Weekly Update
Nov 16- Nov 20

Summary of work:

This week, we finalized our models and generated predictions that will be used in our next stage, which is to build an algorithm for a dynamic stock portfolio selection. Binqi took the responsibility of finalizing the Vector Autoregression model and generating an output file that stores the stock return predictions for 6 months (from 2019-08 to 2021-01). Matt and Isaac focused on the final tuning of the Random Forest model and generated an output file that stores the stock return classification (1 for increase and 0 for decrease) for 6 months (from 2019-08 to 2021-01). Congda was in charge of researching and coming up with the draft of our portfolio selection algorithm.

Classification Prediction Update (Random Forest)

______We have tried random forest classification to see if it classified any better than our previous models. We initially received a 65% accuracy rate on our validation set and were ecstatic about the result, but, to our dismay, we realized having a random split for training and validation would provide an unfair advantage for predicting time periods already trained on by the model based on including the date as a predictor, which results in inideal performance on validation set, and places less emphasis on other predictor variables. We re-trained our model that was independent of time, and still received an accuracy rate of 55% on our validation set. While it does not necessarily predict better than other models, we decided to use this as our final model for its ability to be visualized and explain each of its features' importance.

From this model Matt was able to get predictions of each stock on our test set (2019-08 to 2020-01). We eagerly wait, the magnitude of the performances by each class.

Stock Return Prediction Update (Vector Autoregression)

Based on last week's findings, we decided to move forward with the Vector Autoregression model to predict the stock return since it gives the best results (lowest discrepancy score). In addition, with the fact that the lagging 2 dataset we generated last week generates the highest accuracy, we decided to use a lagging period of 2 in our predictions.

Binqi first created a list of dataframes where each data frame corresponds to one company's observation. She then looped over each data frame and filtered the data to only contain the rows where the company has actual stock prices from 2019-06 through 2019-11 (2 month lagging effect taken into account). This is essential for not only fitting the VAR model, but also crucial to our stock portfolio selection.

During the process of generating 6-month predicted prices for each company ticker, Binqi encountered several obstacles. To start with, since the main goal in the previous modeling stage was to see and test model accuracy, some columns including company ticker (non-numeric) were dropped in order to fit the VAR model. However, our focus this week is to predict stock returns for each company, the company ticker column becomes indispensable. In order to deal with this, Binqi first created an index list to store the rows that are filtered, then she referred it back to the original dataset to find the corresponding company tickers and append it to the dataframe. Another issue Binqi realized was that she initially used predicted monthly stock price to predict the following month's stock price. This introduced a large discrepancy between the actual and the predicted stock price, especially for the farthest prediction (2020-01). After she discussed this with the team, she changed her code so that the following month's prediction is always based on the historical actual prices plus the past month's actual prices. This will be helpful in our portfolio selection stage because our main plan is to predict stock returns for the following month and make stock selections for our portfolio. By the time we make the next month's prediction, we will have the actual stock price of the current month. This largely reduced the prediction error.

One important thing Binqi learned this week is to always split big tasks into smaller tasks. This allows her to look at the bigger picture to clarify her goals before actually dealing with each smaller subtask. In the process of using the prediction model to predict the last 6 months' stock return, the main technique Binqi used was to devised sample codes to predict 1 month's return and then found a way to loop it for completing the whole task. She also validated her codes after each step to ensure code correctness. She ensures that every step is clearly documented and easy for group members to comprehend the logics involved.

Plans for next week:

This weekend, we will work on the stock selection algorithm and come up with a tentative portfolio of at least 30-50 stocks. We will first start with the two output files that are generated by the two models, which include the predicted values in a time span of six months (2019-08 to 2020-01). Our tentative trading strategy is that for each month, we add new "top" stocks from prediction to the portfolio, and drop stocks that have lost money for two consecutive months. For each month, we rebalance the portfolio using portfolio weight optimization. And finally, we calculate the return of the entire portfolio during the six months period.