Predicting Depression through Machine Learning Methods

BeeVee Trade

12/11/2018

```
library(tidyr)
library(dplyr)
library(mosaic)
library(tree)
```

```
## Warning: package 'tree' was built under R version 3.4.4
```

```
library(ISLR)
library(ggplot2)
library(class)
library(foreign)
```

```
## Warning: package 'foreign' was built under R version 3.4.4
```

```
library(readr)
library(gbm)
```

```
## Warning: package 'gbm' was built under R version 3.4.4
```

Introduction

Mental health is an elusive subject of study. It is highly heterogeneous? and has a broad spectrum of symptoms and severity. Major Depressive Disorders (MDD) affects about 40 million adults in the United States, which begs the question: Can we predict whether or not someone will develop a form of depression? We built a model that attempts to predict if someone is depressed based on key variables. We have selected themes that extend from social networks to substance use. The hope is that we will be able to draw a predictive model from within these themes. The issue of MDD has become more predanite in recent years as the rates for diagnosis have increased. Depression affects both the mental and physical state of an individual. It can cause feelings of helplessness, a loss of interest in hobbies, changes in weight and appetite, and in extreme cases can lead to death. It is a delicate subject no matter how it is brought up and is often taboo to discuss. Machine learning techniques will allow us to see if we can determine who is most at risk and then be able to enact preventative measures. Within our own community there are many who suffer from depression and there are many more who have yet to be diagnosed. College students are at high risk of developing depression, as there is an acute combination of high stress, availability to substances, and a possible lack of a support group, which all contribute to poor mental health and poor self-care. Depression is close to many of our hearts, and addressing this issue might help us lead better and happier lives. Therefore, we hope to investigate the question about the determinants of major depressive disorders (social network, environmental, personal, etc., and comorbidities, like eating disorders) as these are relevant for both college students and the general public. Specific questions that can be asked: Are students that engage more time in social activity/ maintaining social network less prone to be diagnosed with MDD? Does increased frequency of substance use increase the chance of being diagnosed with MDD and increase the episode? As a group the key predictors we are interested in are eating disorders, family relationship, social network, personality traits, substance use (tobacco and alcohol), and suicidal thoughts and tendencies.

Data

We used a large joint dataset of three nationally representative surveys: the National Comorbidity Survey Replication (NCS-R), the National Survey of American Life (NSAL), and the National Latino and Asian American Study (NLAAS). This Collaborative Psychiatric Epidemiology Surveys (CPES) was initiated in recognition of the need for contemporary, comprehensive epidemiological data regarding the distributions, correlates and risk factors of mental disorders among the general population with special emphasis on minority groups. Its data collection covers a total of 252 geographic areas across United States and contains a total of 20013 observations and 5543 variables. All variables are also crosswalked under each category of mental disorder.

Data Cleaning and Methods

From the numerical summaries of these variables from the Suicide subset of the data we can see that there are some categorical variables, but most are numerical that depend on a larger range of numbers, and in this case the numbers represent the ages of the people involved in the studies. With our data we learn that most people said no to seriously committing suicide both in the last 12 months and at all. But from the other variables we can see that the people that have thought about committing suicide had these thoughts in their 20s. There are many NA values, so the summaries don't represent all of the observations from the data.

We also realize that we need to convert some numeric values to binary categorical values, for instance, there are two numbers, 1(no) and 5(yes), are used for indicating ICD alcohol dependence (lifetime). But we believe it will be confusing to use it as numeric and we want to convert it into a categorical variable. This is important to make sure about our consistency in treating variables with two categories as categorical, and treating variables with multiple categories as numeric. There is a huge inconsistency in the values available for observations used in different categories. For instance, it is hard to compare the data from variables under category eating disorder with variables from category family cohesion due to two datasets being actually collected from two separate surveys. Even though these two datasets were compiled together, it shows that the observations from one survey do not necessarily include variables from another survey. Therefore, we believe that it is essential for us to go back to our cross-walk original dataset again to select variables that are all from the same survey, instead of the collaborative survey that contains observations from other surveys.

The same problem is also shown when we tried to use tree decision as our main model. We were very excited to use tree decision model, however, our model turned out to be a disaster. We only got one node and aren't sure the cause. We believe that it might be due to missing values in ou data. Our data worked well with KNN, mainly because KNN does not have any assumptions. The only obstacle was that we were required to filter out observations that had any missing values, however this still left us with over 1,500 for the method. As for logistic regression, our data met the assumption that the response variable is binary. We also tested on large enough data (a sample of 1670). However, we are concerned about potential multicollinearity, such as seriously contemplating suicide and actually attempting suicide. It may suffice that we pick only one of these variables.

We believe that we are having major issues with the variables we are using right now. Many numeric variables containing only two values should be converted into categorical variables. And we need to revisit our initial category and variable selection by making sure all variables are from the same survey, instead of the collaborative survey. Even though it is painful to acknowledge this mistake we made earlier on about variable selection, we have learned a lot about reading and understanding complex collaborative survey datasets. And then we will try tree decision or logistic model again

However, it was at the very late stage of our data analysis did we realise that we have imported the SPSS dataset in the wrong format due to the wrong package. Therefore, it was no surprising that we were frustrated with data class conversion because all of them were in the wrong class. Thus, there was always an error here and there with variable selection and conversion. But this is a good learning experience in the future for us to always check the normality of the data structure before deciding the next step.

```
newdata= read.csv('newdata.csv')
#filter variables with less than 25% missing values
dat.25 <- newdata[, -which(colMeans(is.na(newdata)) > 0.25)]
```

```
#converting the response variable into either 1 or 0 and remove observations with missing values
dat.25$X <- NULL
dat.25$V07657 <- NULL
dat.25$V07876 <- NULL

dat.25$V07655 <- as.numeric(dat.25$V07655)
dat.25$V07655[dat.25$V07655 == "2"] <- 0

dat.25 <- dat.25 %>%
  filter(complete.cases(.))
```

Results

When we first ran our decision tree model we found that we had created trees with no prediction values as all the nodes are zero. This is what lead to the second wave of data cleaning. After we cleaned our data for the second time, we found through the decision tree model that there were three main predictors of endorsed depressive 12 month episode were number of years school the mother completed, religious preference, and highest grade of school or college completed. We found that the individual tree models predicted with 59% overall accuracy rate, a 49 % accuracy rate for the model to predict if someone will be depressed, and 62 % accuracy rate for the model to predict if someone will not be depressed. However when we ran the boosting trees our model predicts with about 60% accuracy when using test data.

```
#because the orignial percentage of depressed cases in the overall dataset is very small (only 8%).
#we restore the balance by adding one more condition: whether the person has any serious suicidal thoug

#the new percentage is 20%

dat.25 <- dat.25 %>%
  filter(dat.25$V01993 == "YES")
dat.25$V01993 <- NULL

#there are 319 out of 1597(20%) are diagnosed with depression
```

Using classification tree

```
#turning logi and numeric into factor

dat.25$V03221 <- as.factor(dat.25$V03221 )
dat.25$V03223 <- as.factor(dat.25$V03223 )
dat.25$V03224 <- as.factor(dat.25$V03224 )
dat.25$V03225 <- as.factor(dat.25$V03225 )
dat.25$V03226 <- as.factor(dat.25$V03226 )
dat.25$V03227 <- as.factor(dat.25$V03227 )
dat.25$V03228 <- as.factor(dat.25$V03228 )
```