BUFN 650 – Group Project

Group size You may work in groups of up to 6 people on this project.

Computations Please use Python and Google Colab (or Jupyter) and upload your *.ipynb* file on Canvas with your submission.

Deliverables Please submit your slides and a Colab notebook with detailed comments and analysis on Canvas. You will have to prepare a 5-7 minute presentation of your findings. You should prepare **no more than five slides** for your presentation. The presentations will take place on **Wednesday, December 7, regular class time**. Please sign up for your presentation slot *here*.

Setup

You work for a hedge fund. Your boss asks you to create a good model for forecasting excess returns on stocks. You obtained detailed data on a large cross-section of stocks, as well as many characteristics and accounting ratios for these stocks. All these data are available in the *characteristics_anom.csv* file in the Google Drive folder (it is a large file!). If you are curious about what these characteristics are, have a look at the accompanying *anom_doc.pdf* file for definitions.

Note that all characteristics are cross-sectionally appropriately normalized, so you do not need to standardize anything. Returns on every stock are provided in the variable *re*. All characteristics are lagged and shifted appropriately, so by simply regressing the column of returns (re) on all characteristics ($size$—$shvol$), you are estimating the following *panel regression*:

$$r_{i,t+1} = a + b_{\text{size}}\text{size}_{i,t} + b_{\text{value}}\text{value}_{i,t} + ... + b_{\text{shvol}}\text{shvol}_{i,t} + \epsilon_{i,t+1}, \tag{1}$$

where $r_{i,t+1}$ is excess return on a stock i from time t to $t+1$, and $x_{i,t,\cdot}$ are stock characteristics measured on or before date t. This is a panel data set; you may assume that all coefficients are constant across stocks and time.

In the file, stocks (i) are referenced by the variable *permno*, time (t) is referenced by *date*. Therefore, each observation can be uniquely indexed by the pair *(permno, date)*.

You may use the Colab notebook template I provided to load data from this file, which is located in your shared Google Drive folder.

Questions

Please split the sample into train and test parts. **Start the test sample in 2005. Make sure you do not shuffle observations! Do not use the test sample for anything but reporting the results!** For any parameter choice, cross-validation, etc., use the train sample only.

The ultimate goal of the exercise is to understand what variables can help forecast excess returns of any stock. **Please be creative!** Although I have provided you with the data, there are many ways you can manipulate the data to improve your model. For instance, you may think non-linearities might matter and use various non-linear methods, or include polynomials of X as your regressors for linear methods. You may formulate the task as a regression problem or classification problem (forecast whether stocks will go up or down next month). You may decide to work at an annual frequency by cumulating monthly returns to annual ones, or consider predictors farther back in time, or construct moving averages of predictors (be mindful of the fact that this is a panel data set and all such adjustments should be done separately for every stock). You may try many different machine learning techniques or even combine them to improve your results (i.e., use bagging or boosting techniques).

Your boss wants you to be precise in answering the following questions:

1. Explain any data manipulations you may have performed, if any. Which predictors do you use? Did you expand the set of predictors? Explain how and why if you did.

2. Which machine learning methods and techniques have you considered? At the minimum, please try at least four different methods, such as OLS, ridge, lasso, elastic net, random forests, principal component regression, neural networks (deep learning), or any other ones.

3. Compute various performance metrics in your train/validate/test data, such as R^2, MSE or RMSE. Which methods seem to work best in sample and out of sample?

4. Pick the best model and evaluate its performance in the train sample. Remember that if you want to pick the single best model/method among these and evaluate its performance, you need to make this choice based on a validation sample (NOT a part of the test sample), or use cross-validation in the train sample.

5. How much predictability do you find? Which variables are the most important ones? Would you start a hedge fund based on predictability you have discovered?

This list is not exhaustive. Feel free to add items if you find interesting results.