Correlation One: Machine Learning Challenge

Siddharth Chakravarty

Overview

Problem statement

Approaches

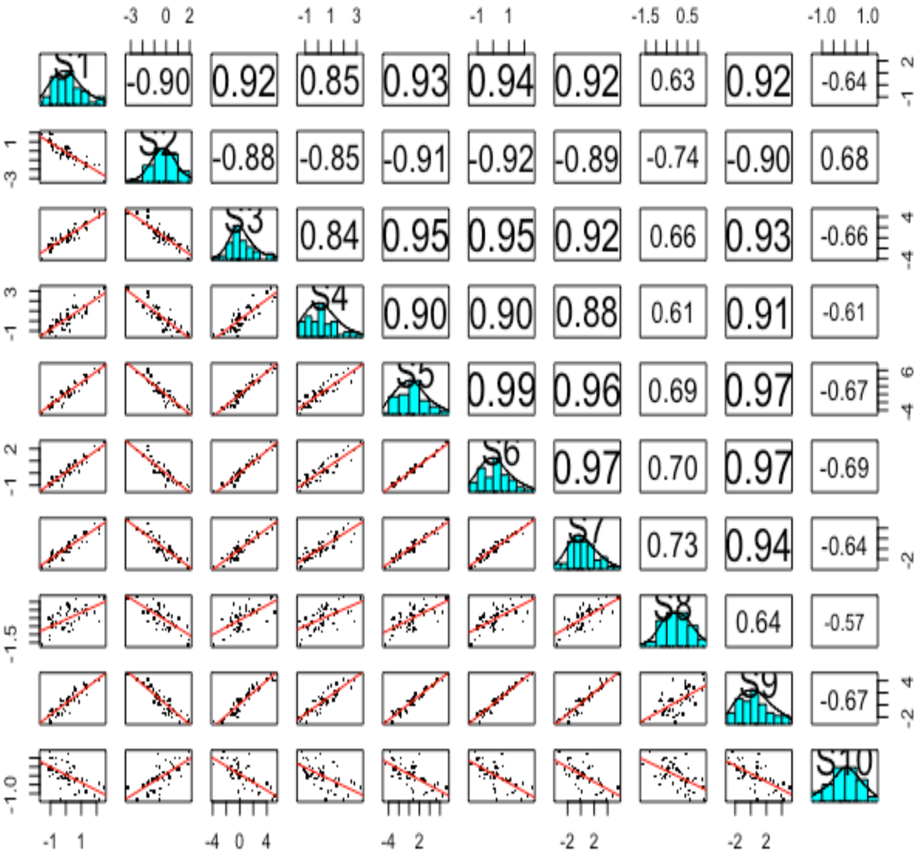
Model performance

Result and Summary

Problem statement

- **Predict performance of stock S1 that trades in US as part of S&P500 as function of S2….S10 stocks that trade on Nikkei.**
- There are three ways to predict stock performance
  - Fundamental approach -examines the economic factors that drive the price of stock
  - Second approach - traditional technical analysis to anticipate what others are thinking based on the price and volume of the stock.
  - Quantitative technical analysis -is the third approach to predicting market direction
- The quantitative approach uses several methods for predicting stock performance, depending upon the position of the equity, type of equity and method trading- Long, Short, High Frequency, Options, Stock or Commodity, etc.
- Several methods are available – Regression, Classification, Vector machines, Neural networks, KNN and other methods are constantly evolving.
- But no one model suffices the complexity of the problem, and typical quant fund will use a combination of algorithms to predict stock prices.
- For this exercise the third approach is used to predict S1 performance. Two machine learning algorithms are used evaluated– Regression and SVM to predict S1.

Assessing correlation and co-linearity



	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	1	-0.9	0.92	0.85	0.93	0.94	0.92	0.63	0.92	-0.64
S2	-0.9	1	-0.88	-0.85	-0.91	-0.92	-0.89	-0.74	-0.9	0.68
S3	0.92	-0.88	1	0.84	0.95	0.95	0.92	0.66	0.93	-0.66
S4	0.85	-0.85	0.84	1	0.9	0.9	0.88	0.61	0.91	-0.61
S5	0.93	-0.91	0.95	0.9	1	0.99	0.96	0.69	0.97	-0.67
S6	0.94	-0.92	0.95	0.9	0.99	1	0.97	0.7	0.97	-0.69
S7	0.92	-0.89	0.92	0.88	0.96	0.97	1	0.73	0.94	-0.64
S8	0.63	-0.74	0.66	0.61	0.69	0.7	0.73	1	0.64	-0.57
S9	0.92	-0.9	0.93	0.91	0.97	0.97	0.94	0.64	1	-0.67
S10	-0.64	0.68	-0.66	-0.61	-0.67	-0.69	-0.64	-0.57	-0.67	1

- The pair-wise correlation clearly highlights a strong correlation among the dependent and independent variables.
- In addition there is a strong correlation between the independent variables, that could lead to co-linearity.

- The correlation matrix clearly highlights the correlation among the variables
- The next slide will focus on determining the independent variables required to build the regression model.

Comparing various LM models

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-0.07	0.05	-1.33	0.19
S2	-0.27	0.12	-2.26	0.03*
S3	0.09	0.08	1.07	0.29
S4	-0.04	0.10	-0.43	0.67
S5	0.05	0.11	0.50	0.62
S6	0.27	0.43	0.62	0.54
S7	0.12	0.09	1.35	0.19
S8	-0.21	0.12	-1.77	0.08 .
S9	-0.02	0.11	-0.15	0.88
S10	0.02	0.10	0.20	0.84

LM Model 1

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-0.10	0.06	-1.50	0.14
S2	-0.84	0.08	-10.02	3.20E-13***
S3	-0.12	0.14	-0.87	0.39

LM Model 2

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-0.07	0.05	-1.33	0.19
S2	-0.30	0.10	-2.95	0.005**
S3	0.15	0.06	2.37	0.02*
S7	0.20	0.06	3.18	0.002**
S8	-0.21	0.11	-2.03	0.04*

LM Model 3

- Comparing various LM model confirms that not all independent variables are required to predict S1.
- LM model 3 yields the best performance balance - between Multiple R^2 and Adjusted R^2 values with all least number of independent variables required to predict S1

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	-0.07	0.05	-1.59	0.12
S2	-0.25	0.11	-2.33	0.02*
S3	0.09	0.08	1.24	0.22
S6	0.32	0.26	1.25	0.22
S7	0.13	0.09	1.53	0.13
S8	-0.20	0.11	-1.94	0.06

LM Model 3

Independent variables	Multiple R-squared	Adjusted R-squared
S2 – S10	0.9048	0.8834
S2, S8	0.8064	0.7982
S2, S3,S7 and S8	0.9002	0.8913
S2, S3,S6,S7 and S8	0.9036	0.8926

LM model performance

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

ANOVA

- ANOVA is used to verify the model selection, by comparing LM objects for two nested models.
- The variance tables below highlights the results from the ANOVA comparison

```
Analysis of Variance Table 1

Model 1: S1 ~ S2 + S3 + S4 + S5 + S6 + S7 + S8 +
S9 + S10
Model 2: S1 ~ S2 + S8
        Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1      40 4.0328
47 8.2034 -7   -4.1706 5.9095 9.326e-05 ***
```

```
Analysis of Variance Table 2

Model 2: S1 ~ S2 + S8
Model 3: S1 ~ S2 + S3 + S7 + S8
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1      47 8.2034
2      45 4.2318  2    3.9716 21.116 3.404e-07
***
```

```
Analysis of Variance Table 3

Model 3: S1 ~ S2 + S3 + S7 + S8
Model 4: S1 ~ S2 + S3 + S6 + S7 + S8
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1      45 4.2318
2      44 4.0859  1   0.14588 1.571 0.2167
```

These ANOVA tables indicate that model 3 is the preferred model to use .

Predicting S1 values using LM model

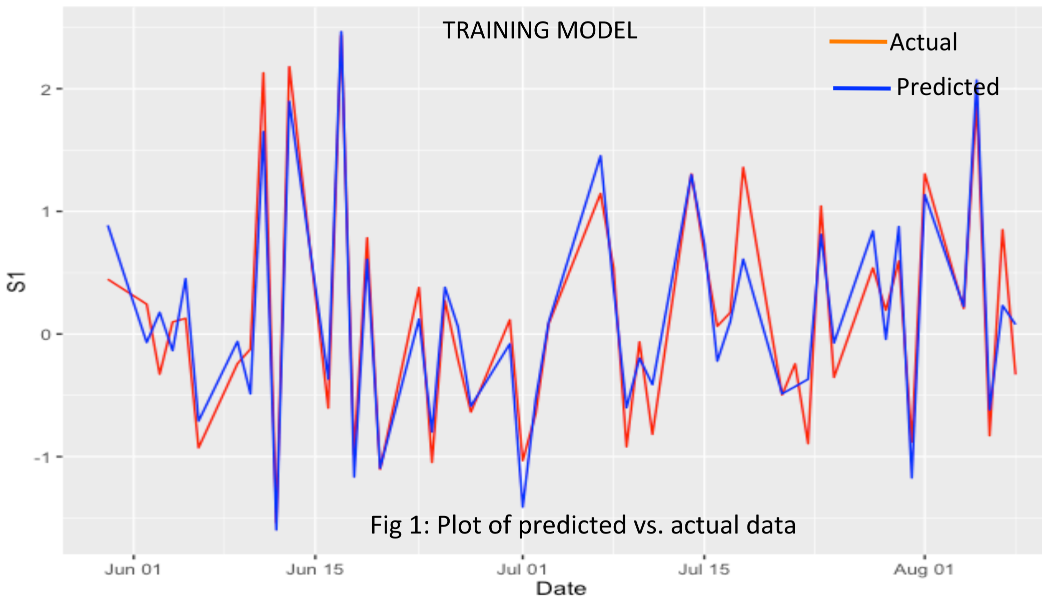


Fig 1: Plot of predicted vs. actual data

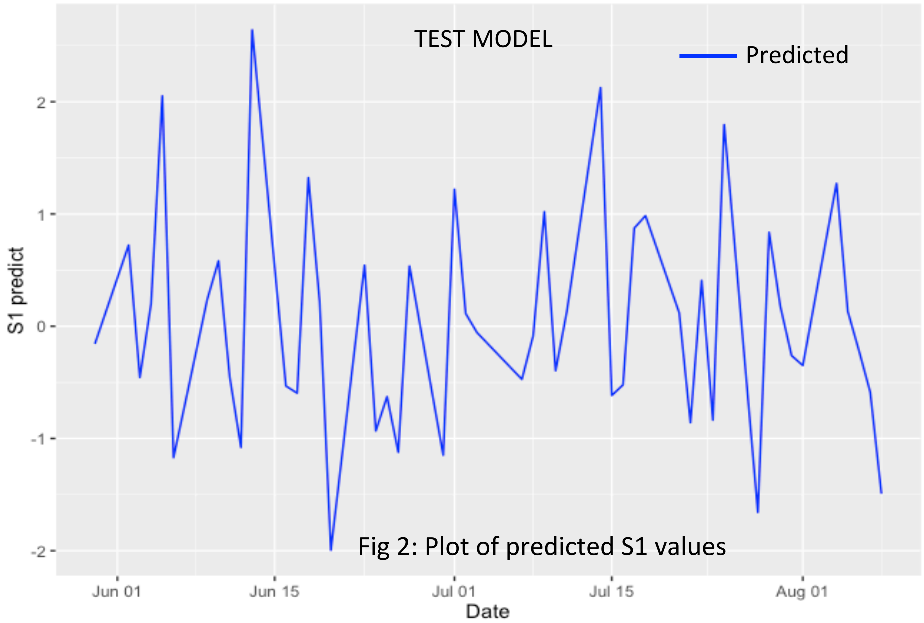


Fig 2: Plot of predicted S1 values

- The training data is first used to test the LM model.
- Fig 1, overall the model shows a good fit between the actual and predicted value.
- The R^2 and RMSE value for LM model are 0.90 and 0.30 respectively
- Output of LM model 3 for test data is shown in fig 2
- **Based on the results, the LM model unsuitable for predicting S1.**

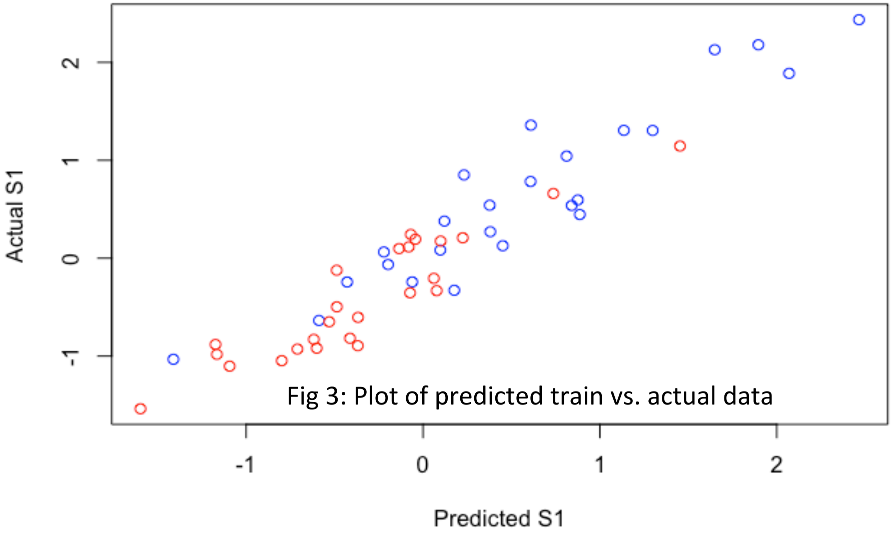
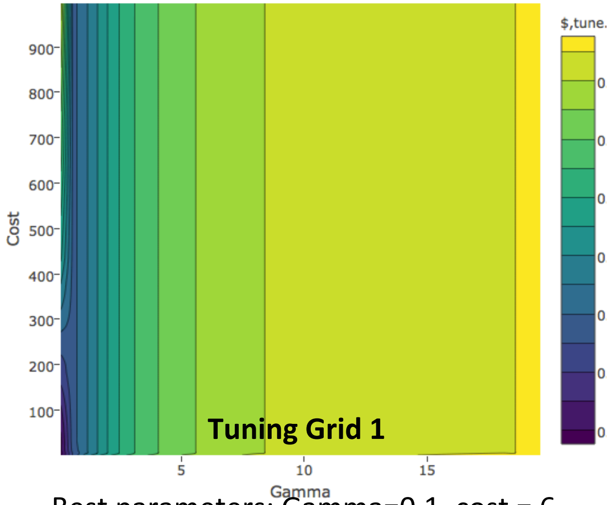


Fig 3: Plot of predicted train vs. actual data

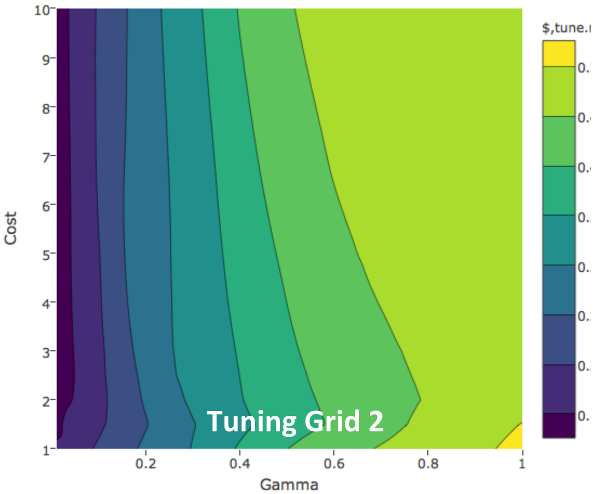
SVM model

- SVM is used to train a support vector machine MODEL. SVM models are used to carry out general regression and classification problems (of nu and epsilon-type).
- The accuracy of a SVM model relies on careful selection of several factors, mainly
  - Model Type
  - Cost
  - Gamma
  - Kernel type
- Of the several classification and regression models available(see below), the eps- regression type is chosen for this problem
  - C-classification
  - nu-classification
  - one-classification (for novelty detection)
  - eps-regression
  - nu-regression
- The tune.svm function is used to select to values of Cost and Gamma.
- Of the various kernel types, the RBF kernel is chosen for this model
- The following slides will discuss the steps and results for building the model.
- The model is built using the e1071 SVM package in R

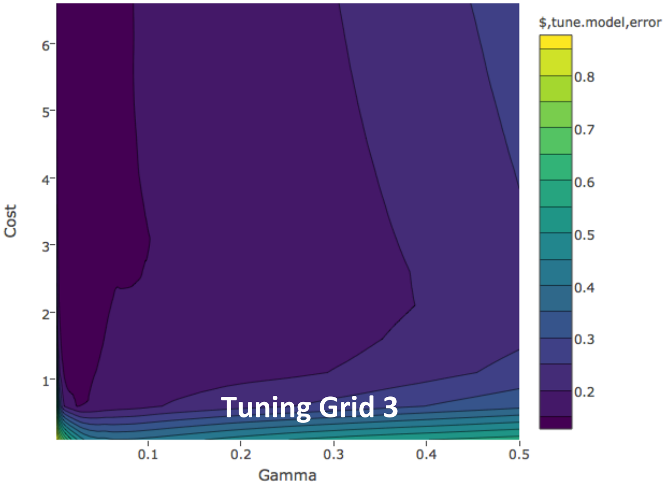
SVM model tuning for cost and gamma



Best parameters: Gamma=0.1, cost = 6
Best performance = 0.817



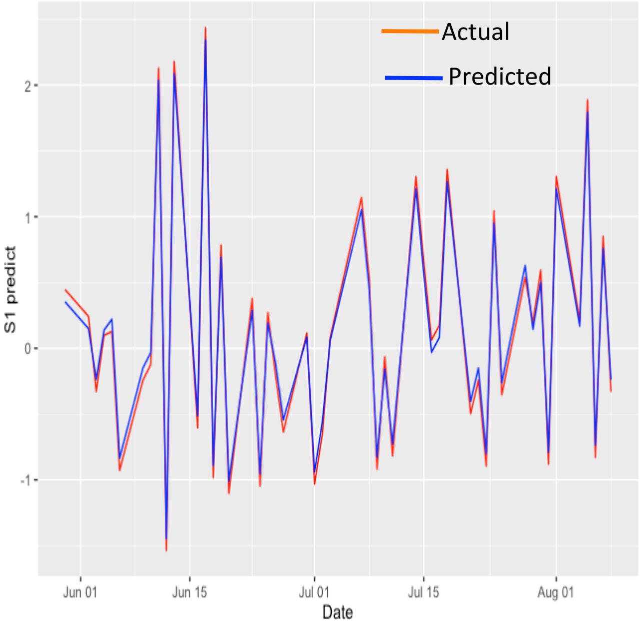
Best parameters: Gamma=0.01, cost = 5
Best performance = 0.881



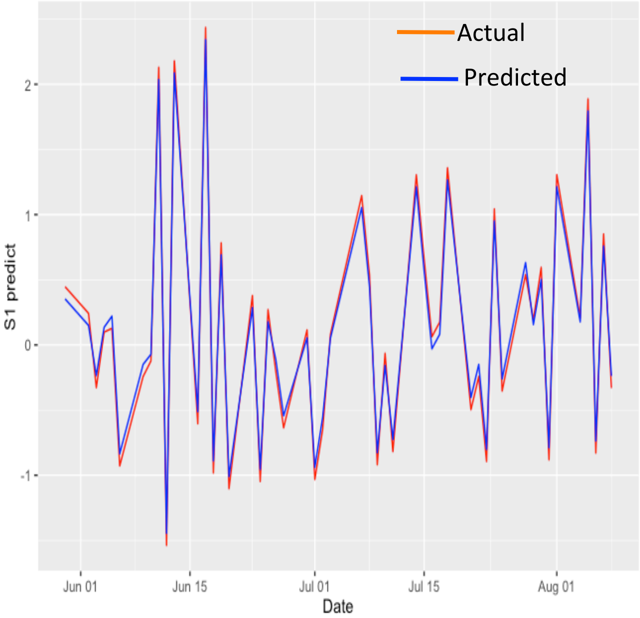
Best parameters: Gamma=0.004, cost = 4.6
Best performance = 0.90

- Cost (C) and Gamma (γ) parameters are the two key parameters for the SVM model
- Initial values of cost and gamma for the SVM model are estimated using the “tune function”, and running a coarse grid search.
- Following the initial coarse grid search, the value of cost and gamma were fine tuned
- The tune function implemented uses a 10 fold cross validation
- Parameter tuning of ‘svm’:
  - Sampling method: 10-fold cross validation
  - Best parameters are determined by plugging in the values for cost and gamma in the training model.

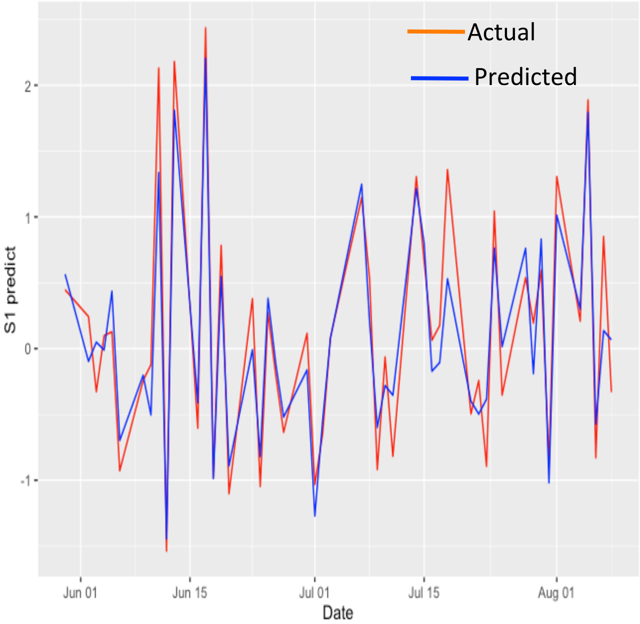
SVM model performance



SVM Model 1: Plot of predicted vs. actual data



SVM Model 2: Plot of predicted vs. actual data



SVM Model 3: Plot of predicted vs. actual data

- Three training models were built using the cost and gamma values obtained from grid tuning
- The performance of model 1 and 2 is quite identical, but model 3 clearly shows deviation of predicted and actual data.
- The final model is chosen by comparing the R^2 and RMSE values, and model performance from grid search.
- **Model 2 is chosen for the final SVM model**

	RMSE	R^2
SVM Model 1	0.08869541	0.9959395
SVM Model 2	0.08853574	0.9959312
SVM Model 3	0.3137158	0.8939649

Summary

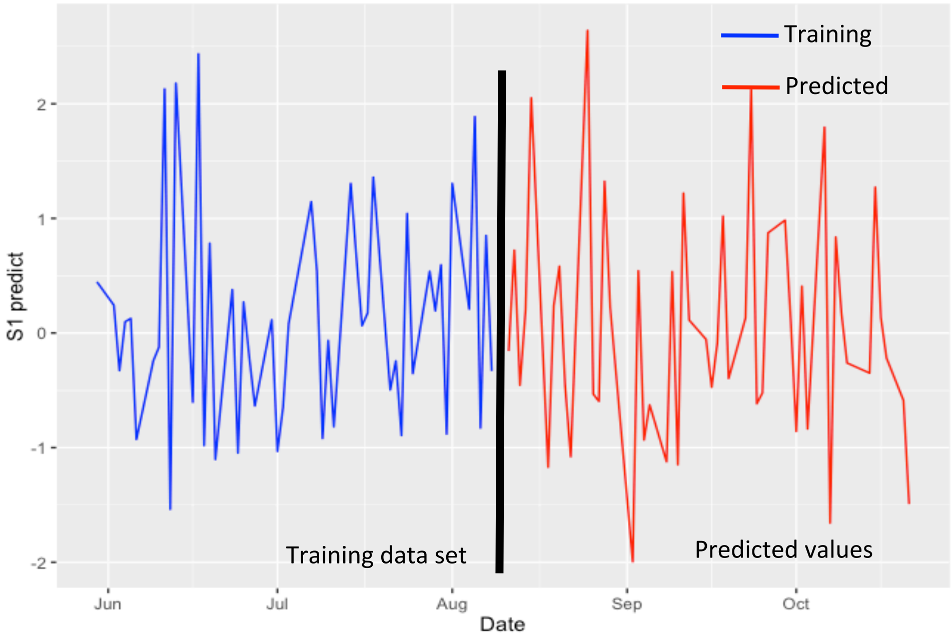


Fig 1: Plot of predicted S1 values using LM model

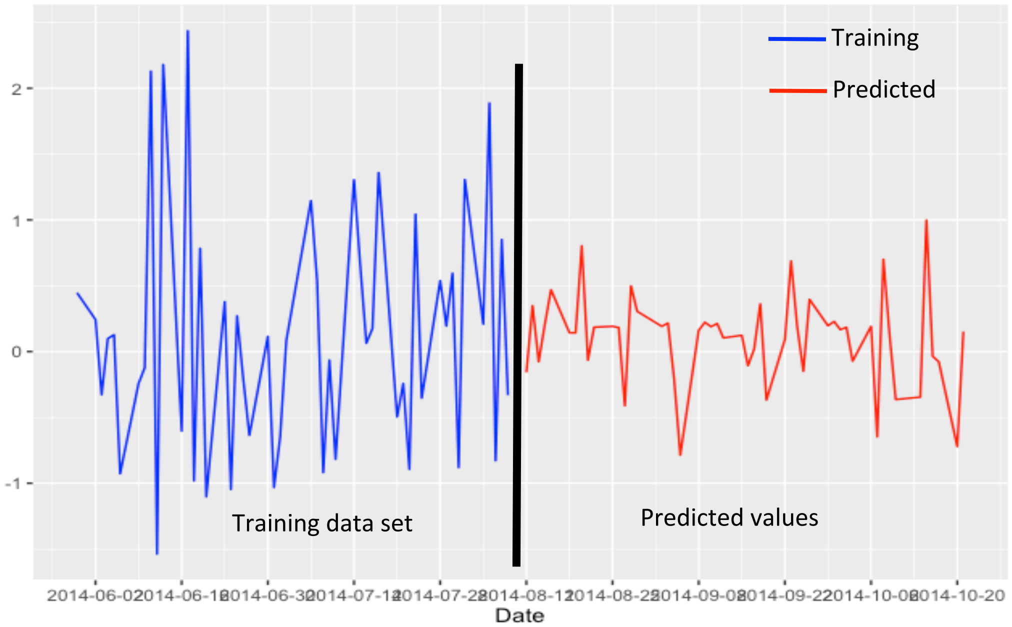


Fig 2: Plot of predicted S1 values using SVM model 2

- Both models predicts very different values for S1.
- The LM model predicts a high volatility in price change with S1 ending on a low note.
- The SVM model predicts reduced volatility and a upward trend for S1.
- **Based on the model performance, R^2 and RMSE values for LM and SVM models, SVM model 2 is used to predict S1**

	RMSE	R^2
LM	0.30	0.90
SVM Model 1	0.08869541	0.9959395
SVM Model 2	0.08853574	0.9959312
SVM Model 3	0.3137158	0.8939649

Questions

1. Which variables matter for predicting S1?(refer to slide 4 and 5)
   - Independent variables S2,S3, S7 and S8 are the most important for predicting S1. They not only have strong correlation with S1, but also have the least degree of co-linearity among them. In addition, the P-value for the LM shows all factors having high degree of significance in predicting the value of S1.
2. Does S1 go up or down cumulatively (on an open-to-close basis) over this period?
   - S1 will go up 11.25% over this period
3. How much confidence do you have in your model? Why and when would it fail?
   - With the given data set, the SVM model should perform well. But it may not perform as well with larger “out of sample” data set, primarily due to the limited size of the training data set used to train the model. While this approach may work in research setting, a field application the SVM model alone won’t suffice. In addition to the quantitative data, market sentiment governed by “qualitative” data also need to be factored in. A more realistic model would be a voting machine learning algorithm, rather than an absolute one.
4. What techniques did you use? Why?
   - I first assessed the correlation and co-linearity among the independent variables and dependent variable S1. This was to primarily reduce overfitting of the model and choose independent variables that contributed most to predicting S1. For the machine learning learning algorithm, I choose a linear regression (LM) and Support vector machine (SVM) models. Both approaches have their pro’ and con’. While LM models are easier to implement, they may work very well in certain circumstances, and if the training set is small it tends to overfit the model. SVM on the other hand is very versatile, and can be used for both classification as well as regression problems. But, it requires several iterations' before a model can be built. In addition as with LM, SVM too can lead to overfitting if the wrong values of cost and gamma are chosen.
   - But overall SVM tends to give better results for most non-linear problems and is the preferred model for many prediction ML algorithms.