MF 850: Advanced Computational Methods

Professor Gustavo Schwenkler Fall 2016

Final project

Due: Monday, December 19, by 10 am via email to Professor Schwenkler (gas@bu.edu)

Assignment

The file "mf850-finalproject-data.csv" contains monthly stock return data for publicly traded companies in the United States during the year 2015. The file also contains fundamentals data characterizing each company in each month, as well as the monthly returns of the S&P 500 index. Your first task consists of constructing forecasts of the monthly stock return of a company based on the characteristics of the company and the market. Your second task consists of constructing predictors of whether a stock will grow or fall over the course of a month based on the characteristics of the company and the market.

To achieve these tasks, you may use any of the machine learning tools we studied in class. Your goal is to construct and train models that can generate accurate forecasts of the one-month stock return of a company, and of the grow-fall indicator. You may construct any models you consider to work well, as long as your model choices are motivated by an empirical analysis of the data which you will summarize in a final paper to be handed to Professor Schwenkler.

Data

Each row of the file "mf850-finalproject-data.csv" contains a realization of the monthly stock return of a company, the monthly return of the S&P 500 during the same month, as well as several indicators of firm fundamentals. Overall, the file has 62 columns:

- The first column "Date" gives the last day of the month in which the stock return is observed
- "RETMONTH_SPX" gives the monthly return of the S&P 500 index during that month, adjusted for dividends, splits, mergers, acquisitions, and other corporate events
- "compid" is an anonymized company identifier
- "Close" is the closing price of the company's stock on the last day of the month
- "Adj_Close" is the closing price of the company's stock on the last day of the month, adjusted for dividends, splits, mergers, acquisitions, and other corporate events
- "Volume" gives the average trading volume of the stock over the course of the month
- "Adj_Volume" gives the average trading volume of the stock over the course of the month, adjusted for dividends, splits, mergers, acquisitions, and other corporate events

- "Industry" gives the industry of the firm
- "RETMONTH" gives the monthly stock return of the firm from the end of the previous month to the end of the current month, adjusted for dividends, splits, mergers, acquisitions, and other corporate events
- All other columns are described at: `https://www.quandl.com/data/SF1-Core-US-Fundamentals-Data/documentation/indicators`

Deliverables

You should deliver to Professor Schwenkler via email at `gas@bu.edu` the latest by Monday, December 19, 10 am, the following items:

(i) A written summary paper of the empirical analysis of the data that motivated your modeling choices. This paper should be at most 15 pages long, and should include all necessary figures, tables, and estimates you computed during your analysis. Any codes you use do not need to be printed out and added to the summary paper. Instead, you should email your codes as a Matlab or R script to Professor Schwenkler.

(ii) A Matlab or R script that includes one function which takes as input all variables of the file "mf850-finalproject-data.csv" except column "RETMONTH", and returns a forecast of the monthly stock return "RETMONTH" based on the inputs. Note that not all the variables may be relevant (you may wish to run some variable selection method). However, your function should still take as input all variables and assign a zero coefficient to the variables that are irrelevant. The file should be self-contained. That is, any libraries that need to be loaded to run the function should be loaded at the beginning of the script. Any special functions that are used in the script should be defined at the beginning of the script. And any variables that are used throughout the script should be properly declared and defined. You should add comments to your script to make it easily interpretable.

(iii) Another Matlab or R script that includes one function which takes as input all columns of the file "mf850-finalproject-data.csv" except column "RETMONTH", and returns a prediction whether a stock associated with the inputs will grow or fall over the course of the month. Note that not all the variables may be relevant (you may wish to run some variable selection method). However, your function should still take as input all variables and assign a zero coefficient to the variables that are irrelevant. The file should be self-contained. That is, any libraries that need to be called to run the function should be called at the beginning of the script. Any special functions that are used in the script should be defined at the beginning of the script. And any variables that are used throughout the script should be properly declared and defined. You should add comments to your script to make it easily interpretable.

Your functions from parts (ii) and (iii) do not need to be consistent with each other. That is, it is OK if your function from part (ii) delivers a forecast of the monthly stock return of a firm that is positive, while your function of part (iii) predicts that the same stock will fall over the course of the a month.

Grading

Your final project will be graded on a scale from 0 (worst possible grade) to 40 (best possible grade) as follows:

- *Accuracy of your stock return forecasts (10 points).* Your stock return forecast function will be tested on a special test data set selected by Professor Schwenkler. The test data will include the same variables as the file "mf850-finalproject-data.csv", but will be out of sample. That is, the test data is not included in the file "mf850-finalproject-data.csv". The test data will include at least 3000 stock return observations of U.S. companies over the course of one month in 2016. The month that will be used for testing purposes will be randomly selected by Professor Schwenkler. Accuracy will be measured through the mean squared prediction error. Professor Schwenkler will rank all forecasts from most to least accurate, and will assign a grade for accuracy according to your achieved ranking. If your stock return forecast script does not run on any of Professor Schwenkler's computers, you will receive a grade of 0 points for this category. So make sure to test your script multiple times on many different computers, to ensure that your script will run properly.

- *Accuracy of your grow-or-fall forecasts (10 points).* Similarly, your grow-or-fall forecast function will be tested on the same test data used to test your return forecast function. Here, however, accuracy will be measured through the ratio of correct forecasts (That is, the fraction of all forecasts for which a prediction of "grow" was made when the stock actually grew, or "fall" when the stock actually fell). Professor Schwenkler will rank all forecasts from most to least accurate, and will assign a grade for accuracy according to a your achieved ranking. If your grow-or-fall forecast script does not run on any of Professor Schwenkler's computers, you will receive a grade of 0 points for this category. So make sure to test your script multiple times on many different computers, to ensure that your script will run properly.

- *Model justification (15 points).* Professor Schwenkler will judge whether your model choices are justified based on the analysis that you describe in your written summary. Models that appear unjustified (or randomly selected) will receive a low model justification score.

- *Model uniqueness (5 points).* Your model will also be evaluated along the uniqueness dimension. If your predictions are based on model that multiple submissions use, then you will receive a low model uniqueness score.

Rules

- You may work on the final project in a team consisting of **4 to 5 people.** Groups of less than 4 people or more than 5 people will not be accepted.
- Each team should hand in one summary paper (and used codes if necessary), one function for part (ii), and one function for part (iii).
- Your team can collaborate with other teams. However, teams that hand in very similar projects will be penalized with a low *model uniqueness* score.