Econ 3750:Big Data Project

Ammar Bin Zulqarnain

This project covers predicting the next day's direction of movement for the index of NYSE based on various sets of initial variables. I found and analyzed the relevant data:

> *Article:* CNNpred: CNN-based stock market prediction using a diverse set of variables, Expert Systems with Applications,Volume 129,2019,Pages 273-285, Link: https://www.sciencedirect.com/science/article/abs/pii/S0957417419301915?casa_token=X4geTnwiPW0AAAAA:glja7HHWTT2byFbUuJtn_Sii2oOKfZtQHTX8wOtXQQbLK7ZxAa2L60LPC0EFBGwlJO8bliq9
> *Authors:* Ehsan Hoseinzade, Saman Haratizadeh

For this prediction task, I had 81 potential predictors of the Closing Price for representing each day of each index. Some of these variables are index-specific while the rest are general economic variables and are replicated for every index in the data set. This set of predictors can be categorized in eight different groups: primitive variables, technical indicators, world stock market indices, the exchange rate of U.S. dollar to the other currencies, commodities, data from big companies of the U.S. markets, future contracts and other useful variables.

This data is from the period of Jan 2010 to Nov 2017. It had 1985 observations, including those with missing values. After omitting them, I was left with 1114 observations. The first 80% of the data is used for training the model and the last 20% is the test data. The value being analyzed was “Close” that is the closing price of NYSE. I dropped the columns “date” and column 59 as it was not identified as the potential predictors

I started the analysis with OLS Multlinear Regression in R. I got the results according to that where some of the potential predictors did not have the coefficients and just had “NA”. I believe this was due to the fact that OLS multilinear regression was not appropriate for such a dataset with the large amount of p as there were multicollinearity issues. Moreover, many variables were not statistically significant. Hence, OLS was not a suitable method for this regression and we need such a measure that accounts for the multicollinearity and could potentially select variable.

I used the two selection methods that were LASSO and ridge to analyze the given dataset to check how the Close was dependent on the potential predictors. For analysis to be replicable, I set the seed to be 1 for both models. I used the “caret” library to perform these operations. This library allowed me to perform 10 k-folds (due to the large number of p) cross validations in LASSO and ridge. I used the train function of the library that allowed me to standardize the

dataset by centering them and dividing its columns with its standard deviation through the argument preProcess. Similarly, this function allowed me to use the optimal lambda value to be used as a penalty for the predictors having multicollinearity. I also used the glmnet package for LASSO and ridge operations that was included in the train function. Example of the train function is ridgemodel= train(Close~., data=trainingSet, preProcess=c("center","scale"), method="glmnet", tuneGrid=expand.grid(alpha=0,lambda=lambda_vector),trControl=ctrlSpec)

After training both models with the training set, I used the testing set(the remaining 20% of the data) to see the accuracy of both models on the unobserved data. I calculated the RMSE and R^2 of both models to see the performance of both models.

The results of both models were very significant as the ridge model used all of its potential predictors, associating coefficients while the LASSO model dropped most of the predictors, leaving “mom”(return of 2 days before), “ROC_5”(rate of change 5 days ago), “ROC_10”(10 days ago), “ROC_15”, “EMA_10”(10 days exponential moving average) as the significant predictors which are the technical indicators.By making the models to predict on testing set, the values of RMSE and R^2 were 37.05261 and 0.9994970 for LASSO respectively and 91.45649 and 0.9956954 for ridge.

To summarize all the analysis and making conclusions, we can see that the LASSO model performs better than the ridge model in terms of RMSE and R^2. The LASSO model suggests that most of the potential predictors were leading to multicollinearity issues and the closing price of the NYSE index can be predicted from its historical stock prices sufficiently. Ridge model was trained and tested in comparison with LASSO to identify which model performs the best. The LASSO model suggests the following equation to determine the Closing Price of NYSE:

$$Close = 9632.7 + 12.39 * mom + 49.23 * ROC\_5 + 27.39 * ROC\_10 + 10.88 ROC\_15 + 1345.75 * EMA\_10$$

Financial markets are considered as the important aspect of the world's economy in which billions of dollars are traded every day. A good prediction of future behavior of markets would be extremely valuable in various areas. The LASSO model can be utilized in making good predictions of the future closing prices of the NYSE index.