STAT 452/652 Fall 2023 Project 1

Owen G. Ward

This project involves a prediction task, and will be graded out of 100 for 10% of the overall course grade. Please read the instructions carefully below.

To complete this project, you must develop a prediction model given a large training set. We will then check the accuracy of your predictions on test data which you do not observe. You can (and should) use any methods covered in this class, up to and including Section 17.

This project must be done independently, and no discussion with other students in the class or anyone else is allowed. Failure to follow this rule will result in a score of 0. You may ask clarification questions of me or the TA but we will not give any feedback on your approach.

The deadline for all parts of this project is midnight November 30th. Submissions up to 12 hours late will be accepted, but will be penalised 5 points for each hour (or part thereof) late.

Instructions

- For this project you will receive a training dataset `training_data.csv` consisting of 10000 observations. This contains approximately 20 numeric predictors `X1,X2,...` and a numeric response `Y`. The goal of this project is to accurately predict `Y` for a test set and describe how you arrived at those predictions. You are given the predictors for this test data `test_predictors.csv` but not the response. You must use the skills from this class to build a predictive model and you will be evaluated in terms of the MSPE between your predictions and the true `Y` in the test set.
- You submission will include the exact predicted values for the response in the test data (see below for further details) and a brief report (max 4 pages) describing your procedure for choosing the model you fit. This report will be worth 50 points of the grade. 50 points will be determined by how good your predictions are.
- You must submit both an rmd file detailing how you got your predictions and a pdf report describing the procedure. **Unlike the homeworks, these do not need to be identical**. Detailed instructions for the report are included below.

- You must also submit your predictions as a CSV file containing **one column**. These predicted values of the response for the test data should be in the **same order with no row numbers and no column header**. The following code should be used to save your predictions to a CSV file.

```
write.table(predictions,
            file_name,
            sep = ",",
            row.names = FALSE,
            col.names = FALSE)
```

- You submit these predictions at this site where you enter your student ID number and upload the CSV. You must get `Submission Successful` to be sure your submission has been saved. If you submit multiple times it will overwrite previous submissions.
- 50 points of the total 100 will be based on how good your predictions are compared to other students in the class. In particular, we will compute the MSPE of your predictions on the test data. If the best prediction in the class is $MSPE_{Best}$ and your predictions have MSPE $MSPE_{You}$ then you will receive

$$\text{Your Score} = 50 \times \frac{MSPE_{Best}}{MSPE_{you}}.$$

- You must ensure your report is **reproducible**. In particular, the `rmd` file you submit must exactly create your predictions. You should use `set.seed` to ensure that when you rerun the script you get the same predictions. We will select submitted `rmd` files from those submitted at random and attempt to recreate your predictions. If we cannot recreate the predictions you submitted with **minimal effort you will receive 0 for the prediction component**.

The report pdf should briefly describe the methods you used and how you chose the model you did. This should contain code to illustrate what you did but this code does not need to run. To include code which does not run you can use the `eval=FALSE` chunk option.

You must address the following questions clearly in your report, and can include code to illustrate how you obtained your answer:

- What models did you try to use for this problem?
- How did you evaluate and compare models? Describe clearly the process you used.
- If the method you used has tuning parameters, how did you choose those parameters?

- Did you identify some predictors which are not important? How many true predictors do you think there are in this problem? Provide an estimate of the number of true predictors.