerous".

Health This section provides information about the respondents current health status. The recent past of the individual's health condition is also enquired. Adult/ youth mental health, adult/adolescent depression are separate sections in the survey all regarding the health status of the respondent. These sections include detailed questions that depict the psychological health of the individual.

Youth Experiences This section is about the school and other activities and asked only to respondents that are enrolled in an education institution. The first questions are about the opinions and overall performance of the students. Then, respondents are asked about substance use in their immediate surroundings. There are numerous other types of questions in this section, it would not be viable to explain them in detail.

Demographics The demographics section consists of questions about education, employment, household composition, military status, health insurance and income.

2.2 Pre-Processing

Before any kind of predictive analysis can take place, various stages of pre-processing are applied which are described and in the following section.

RFD-Score: The observed quantity, a measure for the level of drug abuse, is first to be constructed from features from the data set regarding specific drug abuse. As this is part of the modelling approach, the score is explained in detail in the model section.

Pre-Selection of Features: Although algorithmic feature selection is applied at a later stage, only a subset of the 2.666 features is viable for the desired analysis, prerequisites include: non-sensitivity of information, observability, measurability, bearing predictive value. This pool of variables is then used for feature selection where it

is going to be reduced to just a small subset. Hence the codebook was screened for variables which may turn out as viable predictors in following categories: Demographics, Alcohol & Tobacco use, Health, Mental Health, Social Environment, Youth Experiences, Adolescent & Adult Depression. 399 variables were hand picked during this process.

Test/Train Split: As means of test and train division, scikit-learn train_test_split() is used. Stratification is applied, such that the distribution of RFD Total Scores (see next section) is the roughly equal for both train and test set. The training data is 80% of the original data set and the test data 20%, although the training data set is heavily reduced in the next step. Cross-validation is not used in this project, as working with the scale of dataset and limited computational power wasn't feasible due to time constraints.

Balancing: Test/Train Splitnbalanced (most of the samples have a RFD Score of zero, as most individuals haven't taken hard drugs). Due to this, it can be expected that predictive models won't fit well to the samples that have taken drugs. Thus the training data set is being balanced, meaning under-sampling of the subset of never-taken-drugs samples (RFD=0) which is applied via the imbalanced-learn python package. The 'NearMiss' routine is used, which provides heuristics for non-random undersampling. The under-sampling reduces the train set by 87%.

Imputation: Some of the labels of the categorized features don't bear usable information in regards of the conducted analysis. Examples for that are NaN values or codes stating "bad data", "don't know", "refused" or "blank" as an answer. All of these codes are replaced by NaN values and following simple imputation as implemented by scikit-learn. Multivariate imputation was tried, this imputation heuristic iteratively models missing values as functions of other features. After imputation the data set is left with only applicable codes for the analysis, but as in the consecutive rounds it deviated from selected imputation strategy (most-frequent) to the categorical variables and produced mean values, it was discarded.

3 Model

The following chapter will cover modelling of the observed quantity as well as a description of the applied prediction models.

3.1 RFD-Score

The objective of this project is to predict the probability of hard drug abuse from a set of features that are not directly related to substance use or any illegal activities. However, there is not a numerical feature that refers to degree of drug use in the dataset. Therefore, a substance use score is needed as a first step. In the survey, there are questions regarding the recency, frequency and duration of drug use. Using these features in the dataset; "recency", "frequency" and "duration" scores for each observation is calculated. Finally a "RFD score" is created from these sub-scores.

3.1.1 Recency

The recency score is determined using the categorical answers to "How long has it been since you last used [substance]" questions. The possible answers are: Used within the last month, used within the last year but not in the last month, used more than 1 year ago. The recency scores corresponding to these categories are given as follows:

$$R = \begin{cases} 1 & \text{if } t < 30 \\ 0.5 & \text{if } 30 \leq t < 365 \\ 0.2 & \text{if } t \geq 365 \end{cases}$$

3.1.2 Frequency

The frequency of substance use is simply the total number of days a respondent used the substance during the last year. The values are mapped to 0.3-1:

$$F = 0.3 + 0.7 * \frac{\#\text{days}}{365}$$

3.1.3 Duration

The duration score is calculated as the difference between the year survey conducted and the year in which the respondent used the substance for the first time (yfu).

Then the values are normalized and mapped to 0.3-1. Even though this is not the best measure for a duration score, it at least provides a timeframe and might be meaningful along with recency and frequency scores.

$$\Delta = 2015 - yfu$$

$$F = 0.3 + 0.7 * \frac{\Delta}{\Delta_{max} - \Delta_{min}}$$

3.1.4 RFD-Score

$$RFD = \begin{cases} R * (\frac{6}{5}F - \frac{1}{5}D) & \text{if R} = 2 \text{ and F} \leq 0.35 \\ R * (\frac{6}{5}F + \frac{1}{5}D) & \text{else} \end{cases}$$

The duration score is subtracted when recency and frequency are at the minimum. Minimum recency and frequency scores imply that the respondent did not use the substance within the last year. Therefore, the RFD score should be smaller as the duration increases since it means a longer time the respondent have not used the substance. However, if the respondent have used the substance recently and/or have a high frequency score; the greater duration score might imply a longer period of substance dependency/abuse. Hence, it should increase the RFD score.

3.2 Feature Selection

The original datasets consists of 2666 features. These features are screened manually to eliminate similar/same features as there are encoded, recoded features all holding the same information. After the pre-selection, the number of features is decreased to 399. The main purpose is to extract a handful of features that will help to predict the drug use probability but at the same time will not be directly about drug use. For this purpose, 400 features should be trimmed even more, to have a feasible subset of features that can be easily observed. To achieve this, the XGBoost feature importance algorithm is used. The resulting list of features with their descriptions can be found in table 1.

3.3 Predictive Model

For prediction purposes, the XGBoost Regressor with an logistic objective function and the was chosen. Logistic functions are a natural choice when it comes to predict probabilities, as it possesses suiting properties as being bounded by [0,1]. The model is specified as follows.

$y_i \in [0, 1]$	RFD-Score
X_{ij_k}	Feature, where $k \in [1, N_j]$
N_j	#Levels of feature j

$$\hat{y} = f_{XGBR}(X)$$

Where $f_{XGBR}x$ is a prediction constructed by the XGBoost Regressor, which is an ensemble tree predictor. It relies on gradient boosting to improve results, a method which in a very simplified manner can be described as iteratively training trees on residuals to reduce the prediction error. Providing a mathematical description of the XGBoost Regressor itself would go beyond the scope of this work. Hence it is regarded as a "black box" model here. It was chosen because it a state-of-the-art predictor tends to provide good out-of-the-box results.

	Feature	Description
1	APPDRGMON	Approached by someone selling ill drugs pst 30 days?
2	CIGYR	Past year cigarette use.
3	K6SCMAX	Worst K6 score in past year.
4	IRPINC3	Total income.
5	K6SCMON	A score indicating the level of psychological stress in past month.
6	IRWRKSTAT	Employment status.
7	DIFGETHER	Important for friends to share religious beliefs?
8	CIGAVOID	Tend to avoid places that don't allow smoking?
9	PSYANYYR	Any psychotherapeutics in the past year?
10	SMIPP_U	Predicted serious mental illness probability.
11	RSKYFQTES	Like to test yourself by doing risky things?
12	ARGUPAR	Number of times argued/had a fight with one parent in past year.
13	RSKCOCWK	How much people risk harming themselves phys. and in other ways when they use cocaine once or twice a week?
14	WRKDRGHLP	Any assistance program offered through work?
15	TRQANYYR	Any tranquilizers past year?
16	IRMARITSTAT	Marital status.
17	INHOSPYR	Stayed overnight as inpatient in hospital in past 12 months?
18	OXYCNANYYR	Oxycontin past year use.
19	WRKOKPREH	Would you work for employer who does drug test pre-hire?
20	SMKLSSYR	Smokeless tobacco past year use.
21	WRKSKIPMO	Number of days skipped work past 30 days.
22	STMANYYR	Stimulants past year use. (not misuse)
23	IREDUHIGHST2	Education level.
24	RSKHERWK	How much do people risk harming themselves when they use heroin once or twice a week?
25	IROTHHLT	Health insurance.
26	UADPEOP	Social context of most recent alcohol use.
27	RSKHERTRY	How much people risk harming themselves when they try heroin once or twice?
28	DIFGETCRK	How difficult to get crack if you wanted to find some?
29	WRKNUMJOB2	How many different employers have you had in the past 12 months?
30	YFLMJMO	How do you feel someone your age using marijuana/hash monthly?
31	HIVAIDSEV	Ever told had HIV or AIDS?
32	HRTCONDEV	Ever told had heart condition?
33	WRKSTATWK2	Work situation in past week.
34	WRKDRGPOL	At your workplace is there a written policy about employee use of alcohol or drugs?
35	RSKYFQDGR	Get a real kick out of doing dangerous things?
36	HIGHBPAGE	How old were you when your high blood pressure was first diagnosed?
37	IRWRKSTAT18	18+ employment status.
38	PNRANYYR	Past year pain reliever use. (not misuse)
39	RKFQPBLT	Wear a seatbelt when ride front pass seat of car?
40	WRKDPSTWK	Did you work at job last week?

Table 1: Features sorted by importance

4 Evaluation

In the following section evaluation methods, results and a critical review are presented.

4.1 Measure of Fit

$$R^2 = 1 - \frac{RSS}{TSS}$$

To assess quality of regression results, R^2 metrics is used as the measure of fit, where RSS is the Residual Sum of Squares and TSS the Total Sum of Squares. R^2, also called coefficient of determination gives the explained fraction of variance in the results, with R^2 meaning a model being fully determined. Negative values are possible, if the model does not bear any descriptive value.

4.2 Training Results

In total, four cases are analysed. For once, the chosen RFD-Score is differently constructed, in the first case it is just constructed from Heroine-consumption features and in the second case an aggregate score is used were the single scores of Heroine, Meth and Crack are added up (and still capped at 1), hence given a probability of drug abuse for any of those drugs. The second dimension varied is feature size, which once is the 20 most-important features and the other time the 40-most important features.

Case	R^2
20 features, single score	-0.2315
40 features, single score	-0.2284
20 features, aggr. score	-1.1030
40 features, aggr. score	-1.2025

The only derivable result being here that the single RFD-Score is a less worse measure than the aggregate score.

4.3 Critique

The resulting R^2 scores make it clear that the selected approach does not allow to predict any drug abuse probability, in this section a critical review of possible factors should be provided. Especially, as time constraints didn't allow to greatly deviate and reiterate the various stages of development.

RFD-Score The most obvious factor is the "constructed" nature of our RFD-Score, which is a rough modelling approach of reality but due to the limited scope of the project is not by any means scientifically backed. The main problem being here, that if the score is not modelling reality in a coherent way, the prediction may fail as the assumed and modelled relationship just doesn't exist. A better alternative could have been to rely on "hard" features as a binary classification of drug used during the last month.

Pre-Processing Pre-Processing could have impeded the information level of the constructed data set at different points, e.g. balancing can possibly influence the predictor quality (as well as the chosen under-sampling heuristics). Although the hugest influence is probably the data imputation, as a large amount of cell-values was imputed it may be that the sparse nature of this data set renders it illegible for specific drug abuse prediction without rigorous pre-selection of used samples.

Chosen Model Without time constraints, reformulations of the prediction problem as well as alternative predictive approaches could have been tested on their ability to caption the relationship. Especially neural networks could prove to be an interesting approach as we have a relatively huge sample size and NN have the ability to caption non-linear relationship of unknown form. Still, this would require a decent amount of Network modelling and hyper-parameter tuning.

Feature Importance The acquired features are coherent with the general conception of factors driving drug abuse, hence at this point better results were expected, it seems at this stage that the correlation between the recorded predictors and the RFD-Score is just too weak to predict anything.

5 Conclusion

The chosen approach of mapping drug abuse to a numerical score remains questionable and should be refined using more scientifically-backed approaches. Alternatively it shall be relied on other, directly observable, metrics.

Feature importance provided an interesting insight and is in-line what was expected in advance, still the predictive value of these single features is small, if interaction effects allow for better prediction would require further analysis.

Prediction results using the XGBoost Regressor with logit objective are completely uncorrelated to the observed RFD score, hence the model doesn't fit the modelling approach using RFD-scores.

Thus the task to predict drug abuse probabilities using only non-sensitive information was insightful, but not successful.