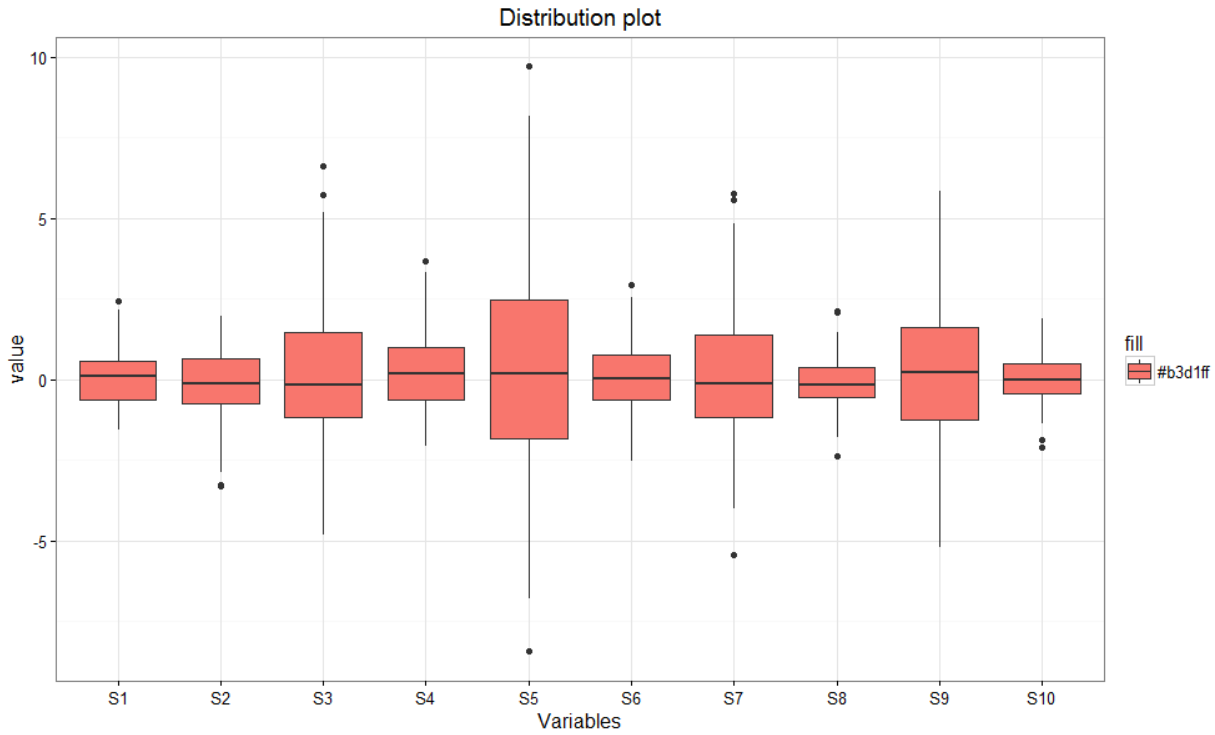
Table of Contents

Problem Statement:

Predict S1 stock values based on stock values S2, S3, S4….S10 traded in Japan. Build machine learning/statistical models to predict S1 trade values.

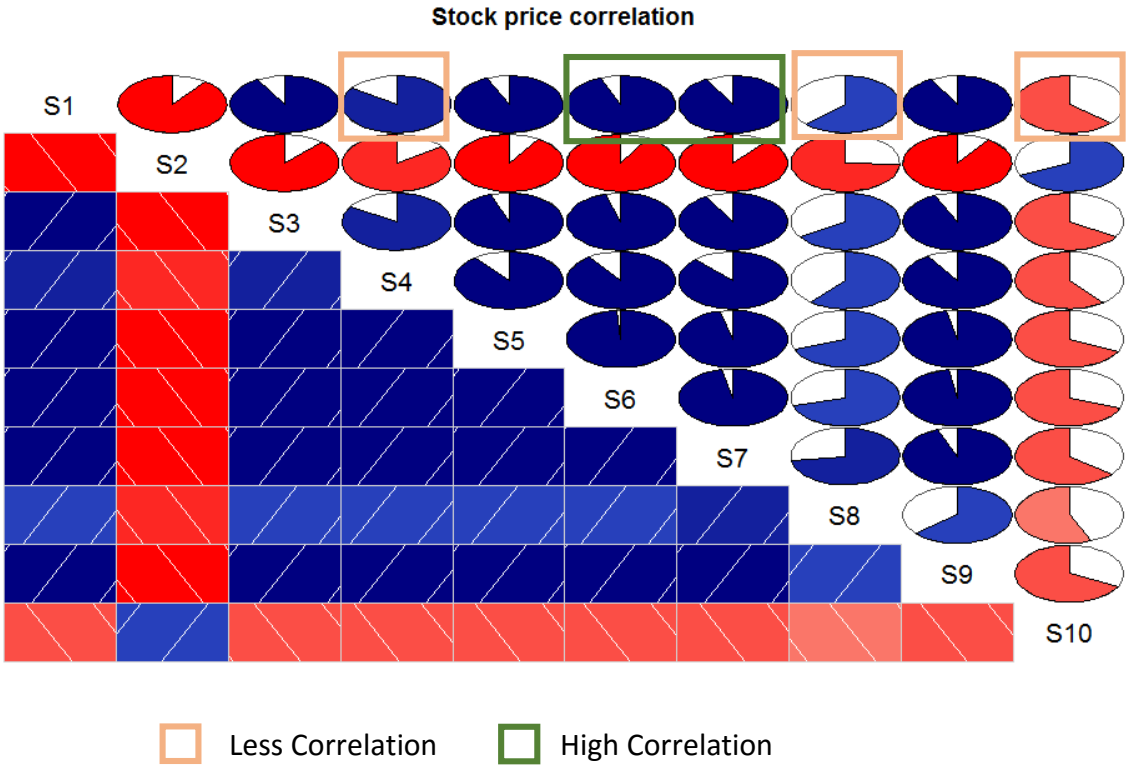
Exploratory Data Analysis

Data Distribution



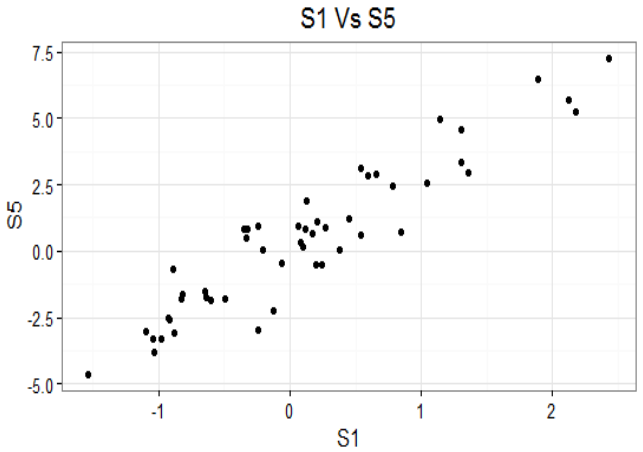
- The distribution of all the stocks are near normal distribution hence need not transform the variables. All the stocks have a median stock value close to zero and they are symmetrical in distribution. Outliers in the data are quite negligible

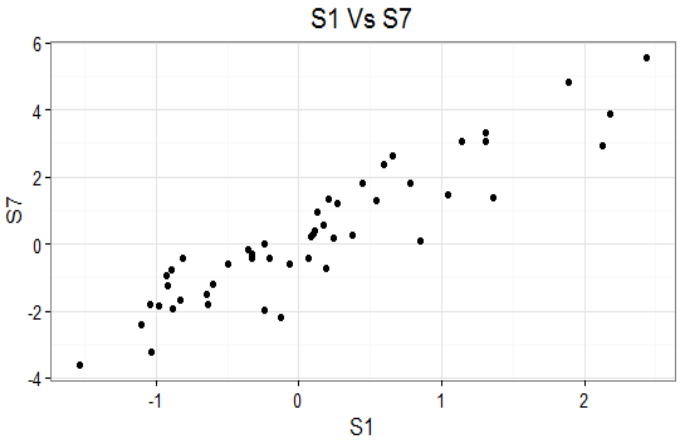
Correlation Matrix



- By Visual Inspection S4, S8, S10 does not seem to have high correlation with S1, we would drop these variables by the support of a model.
- Most of the stocks S5, S6, S7 does have high correlation with Stock value S1 hence linear models like linear regression, SVM would perform better
- Stocks S6, S9 has high correlation with other variables. Hence these variables are multi collinear hence redundant.

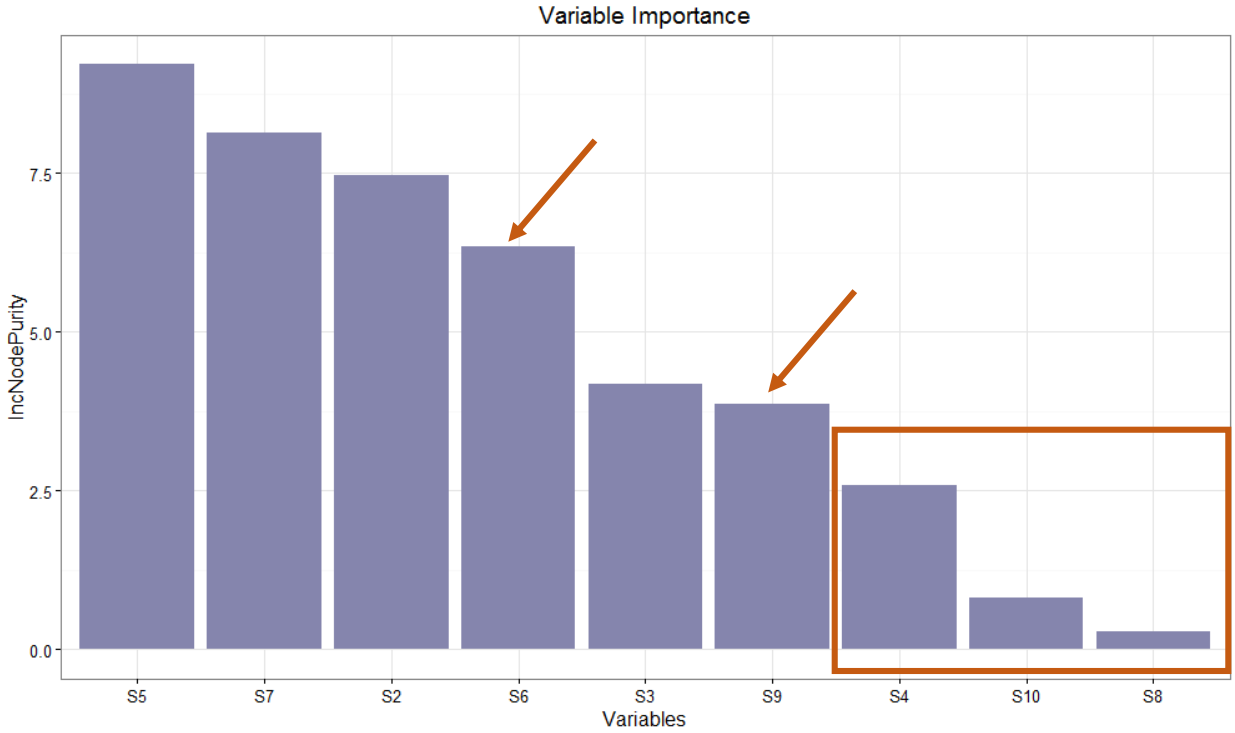
Scatter plots to inspect correlation





Feature Selection:

- Built random forest model on the training data to select the variable with higher relative importance compared to other variables
- Variable importance from random forest



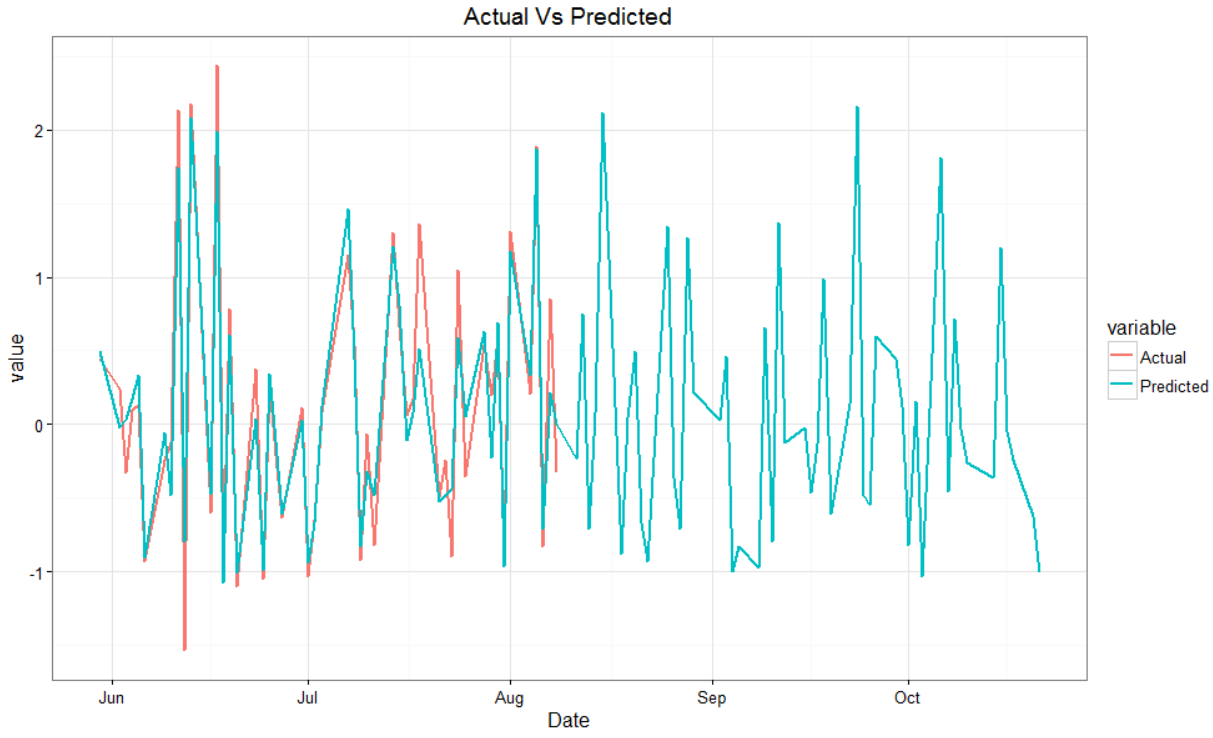
- The measure based on which the optimal condition is chosen is called impurity and for regression trees it is variance. Thus, when training a tree, it computes how much each feature decreases the weighted impurity in a tree.
- As expected, stock values S4, S8, S10 has relatively less importance and S6, S9 are highly correlated with other stocks and relatively less important.
- Hence variables S5, S7, S2 and S3 are considered for the model to predict S1 values.

Modeling:

Support Vector Machine

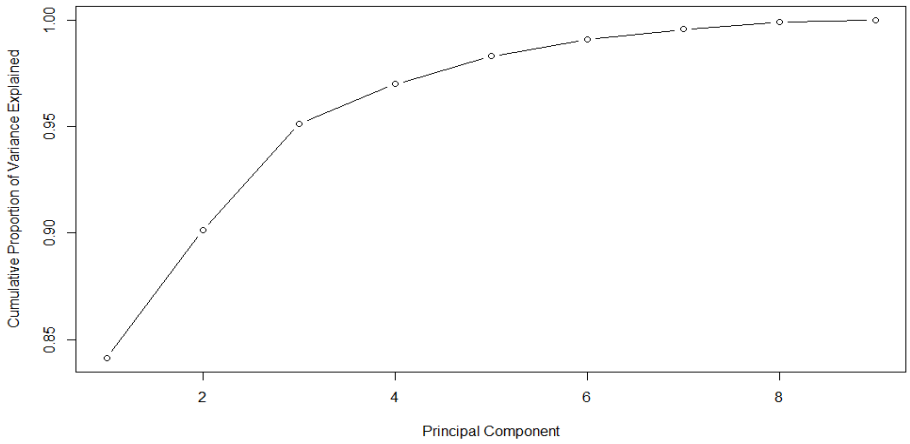
- Divided the Train data of 50 observations into Train_X (40 observations) and Train_Y (10 observations) to validate the model
- Sum of absolute deviations on Train_Y is **3.10** and on complete train data is **10.63**

SVM Model Prediction



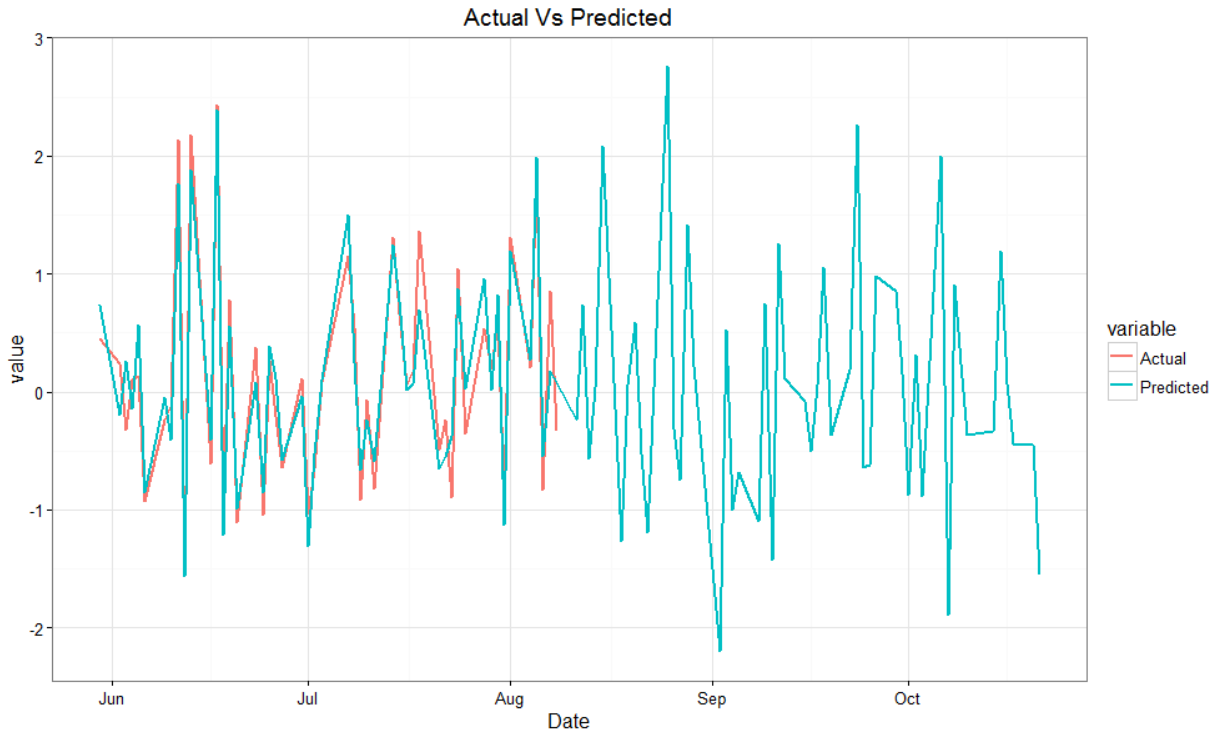
Principal Component Analysis

- Divided the Train data of 50 observations into Train_X (40 observations) and Train_Y (10 observations) to validate the model
- Considered 5 principal components where most of the variance is explained



- Built a linear regression model on top of the selected principal components
- Sum of absolute deviations on Train_Y is **2.9** and on complete train data is **12.17**

PCA Model prediction



Results:

Sum of absolute deviations on the train and test splits for the model:

- Support Vector Machine (SAD – 3.10)
- Principal Component Analysis (SAD – 2.9)

Question1

Top Features for predicting S1

- S5
- S7
- S2
- S3

Question2

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

Based on the model predictions S1 goes up cumulatively over this period

Question3

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

Linear regression on the principal components gives an Adjusted R-squared of 88.9 % which signifies the model explains 89 % variability of the response data around its mean

Limitations:

- PCA relies on linear assumptions
- Directions with largest variance are assumed to be the most important by adding more data the principal components may change.

Question4

What techniques did you use? Why?

Feature Importance: Random Forest

- Random Forest is a tree based model has relatively good accuracy, robust and explanatory. when training a tree, it can be computed how much each feature decreases the weighted impurity in a tree.
- For a forest, the impurity decrease from each feature can be averaged and the features are ranked according to this measure.

Prediction: Principal component Analysis

- Principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set with a motive to capture as much information as possible
- Performs better when most of the variables are correlated

Relevant Experiments:

1. SVM model was built on the dataset and used grid search technique to tune the parameters. Eventually PCA was the winning model compared to the best SVM model
2. This small dataset is split into train and test set. Repeated cross validation was used in order to estimate the models performance on out of sample data.