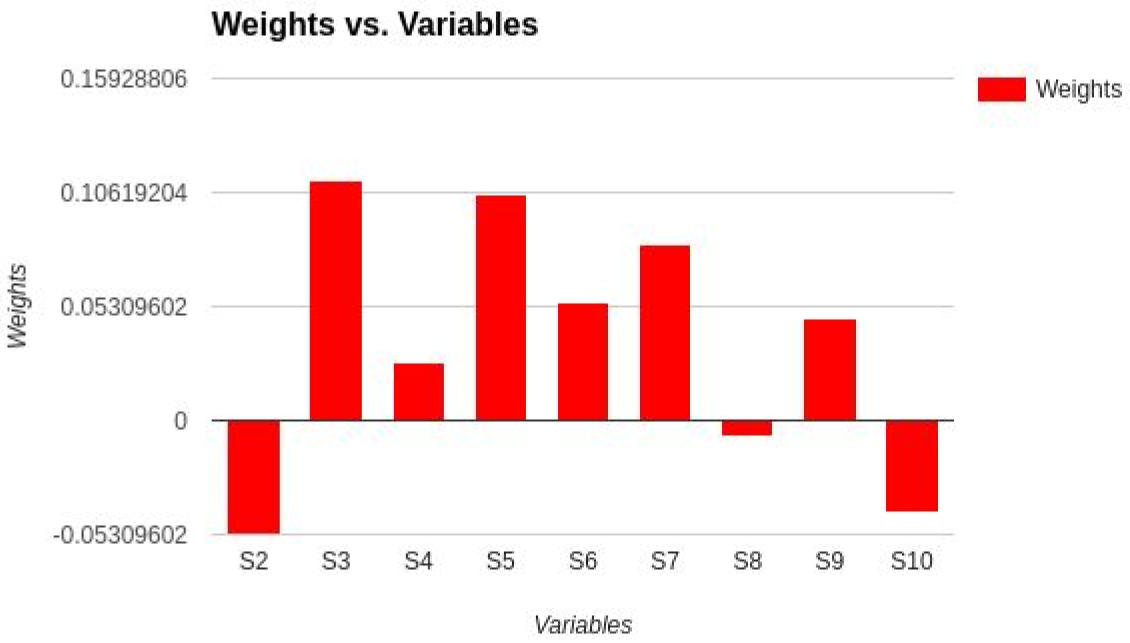
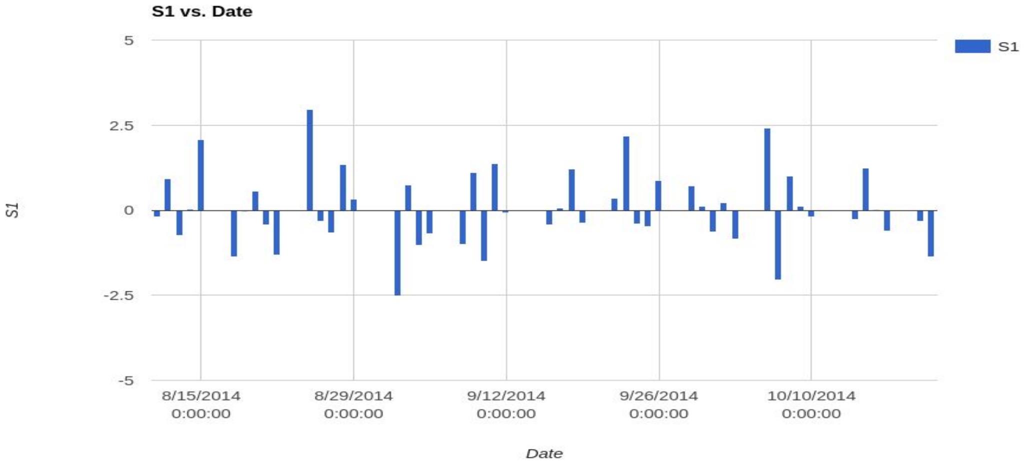
1. Which variables matter for predicting S1?



Importance of the variable can estimated from the absolute value of the coefficient. From the graph above, we can estimate that S3 and S5 were the best variables.

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



It goes up cumulatively by average of 0.03

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

3.1. The model is confidence is calulated using p-value, specifically P-Value from Pearson(R) Calculator.

3.2. The R Value = .885 for sample N = 18

3.3. P value is 0.0003, which is << 0.05, so we can say 95% confidence that model has good prediction.

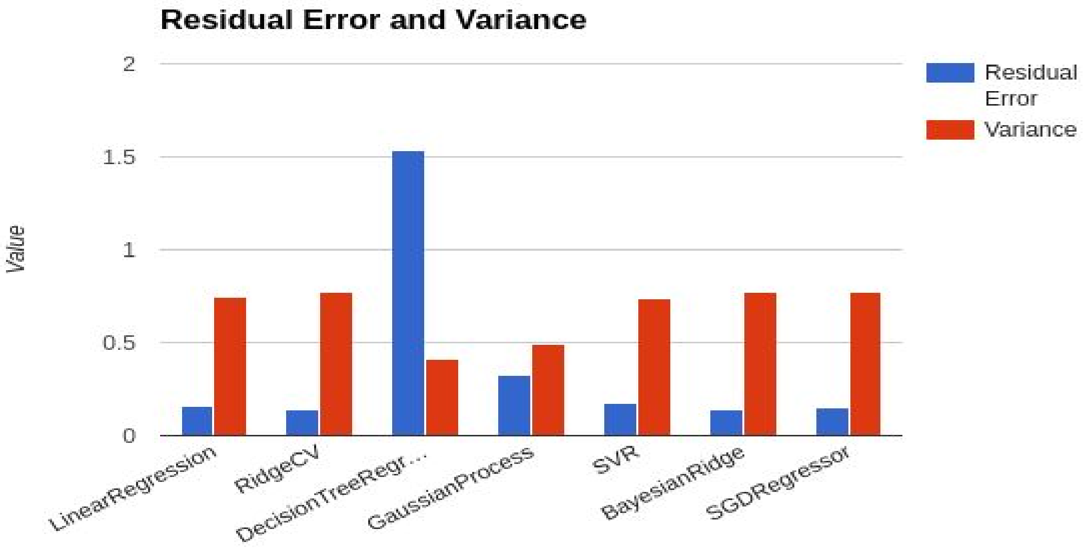
4. What techniques did you use? Why?

4.1. Data Preprocessing : Read the data from CSV, broke into labeled and unlabeled depending on S1. Then split the data into ⅔ for training and ⅓ for test sets.

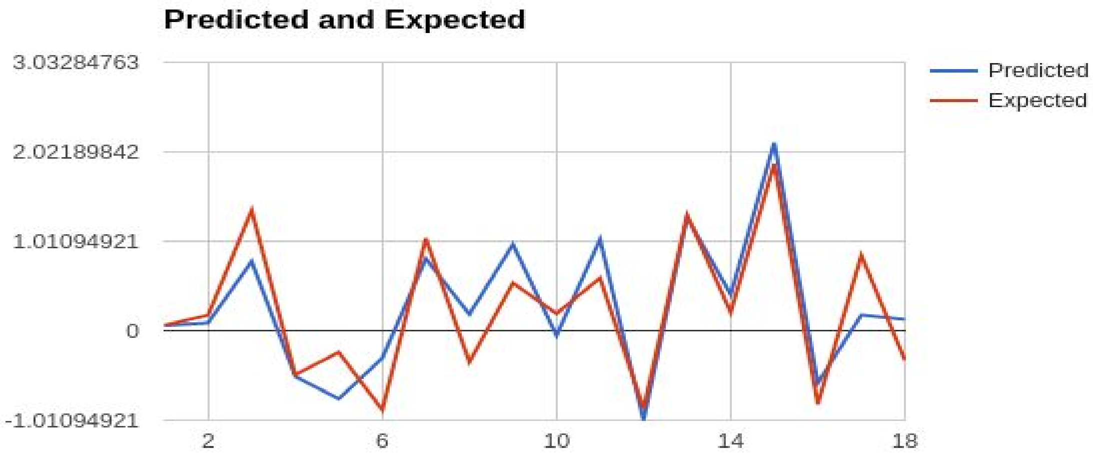
4.2. Model Selection : I used 7 models namely,

4.2.1. LinearRegression,
4.2.2. Ridge Regression,
4.2.3. DecisionTreeRegressor,
4.2.4. GaussianProcess,
4.2.5. Support Vector Regression,
4.2.6. BayesianRidge,
4.2.7. SGDRegressor

4.3. Best Model : Out of these models, Ridge Regression had lowest error and highest variance. More importantly the parameters were tuned using cross validation to reduce over fitting. Refer the chart below.

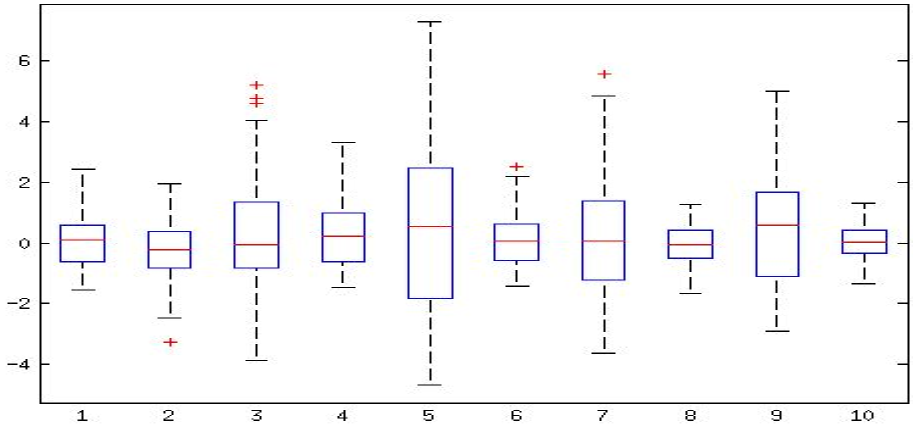


4.4. Visualization : Predicted values vs Testvalues



Summary : Other things to be tried

1) Feature subset selection
2) Validation set for parameter tuning
3) Clustering of data for visualization
4) Box plot of input data showing variance


6
4
2
0
-2
-4
1
2
3
4
5
6
7
8
9
10