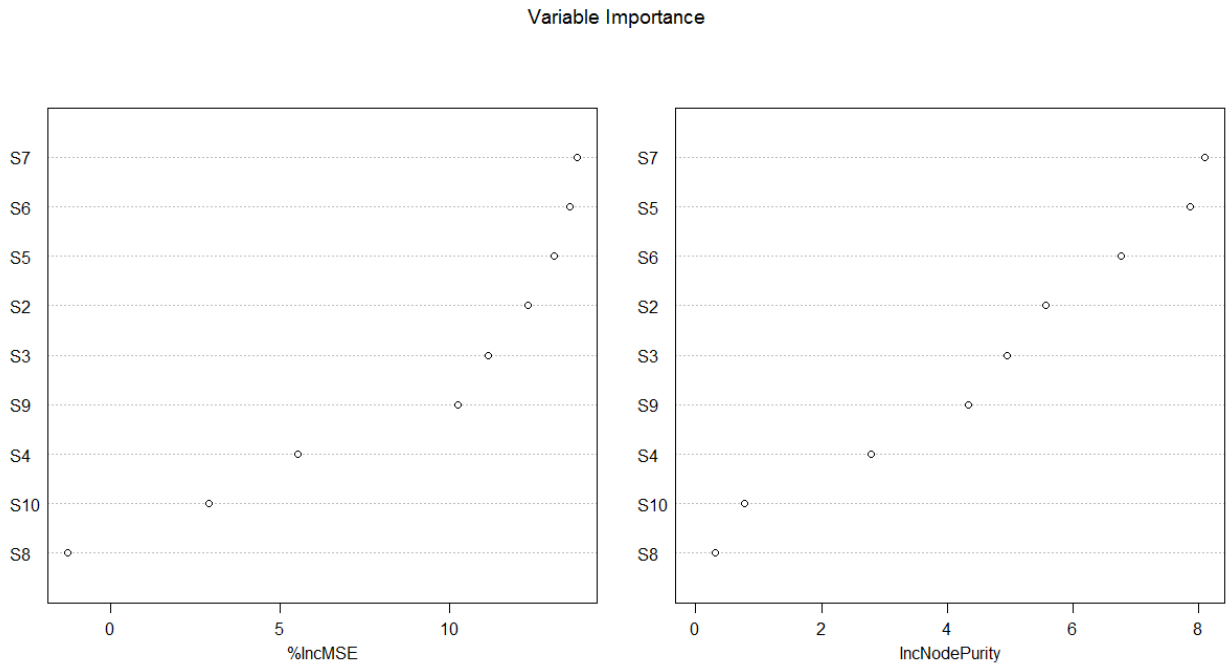
Objective: Build a model t forecast S1 stocks based on the changes in other S2, S3, S4, and etc. stocks. The dataset consists of the simulated daily open-to-close changes of a set of 10 stocks. Please find my below analysis for forecasting S1 stocks

(1) Which variables matter for predicting S1?

- Variables that will help us explaining the variance in S1 stocks are S7, S6, S5, and S2.
- We used Random Forest to help us identify the importance of variables in the model which will help us better predict S1.
- As you can see from the below graph, the variables at the top of the chart are highly correlated to S1 and hence will help in better explaining the predicted value of the stock price changes.



(2) Does S1 go up or down cumulatively (on an open-to-close basis) over this period?

- Based on the predicted values that are generated from the Random Forest model, we can observe the cumulative value over the period from 8/11 – 10/21 by calculating the sum of predicted values.
- SUM (Value) under predictions.csv = 1.001 which is a positive value.

- Hence, we can state that S1 goes up cumulatively on an open-to-close basis over the period

(3) How much confidence do you have in your model? Why and when would it fail?

- In a 95% confidence interval whether the predicted values are either positive or negative defines the confidence of the model
- I this case 30/50 observations are either positive or negative, hence there is a confidence of 60% that the model will not fail
- However, when the 95% confidence interval holds a zero, i.e. when the predicted value lies in the range of positive and negative value it states that the model fails for such values
- These metrics are drawn from the "Overall_Predict.CSV" file which holds columns for Predicted Mean, Predicted Low (0.025) confidence, Predicted high (0.975) confidence interval, and predicted standard error.

(4) What techniques did you use? Why?

- The "stock_returns_base150.csv" data provided contains total of 100 observations, out of which 50 observations are used to train the data
- As we know this sample is very less to do a good prediction or forecasting the changes in the stock price of S1
- Hence, we have implemented 10 Cross fold validation technique to avoid overfitting of the data when training the prediction model
- Models selected and compared for prediction are Random Forest (Regression) and Elastic net Regression
- Elastic Net creates a regression model that is penalized with both the L1-norm and L2-norm. This has the effect of effectively shrinking coefficients (as in ridge regression) and setting some coefficients to zero

- This helps in reducing the effect of correlation within the data and, helps in identifying the important variables to build and effective predictive model. Please find below correlation plot for the data
- The 10 fold cross validation method is also used for comparing the accuracy of the models by calculating the absolute difference in the predicted and actual values.
- In our case, Random forest had better result compared to Elastic Net Regression