Analysis

(1) Which variables matter for predicting S1?

Variables Useful for predicting S1 were: S7, S5, S2, S3 and S6 (in decreasing order of relevance for predicting the S1).

The variables were selected based on 2 machine learning methods namely:

1. recursive feature elimination
2. random forest regressor.

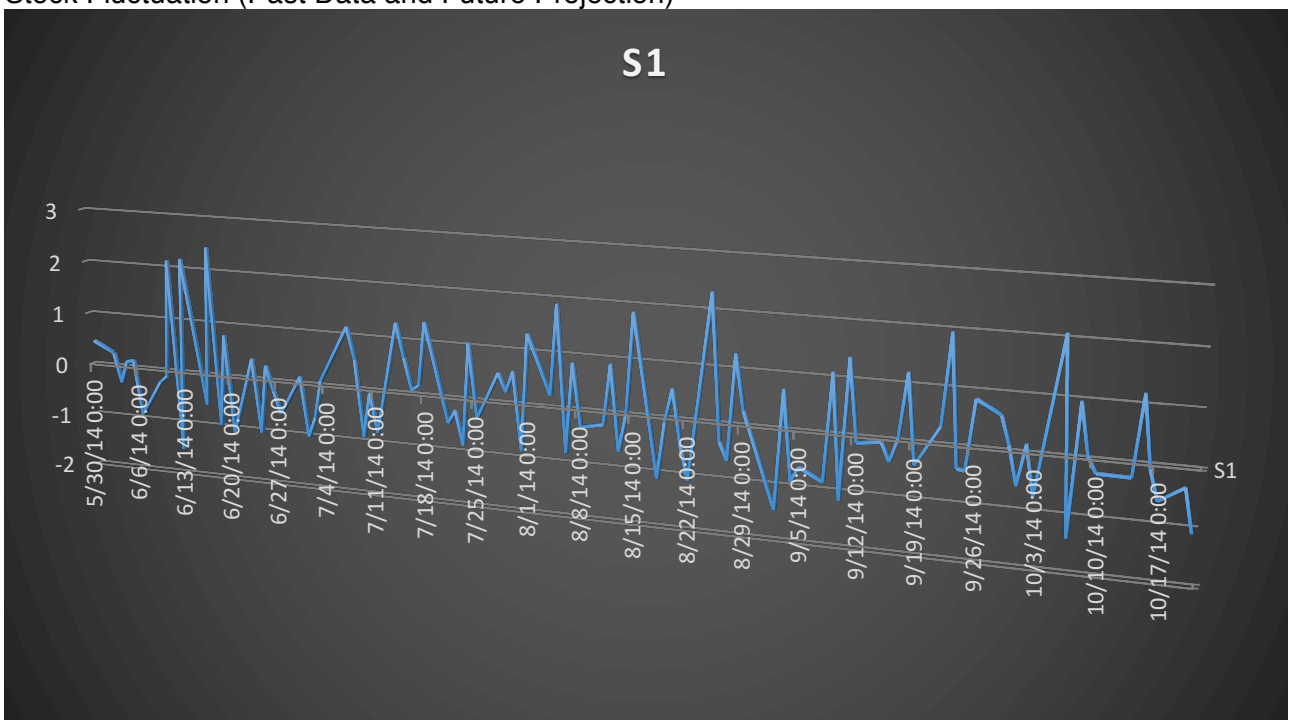
I confirmed selected features using the correlation matrix also, as the selected features were in correlation with S1 and also not in much correlation with other features. Features which are correlated among themselves are bad for models.

Along with feature selection algorithm mentioned above I also trained a model using regularization which automatically selects the best features for the task.

At the end I end up using regularization based method for producing output as it was generalizing well with cross-validation data set.

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

Stock Fluctuation (Past Data and Future Projection)



According to our analysis the stock price of S1 would go up by rate of 0.05 in the projected period.

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

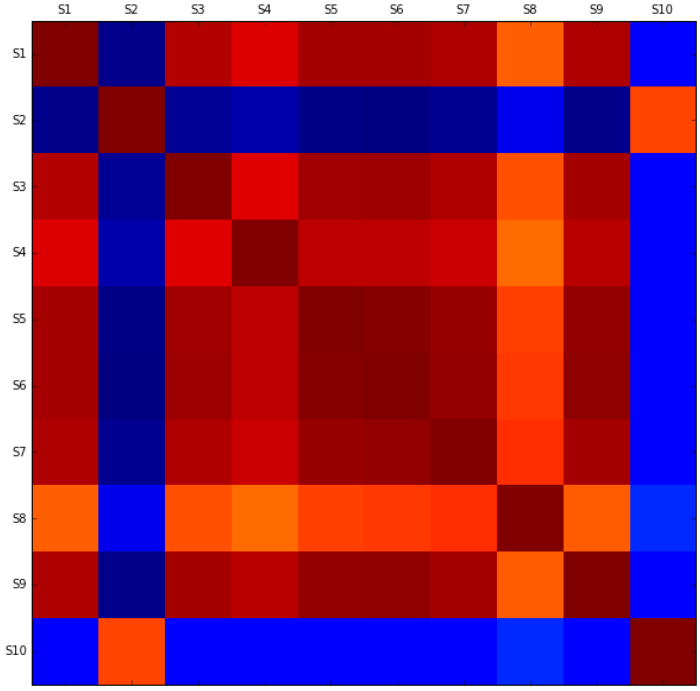
The model used for the projection of rate of change of stock is tested on the cross validation set (3-fold validation) and was generalizing pretty well on the cross-validation test. On our analysis 82% of times the model predicted the correct result. We did not factor in the sentiments of the stock used (S2-S10) for this projection. Hence a the model would fail if there is sudden change in the stock price due to some unexpected incidents at the parent company.

(4) What techniques did you use? Why?

- For this regression/prediction task we would be using Epsilon-Support Vector Regression (SVR) with an RBF Kernel. This model performs good with less training data, in our case we had dataset of 50 data points for training.
- For Parameter selection (for SVR) we tried various parameters using Grid Search function of Sci-kit learn, which allowed to find the best set of parameters for SVR.
- The features used for predicting the stock S1 were selected based on 2 machine learning methods namely:
  1. recursive feature elimination
  2. random forest regressor.

  We verified the selected features using the correlation matrix also, as the selected features were in correlation with S1 and also not in much correlation with other features. Features which are correlated among themselves are bad for models.

[Legend : Red -> Blue, decreasing order of correlation]



Along with feature selection algorithm mentioned above I also trained a model using regularization which automatically selects the best features for the task.

At the end I end up using regularization based method for producing output as it was generalizing well with cross-validation data set.

- Along with this report I have attached an Jupyter Python Notebook (compatible with python2.7) which can be used to reproduce the result. Dependencies for code would be : Scikit-learn, numpy, scipy, matplotlib, pandas and standard python2.7 libraries.