IE 7280 Spring 2023 Course Project

Data

In this project, we have a multiple linear regression problem. The training dataset contains 999 observations, where the response variable is Y (last column) and the regressors are $X1, \dots, X11$.

Project Instructions

Download the data. Try fitting as many models as possible, ranging from basic models (e.g., linear regression on subsets of regressors) to more complex models (e.g., considering transformation of regressors like polynomial regression). After trying different models, pick your best two. Though you do not have to, you are welcome to try methods not covered in the class (e.g., LASSO, Ridge regression, etc.).

For the two models, report your technical analysis, the training error (e.g., SSE), and an approximation of the test error using a technique you see fit in **3-pages including the figures and tables**. For reporting the performance, be specific about whether you are using the training or testing SSE, MSE, or you are relying on R^2, adjusted R^2, etc.. Your comparison metric should be clear from the context.

In total, you will have **3-pages** of technical analysis for the two models you worked on. After that, pick your best model (out of the two) and explain why it is your best candidate in **1-page**. By April 10th 9:00pm, you need to submit the technical analysis of your models (total **4-pages**), as well as the code for your best model. Then, we provide the test data and **you run the exact same code you submitted to us** on the test data to calculate your test error (SSE on test data). Briefly report that result (you will not need more than **1 page**). If you want to modify your code to achieve a better test error, you have the option to do so, in which case you need to report the changes as well as the improved rate in **1 additional page**. By April 20th, 9:00pm, you need to add your test results, your optional modification, and your executive summary to the previous report, and submit the final report on Canvas. If you have a modification upon your best model, you must submit the code for that as well on April 20th.

The final version of your report should have the following structure:

1. Executive summary (at most **1 page**, due April 20th) page 1
2. Technical report of models 1 & 2 (at most **3 pages**, due April 10th) page 2-4
3. Comparison of the 2 models and specifying the best model you want to use for testing (at most **1 page**, due April 10th) page 5
4. Evaluating test points only on your **best** model and reporting the test error (at most **1 page**, due April 20th) page 6
5. (Optional) Steps to improve your best model **or** introducing a better model (at most **1 page**, due April 20th) page 7

Evaluation of the project: The project has 15 points + 1 bonus point:

- Executive summary (2 points)
- Technical analysis of models 1-2 (8 points)
- Reasoning for choosing the best model (1 point)
- Evaluating your test result (2 points)
- Competition: the group with lowest test error will receive (2 points). Other groups will either receive 1 point or 0 point (it is also possible that the competition has multiple winners).
- (Bonus) the improvement upon your best method (1 point)

Rules

1- The results **must be reproducible.** If we run the code of your best model on test points and get a different test error from what is reported, you will not get points from the competition and lose the points from the technical analysis of the model. The same rule applies to modified code (if submitted).

2- Your code should be accompanied by a read me file clearly describing how to run your model on a given dataset. If you use .R, this read me file can be short simply stating the packages/libraries you use.

3- The report should be concise and to-the-point. The font size should be at least 11, and exceeding the suggested page limits will result in point deduction.

4- **Late submissions are penalized at the rate of 0.5 point per hour**.

5- Make sure to write a concise executive summary. This is a **non-technical** summary of your analysis for a broad audience. Do not include any graphs or tables in the executive summary.

6- In the last paragraph of the executive summary, **the contribution of each individual in the project must be clarified**. Evidently, if a group member does not contribute to the project, s/he will not receive any points from the project.

7- Please note that **not submitting the code of your best model on April 10th results in losing at least 4 points as we cannot evaluate the performance of the model "after" posting the test data.**

8- Every individual in the group must submit the report/code, e.g., if a group has 4 members, all 4 individuals submit the same materials on canvas.

9- You can use any software that is more convenient for you (e.g., R, SPSS, SAS). If you use a software that is **not reproducible**, you **must report the exact parameters** of your best model (due April 10th), so that we can check that for test data.