Drug Data Sets

Deepak Ghimirey

Project Category: Statistical Learning/Machine Learning

1 Introduction

Drugs have been a major factor in people's lives for a long time. There are several reasons that people use the drug, it may be for health-related issues, or experimental uses. There are two parts to this project, one is to reproduce the similar results from the article, "How Baby Boomers Get High" (fivethirtyeight.com). The second part of the project aims to look at the correlation of alcohol uses among participants with college degree and non-college degree. Also, it aims to develop a classification model using Support Vector Machine(SVM) to use the information in train.csv to classify the level of alcohol consumption of participants in the test data.

2 Related work

Existing report by Fivethrityeight states there has been "higher rates of illicit drug use, drug-related hospital admissions and overdose among baby boomers". It also mentions the wall street article and states "Woodstock mentality" that's making them carry out their young habits. In the article, it shows that very good margin of baby boomers (age 50-64) consumes drugs like alcohol and marijuana, and less of other drugs. The main reason behind is marijuana is not a hard drug as listed on the article. The highest percentage of drug consumed was Marijuana and the least used was Heroin. In the article, they have implemented a simple bar graph to show their findings.

This Project intends to further analyze the similar data set used in the article to analyze the consumption of alcohol by implementing method known as SVM; to use the information in train.csv to classify the level of alcohol consumption of participants in the test data. And use a simple bar plot to analyze the alcohol consumption among participants with college degree vs non-college degree.

3 Dataset

The main data source for this study was from a 2012 survey on drug use from the federal Substance Abuse and Mental Health Data Archive (SAMHDA). It has several features that can be used to analyze various related to issues and contribute the new findings to knowledge about the drug uses.

3.1 Sentiment Analysis on the Reviews

The reviews of data was analyzed in R Studio. Many of methods for data cleaning and masking was applied to this dataset. One of the important features “alcohol” was changed to 0 (never used alcohol) or 1(have used alcohol). The value of 0 was assigned to variables less than “CL4” and 1 to value greater than or equal to “CL4”.

3.2 Feature Selection for SVM

After data cleaning and preprocessing, the data set was divided into train(dimensions, 1570 by 13) and test (dimensions, 315 by 13). The data set was provided with names correlating to its descriptions given in the separate word document. Then the feature alcohol was picked to do the classification using 0 or 1 to denote it’s given value. In the data set, for the feature alcohol, the categories were the following:

Table 1.1

- CL0 Never Used
- CL1 Used over a Decade
- CL2 Used in Last Decade
- CL3 Used in Last Year
- CL4 Used in Last Month
- CL5 Used in Last Week

So, to do classification, I assigned the value of 0 “CL0” – “CL3” and 1 to “CL4” – “CL6”.

3.2.2 Feature Selection for analysis of consumption of alcohol to participants with college degree vs non-college degree.

And did the same for people with college degree (assigned value of 1) and non-college degree (assigned value of 0). The categories for education of the participants were the following:

Table 1.2

- -2.43591 Left school before 16 years 28 1.49%
- -1.73790 Left school at 16 years 99 5.25%
- -1.43719 Left school at 17 years 30 1.59%
- -1.22751 Left school at 18 years 100 5.31%
- -0.61113 Some college or university, no certificate or degree 506 26.84%
- -0.05921 Professional certificate/ diploma 270 14.32%
- 0.45468 University degree 480 25.46%
- 1.16365 Masters degree 283 15.01%
- 1.98437 Doctorate degree 89 4.72%

So, to analyze the alcohol consumption of participants who have college degree to non-college degree. I made the following changes to above Table 1.2,

- Left school before 16 years -> 0
- Left school at 16 years -> 1
- Left school at 17 years -> 2
- Left school at 18 years -> 3
- Some college or university, no certificate or degree -> 4
- Professional certificate/ diploma -> 5
- University degree -> 6
- Masters degree -> 7
- Doctorate degree -> 8

Then to I filter the data set with feature alcohol and education as follows, with classification value(cv) of 0 and 1 to filter the data.

Table 1.3

- Consumed Alcohol(cv:1) and Have Education(cv:1)
- Consumed Alcohol(cv:1) and No Education(cv:0)
- Not Consumed Alcohol(cv:0) and Have Education(cv:1)
- Not Consumed Alcohol(cv:0) and No Education(cv:0)

4.1 Methods

To develop a classification model, Support Vector Machine (SVM) to use the information in train.csv to classify the level of alcohol consumption of participants in the test data. In order to train the model, I used the "tune" function in the library "e1071". Then after using different types of the kernel such as liner, polynomial and radial, I got the best cost from list of "0.01, 0.1, 1, 5,10". Then with the best tuning parameter, I applied SVM method to each of following kernel to get the test error and train error.

4.2 Bar Plot

To develop a model for participants consuming alcohol with college degree and non-college degree, a simple bar plot was implemented. To do this, I counted the for the occurrence of description shown in Table 1.3 and made a bar plot using package ggplot2.

5.1 Results and Discussion

Mean squared error were used to evaluate the accuracy of the model for the SVM. The sample was trained on dimensions, (1570 by 13) and tested on dimensions, (315 by 13). After

implementing the tuning function, the best cost was selected as shown below for each of the kernel. Below you can the results in Table 1.4.

Table 1.4

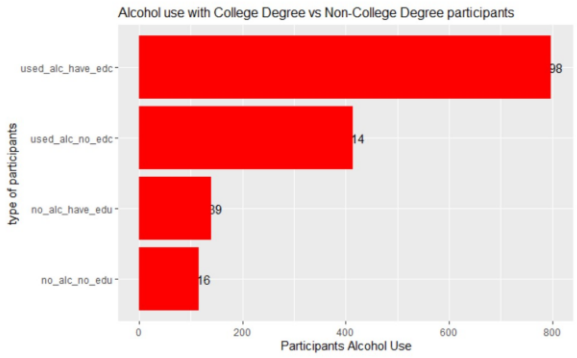
Model Name	Kernel	Cost	Training error	Testing error
SVM	Linear	.01	.1783439	.1714286
SVM	Polynomial	.1, degree: 3	.177707	.1714
SVM	Radial	.01	.177707	.1714

As you can see from above there is not much a different than using either of the method Radial and Polynomial. Both seems to be good but not by much.

5.2

Below tables shows the correlation of alcohol consumption of college degree holders vs non-college degree.

Figure 1.1

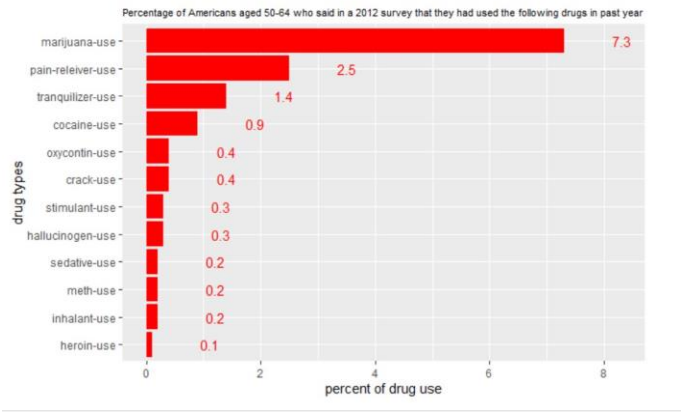


As you can see, people who have college degree have used consumed alcohol more frequently than other groups. From above plot, we can observe that, people with no education have consume less alcohol, irrespective to what my hypothesis as I began this analysis.

5.3 Reproducing FiveThirtyEight Results

How Baby Boomers Get High, FiveThirtyEight

Figure 1.2



The above bar plot is part of assignment to reproduce the plot from the website, FiveThirtyEight," How Baby Boomers Get High". As stated in the article, it suffice to say that baby boomers consume Marijuana the most out of all the drugs shown in the plot. It seems to draw more conclusion that, Marijuana use is more prevalent because people can buy it anywhere and it is not categorized as schedule 1 drug. Whereas heroine was used the least because it's hard to get, and it's classified as schedule 1 drug, and those who use it can be imprisoned for a long time.

6 Conclusions and Future Work

The project used SVM to train and test the accuracy of the model. In future, it would be great idea to do the same with several other classification method such as logistic regression, decision tree, random forest, cross validation, and many more to get the best accuracy for the model. Also, for future instances to analyze alcohol, we can look at different types of features and see how it relates to people drinking alcohol. The most interesting feature that I think to look at next time to analyze is how impulsiveness/personality traits correlates to use of alcohol.

Resources

For the outline of the Project:

- gitlab.com/hoorir/cs229-project.git

Reproducing the results/bar plot:

- https://fivethirtyeight.com/features/how-baby-boomers-get-high/

Implementing SVM method:

- https://www.geeksforgeeks.org/classifying-data-using-support-vector-machinessvms-in-r/

Drug Data Set:

- https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29