Big Data in Finance II Group Assignment

Instructions: Please submit both your code, output such as graphs and tables, and written answers to the conceptual questions. We strongly recommend that you submit everything as one Jupyter notebook. Use markdown cells to include written answers. At the beginning of the answer to each separate question, place a markdown cell that states the question.

You are welcome to use any software for data analysis, although Python is recommended. You can use the numpy code that we developed in class from scratch, or you can use a standard package such as Keras to fit neural networks. The former gives you more flexibility, while the latter sometimes allows you to iterate over different models more quickly. We leave it up to you to pick you preferred tool.

Grading criteria: Many of the tasks in this assignment are open-ended and there is no single right answer. This is deliberate – we will not grade you exclusively on the results of your algorithm. Rather, we will award high marks to groups who set up the problem carefully, follow a rigorous process, document their reasoning, and give good economic intuition in their answers.

Questions:

1. In the data used by Gu, Kelly and Xiu (RFS 2019 – provided in class), use a similar procedure to theirs to predict stock returns with neural networks. Start by finding a suitable baseline configuration, and use a validation procedure to pick optimal hyperparameters for three neural network models: One with 2 hidden layers, one with 3 hidden layers, and one with 4 hidden layers.
2. Use test data to get an idea of the out of sample performance of each model. Convert the standard MSE metric for out of sample performance to the "R2 out of sample" metric that was discussed in class. Compare your results to those in Gu-Kelly-Xiu and comment on the differences.
3. Pick the model that performs the best out of sample, and interpret its output by doing the following analysis of variable importance:
    a. First, for all stock characteristics, get variable importance by setting one predictor at a time to zero and finding the decrease in out of sample R2. Show a table of the 10 most important variables according to this measure, and give an economic interpretation.
    b. Second, get a measure of the joint importance of all our "macro predictors" (i.e., those taken from Welch and Goyal 2008), by setting them all to zero and finding the decrease in out of sample R2. Comment on how important macroeconomic variables are relative to stock characteristics in predicting returns.
    c. Repeat the two steps above, but by using a measure of the sensitivity of predictions to each input variable, as outlined in the lectures.
4. Fit a penalised linear model (LASSO) to the same data, using validation data to pick the best penalty (e.g., you can use the "sklearn" package in Python to do this easily). Compare its test data performance to the neural network.
5. Suppose somebody tells you to collect 10 more micro or macro variables that can predict returns and are not in our current dataset. How would you choose those variables, based on the intuitions you have gained in this project?