FTEC5580 Project 1

Due 11:59pm, Mar 14, 2022

Instructions:

- Prepare a single Jupyter notebook (with code, results, your explanations and interpretations) and submit it in Blackboard. You CANNOT submit more than 2 times.
- Late submission incurs a penalty: 10% for submission on the first day after the deadline, 20% for submission on the second day after the deadline. Submissions made on the third day after the deadline and thereafter are NOT accepted.
- Name your Jupyter notebook as "last name-first name-P1", e.g., Li-Lingfei-P1. **Please follow the naming convention strictly.**
- You must work on the project independently.
- You can only use R as the kernel in Jupyter for this project.
- The TA responsible for grading this project is *WANG Boyu*.

In this project, we are interested in predicting monthly returns of stocks using linear regression models and boosted trees. Consider the following formulation:

$$R_{i,t+1} = f(\boldsymbol{X}_{i,t}; \boldsymbol{\beta}) + \epsilon_{i,t+1},$$

where

- $R_{i,t+1}$ is the return of the i-th stock in month $t+1$.
- $\boldsymbol{X}_{i,t}$ is the vector of covariates of the i-th stock in month t.
- $\epsilon_{i,t+1}$ is the error for the i-th stock in month $t+1$. The errors of different stocks and in different months are assumed to be i.i.d.
- $\boldsymbol{\beta}$ is the vector of parameters in the model and it is important to note that it is the same for all stocks. Therefore, one needs to pool observations of different stocks in different months together to estimate $\boldsymbol{\beta}$.

1 Data

Data is collected from CRSP, CompuStat and WRDS Beta Suite. All three databases can be accessed through Wharton Research Data Services (WRDS). The data contains the following types information:

- Ticker: the tickers for 30 stocks (current constitutents of the DJIA index with DOW replaced by C).

- Month: From Dec 2009 till Dec 2019. We use covariates of the current month to predict the return of the next month.
- Covariates: There are 49 covariates in total, including the current-month return and measures constructed using the price data, and beta, alpha, idiosyncratic and total volatility of a stock as well as financial ratios. You should read "explanation of some variables.xlsx" and "WRDS_Industry_Financial_Ratio_Manual.pdf" for explanations.
- Response variable: RETN, which stands for return of the next month.

For each stock, there are 120 observations and thus the observations of all 30 stocks combined is 3600.

The initial data downloaded from the databases contain missing values. To deal with them, we follow a common practice to use the average value of the covariates for all other companies in the same month as a replacement of the missing values. That is, if covariate X_j of Stock i is missing in Jan 2010, then the average value of X_j of the other 29 stocks in Jan 2010 is used. After data processing, we obtain the file "data.csv", which is ready for use.

2 Tasks

(1) Perform ordinary least squares (OLS) regression with all 49 covariates. Does the linear regression model provide a good fit and what is the test MSE estimated by the 10-fold CV?

(2) Try regression with LASSO, PLS and boosted trees. What are their test MSEs estimated by the 10-fold CV? Do you observe significant improvements over OLS?

(3) Which covariates are significant for predicting the return of the next month?

(4) From all your regression results, do you think returns of the next month are predictable using the covariates given here? Discuss the implications of your results for the semi-strong efficient market hypothesis.

Remark: The 49 covariates are on different scales. You need to standardize them before using them in any regression and the boosted tree model. Standardization means dividing a covariate by its sample standard deviation.