Assignment-1 Report

Zhaocheng Li (V00832770)

Instructor: Xuekui Zhang | STAT454 | Fall | UVic

This is a research on the relationship between HIV drug resistance and genetic mutations.

And this assignment is focusing on sparsing data set and analyzing with multiple regression models (e.g., standard linear regression model, lasso and other elastic net regression models) in R.

Code is stored as R markdown compressed along with this report. We are apply professor's sample code in it. And all the command executing required statistics is commented out for code clarity.

Answer the following questions of interests with statistics we get.

Question set:
[Notice: meanings of statistics]
[**MSE**: the mean squared errors, stand for average squared differences between predicted values and observed values]
[**Average Bias**: stands for the average differences between expected values and true values]
[**Variance**: basically same use.]

1. Which regression model works consistently better than the others?

- Solution: First of all, estimation with the stacking process worked sufficiently better all the time than estimation without it. Given the expert data set (relatively small one), there is not much difference of efficiency between standard stacking and cross validation improved stacking. Overall, the standard stacking works even better a little. Those observations above are all based on three types statistics we get. Normally, we tend to believe that smaller values of MSE, average bias, and variance will give us better way of estimation.

2. Does stacking really always do something? Explain if not.

- Solution: From the dataset, the values of all MSE has been decreased significantly after applying stacking. As we mentioned in previous question, the stacking process really efficiently improve the quality of predictions. This is because we want to multitasking on 5 different drugs and assess the estimation performance overall. Stacking would help us consider the factor of potential relationship among drugs (sample types) into account. That would give us a more reasonable and accurate result, especially when there actually exist such relationship between drugs. **Logically it would not help much if such relationship failed to exist.**

3. Does improved stacking really improve the performance of predictions? Try to explain why.

- Solution:Since values of MSE, and other two stats do not different significantly, then in my main focus (with the expert dataset), the improved stacking (cross validation like) did not seem to help a lot, comparing with the performance with standard stacking. The reason is that, first of all the cross validation we use inside the stacking function is not technically a great cross validation procedure, it is just cross validation like. It influence the performance more or less.

4. Does number of predictors really affect your prediction? (Complete set versus expert set)

- By comparing the mse, average bias and variance of dataset complete and expert, they do not reveal a big improvement from one to another. Hence in my research, more predictors do not lead to a more accurately predictions.
- In data analysis, the quantity of data is not the most important part to impact the quality of predictions. But the the proper model and quality of data are. After analyzing with both complete set and expert set, we observed that the difference of quality of predictions between two scales is not as obviously as such difference between difference analysis procedures (i.e., proper model + proper way, stacking, no stacking, or improved stacking). Therefore, more predictors is not necessarily give us more accurate predictions.

5. Anything else you'd like to from your findings.

- We can discover a lot from our findings apart from what have already mentioned. We see that throughout all statistics in no stacking process, those error associated statistics becomes obviously large in elastic net model with 80%, 90% proportion than other regression models, it means these two regression models performs especially not good under no stacking environment, After stacking, they perform just as good as other regression models.

- Secondly, we see the standard linear regression model does not take stacking process as a factor to influence its estimation. Because we get the same values for any statistic before and after stacking, or improved stacking.