IEOR E4650: Business Analytics
Fall 2017 Professor Jacob Leshno

Assignment 2: Model Selection

Due date: October 6, 11:59pm

Attention: Please prepare two files for each homework assignment: a .pdf file for your answers including relevant figures, and a .R file for your relevant R scripts. File names should be `Last_First_hw.pdf` and `Last_First_hw.R`, e.g., `Obama_Barack_2.pdf` and `Obama_Barack_2.R`. Your submissions must be based on your own original work. Late submissions will not be accepted.

In this problem you we will revisit the financial analytics session and apply model selection tools. The file `ibm_return.csv` contains the return of the IBM stock over the course of a year, together with variable that encode lagged returns.

1. Open the file `ibm_return.csv` in R and use the command `summary` to print a summary of the data.

2. Divide your data into two parts: a training set (75%) and a test set (25%). Note that we cannot split the data randomly, why?

3. Create 4 validation tests where you use 4 months of data to fit the model and then measure the performance on the following month. For each, use best subset selection to find the best model. Consider subsets of sizes from 1 to 8. Which subset size is best? What is your final model?

4. On the same 4 validation tests, use lasso regression to find the best model. Consider the values 0, .001, .01, .1, 1, 10, 100, 1000 for λ. Which choice of λ is the best? What is your final model?

5. Pick one of the two models final models from the previous two questions. What is the MSE of your model on the test data? How does that compare to the MSE on the validation tests?

6. Create a trading strategy from the model you picked. Start with \$1 of investment and every day select to go either long or short according to the prediction of the model. What is the return of your trading strategy on the test data? Based on the results, should you invest using this strategy?