**Weekly Update**

**Nov 9- Nov 13**

Summary of work:

This week, we started building our prediction models based on what we discussed last week. Matt and Isaac were responsible for the classification model in which they try to optimize the accuracy when classifying whether the stock return will increase (outcome 1) or decrease (outcome 2). Binqi and Congda focused on trying out different models and chose one with the optimized predictive power in forecasting the actual value of the stock price change. In addition, we also developed metrics to evaluate the accuracy of these models.

Classification Prediction

Our classification models were mainly explored through SVM’s but we explored with Random Forest models as well as Linear Discriminant Analysis. To reiterate, our goal with these models is to predict which companies will have an increase in their monthly return for a given month, also depending on the number of months we decide to lag the effects of these behaviors.

Hyperparameter tuning took an extensive amount of time, and training the data took much computer resources. To get around the issues, Matt and Isaac set up jupyterlab remote environment, and ran their program in it. Many different combinations of types of kernels, gammas, and cost were tested. They concluded that the “Sigmoid” function is the only kernel that results in balanced decisions, and low gammas and cost results in best accuracy due to high number of overlaps of two different populations.

However, the accuracy of these models were on average 55% for 5 fold cross validation. We would like to see this accuracy increase as the data of increasing monthly return consists of a similar proportion/ We would like to investigate applying time series analyses to the data to see if we can resolve some underlying bias and increase accuracy. This might involve Vector Autoregression or other time series based models, similarly to what the regression group did this week.

Stock Return Prediction

The first models that Binqi and Conda built were regression models. More specifically, they tried to use Linear Regression and Random Forest Regressor to predict the actual amount of stock price change. For the Linear Regression Model, they first utilized VIF (Variance Inflation Factors) and correlation matrix to see if there is a multicollinearity problem in the variable selection. After they found out that there is no multicollinearity in the predictor variables, they split the data into train and test sets (took the last three months as our test dataset) and used the predictors to fit a regression model in the train dataset. Based on our validation metrics, where the discrepancy is defined as ‘100 \* (predicted - actual) / actual’, the discrepancy in the final prediction is more than 100, which is not very ideal. For the Random Forest Regressor, we used Cross Validation and GridSearchCV with the sklearn package of Python to tune 720 different fits of hyperparameter. We found the best combination of hyperparameter to be: 'bootstrap': True, 'max\_depth': 10,  'max\_features': 'sqrt',  'min\_samples\_leaf': 10,  'min\_samples\_split': 5,  'n\_estimators': 500. However, the discrepancy using the Random Forest Regressor is calculated to be around 100%. Therefore, both models’ results indicate that the prediction of stock return is quite unreliable using these two models.

After Binqi and Conda did more research on the subject, they soon realized the problem with Random Forest Regressor. There is a key drawback of Random Forest Regressor: it cannot extrapolate data outside the range of the current data. For example, if the largest stock return in the training dataset is 10%, then the highest predicted stock return in the test set will never exceed this value no matter how high it actually is. Stock returns generally have a positive trend, this cannot be captured using either the Linear Regression or the Random Forest Regressor.

Considering the special characteristics of stock returns, Binqi and Congda decided to try time series models and see if they generate better results than regression models. In order to fit time series models, it makes more sense to use the absolute change in the predictor variables instead of the percentage change we used for the regression models. Therefore the first thing they did was to recreate data tables with different lagging periods, ranging from 0 to 5 months. The time series model we decided to use is VAR or Vector Auto Regression. This model is good for dealing with multivariate time series since it considers not only the historical values but also the dependency on other variables when predicting for one variable (in our case is the monthly stock return). Using this new model, we calculated our discrepancy again and found it to be much lower than the regression models we previously built, indicating much higher predictive power in future stock returns. We applied this VAR model to 6 different lagging possibilities (0 - 5 months) and found that the optimal lagging period is 2-month, whose discrepancy calculated was about 18.65%.

Plans for next week:

Starting next week, we will try to implement both the SVM model and the VAR model in the dataset and start to build our trading strategy. The general idea is that we will first use the SVM model to predict and filter a set of companies that are going to have increased stock return, then we will use the VAR model as our second benchmark to select companies that generate not only positive but also high stock returns. After that, we will set up a set of rules for picking stocks and building a portfolio. Our preliminary strategy is to make predictions month by month, add new predicted to be strong stocks to the portfolio (with high expected return), and drop proven to be bad stocks (losing money for two consecutive months) out of the portfolio.We would choose the last six month of our data set to be our out-of-sample data and run our trading strategy on that. We would get the return on our portfolio during that period and to see if the strategy is effective or not. We plan to arrange another meeting with Cary before Thanksgiving to show him our result.