Analytathon 1

Predictive Analytics of Soft Drinks Production

Team 5
Bhilare,Samira
Ismail,Anjum
McInerney,Niall
Wang,Lijin

Presentation Summary

Part 1 (Niall)	Part 2 (Lijin)	Part 3 (Samira)	Part 4 (Anjum)
- Introduction - Data Wrangling - Visualisation - Lagging	- Ascertaining relationship between target variable and 3 quality variables	- Predictive Modelling Design Process - Linear Regression - Logistic Regression - Decision Tree	-XG Boost - Splitting/ Training/ Predicting - Accuracy and Confusion matrix - Conclusions

Introduction

THE PROBLEM

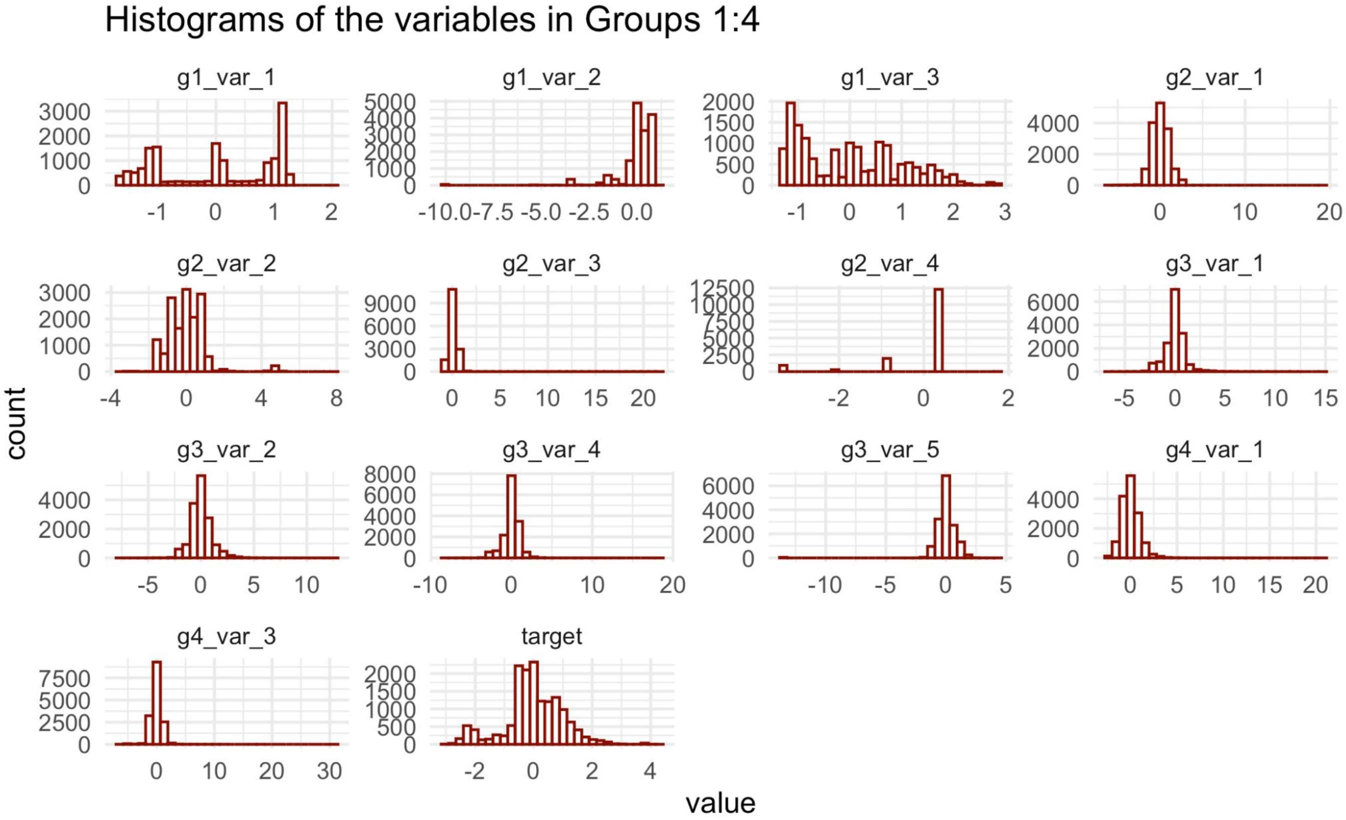
Incorporation of predictive analytics to predict quality of product at several stages of the production process.

-Predict variable *g4_var_2* using the preceding measurements in the process(optimal=-0.8574?)

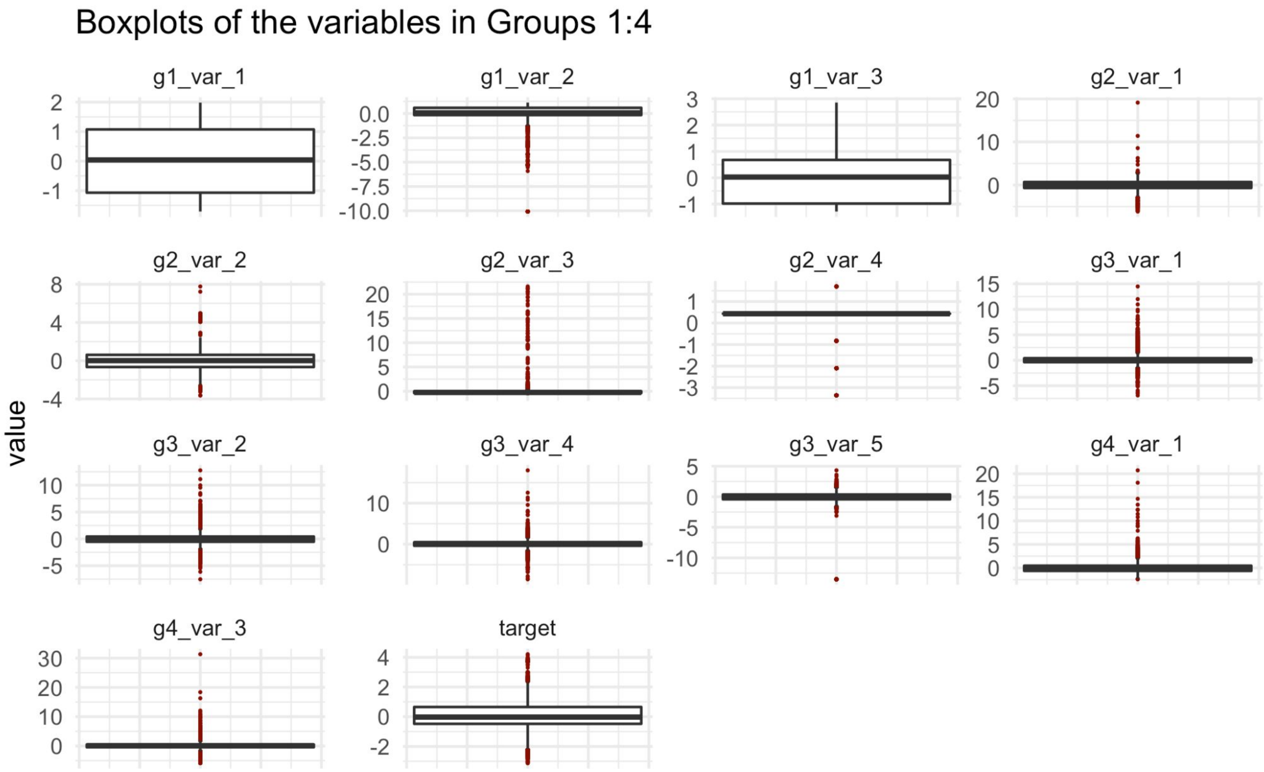
- Ascertain the relationship between *g4_var_2* and *g6_var_2*, *g6_var_3* & *g6_var_4*.

- Comment on model implementation.

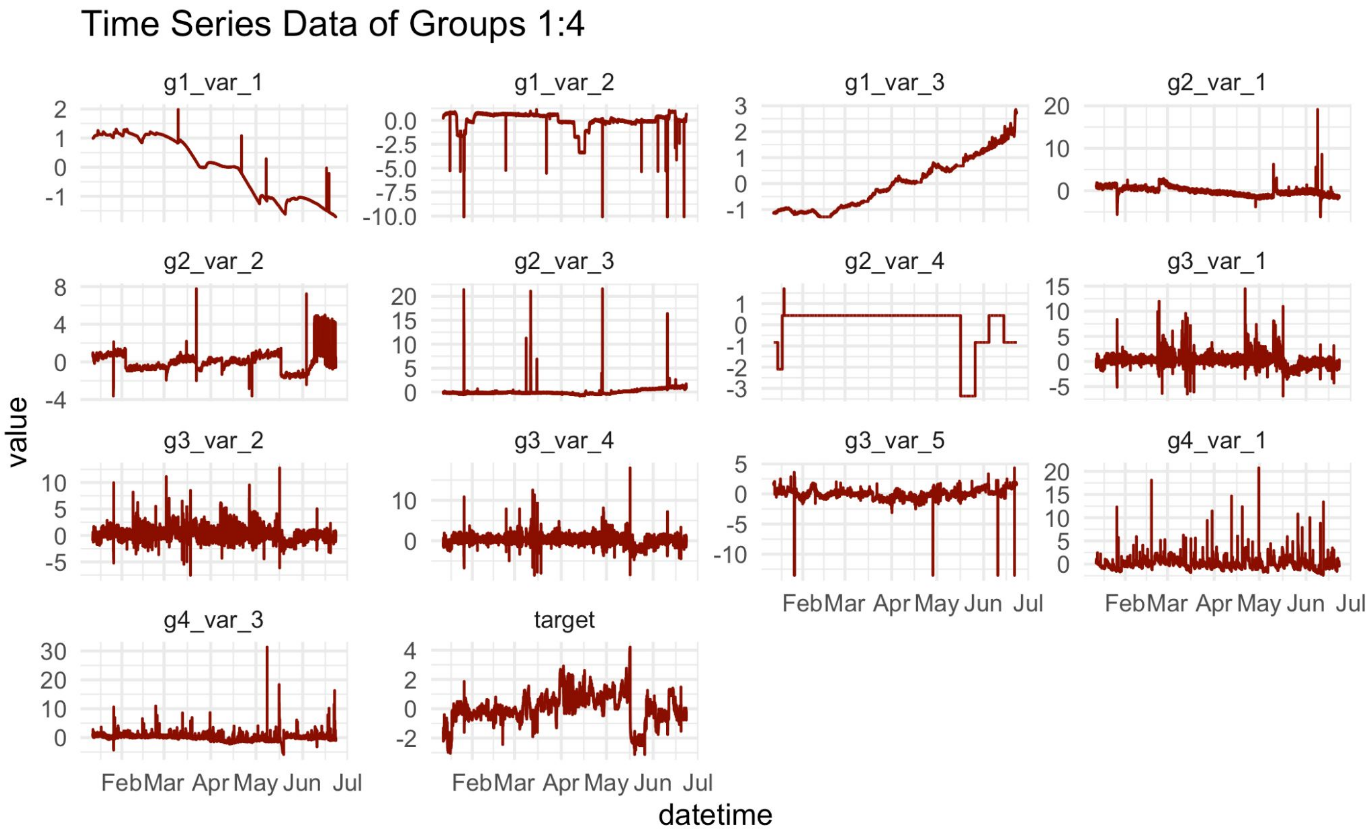
EDA Visualisation (1)



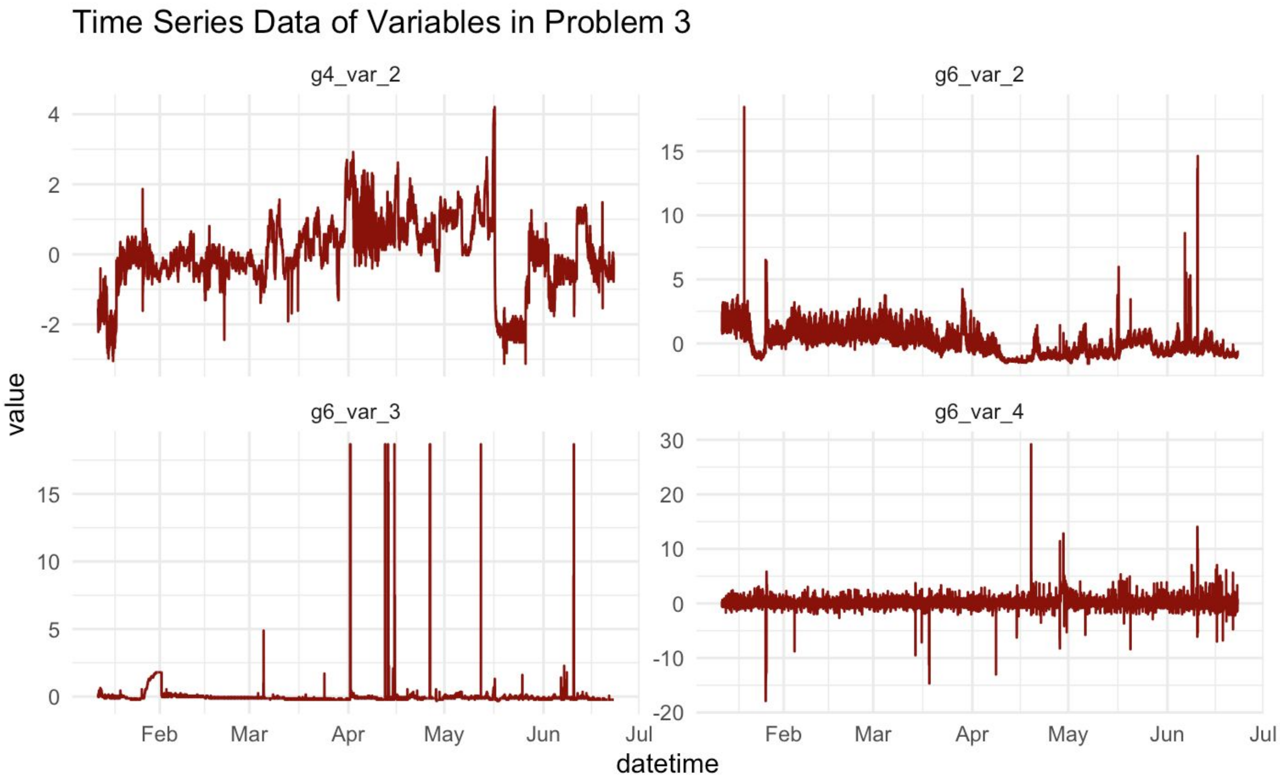
EDA Visualisation (2)



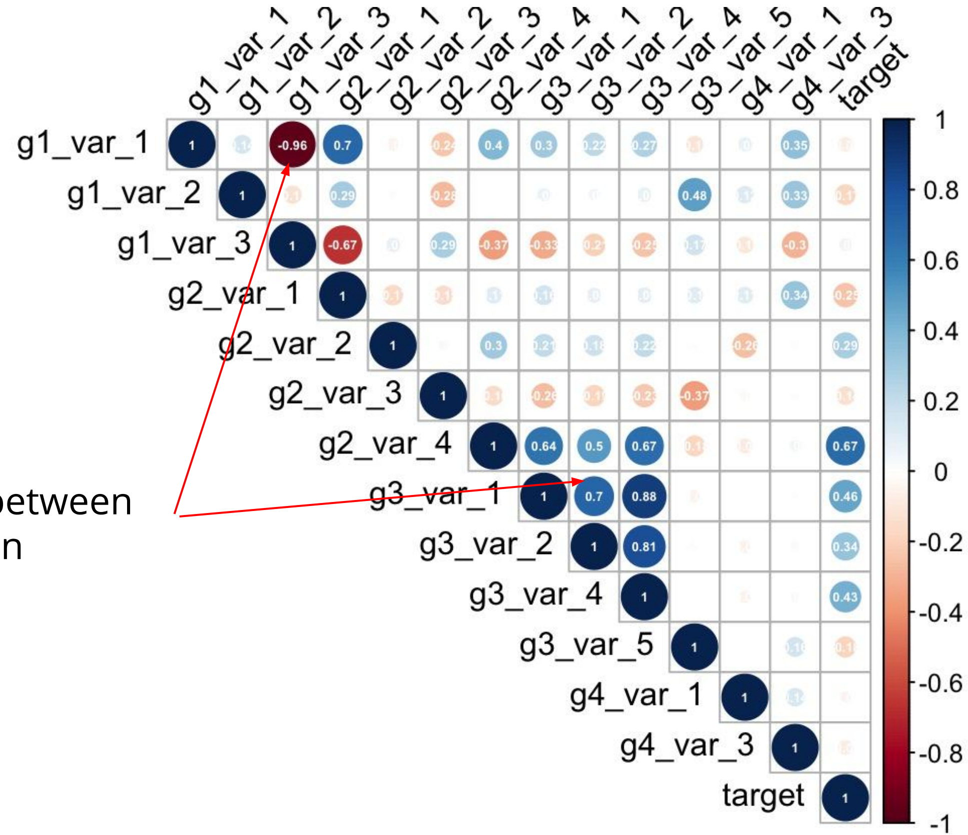
EDA Visualisation (3)



EDA Visualisation (4)

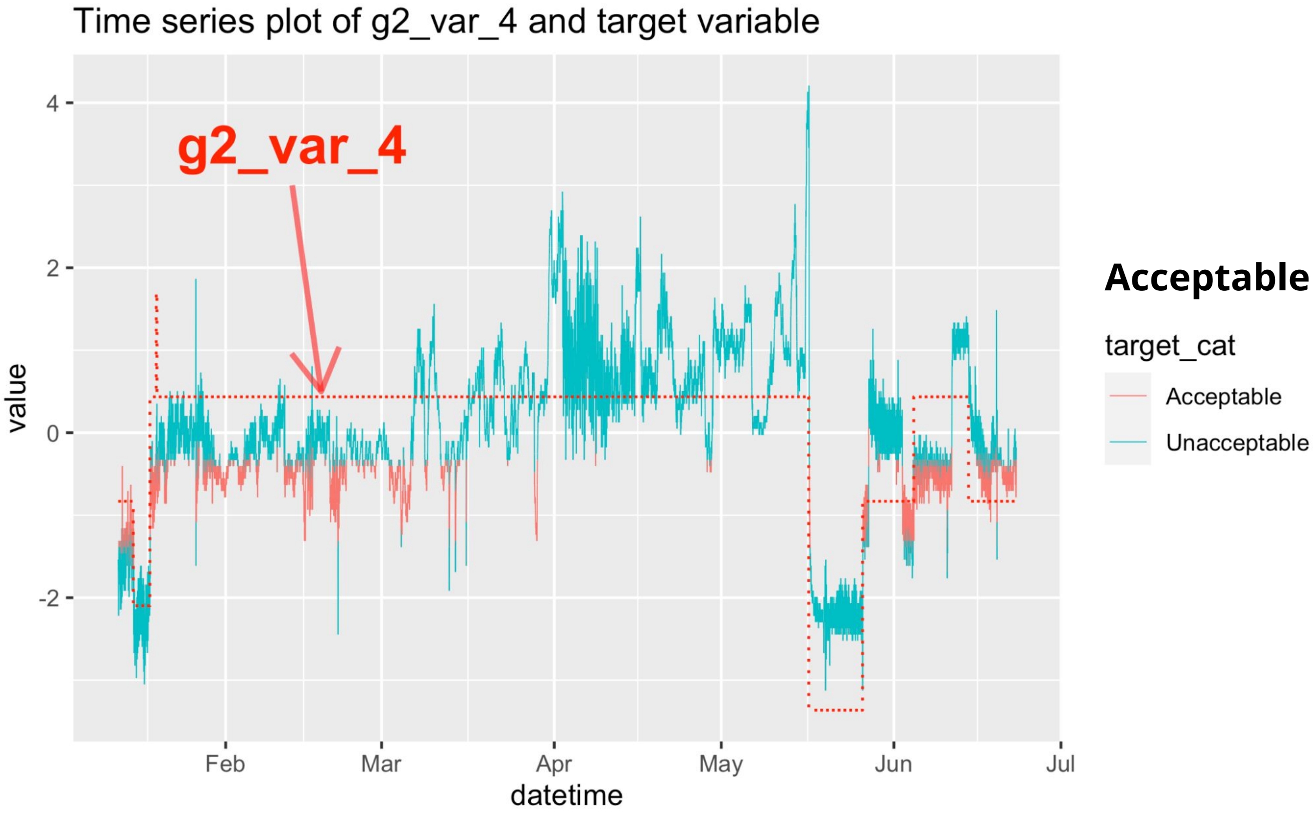


EDA Visualisation (5)



Correlations between the variables in Groups 1:4

EDA Visualisation (6)



Acceptable = Optimum ± a

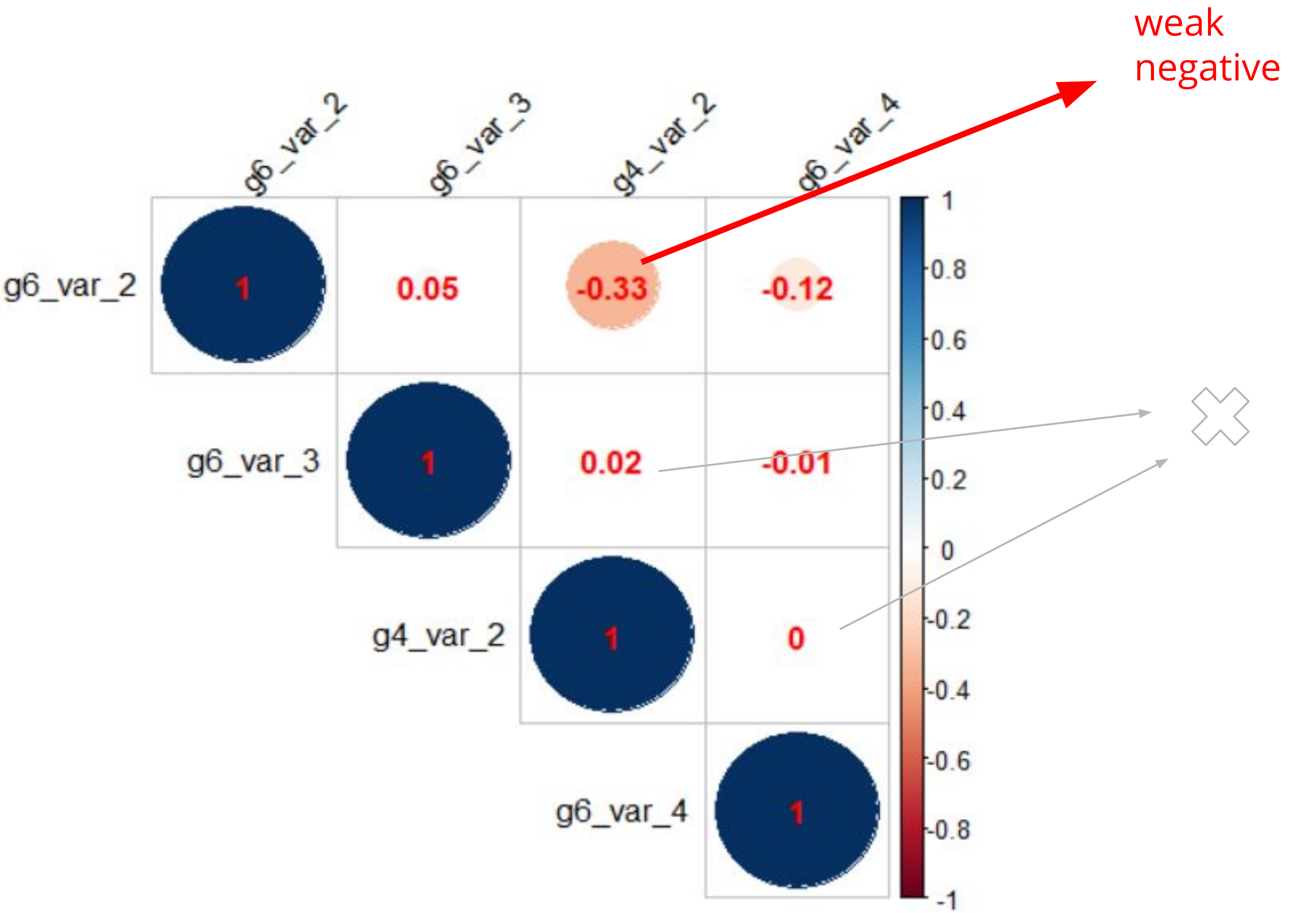
EDA Lagging

Group 1 -60mins　　Group 2 -45mins　　Group 3 -30mins　　......

	datetime	g1_var_1	g1_var_2	g1_var_3	g2_var_1	g2_var_2	g2_var_3	g2_var_4	g3_var_1	g3_var_2	g3_var_4	g3_var_5
1	2021-01-12 16:45:00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	2021-01-12 17:00:00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	2021-01-12 17:15:00	NA	NA	NA	NA	NA	NA	NA	-0.50900215	-0.54027739	-0.71315316	1.762910104
4	2021-01-12 17:30:00	NA	NA	NA	1.2316531	0.85598783	-0.056986420	-0.8310345	-0.37862768	-0.41878562	-0.54296184	1.702068663
5	2021-01-12 17:45:00	0.9931657	0.2628093	-1.113138	1.2316531	0.85598783	0.004010301	-0.8310345	-1.09568725	-0.90475269	-1.39391847	1.762910104
6	2021-01-12 18:00:00	0.9931657	0.2192224	-1.127243	1.0842359	0.85598783	-0.117983141	-0.8310345	-0.83493832	-0.84400681	-1.13863148	1.762910104
7	2021-01-12 18:15:00	0.9931657	0.2854100	-1.127243	1.0842359	0.91665895	-0.117983141	-0.8310345	-0.90012555	-0.90475269	-1.22372715	1.580385780
8	2021-01-12 18:30:00	0.9931657	0.3144680	-1.144874	1.0065069	0.61330332	-0.239976582	-0.8310345	-0.37862768	-0.35803974	-0.54296184	1.641227221
9	2021-01-12 18:45:00	0.9931657	0.2272940	-1.144874	1.1163997	0.79531670	-0.239976582	-0.8310345	-0.31344044	-0.29729386	-0.45786617	1.823751545
10	2021-01-12 19:00:00	0.9931657	0.3144680	-1.123717	1.1163997	0.85598783	-0.117983141	-0.8310345	-0.50900215	-0.47953151	-0.71315316	1.580385780
11	2021-01-12 19:15:00	0.9931657	0.2272940	-1.123717	1.0467116	0.85598783	0.004010301	-0.8310345	0.53399360	0.61389439	0.64837746	1.641227221
12	2021-01-12 19:30:00	0.9931657	0.2837957	-1.123717	1.1941287	0.91665895	-0.117983141	-0.8310345	-0.24825321	-0.35803974	-0.37277051	1.519544339

Ascertain the relationship

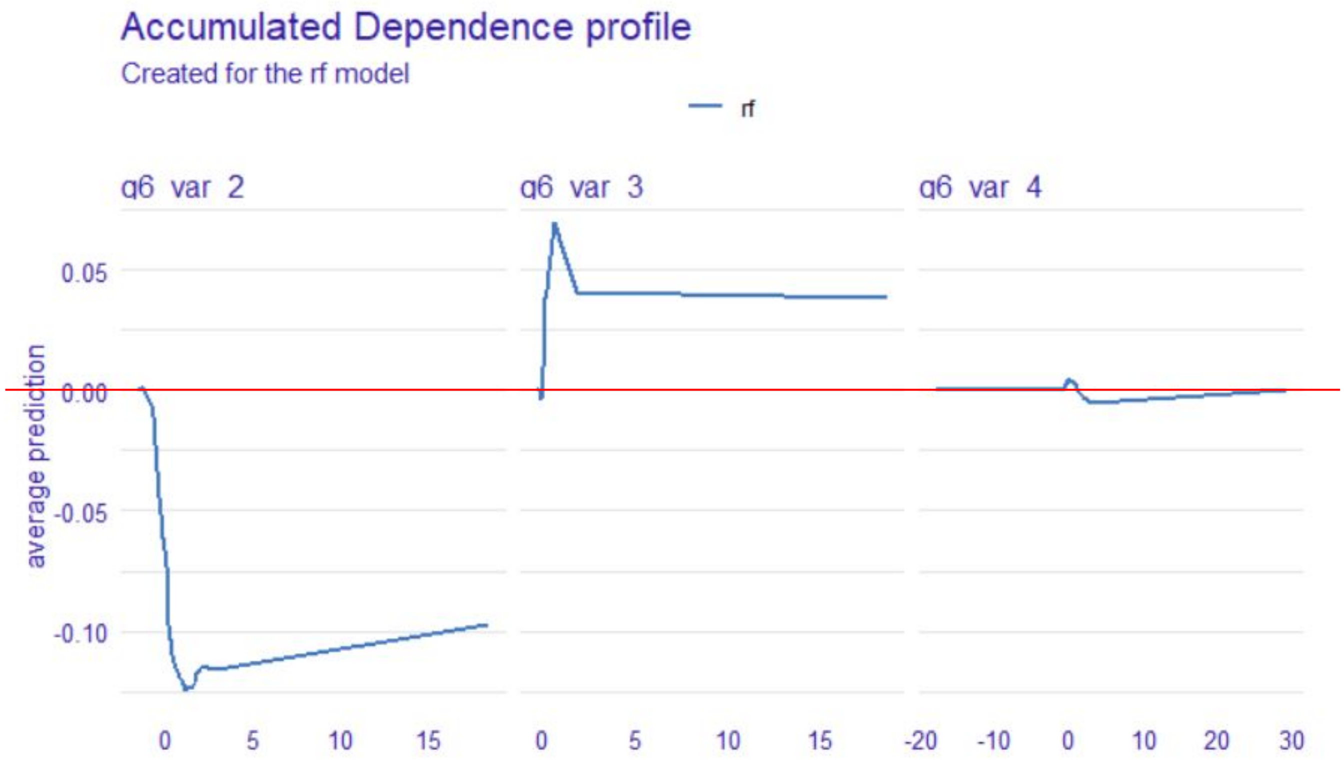
Correlation



Generalised Linear Models (GLM)

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  9.61726    1.11674   8.612   <2e-16 ***
g6_var_2    -0.50746    0.53146  -0.955    0.340    
g6_var_3     0.39745    2.12296   0.187    0.851    
g6_var_4    -0.03948    0.39698  -0.099    0.921    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Accumulated Local Effects (ALE)

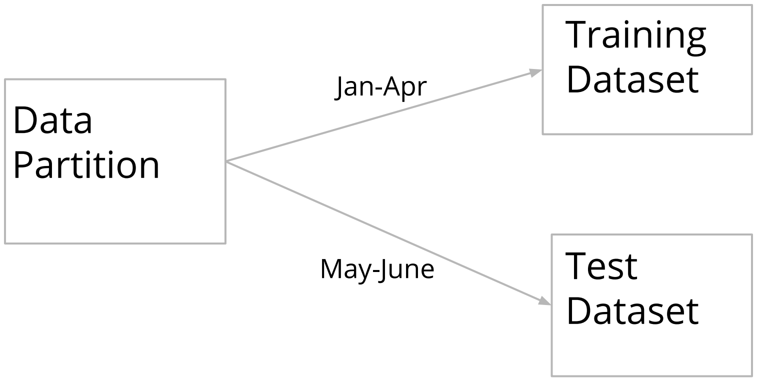


Predictive Modelling Techniques

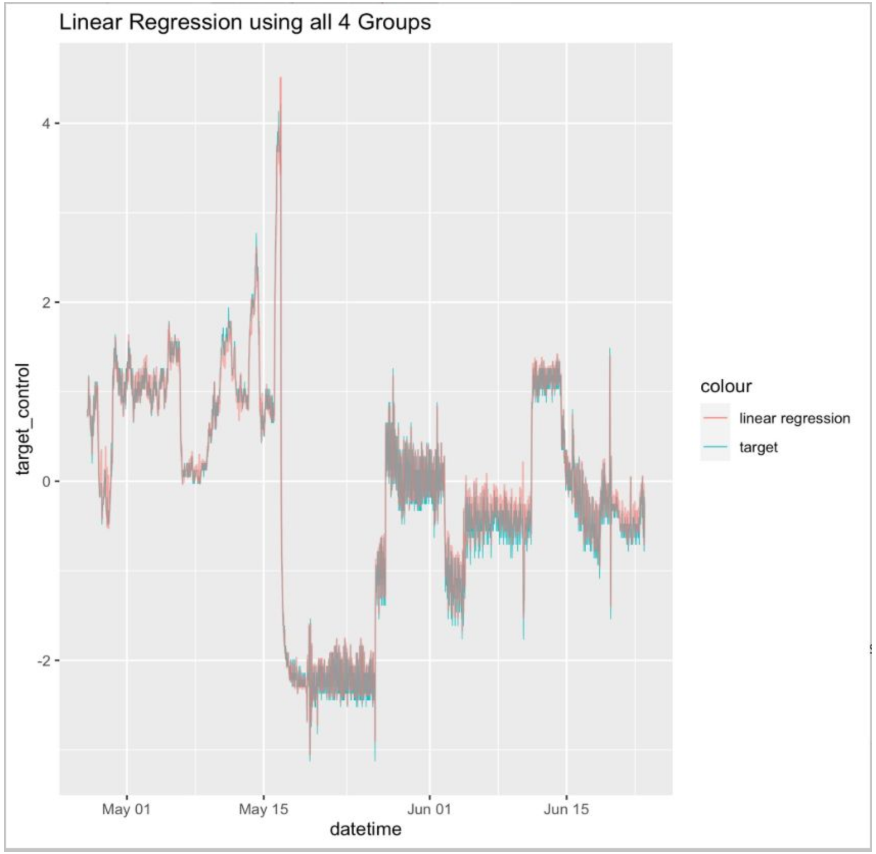
Overview of Predictive Modelling Design Process

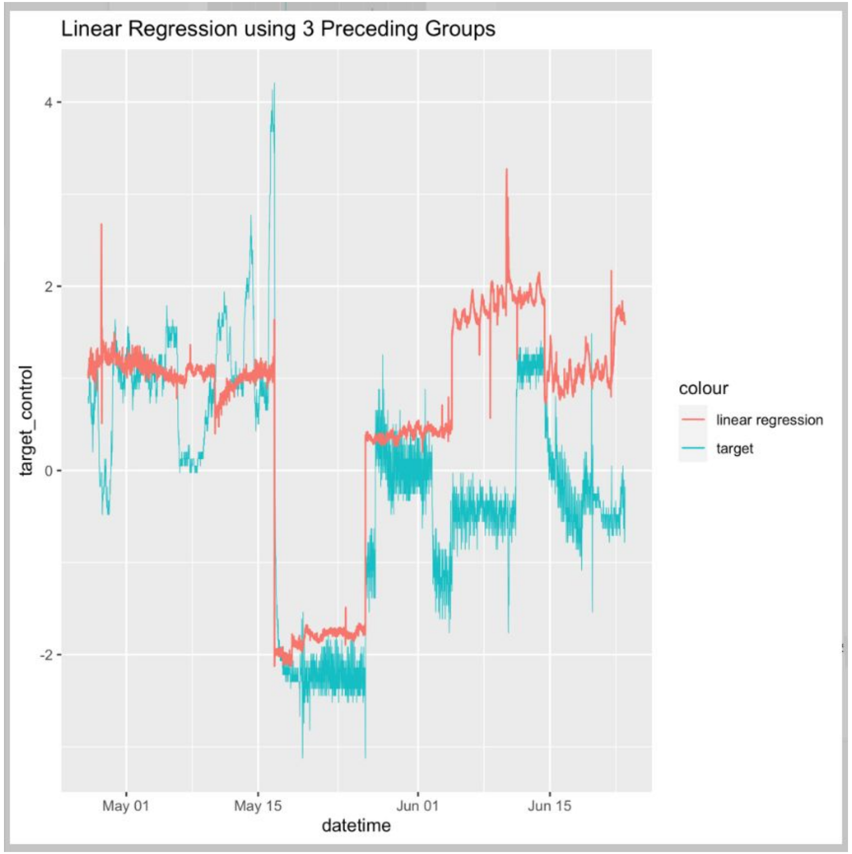
Multiple Linear Regression Model : LM

- Multiple linear regression is a linear approach to modelling the relationship between a scalar response *('g4_var_2')* and multiple explanatory variables *('g1_var_1', 'g2_var_1', etc).*
- Data Partitioning based on 'datetime' variable - considering the month wise split



Visualisation for Linear Regression Model : LM

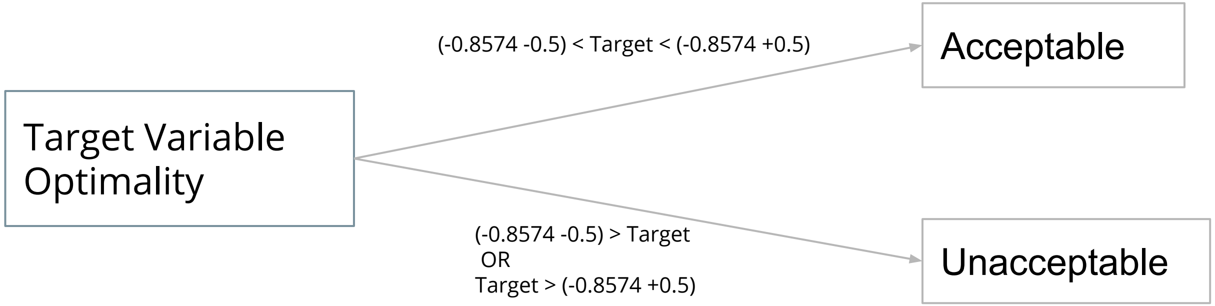




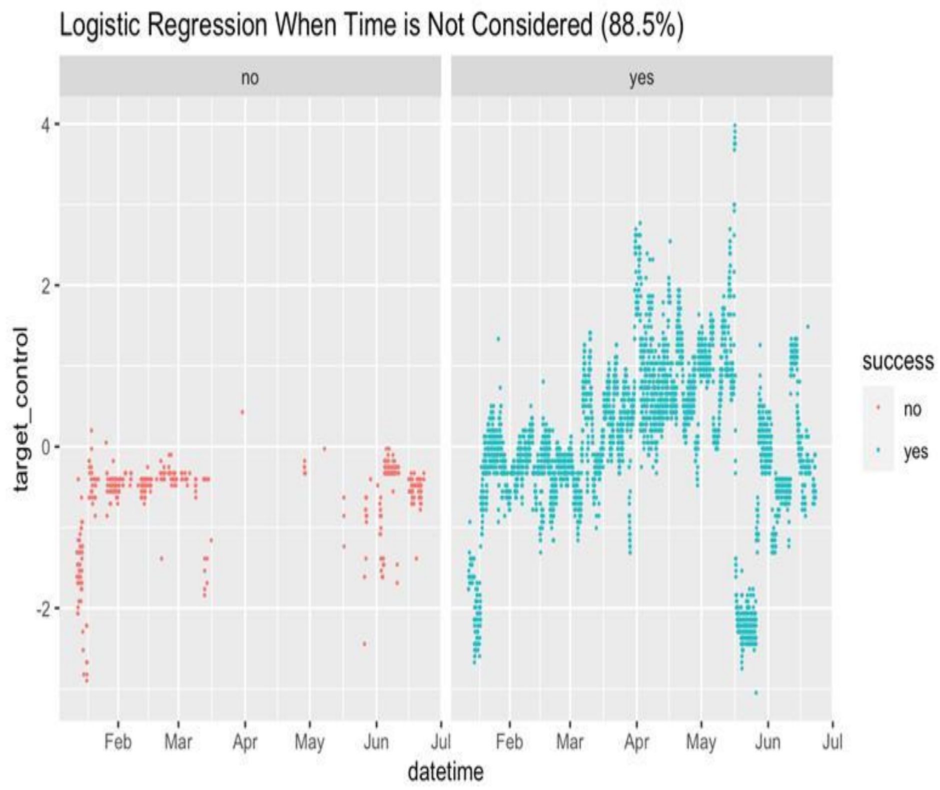
Logistic Regression Model : GLM

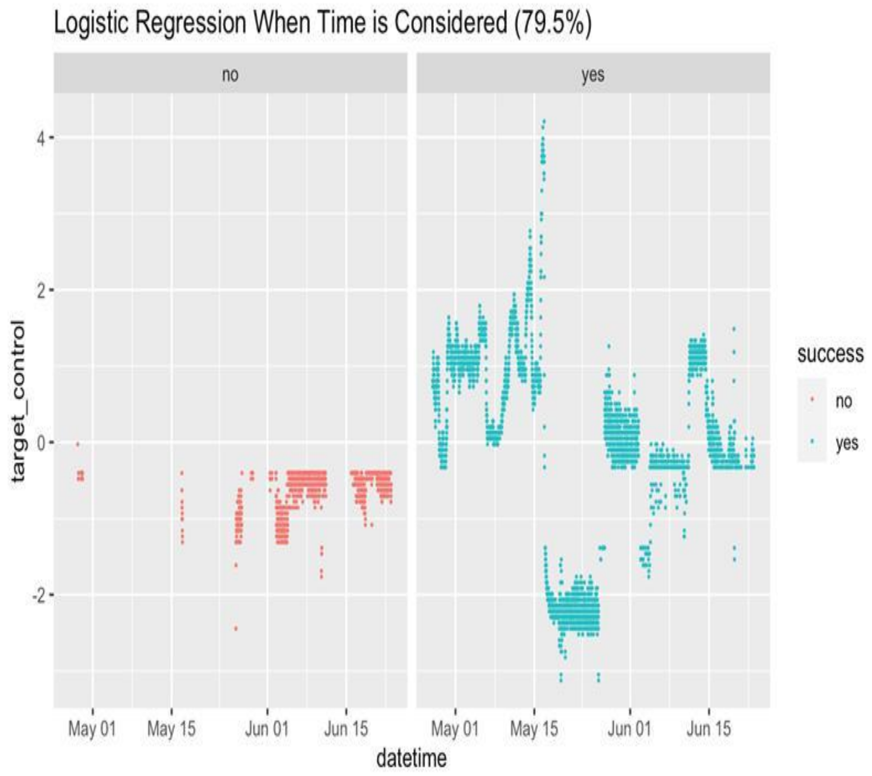
- Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable
- Classification of the target variable (*g4_var_2)* into bins for conversion from a continuous numeric to binary categorical variable (*Acceptable/Unacceptable*) to check its optimality.

Considering Optimal value for target variable= -0.8574



Visualisation plots for Logistic regression model :GLM





Statistics for Logistic Regression model

Statistics when time is NOT considered while data partitioning

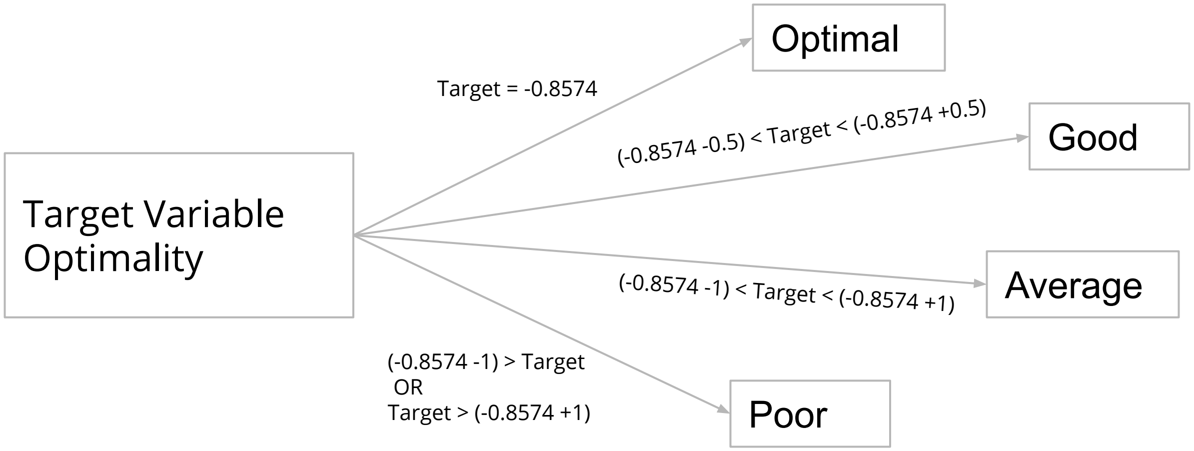
```
Confusion Matrix and Statistics

              Reference
Prediction     Acceptable Unacceptable
  Acceptable          635          182
  Unacceptable        328         3497

               Accuracy : 0.8901
                 95% CI : (0.8808, 0.899)
    No Information Rate : 0.7925
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6461

 Mcnemar's Test P-Value : 1.356e-10

            Sensitivity : 0.6594
            Specificity : 0.9505
         Pos Pred Value : 0.7772
         Neg Pred Value : 0.9142
             Prevalence : 0.2075
         Detection Rate : 0.1368
   Detection Prevalence : 0.1760
      Balanced Accuracy : 0.8050

       'Positive' Class : Acceptable
```

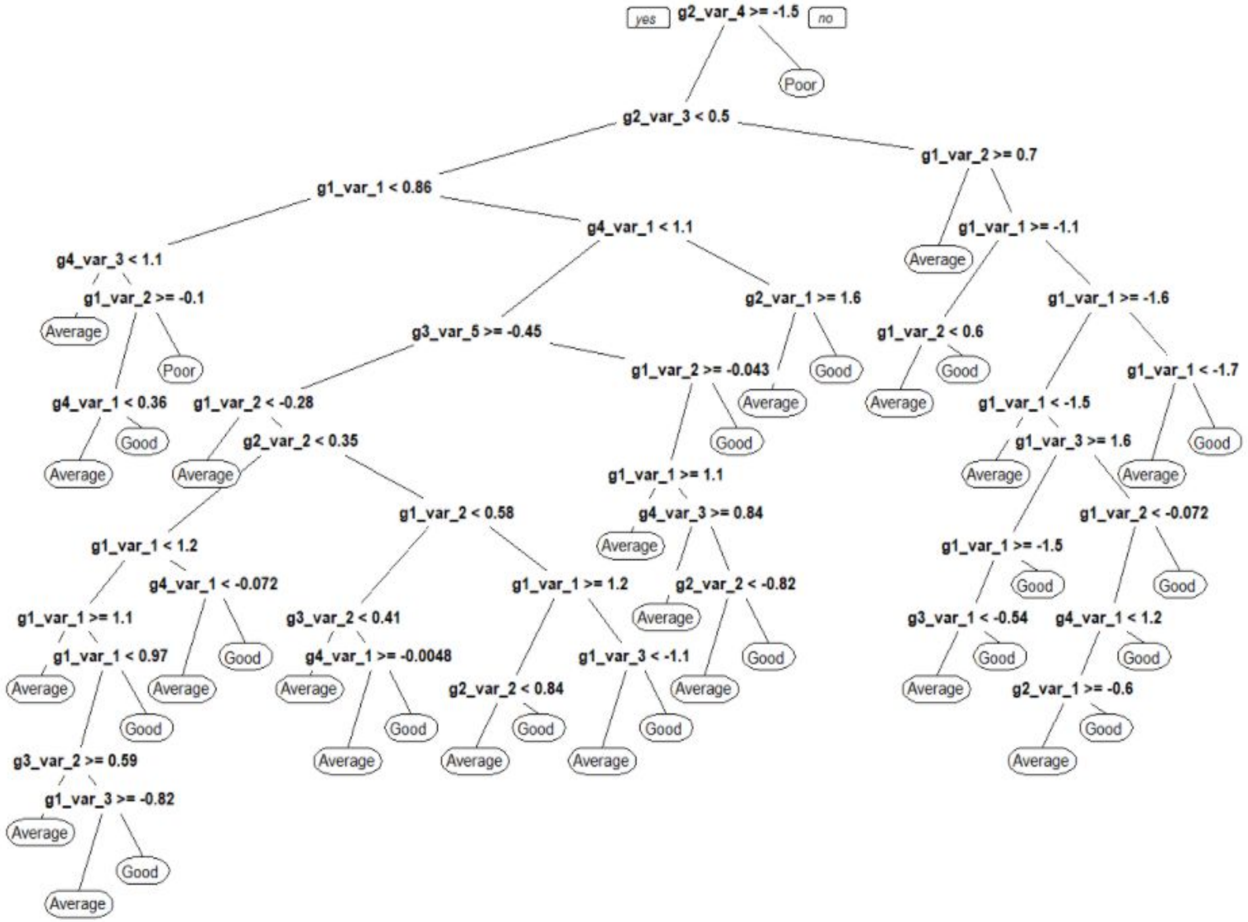
Statistics when time is considered while data partitioning

```
Confusion Matrix and Statistics

              Reference
Prediction     Acceptable Unacceptable
  Acceptable           48            9
  Unacceptable       1105         4312

               Accuracy : 0.7965
                 95% CI : (0.7856, 0.8071)
    No Information Rate : 0.7894
    P-Value [Acc > NIR] : 0.1006

                  Kappa : 0.0607

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.041631
            Specificity : 0.997917
         Pos Pred Value : 0.842105
         Neg Pred Value : 0.796013
             Prevalence : 0.210632
         Detection Rate : 0.008769
   Detection Prevalence : 0.010413
      Balanced Accuracy : 0.519774

       'Positive' Class : Acceptable
```

Decision Tree

- A decision tree as a predictive model are used to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves).
- Using classification tree design for the target variable (*g4_var_2*) classified into 4 categories based on the optimality of the values





- *g2_var_4* plays a vital role in predicting category for *g4_var_2*
- Decision Tree can be used to detect spike in the *g4_var_2* value

Decision Tree Model Observations

Confusion Matrix for Training Dataset	Confusion Matrix for Test Dataset

Confusion Matrix for Training Dataset

```
                model_pred_train
                 Can be Improved Discard Good Quality
Can be Improved             7875     102          557
Discard                      218    1032            9
Good Quality                 589      11         1991
```

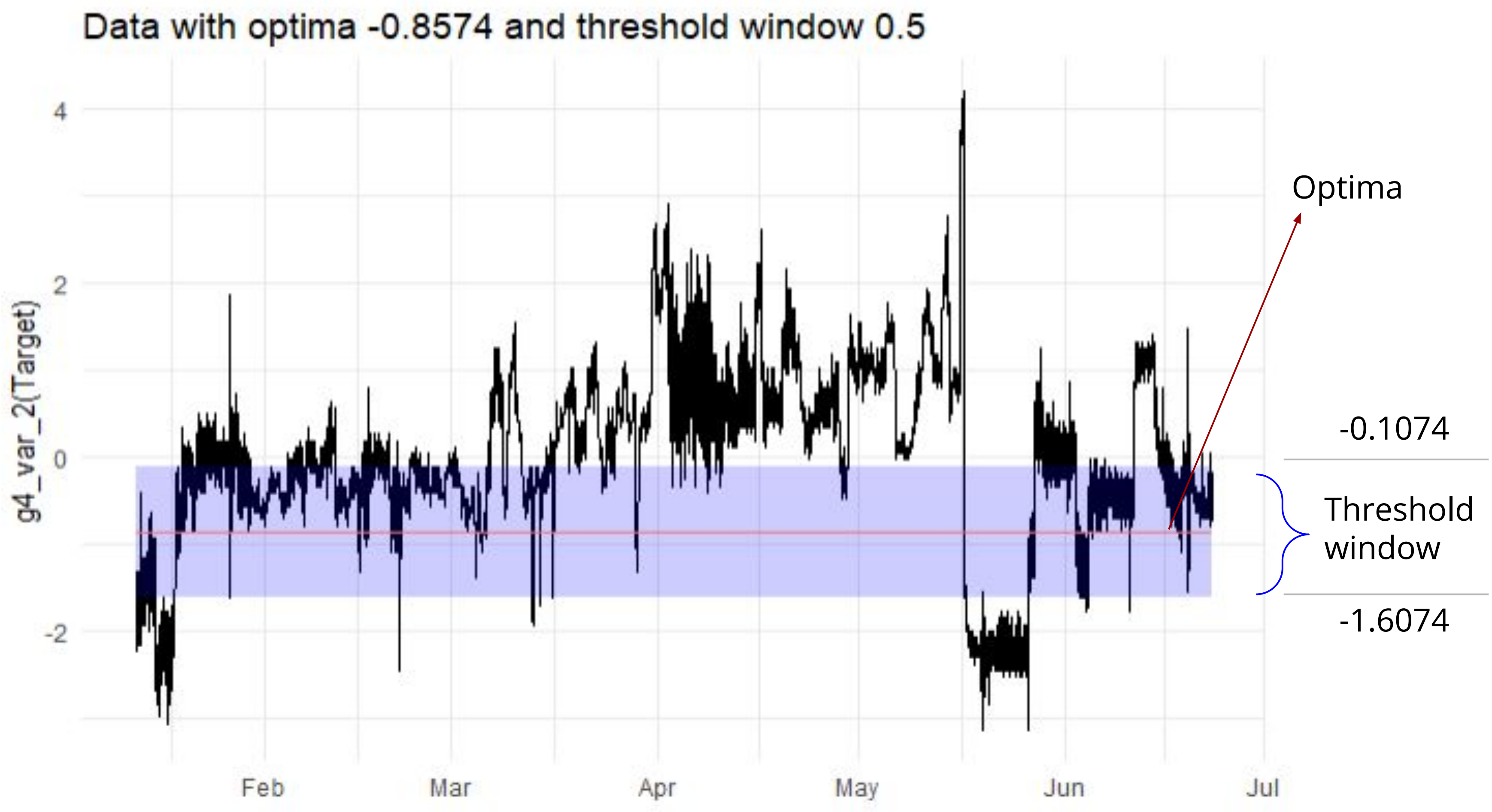
Confusion Matrix for Test Dataset

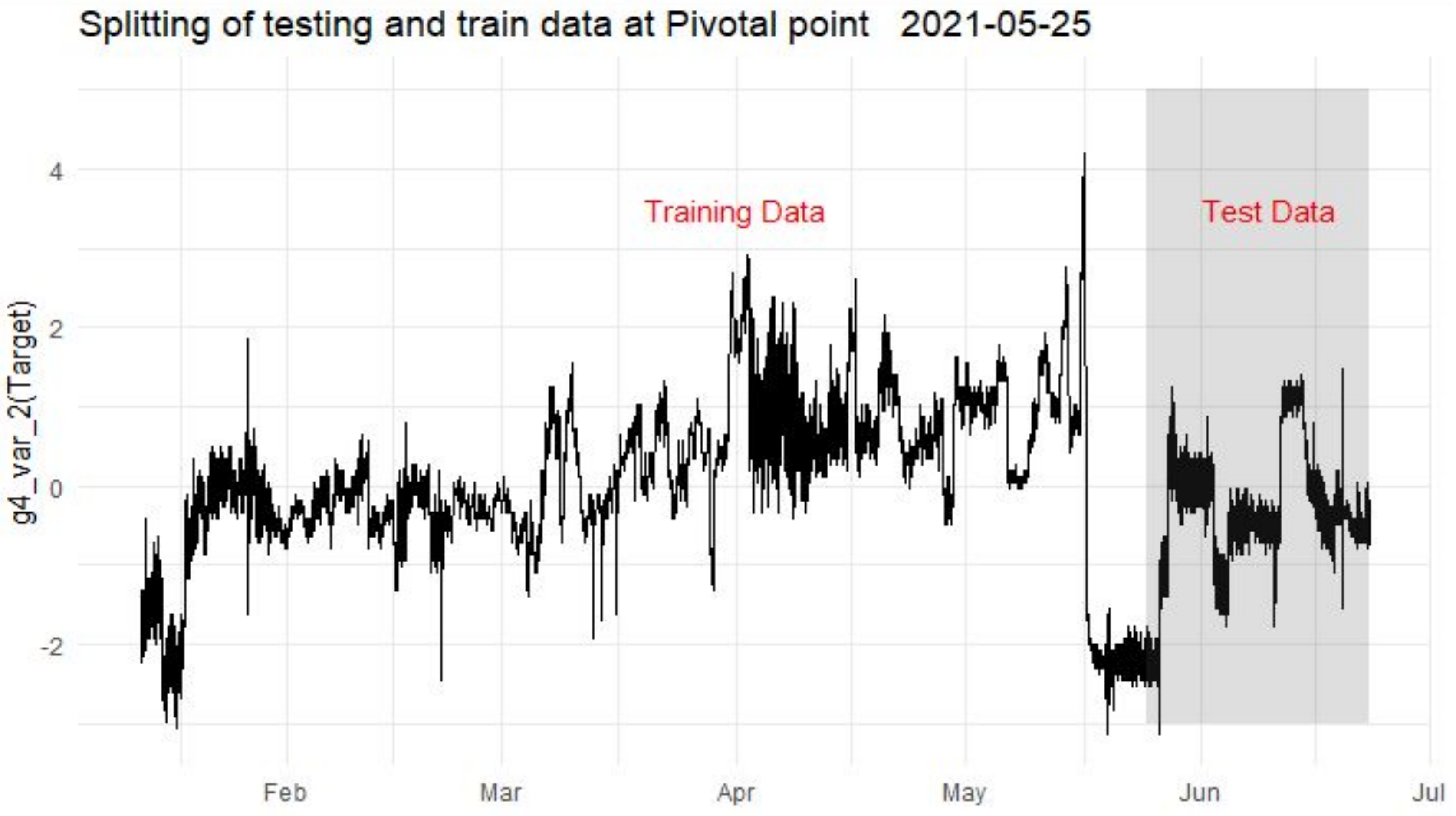
```
                model_pred_test
                 Can be Improved Discard Good Quality
Can be Improved             1951      30          176
Discard                       53     264            1
Good Quality                 168       2          449
```

Model Accuracy for Training and Test Dataset

```
"Training Accuracy:  0.88000645994832"
"Testing Accuracy:  0.861021331609567"
```

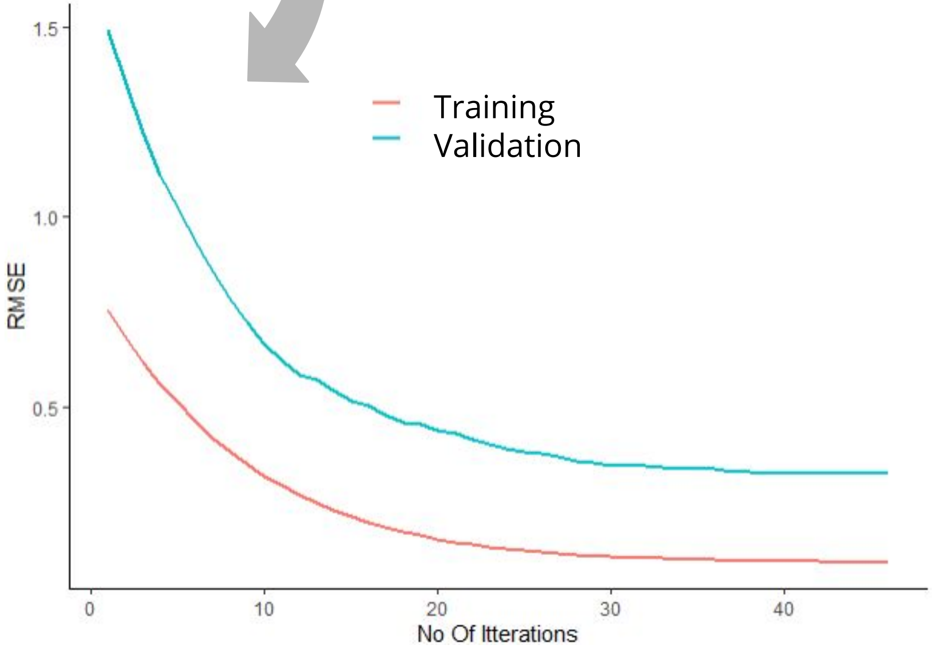
Machine Learning model


Data with optima -0.8574 and threshold window 0.5
g4_var_2(Target)
4
2
0
-2
Feb
Mar
Apr
May
Jun
Jul
Optima
-0.1074
Threshold window
-1.6074


Splitting of testing and train data at Pivotal point 2021-05-25
Training Data
Test Data
g4_var_2(Target)
4
2
0
-2
Feb
Mar
Apr
May
Jun
Jul

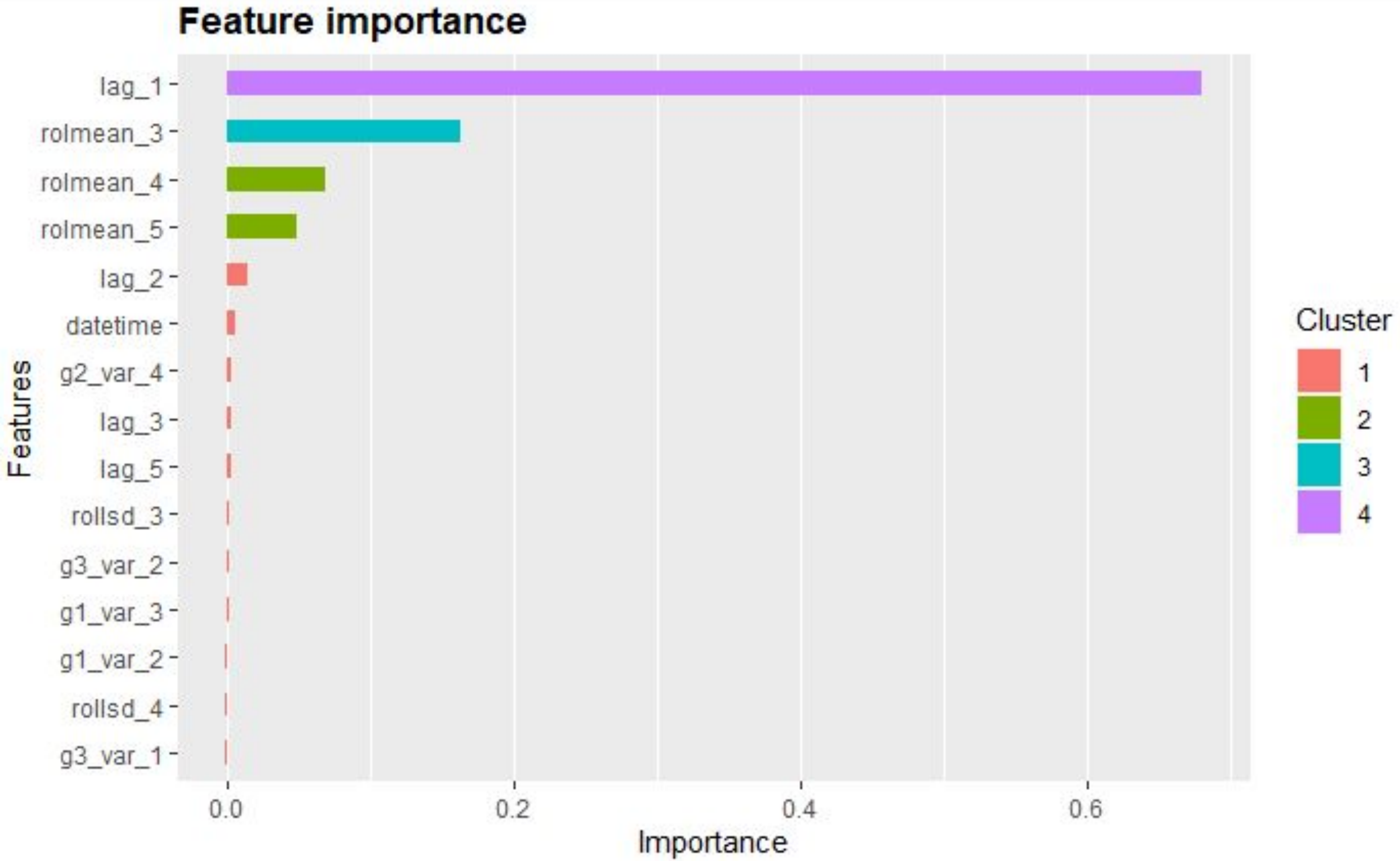
XGBoost

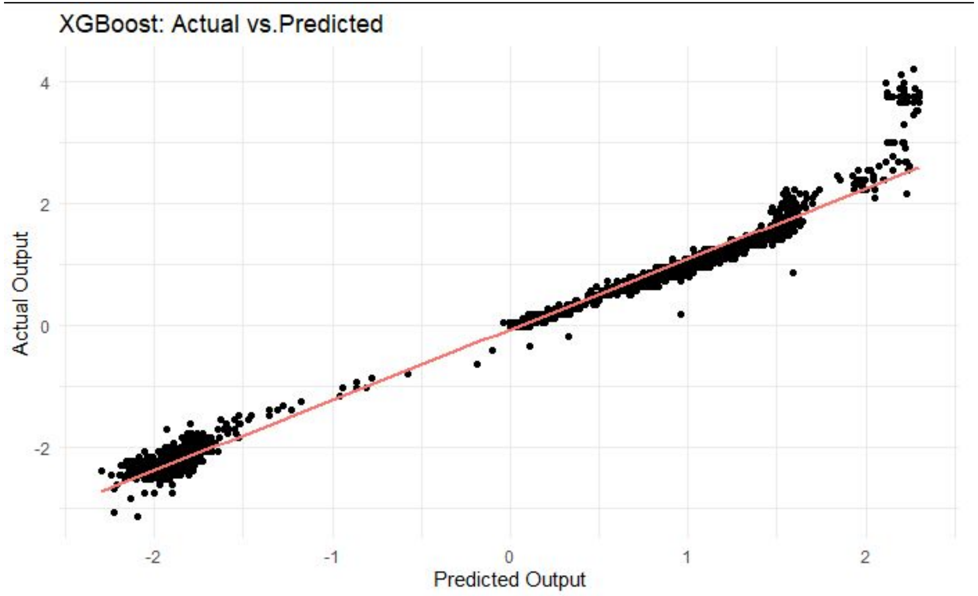
```
[2]	train-rmse:0.751890	val-rmse:1.489582
[3]	train-rmse:0.681327	val-rmse:1.347145
[4]	train-rmse:0.616380	val-rmse:1.217204
[5]	train-rmse:0.559867	val-rmse:1.111690
[6]	train-rmse:0.509248	val-rmse:1.021783
[7]	train-rmse:0.461918	val-rmse:0.933983
[8]	train-rmse:0.419559	val-rmse:0.857485
[9]	train-rmse:0.381708	val-rmse:0.782804
[10]	train-rmse:0.347746	val-rmse:0.720342
[11]	train-rmse:0.317656	val-rmse:0.664537
[12]	train-rmse:0.291722	val-rmse:0.622135
[13]	train-rmse:0.267547	val-rmse:0.582448
[14]	train-rmse:0.247103	val-rmse:0.572553
[15]	train-rmse:0.227950	val-rmse:0.543032
[16]	train-rmse:0.211035	val-rmse:0.514931
[17]	train-rmse:0.196808	val-rmse:0.504335
[18]	train-rmse:0.183455	val-rmse:0.480077
[19]	train-rmse:0.171576	val-rmse:0.459751
[20]	train-rmse:0.161820	val-rmse:0.455947
[21]	train-rmse:0.152396	val-rmse:0.436275
[22]	train-rmse:0.144904	val-rmse:0.429259
[23]	train-rmse:0.137788	val-rmse:0.414120
[24]	train-rmse:0.131429	val-rmse:0.401154
[25]	train-rmse:0.126523	val-rmse:0.390163
[26]	train-rmse:0.121642	val-rmse:0.380869
[27]	train-rmse:0.117437	val-rmse:0.375960
[28]	train-rmse:0.113841	val-rmse:0.367536
[29]	train-rmse:0.110965	val-rmse:0.356714
[30]	train-rmse:0.108481	val-rmse:0.351584
[31]	train-rmse:0.106306	val-rmse:0.344922
[32]	train-rmse:0.104259	val-rmse:0.344132
```

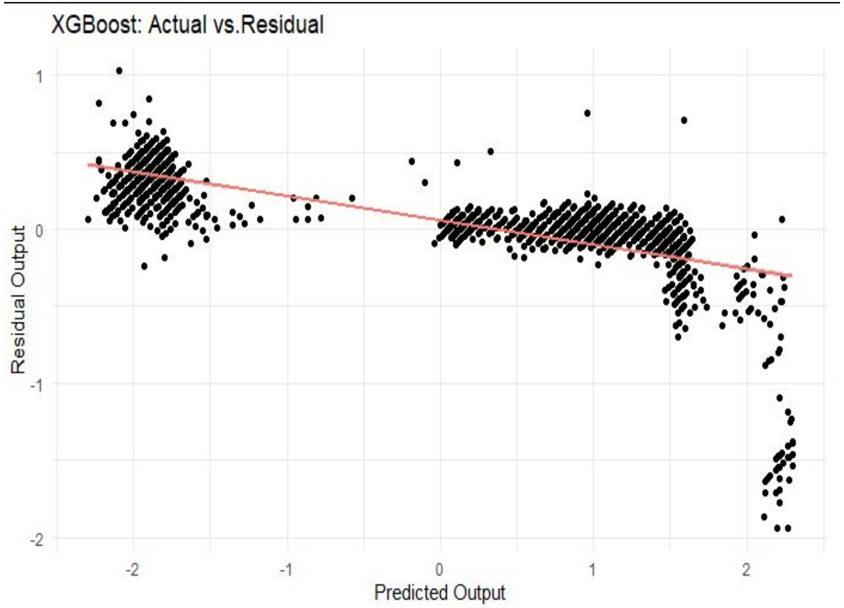


Best Score for Training data: [1] 0.09169

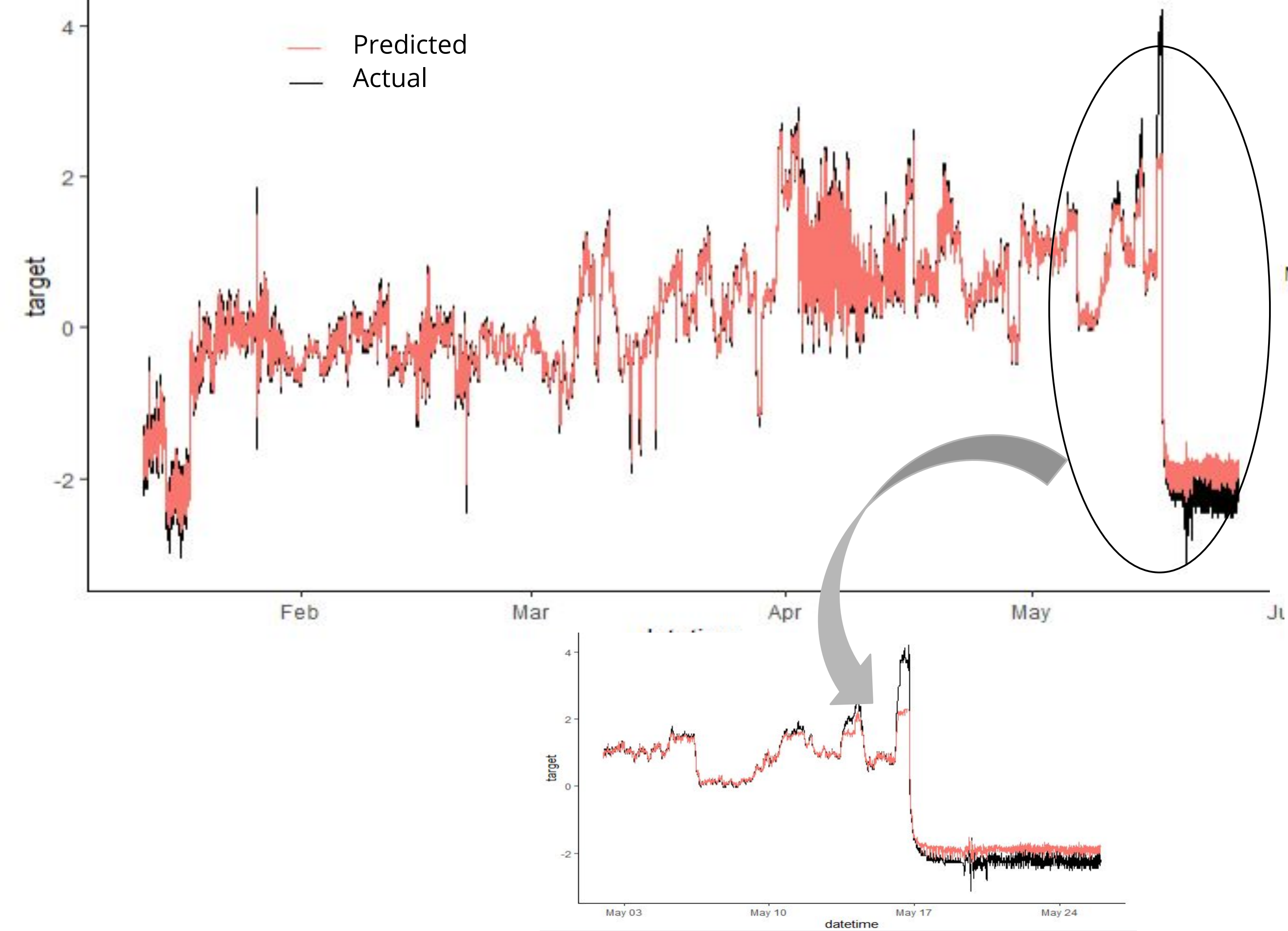
Best Score for Validation data: [1] 0.32312


Feature importance
lag_1
rolmean_3
rolmean_4
rolmean_5
lag_2
datetime
g2_var_4
lag_3
lag_5
rollsd_3
g3_var_2
g1_var_3
g1_var_2
rollsd_4
g3_var_1
Features
0.0
0.2
0.4
0.6
Importance
Cluster
1
2
3
4

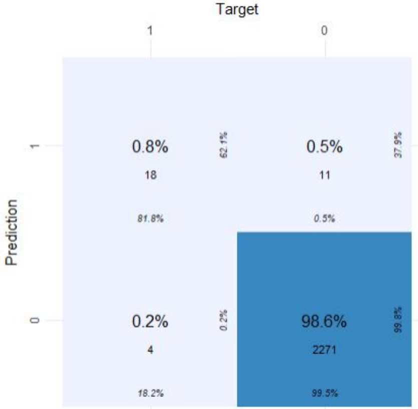


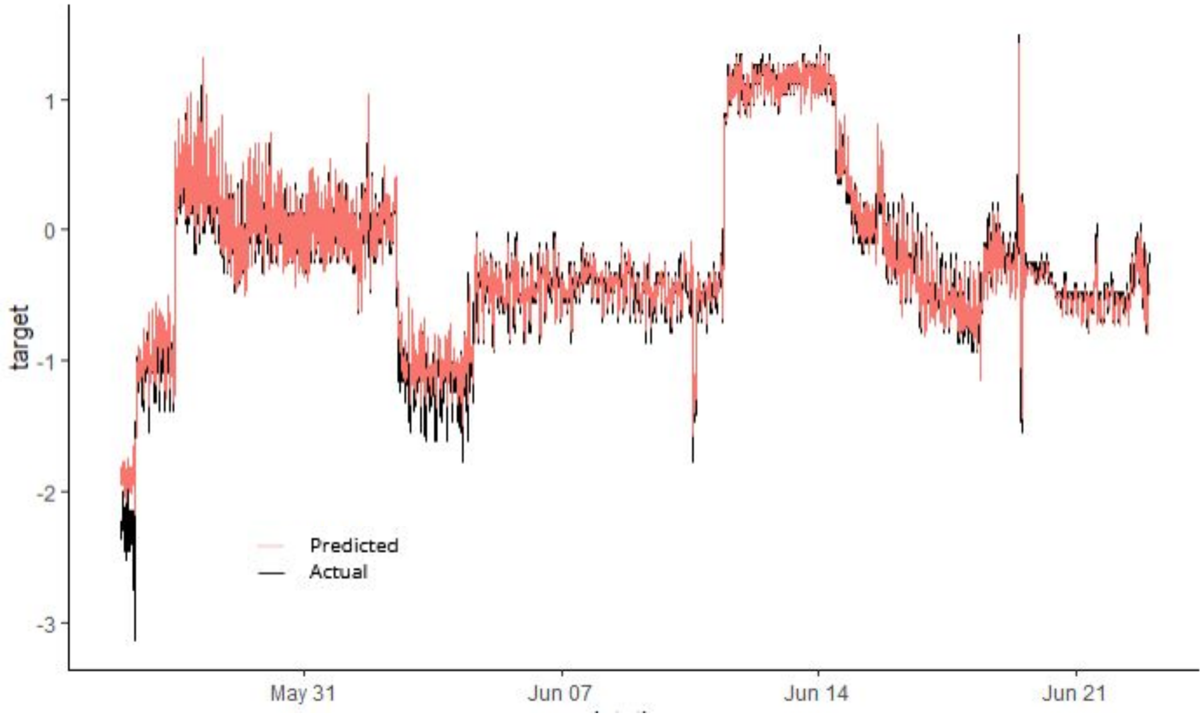



```
[1] "Total Sum of Square 4886.41"
[1] "Sum of Residual Square 240.55"
[1] "Root mean sqaure error 0.1"
```



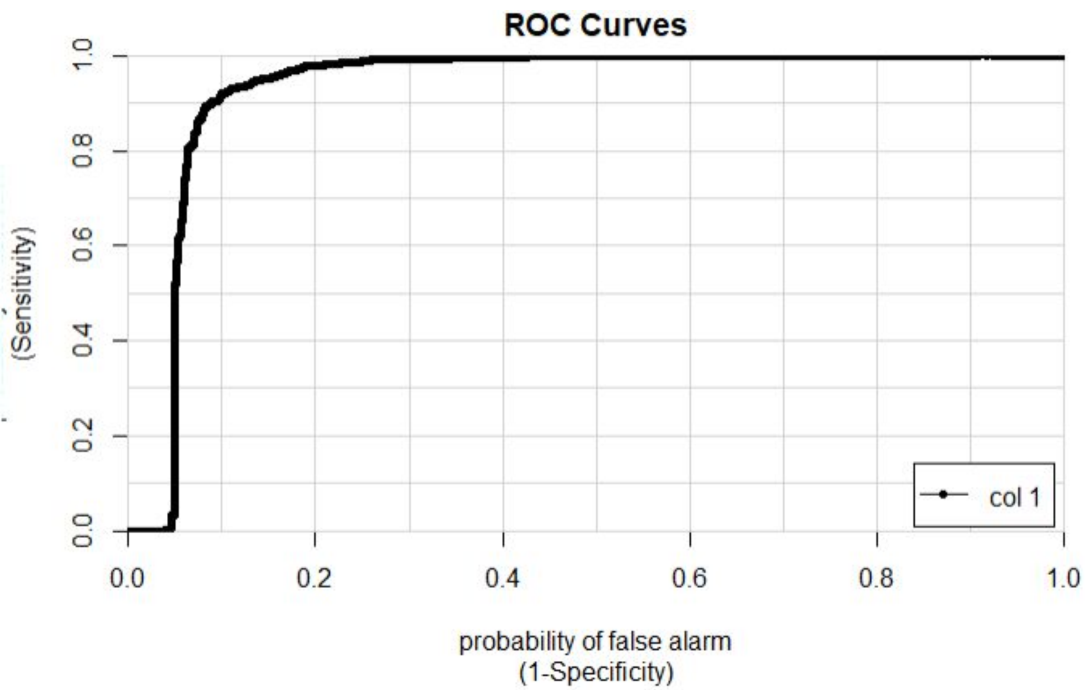

```
Accuracy : 0.9935
95% CI : (0.9893, 0.9964)
No Information Rate : 0.9905
P-Value [Acc > NIR] : 0.07591

Kappa : 0.7027

Mcnemar's Test P-Value : 0.12134

Sensitivity : 0.9952
Specificity : 0.8182
Pos Pred Value : 0.9982
Neg Pred Value : 0.6207
Prevalence : 0.9905
Detection Rate : 0.9857
Detection Prevalence : 0.9874
Balanced Accuracy : 0.9067

'Positive' Class : 0
```






```
                  Accuracy : 0.9277
                    95% CI : (0.9173, 0.9373)
       No Information Rate : 0.6186
       P-Value [Acc > NIR] : < 2.2e-16

                     Kappa : 0.8451

    Mcnemar's Test P-Value : 1.189e-05

               Sensitivity : 0.8750
               Specificity : 0.9603
            Pos Pred Value : 0.9314
            Neg Pred Value : 0.9257
                Prevalence : 0.3814
            Detection Rate : 0.3337
      Detection Prevalence : 0.3583
         Balanced Accuracy : 0.9176

          'Positive' Class : 0
```

```
"Total Sum of Square 1105.79"
"Sum of Residual Square 69.79"
"Root mean sqaure error 0.03"
```



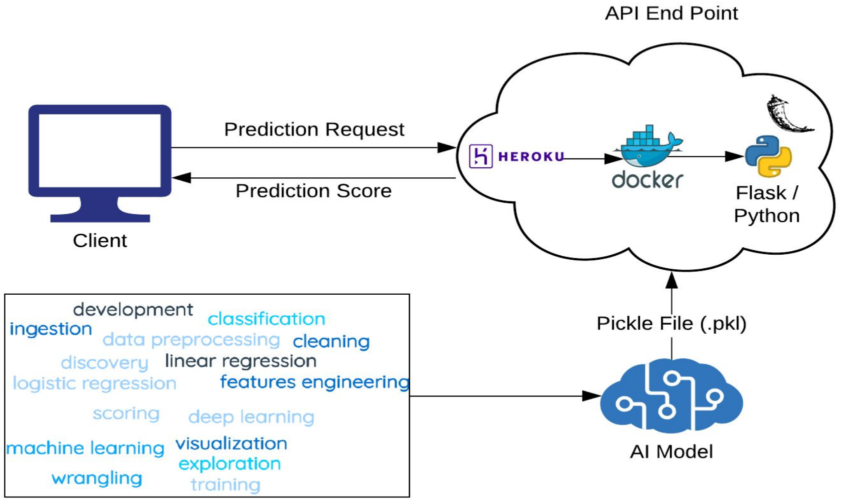
Why XGBoost?

- Parallel Computing
- Regularization
- Enabled Cross Validation(No external packages such as caret and mlr to obtain CV results)
- Tree Pruning
- Missing Values

Model Deployment

The workflow can be broken down into following basic steps:

1. Training a machine learning model on a local system.
2. Wrapping the inference logic into a flask application.
3. Using docker to containerize the flask application.
4. Hosting the docker container on an AWS ec2 instance and consuming the web-service.



There are multiple factors to consider when determining how to deploy a machine learning model. These factors include:

- how frequently predictions should be generated
- whether predictions should be generated for a single instance at a time or a batch of instances
- the number of applications that will access the model
- the latency requirements of these applications

Thanks for watching

Team 5

Bhilare,Samira

Ismail,Anjum

McInerney,Niall

Wang,Lijin