Machine Learning Test

Joe Sieber
joeksieber@gmail.com
(405)850-0090
7/9/2016

Table of Contents

Introduction

This simulated data was provided by Correlation One and represents 10 stocks. These stocks are labeled as S1 through S10. The first stock, S1, is a US stock that is within the S&P 500. S2 through S10 represent stocks that are traded with the Nikkei Index. Each data point represents the daily open-to-close changes. The Nikkei traded stocks have 100 days worth of information and the S1 stock has 50 days of data. The goal is to create a model that will help predict the remaining 50 data points of S1.

Models and data exploration was completed using R. Other languages, such as Python, could have also been used which would result in similar outcomes.

4 questions specific questions were asked. They are:

1. Which variables matter for predicting S1?
2. Does S1 go up or down cumulatively (on an open-to-close basis) over this period?
3. How much confidence do you have in your model? Why and when would it fail?
4. What techniques did you use? Why?

The section heading indicates where each question is answered. However, to make the report easier to read and follow they were not answered in order.

Exploratory Analysis

Variables for Predicting S1 (Question 1)

The data was explored to determine if there were any immediate trends that could be seen. This helped to know where to focus effort. General statistics of the data was looked at first. Table 1 shows these results. The dates ranged from 5/30/2014 to 10/21/2014. Weekends were not included.

	S1	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**	**S9**	**S10**
Min	-1.538	-3.249	-3.860	-1.457	-4.659	-1.416	-3.619	-1.651	-2.892	-1.351
Mean	0.118	-0.251	0.352	0.297	0.453	0.161	0.278	-0.037	0.493	0.009
Max	2.436	1.953	5.201	3.320	7.276	2.503	5.576	1.273	4.989	1.326
Std dev.	0.930	1.063	1.880	1.137	2.801	0.903	2.013	0.641	1.975	0.645

Table 1 - Basic Statistics of Data

All of the features were also plotted against each other to help see any trends. Figure 1 shows this plot. Several features have a clearly visible linear trend compared against S1. Others, such as the date and S8 do not have as clear of a trend.

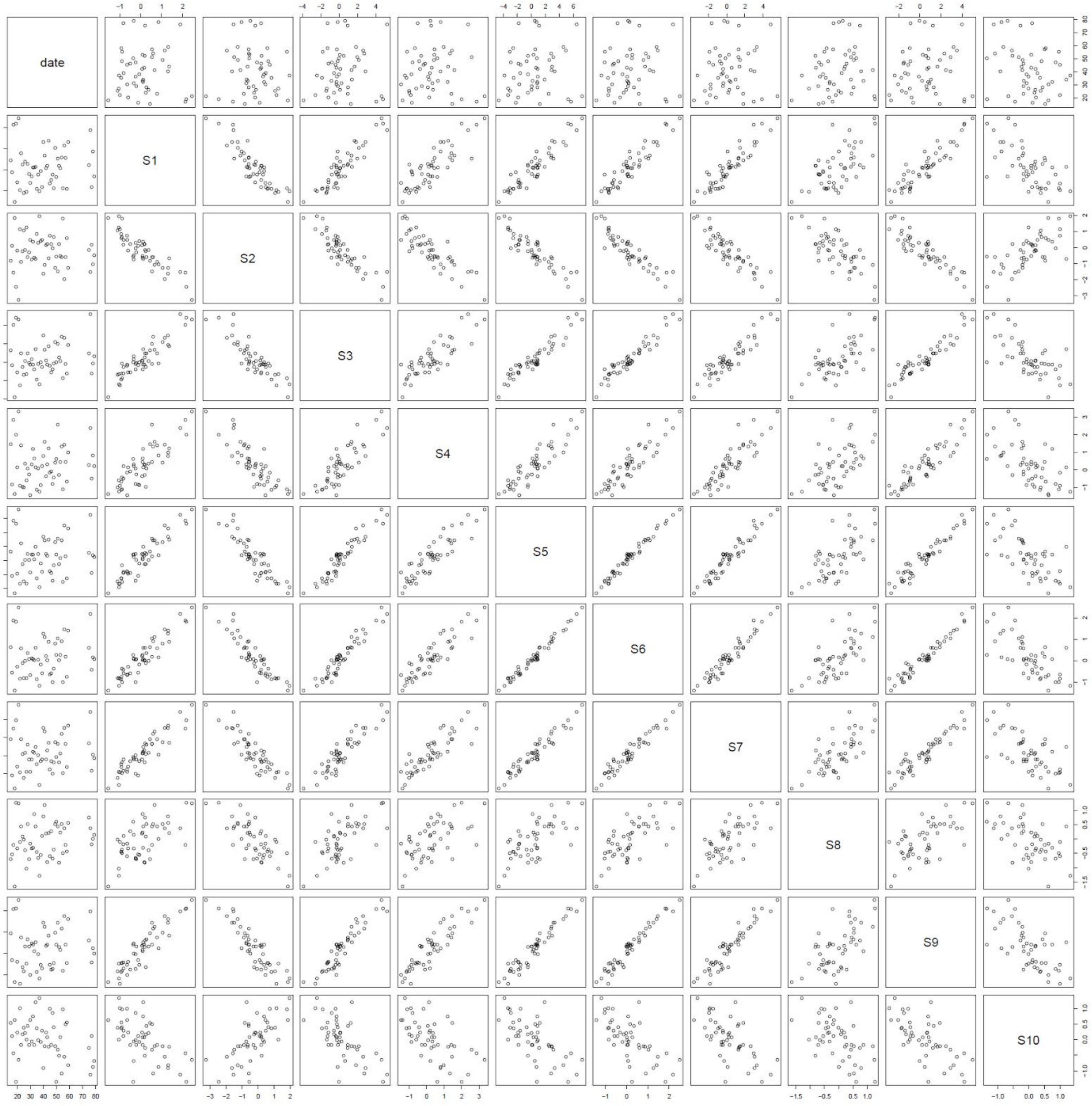

Figure 1 - Pairs Plot

Creating a basic linear regression model on the raw data, the importance of each feature can be determined. The top 5 most important features for this model in predicting S1 are shown in Figure 2. For linear regression, the absolute value of the t-statistic for each feature is used to determine which features are important. The final model might have different features which it finds important. This will be discussed later with the final model.

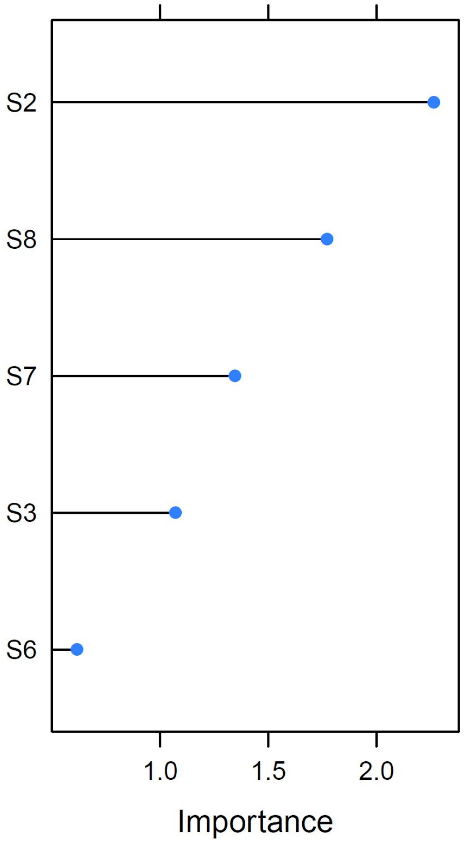

Figure 2 - Linear Regression Variable Importance for Predicting S1

S1 Over Time (Question 2)

The data represents the change from each day. Figure 3 shows what the change was for the first 50 days. There does not seem to be a trend over time for the daily changes.

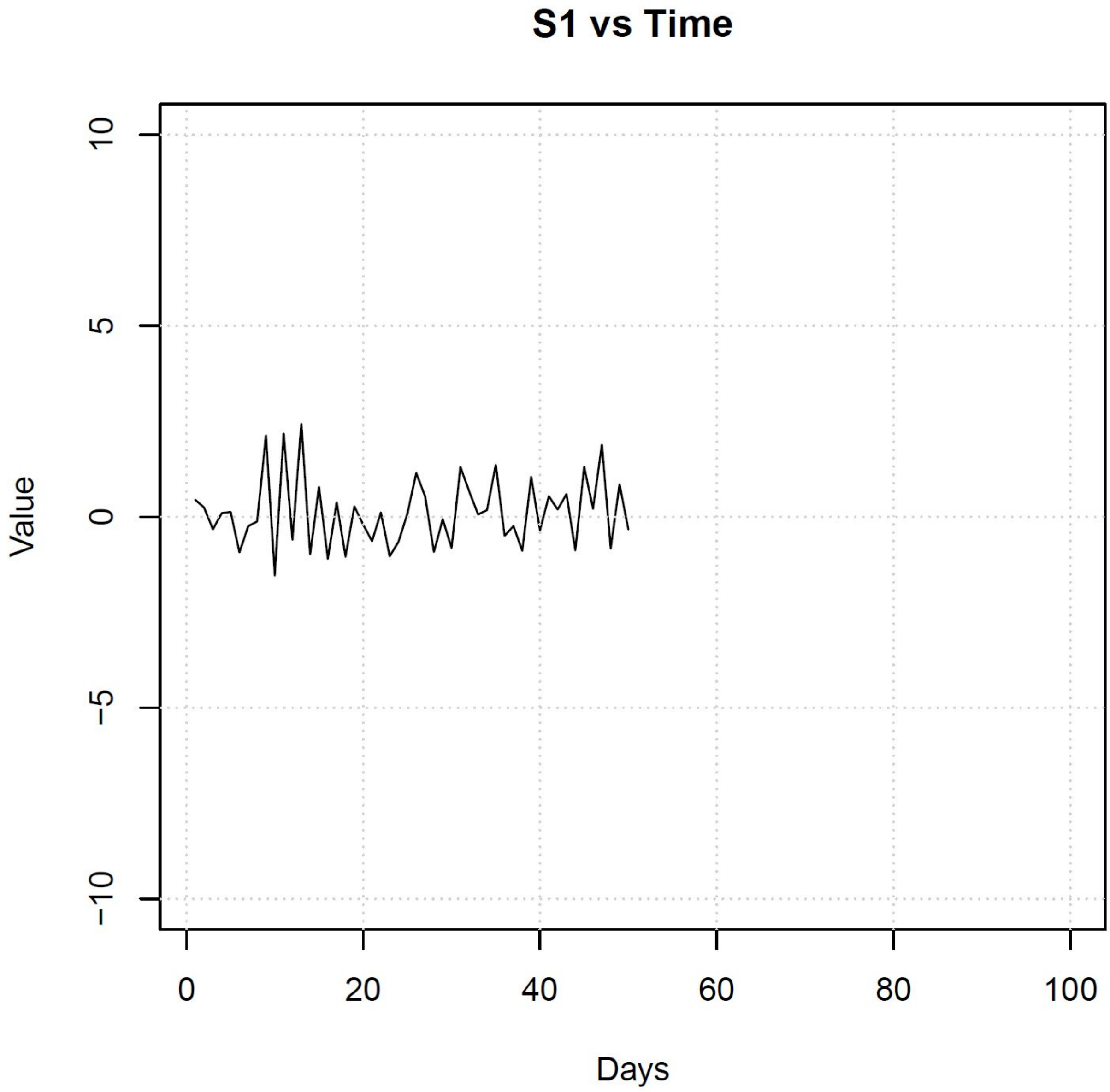

Figure 3 - S1 vs Time

Having the starting price of each stock start at 1 and then taking the daily change to determine the cumulative change is shown in Figure 4. All 100 days are plotted for S2-S10 and 50 days for S1. Several of the stocks do seem to trend together. This is expected since there appeared to be such a strong correlation shown in the pairs plot, Figure 1. S1 does fluctuate over the 50 days but the general trend is up.

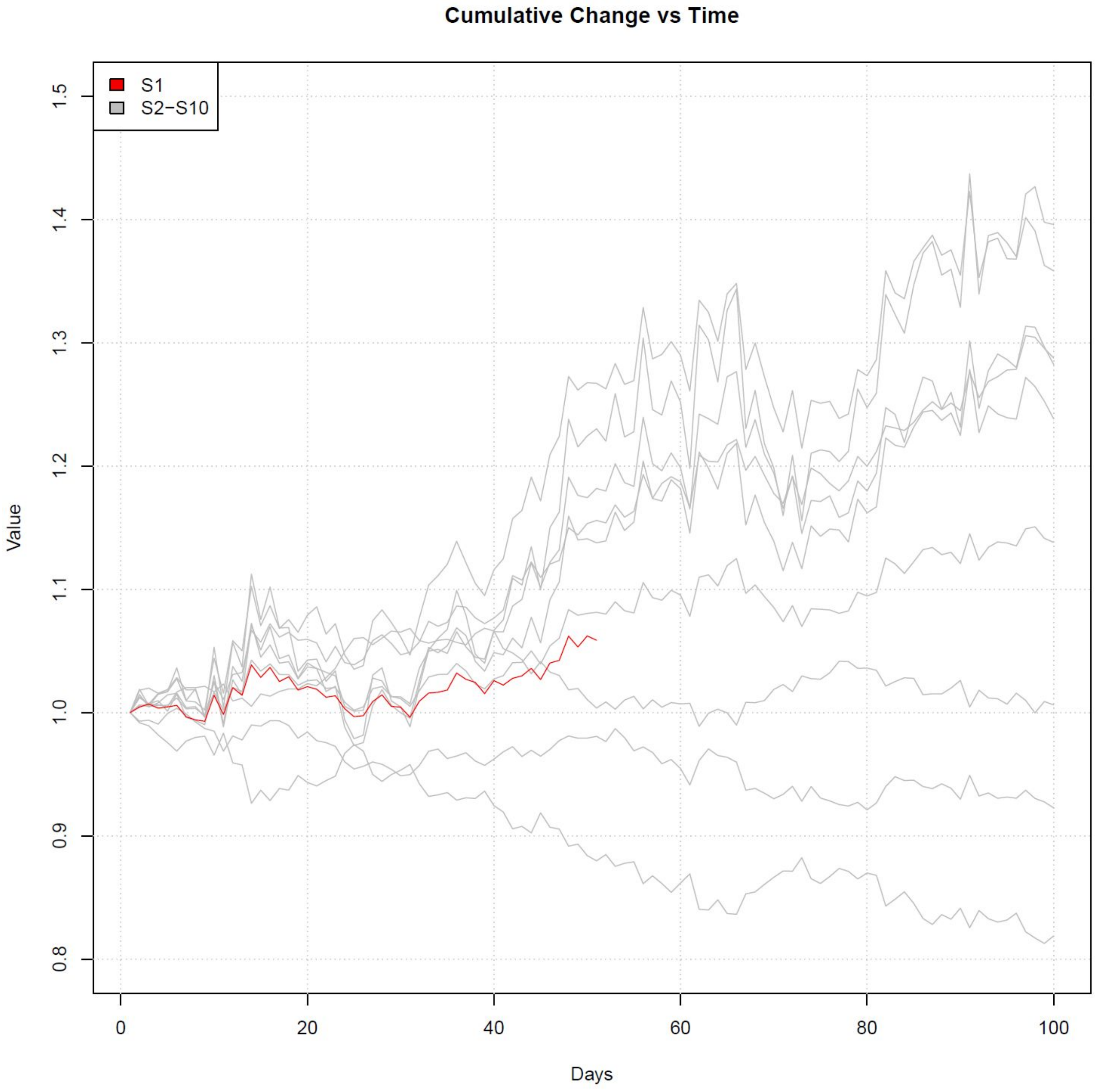

Figure 4 - Cumulative Change vs Time

Creating Final the Model

Techniques Used (Question 4)

Three different models were used. These models were linear regression, random forest, and gradient boosting. Since there are only 50 data points used, overfitting the model has to be considered. These three model types were selected because they are, in general, less likely to overfit the data compared to other models. Three different data sets were created to train these three different models. The first dataset was to just use the raw data without the dates. That dataset was 50 rows with 9 features for S1 The second was to use the raw data with its previous two history points. This resulted in a dataset that was 48 rows and 29 features for S1. The final dataset was to scale each stock between -1 and 1. This resulted in a dataset that was 50 rows and 9 features for S1. Table 2 shows an overview of these datasets.

	Dataset	Training Examples (Rows)	Features
1	Raw Data	50	9
2	With History	48	29
3	Scaled	50	9

Table 2 - Training Datasets

Dataset 1 was used to create a quick benchmark. Since there was already a trend seen in the exploration of the data these results might be enough to generate a reasonable performing model. Dataset 2 was used because the previous data points might help predict the current value of S1. Dataset 3 was used because some of the features might be influencing the models too strongly and not allowing the smaller features to be represented.

For gradient boosting and the random forest, a grid search was used to find the best model. The best model was determined by the R^2 value. With larger datasets, the data would be split into a training set, a cross validation set, and a test set. This small dataset was already split into a training set and a test set. With only 50 points it was not desirable to split that into a training set and a cross validation set. Repeated cross validation was used in order to estimate the model's performance on out of sample data.

Final Model

With the 3 models and the 3 datasets, 9 models, excluding the models generated in the grid search, were considered. Table 3 shows the results of each of these models. The best model will be selected based on the lowest RMSE value.

Model	Dataset	R^2	RMSE
LM	1	0.869	0.352
GBM	1	0.901	0.692
RF	1	0.893	0.322
LM	2	0.785	0.519
GBM	2	0.896	0.661
RF	2	0.895	0.343
LM	3	0.891	0.177
GBM	3	0.859	0.235
RF	3	0.912	0.164

Table 3 - Model Results

The random forest model using the 3rd dataset will be selected to make the final predictions. The variables that this model find important are shown in Figure 5. These are different from the linear regression because it is being measure differently, it is a different model and might find things that the linear regression did not, and the final dataset was different.

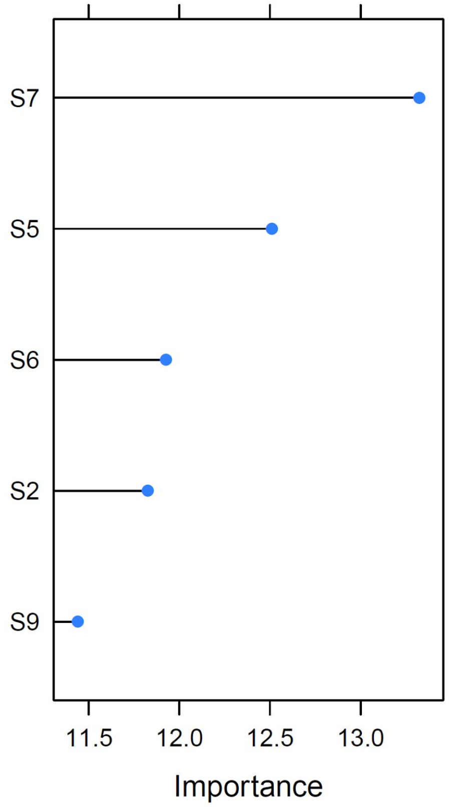

Figure 5 - Final Model Variable Importance for Predicting S1

Confidence in the Model (Question 3)

The 95% confidence interval, CI, was predicted for each predicted value using the final model. The mean CI is -0.166 with a standard deviation of 0.458. The smallest CI was 0.028 on the 10/20/2014 prediction for and the largest was 0.869 on the 8/22/2014 prediction.

Possible Future Improvements

Given more time, the model could be refined by looking into the dataset and determining if it really is simulated or not. The S&P 500 could be scanned over various time frames from historical data to see if it matches this dataset. If nothing is found to match the S1 dataset, then it could be determined that is really is simulated and further research could be focused on how the dataset was constructed. Even if randomness was put into the simulated data, that method could be determined given more time.

More research could also be done into finding out better methods for using the historical data provided for that stock to make predictions. Models were tried, with limited success, using the previous 2 data points to help better predict S1. Researching how the previous data points affect the S1 value and creating new features, feature engineering, would be additional steps.

Gathering more data will also help any model perform better. Even the worst performing linear regression model would most likely outperform the best model with 10 times more data. If more data was available then more advanced machine learning techniques could be used to possibly find deeper insight into the data.

The code can also be optimized with more vectorization and creating functions to streamline the code. This would results in some of the code that is called multiple times to be condensed down to just a function call.

Appendix A - Additional Resources

Github page for all code, plots, and report:

https://github.com/J-Sieber/CorrelationOne-MachineLearning