Report on BRFSS dataset

There are 3 files that 4 files that I uploaded:

- elastic_net.R: perform all the elastic net models on depression and binge drinking
- helpers.R: contains all the function that makes the elastic_net.R looks a little bit cleaner
- download_files.R: download the BRFSS 2011-2015 and extract it in our system
- decode_variable.R: scrape through the `2015 codebook.pdf` file and create a `code_book` variable to get the meaning out of variables names.

I have cleaned the code a little bit by creating a `helpers.R` file to contain most of our functions. The functions in `helpers.R` includes:

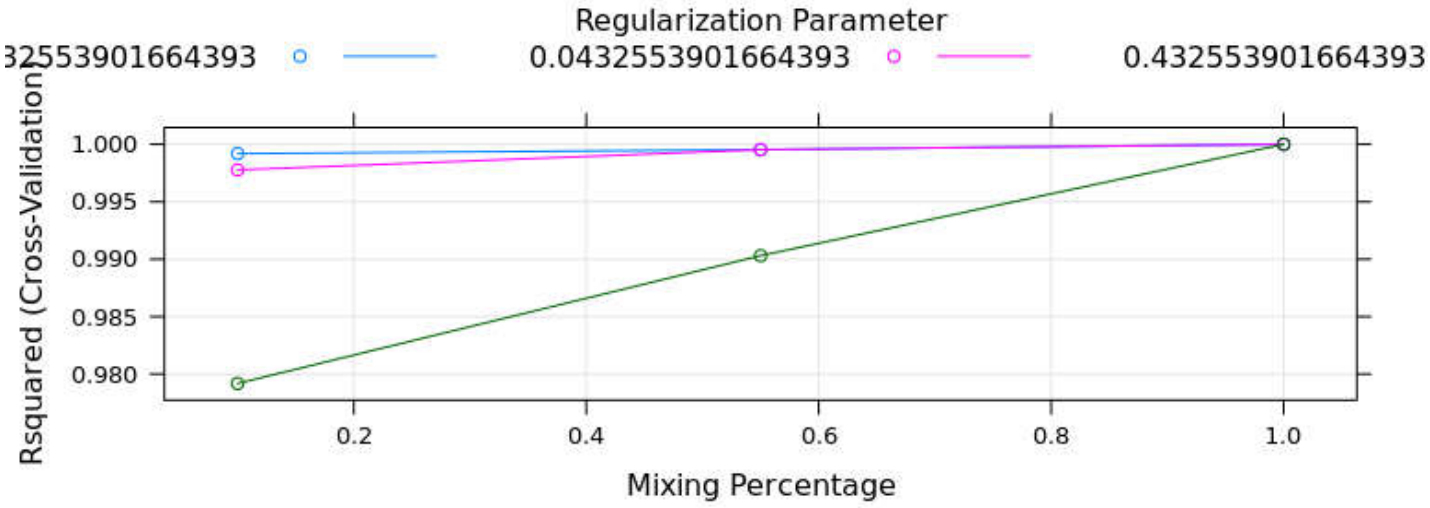
- elastic_var (Global Variable): all the relevant variables used for elastic net models on depression and binge drinking
- find_me: find the meaning of variables by name
- clean_data_depression: clean data for elastic net model on depression
- clean_data_binge: clean data for elastic net model on binge drinking
- var_important: based on **caret** package, finding percentage importance of variables in the model
- chart: draw the chart of variables' levels of importance
- download_file: used in `download_files.R` to download files into the system.

Analysis:

Our dataset in 2011 misses a lot of variables, so we have to use our 2013 dataset as our train set and the 2015 dataset as our validation set

1. Days of derpession against variables

We fit elastic net model and finding the best tuning **alpha** and **lambda** using R-squared

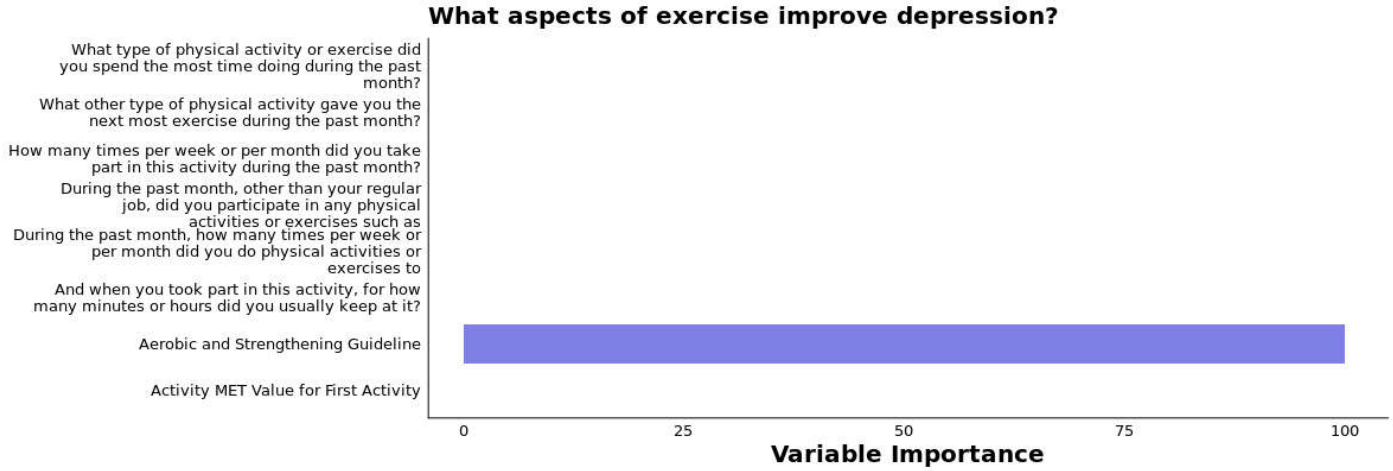


The best tuning parameters are: $\alpha = 1\ \lambda = 0.4325539$

So the optimal model on the 2013 dataset is a LASSO model with $\lambda = 0.4325539$

When we apply this model on the 2015 dataset, the $MSE = 0.2050306$, which is pretty good

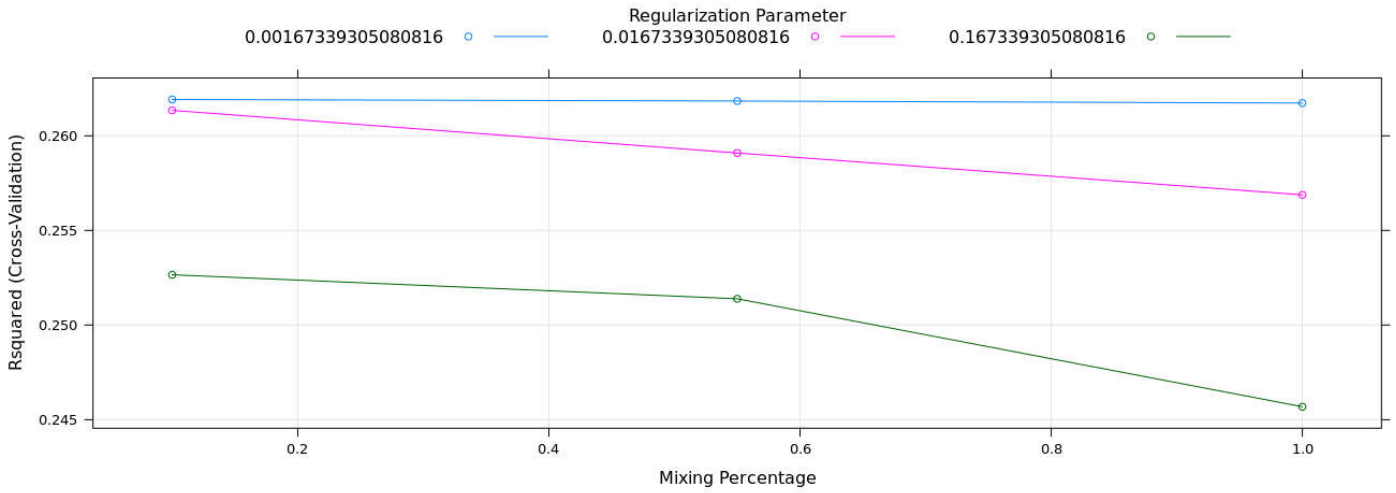
We use the `caret` package to find the importance of variables to our model



So the Aerobic and Strengthening guideline dominates all other variables, approximated 100%.

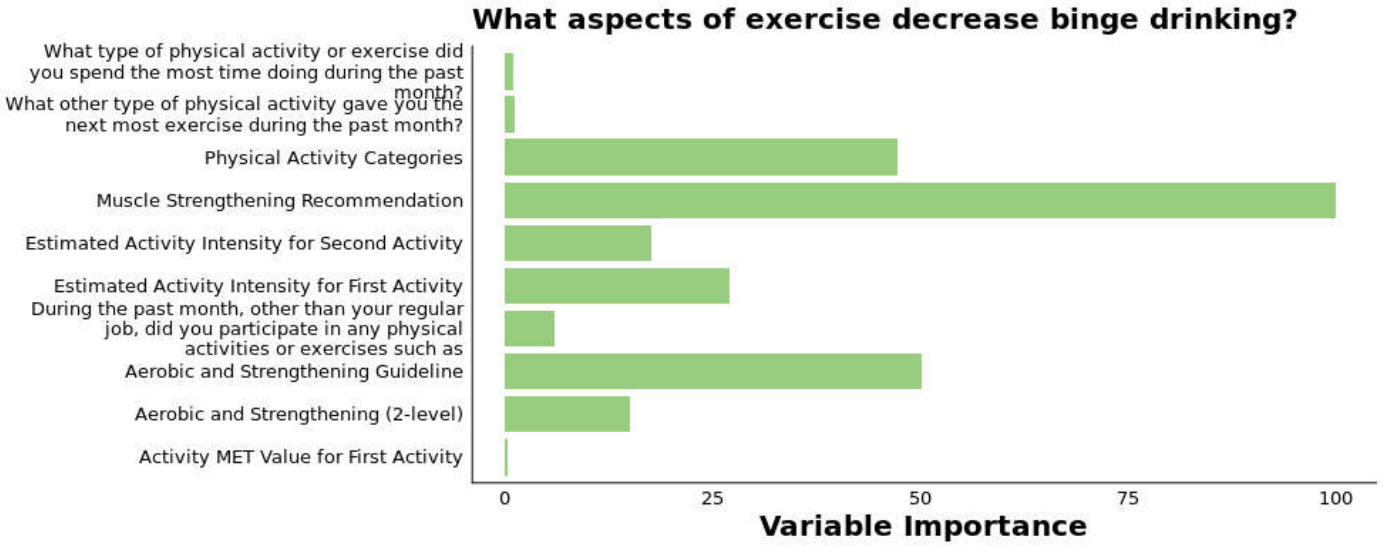
Binge drinking against all other variables:

1. Amanda's approach: elastic net of `X_RFBING5` against the other variables



The best tuning parameters are: $\alpha = 0.1$ $\lambda = 0.001673393$. So it approximates Ridge penalty.

Plotting the variable importance chart:



"

I'm not very sure about this part, since I copied Amanda's code to find the importance of variables. Obviously, the total does not sum up to 100 in this case. Is it something wrong with the original code or is it supposed to be intepreted as some sort of *confidence* of the variables

When I use this model and apply on the 2015 data, the $MSE = 2.491$.

2. My approach

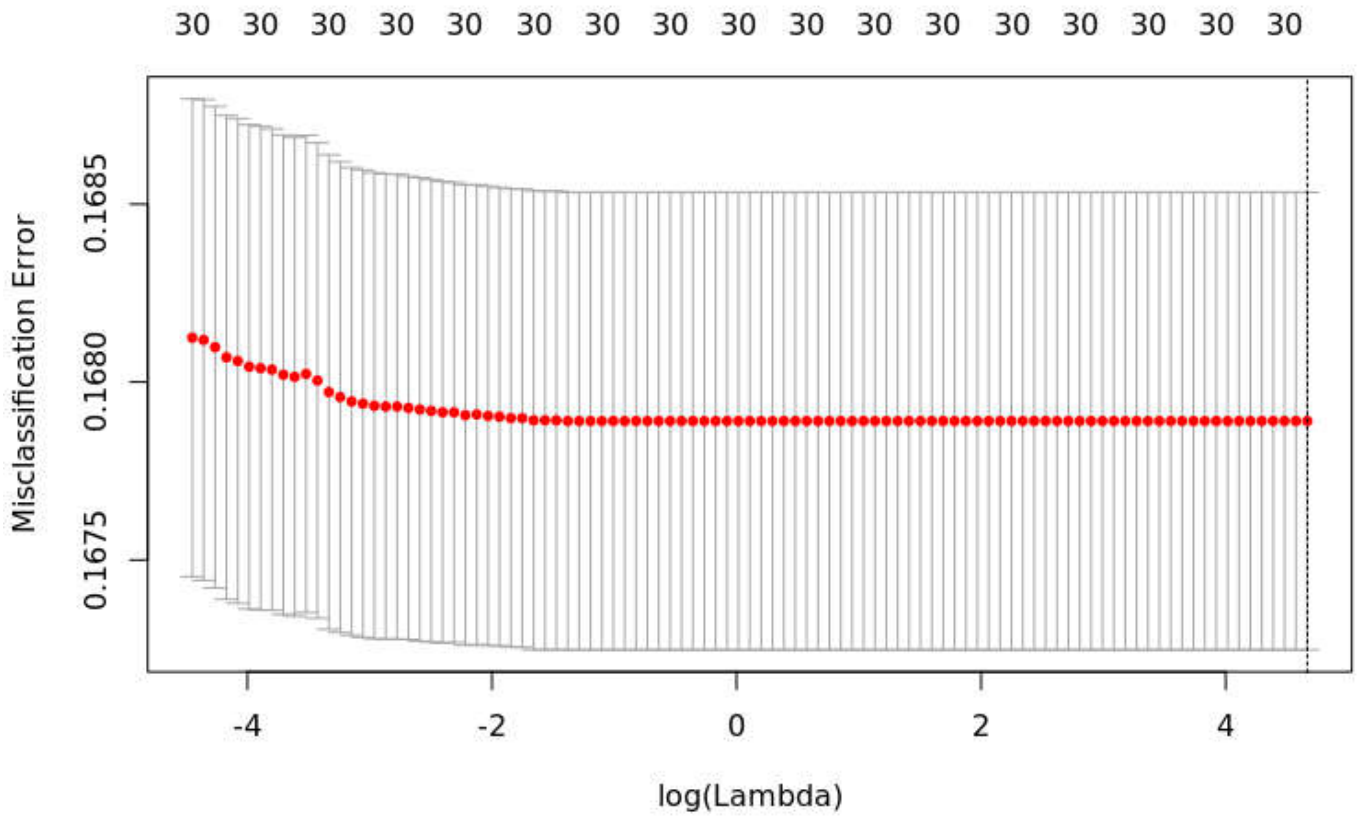
The way I look at this question is that it does not seem to be a regression model as much as a classification question. Variable `X_RFBING5` (Binge drinkers (males having five or more drinks on one occasion, females having four or more drinks on one occasion)) only has 3 factors:

- 1: No
- 2: Yes
- 9: Don't know/Missing

So it makes a lot more sense to approach this question as classification. I uses Multinomial Regression with both Ridge and LASSO penalty:

a. Multinomial Regression with Ridge Penalty

Using cross validation to apply the multinomial logistic regression with Ridge penalty to tune the lambda value:



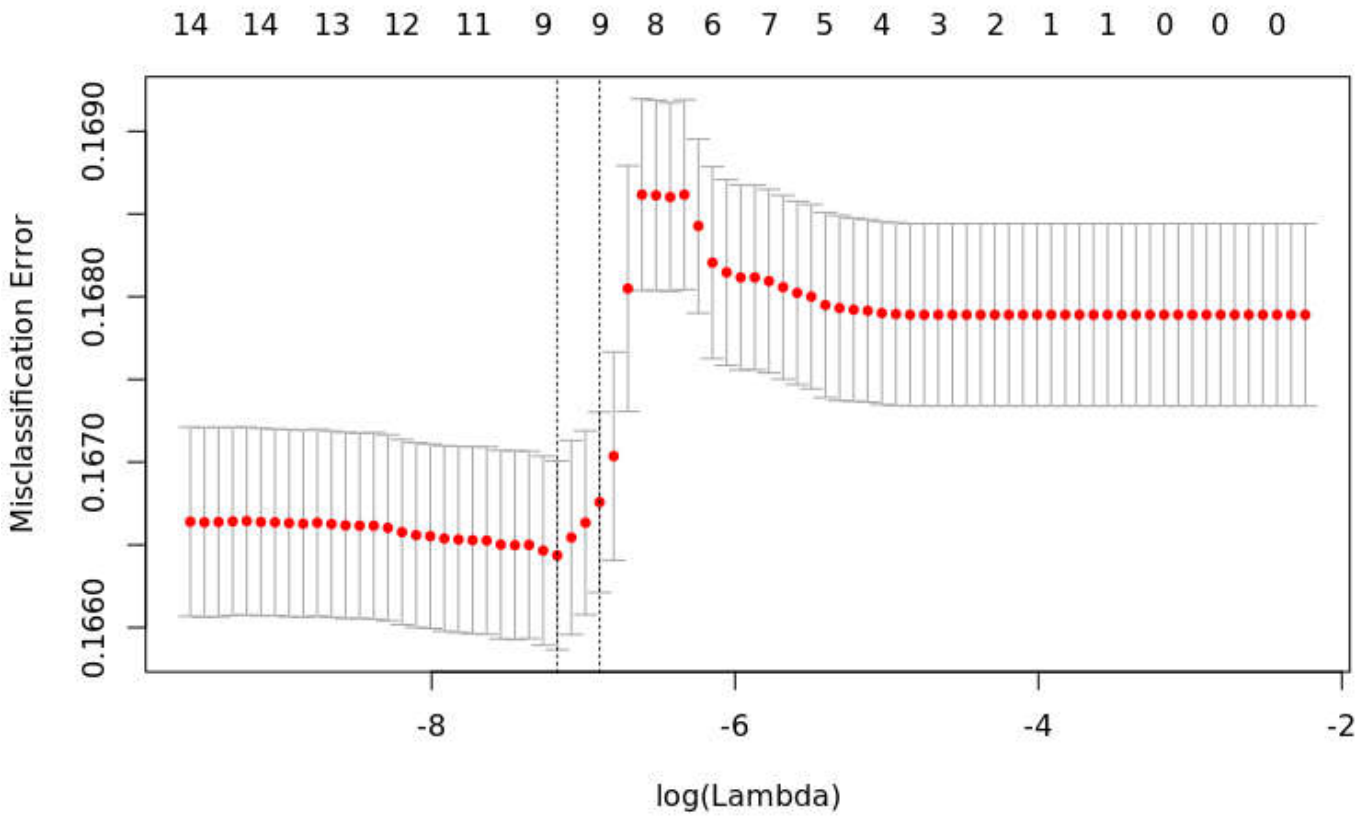
"

The optimal lambda (the value at which misclassication rate is minizized) is $\lambda = 106.3541$

When we apply this model on the 2015 dataset, the misclassification rate is **17.26%** , which is reasonably good.

b. Multinomial Regression with LASSO Penalty

Using cross validation to apply the multinomial logistic regression with Ridge penalty to tune the lambda value:



"

The optimal lambda (the value at which misclassication rate is minizized) is $\lambda = 0.0007679625$

When we apply this model on the 2015 dataset, the misclassification rate is **17.20%** , which does not seem to be a huge improvement from our RIDGE model

However, there is a significant reduction in the number of predictors when we use LASSO. At optimal Lambda, the model uses 26 unique variables.