The purpose of this project is to discover signals that predict the returns of stocks. The dataset consists of the daily open-to-close changes of 10 stocks in which stock S1 trades in the U.S. as part of the S&P 500 Index, while stocks S2, S3, …, S10 trade in Japan, as part of the Nikkei Index. The following steps have been taken to preprocess and model the dataset to predict variable S1.

1. Creating new set of variables to capture the dynamics of the stocks over time.

New variables are defined and added to the feature set because the stock values not only depend on the current day, they may also depend on the stock values in the past days. The following variables were added to the feature set:

- 1-day and 2-day lags: Shift the daily stock returns one and two days and add them to the features.
- Moving average of the last three days: Average stock returns of the past three days.

2. Handling missing values

The new variables have missing values for the first few days because these variables are calculated based on previous days and for the first days, the value for previous days are not available. Missing values were replaced with the next available value in the dataset. For example if for a column the value for the first two days are missing, then these values are replaced with the value for the third day.

3. Building the models

After preparing the dataset, three regression models were developed to predict S1: 1) Support Vector Regression (SVR), 2) Kernel Ridge Regression (KRR) and 3) Random Forest (RF). For tuning the parameters of the models, 5 fold cross validation was used. The evaluation score was Mean Absolute Error. The best parameters were selected among C = [0.1, 1, 10, 100] and gamma = [0.01, 0.1, 1, 10]. Later, the three models were compared based on their MAE. Support Vector Regression were selected as the model with the smallest MAE. Finally, the test set were predicted using the best model and the results were written in the .CSV file.

1) Which variables matter for predicting S1?

Based on the Random Forest model, the three variables which matter the most in predicting S1 are S7, S6 and S2. In order to validate if these variables are correlated to S1, the correlation plot is shown below for all the three variables versus variable S1.

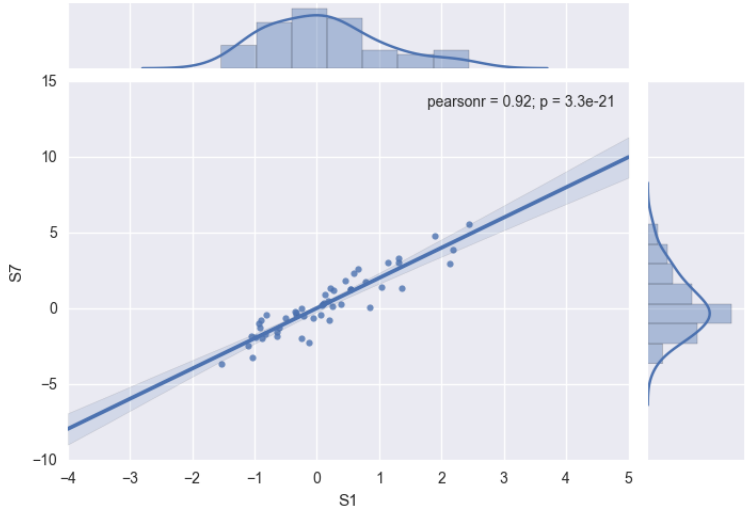

Figure 1) Correlation between S1 and S7

As can be seen from figure 1, the variables S1 and S7 are highly correlated with a Pearson coefficient of 0.92 and a p-value of 3.3e-21.

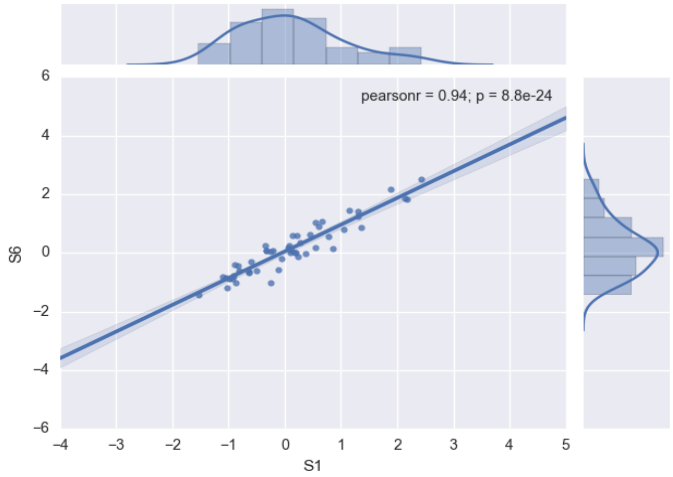

Figure 2) Correlation between S1 and S6

Similar to variable S7, variable S6 is also correlated to variable S1 with a Pearson coefficient of 0.94 and a p-value of 8.8e-24. Finally, the correlation of S2 and S1 are plotted in figure 3.

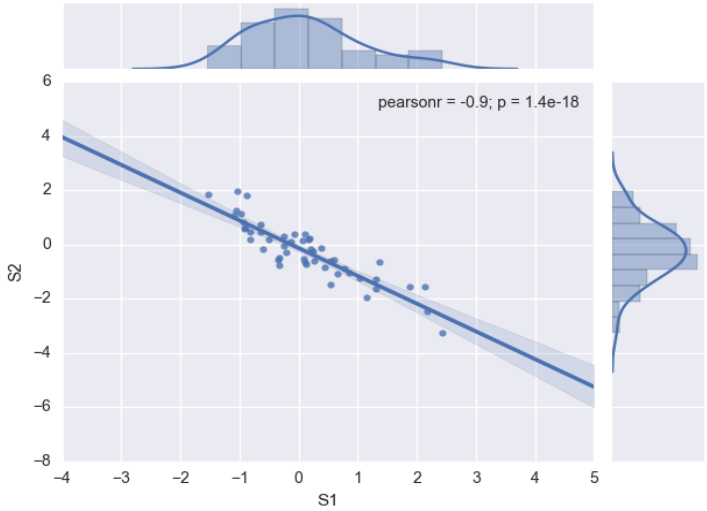

Figure 3) Correlation between S1 and S2

Variable S2 is also highly correlated with S1.

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

Using the data from both training set and testing set and using the following formula:

$$Cum\_Return\_Rate = \prod_{i=1}^{n} \frac{(100 + Return\_Rate_i)}{100}$$

where *n* is the total number of days. If the Cum_Return_Rate is more than 1, then S1 goes up and if it is less than 1, it goes down. This is based on the assumption that the opening stock value for each day is equal to the closing stock value of the day before. Using the predictions from our model for the 50 testing set and the existing values for S1 in the training set, Cum_Return_Rate was calculates as 1.107717. This number shows that S1 goes up over this period of time from 05/30/2014-10/21/2014.

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

In order to show the confidence of the model, R squared, Kendell's coefficient and the p-value of Kendell's coefficient were used to measure the correlation of the predicted values and the true values on the training set. The results were R^2 = 0.9162, Tau Kendell = 0.8001 and the p-value = 2.2882 e-16. As can be seen from the results, the model is reliable because the predicted values and the true values are highly correlated based on R^2 and Kendell. In addition, the Mean Absolute Error (MAE) were also computed for each model. MAE is also small for all three models which shows the models are working well. (MAE for SVR: 0.2733, MAE for KRR: 0.3618, MAE for RF: 0.3035)

The model will fail if the number of features gets larger and the number of examples in the dataset becomes smaller. In this scenario overfitting is hard to avoid and the model becomes so sensitive to the outliers.

4) What techniques did you use?

Three well-known regression models were used to predict S1: SVR, KRR and RF. These models were selected because they are non-linear models and hence they can handle the nonlinear relationships among the input variables and the output variable. The parameters of the models were optimized using the 5 fold CV on the training set. Later, the three models were compared in terms of Mean Absolute Error (MAE) and computational times using leave-one-out CV. Leave-one-out CV was used because the dataset has only 50 examples and is very small. It turned out that Support Vector Regression gives the best MAE, while KRR is the fastest method. As the final model, SVR was used to predict S1.