# Final Individual Project

Your task in this project is to build a predictive model or algorithm of the response Y based on the predictors X1 through X140. Your training sample (n=5200) is in a 5,200 by 141 data frame stored with this R command:

save(Data.train,file="Data.train.RData").

Your testing data consist of another (new) data set of 5,200 observations on the same 140 predictors stored with this command:

save(Data.test,file="Data.test.RData").

There is no information available about the meaning of the data. You need to deliver the following items:

1. A file named “YourLastName\_Final\_Report” (of course, replace “YourLastName” with your name)), which should contain a brief description (up to 3 pages of single-spaced text in a Word file, if possible) of what you were doing in this project to get the best predictions. Also, include your guess (one number) about the MSE that you will get on the test dataset. It will be interesting to see how close your number will be to the actual MSE on this test dataset. Theoretically speaking, the best possible MSE is around 4, but you are more likely to get somewhat larger values. Values up to 10 are still very good, but much larger values are not so good, although might be acceptable. On the other hand, a straightforward linear regression gives the MSE above 300,000. So, as a minimum, you should get MSE significantly below 300,000. Your MSE on the test dataset will only partially determine your grade. You will also be graded on your effort and variety and depth of your approaches.
2. A file with R code (named “YourLastName.Code.R” (replace “YourLastName” with your name)), which calculates your final predictions. This code should take input from files "Data.train.RData" and "Data.test.RData" from the working directory and output the “YourLastName.Predictions.csv” file described in the next point. This code should not include EDA or CV or other preliminary work. It should just use your final predictive model or algorithm. Since your assessment of the test MSE will likely depend on CV, you can just use something like “MSE<- 2000” in your code using the previously calculated value (here 2000). You should test your file as follows:
   1. Delete all objects in your workspace
   2. Run source(“YourLastName.Code.R”)
   3. This should run all your code and produce the file “YourLastName.Predictions.csv” in the working directory
3. Provide a csv file (named “YourLastName.Predictions.csv” (of course, replace “YourLastName” with your name) with one column containing the following information:
   1. Your last name in the first line
   2. Your guess (one number) about the MSE on the test dataset in the second line (don’t write anything like “MSE=…”, only just one number)
   3. Predictions for Y’s in the test dataset (in the order of observations as given) in the following lines (one value per line)

This file should look like the posted template “Bajorski.Predictions.csv”, but with many more values, of course. Double check your format by opening your file. You will lose points for incorrect format.

Submit the three files to the Dropbox (without zipping) by 5pm on December 14, 2021.

You are supposed to work on your own on this project without discussing it with others. Each person has a somewhat different coding style, so it is easy to tell, if people were collaborating on the project. If I see any collaboration on this project, both students involved in the collaboration will receive zero score.