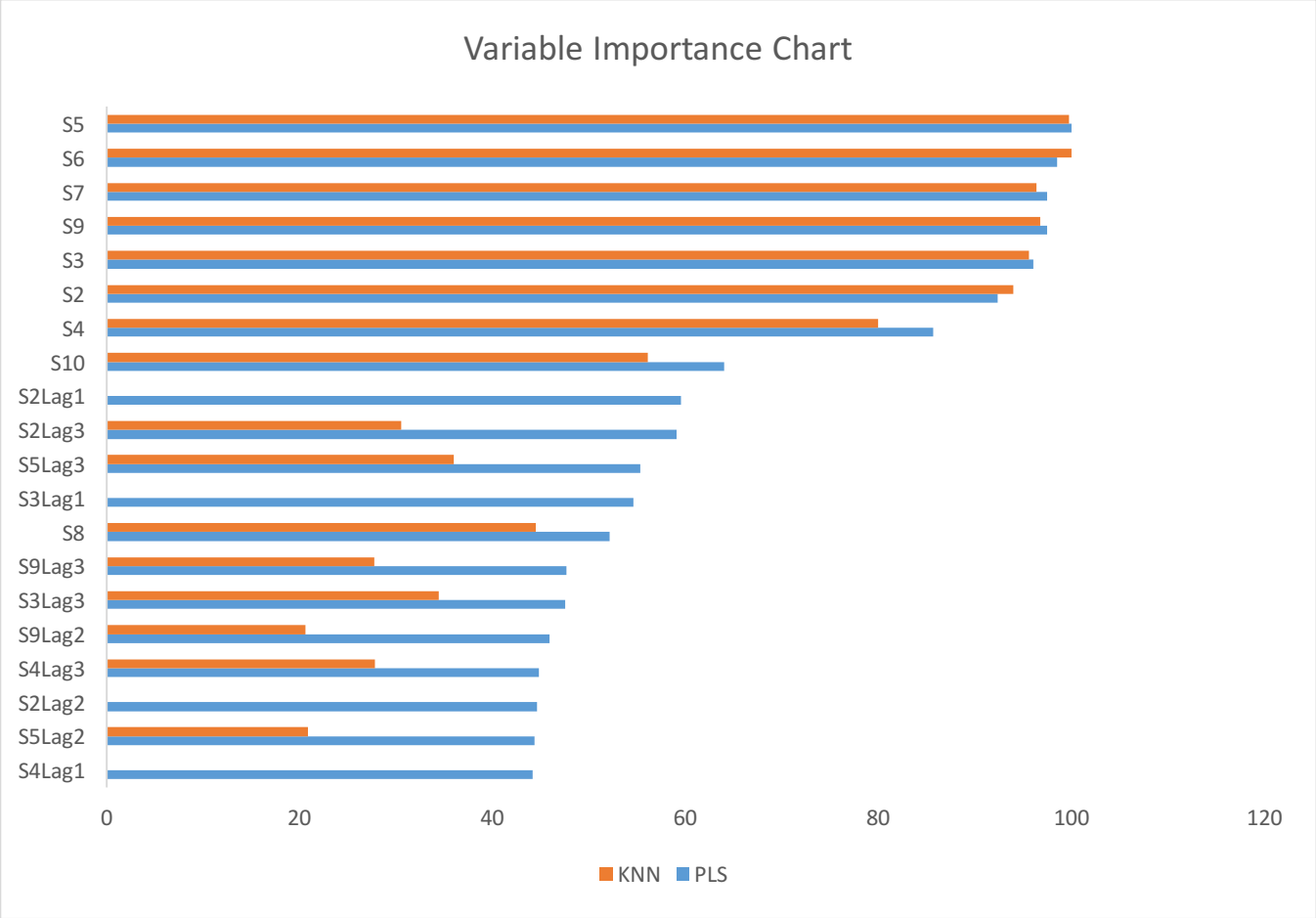
Correlation One: Machine Learning Challenge

Submitted by
Avinash Kamath

(1) Which variables matter for predicting S1?

We notice that from Variable Importance tables of all the 2 models: PLS & KNN are very close. The variable S5, S6, S7, S9 and S3 are the most important variables that can predict S1. Among the lagged variables, S2, S5 and S3 also seem be important based on these 2 models.

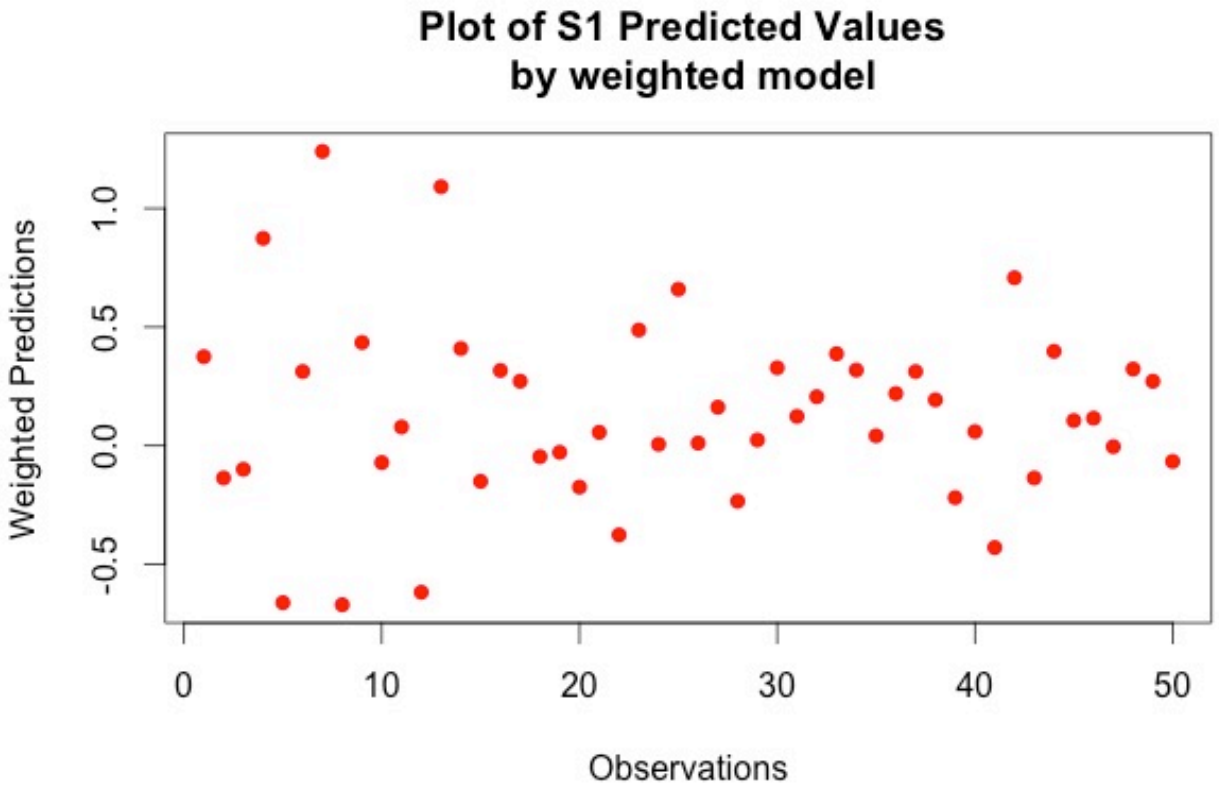
The summarized variable importance table given below:



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

Just as in the train period, over the test period too, S1 fluctuates up and down. There is no specific noticeable movement in S1 that can be generalized.

The graph below indicates the movement of S1 during the 50-day period:



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

The model has to do fairly well considering that the predictions are derived out of an ensemble model that is a weighted average of PLS, KNN & RF models. Averaging the predictions reduces over-fitting. The highest weightage is given to the RF model since it is less prone to over-fitting compared to the PLS which is a regression based model and KNN which is a non-parametric model.

Since the model is based on a rolling forecast method with a horizon of 1 day, the error in the forecasts may build up over period and may not be efficient in the long run over large forecasting periods. The given forecast period is only 50 days and with the weighted averaging method of forecasted values, the final model is expected to work fine.

(4) What techniques did you use? Why?

Rolling Forecast Approach:

Since it is a time series data, a rolling forecast was used to predict S1 values for each period. This was done by creating time slices from the Caret package. A fixed window rolling period for 15 days was used to predict the S1 price on the 16^{th} day for PLS model. For KNN and RF models, a non-fixed window rolling period was used to predict the S1 prices for each day.

This model, considers a different model to predict the S1 price on each day. The flaw in the model may be that the predictions are considered as true values for future forecasts. This may lead to building up of errors and is hence prone to faulty values in the long run.

Model Evaluation Metric:

Since the objective is to reduce sum of absolute deviations, a custom metric was built using the Caret Package so that the best model is chosen to be one that gives the smallest value of Sum of Absolute Deviations. Creating this custom metric aligns the objective with the required objective from the test problem.

Models Selection and weightage:

3 different models were used and a weighted average was taken between them to reach at the final prediction value.

Partial Least Squares model: PLS model form the 'pls' package generalizes and combines principal component analysis and multiple regression.
PLS model should work fairly well because of 2 reasons:

a. With addition of the lagged variables, a total of 36 independent variables were available for utilization over a 50 observations. A principal component regression works well when the number of predictors are higher or closer to the number of observations.
b. Since we have multiple lagged variables (S2 to S10 lagged were lagged by -1, -2 and -3 days), we expect the independent variables to exhibit multicollinearity. PLS models have been observed to work well when there exists multicollinearity among variables.

PLS model is weighted at 30% of the final forecast as it is expected to do much better than the KNN model but is very much prone to overfitting as compared to Random Forest model.

K- Nearest Neighbor model: KNN model is a non-parametric model that selects the best of the values dependent on how the independent variables were distributed during

the learning process. Since the stock market prices are well correlated among each other, their movements can be determined based on how the variables are located. KNN model can follow a deterministic approach of forecasting based on the historical prices among all the stock prices during the period.

KNN model is basically introduced in order to mitigate the risk of over-fitting. The weightage given to the KNN model is only 10% - enough to guide the prices towards the right direction based on the historical movement.

Random Forest model: Although random forest model is rarely used in a time-series model, it is expected to work well in a rolling forecast approach. Additionally, since the forecast horizon is just 1 day, the random forest model is expected to do considerable well out of the above to approaches. Random forest is also rarely prone to overfitting. Any risk of overfitting that arises due to the tuning of parameters is expected to mitigated when the predictions are weighted and averaged to determine the final forecasted S1 price.

As expected, from the SAE values, random forest works better out of three models and hence is awarded with the highest weightage of 60% among the models.

RF tuning result is given below: