Stock_Prediction

DMwR MSBA

9/12/2019

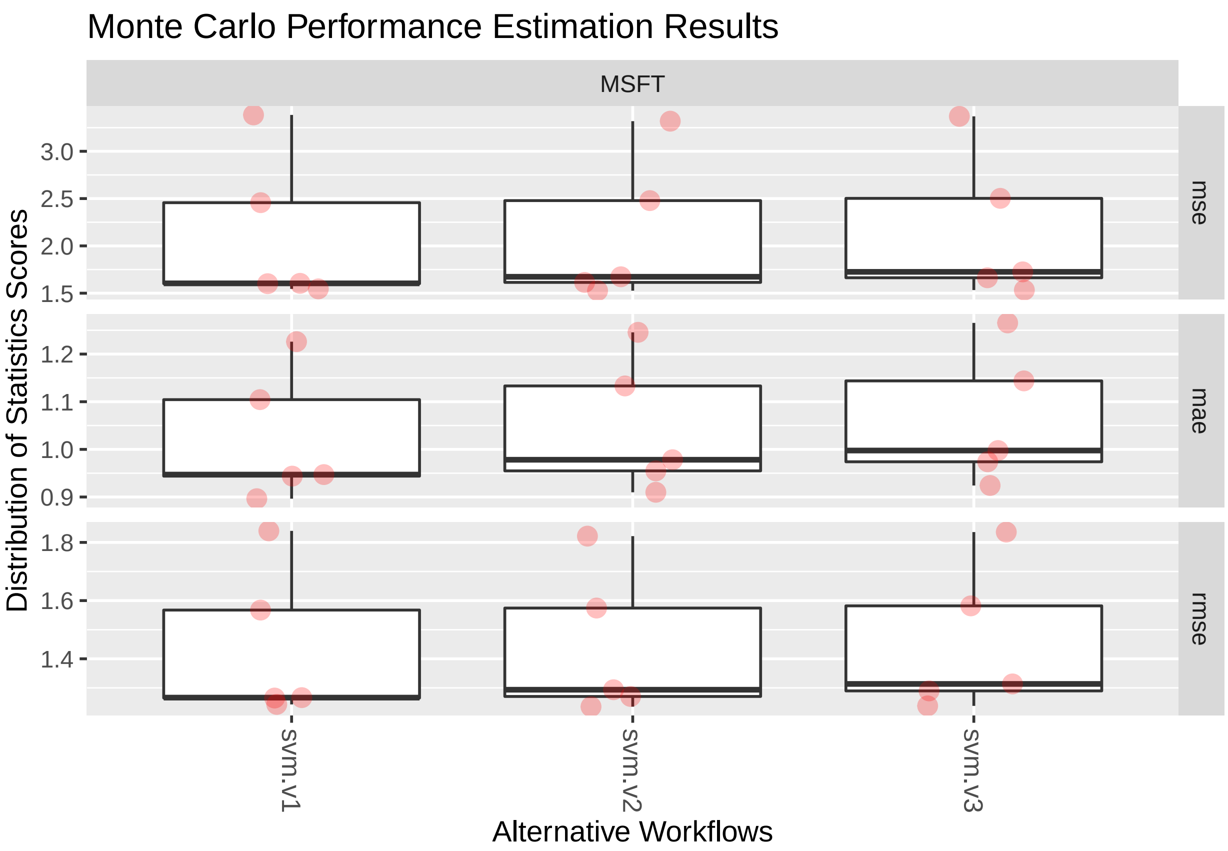
Assignment Objective

For this Predictive Portion, we have considered, Microsoft (MSFT) as a sample stock and we have considered Support Vector Machine (SVM), Multiple Linear Regression (LM) and Multivariate Adaptive Regression Splines (MARS) as predictive models. We have tried using other models such Random Forest and Gradient Boosting Model but eventually did not consider them because of their low prediction power as evident from the Actual Vs. Predicted chart that we have developed as a part of this prediction exercise. Also the stock closing price (which we are trying to predict) is highly nonlinear in nature and SVM and MARS are able to handle nonlinear nature of the underlying data. MARS is a non-parametric regression technique that models nonlinearities and interactions between variables so we think it is effective in predicting stock prices which is nonlinear and depends on multiple market factor interactions.

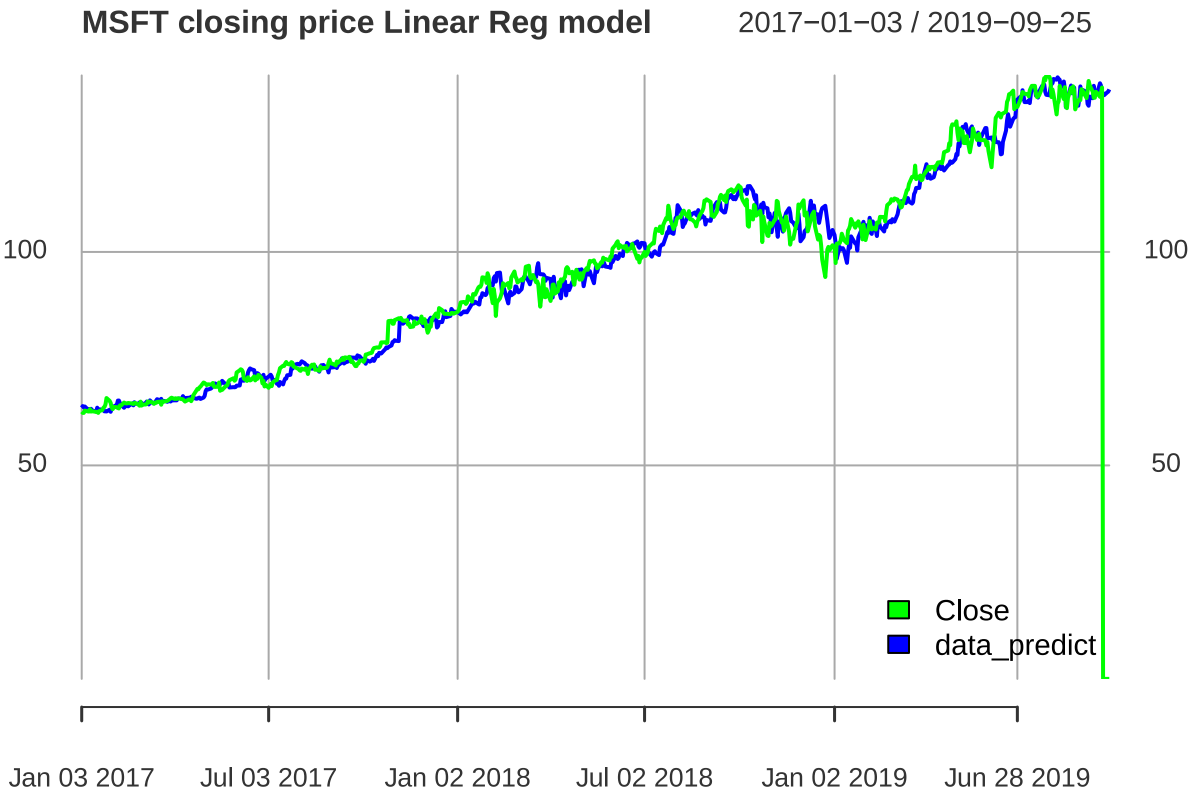
Pre Processing Steps

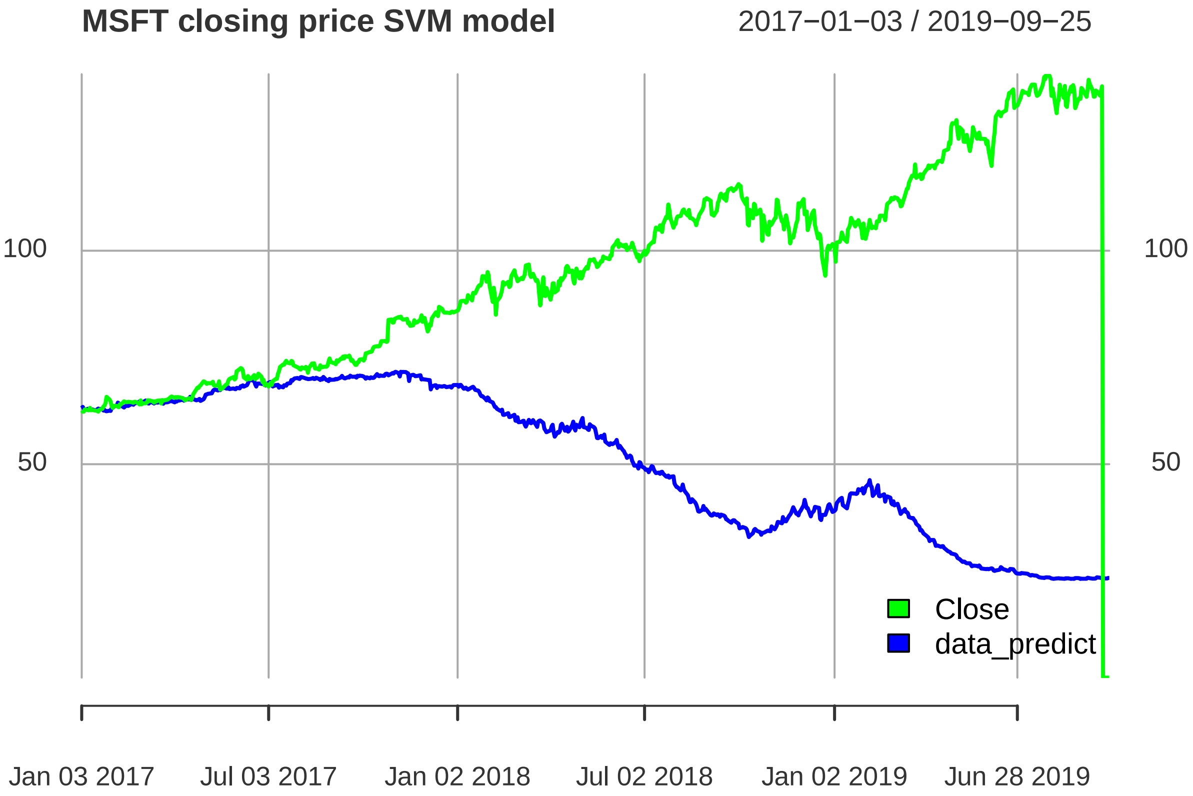
The Stock Detail has the following information for each trading day: Open Price, Close Price, High Price, Low Price, Volume and Adjusted. Since it is a time series data we suspect that there would auto correlation. To minimize the effect of auto correlation, we have introduced lag variables for last 7, 14, 21 and 28 days. Note we are trying to predict the closing price of the stock for the next 5 trading days.

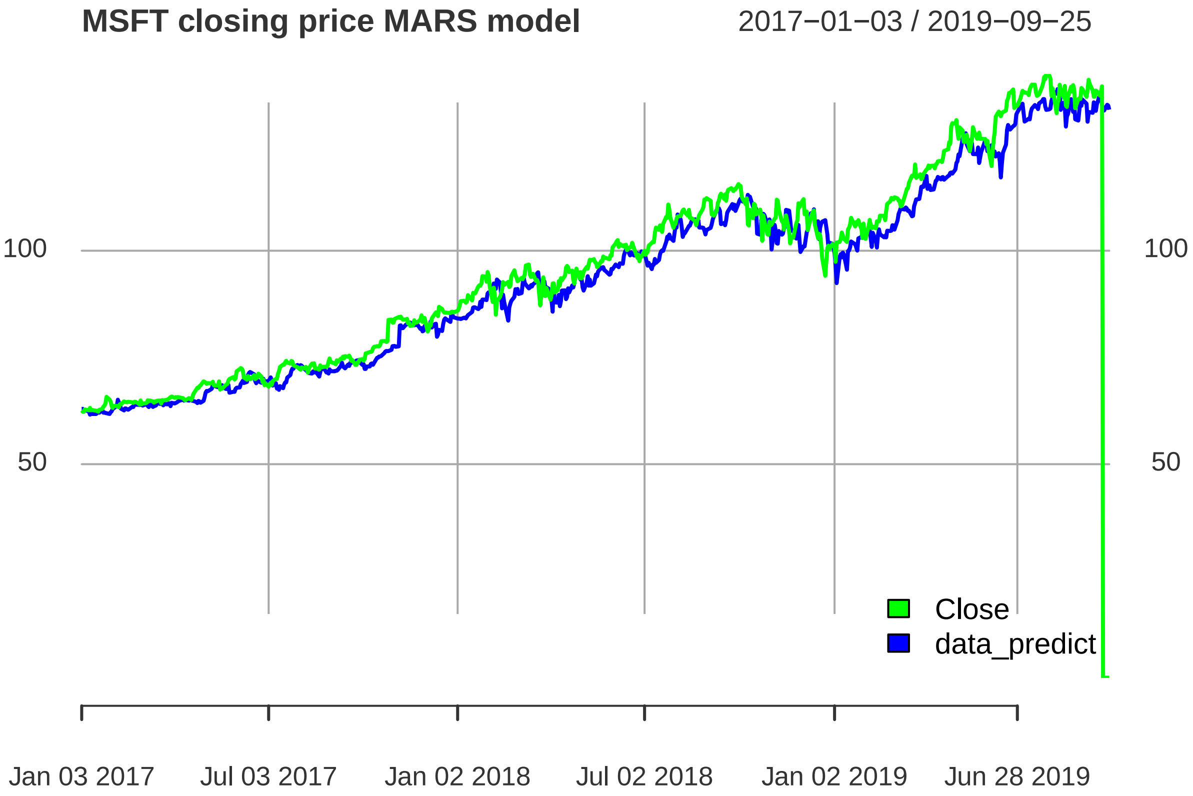
We have close to 30 years of test data starting from the 1st of Jan 1990. We are took the Training to Test split as 9:1. Meaning we will train our models (SVM, LM and MARS) on 27 years of training data and validate the models (SVM, LM and MARS) on 3 years of data. So our training and test data are distinct to avoid any possible 'overfitting' scenario.


Monte Carlo Performance Estimation Results
MSFT
Distribution of Statistics Scores
3.0
2.5
2.0
1.5
1.2
1.1
1.0
0.9
1.8
1.6
1.4
mse
mae
rmse
svm.v1
svm.v2
svm.v3
Alternative Workflows

Our results illustrate that Model 2, Linear Regression, is the best option. We select the parameters from the best workflow with the lowest mse, mae, rmse values and use that variant to train. Linear Regression does not have any 'variants' and therefore there is no performance estimation. The overall model is significant with F statistic ~ 0. The Adjusted R Square of 99.14%, which shows that 99.14% of the variability in the Stock Closing Price can be explained by the predictors.




MSFT closing price SVM model
2017−01−03 / 2019−09−25
100
50
100
50
🟩 Close
🟦 data_predict
Jan 03 2017
Jul 03 2017
Jan 02 2018
Jul 02 2018
Jan 02 2019
Jun 28 2019


MSFT closing price MARS model
2017−01−03 / 2019−09−25
100
50
100
50
🟩 Close
🟦 data_predict
Jan 03 2017
Jul 03 2017
Jan 02 2018
Jul 02 2018
Jan 02 2019
Jun 28 2019