

DATA621_Homework4_JR

Jeyaraman Ramalingam

5/5/2021

Contents

Overview	2
Data Exploration	2
Wine Training Data	2
Sample	2
Input Dataset Summaries	3
Missing Data Check	3
Wine Evaluation Data	4
Sample	4
Input Dataset Summaries	4
Missing Data Check	5
Findings	6
Data Preparation	6
Training Data - Missing Data Re-test	6
Training Data - Summary	7
Training Data - Histograms	8
Training Data - Box Plots	9
Training Data - Skewness Report	9
Training Data - Correlation Report	10
Evaluation Data - Missing Data Re-test	10
Evaluation Data - Summary	11
Evaluation Data - Histograms	12
Evaluation Data - Box Plots	13
Evaluation Data - Skewness Report	13
Evaluation Data - Correlation Report	14
Data Models	14
Model Preparation	14

Poisson Regression Model	14
Poisson Regression Model Prediction Metrics	15
Zero Inflated Poisson	16
Vuong Test	17
Zero Inflated Poisson Prediction Metrics	17
Negative Binomial	18
Negative Binomial Prediction Metrics	19
Linear Regression Model (All Variables)	20
Linear Regression (All Variables) Prediction Metrics	20
Linear Regression Model (STARS)	21
Linear Regression (STARS) Prediction Metrics	21
Step Wise Linear Regression Model	22
Stepwise Linear Regression (STARS) Prediction Metrics	23
Linear Regression (Derived Variable)	23
Linear Regression (Derived Variables) Prediction Metrics	24
Model Selection	24
Evaluation Data Prediction	25
Conclusion and Output	25

Overview

In this homework assignment, you will explore, analyze and model a data set containing information on approximately 12,000 commercially available wines. The variables are mostly related to the chemical properties of the wine being sold. The response variable is the number of sample cases of wine that were purchased by wine distribution companies after sampling a wine. These cases would be used to provide tasting samples to restaurants and wine stores around the United States. The more sample cases purchased, the more likely is a wine to be sold at a high end restaurant. A large wine manufacturer is studying the data in order to predict the number of wine cases ordered based upon the wine characteristics. If the wine manufacturer can predict the number of cases, then that manufacturer will be able to adjust their wine offering to maximize sales.

Data Exploration

Wine Training Data

	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide
Sample	3	3.2	1.160	-0.98	54.2	-0.567	NA	NA
	3	4.5	0.160	-0.81	26.1	-0.425	15	15
	5	7.1	2.640	-0.88	14.8	0.037	214	214
	3	5.7	0.385	0.04	18.8	-0.425	22	22
	4	8.0	0.330	-1.26	9.4	NA	-167	-167
	0	11.3	0.320	0.59	2.2	0.556	-37	-37

Input Dataset Summaries

```

##      TARGET      FixedAcidity      VolatileAcidity      CitricAcid
##  Min.   :0.000  Min.   :-18.100  Min.   :-2.7900  Min.   :-3.2400
##  1st Qu.:2.000  1st Qu.: 5.200  1st Qu.: 0.1300  1st Qu.: 0.0300
##  Median :3.000  Median : 6.900  Median : 0.2800  Median : 0.3100
##  Mean   :3.029  Mean   : 7.076  Mean   : 0.3241  Mean   : 0.3084
##  3rd Qu.:4.000  3rd Qu.: 9.500  3rd Qu.: 0.6400  3rd Qu.: 0.5800
##  Max.   :8.000  Max.   :34.400  Max.   : 3.6800  Max.   : 3.8600
##
##      ResidualSugar      Chlorides      FreeSulfurDioxide TotalSulfurDioxide
##  Min.   :-127.800  Min.   :-1.1710  Min.   :-555.00  Min.   :-823.0
##  1st Qu.: -2.000  1st Qu.: -0.0310  1st Qu.:  0.00  1st Qu.:  27.0
##  Median :  3.900  Median :  0.0460  Median :  30.00  Median : 123.0
##  Mean   :  5.419  Mean   :  0.0548  Mean   :  30.85  Mean   : 120.7
##  3rd Qu.: 15.900  3rd Qu.:  0.1530  3rd Qu.:  70.00  3rd Qu.: 208.0
##  Max.   :141.150  Max.   :  1.3510  Max.   : 623.00  Max.   :1057.0
##  NA's   :616       NA's   :638     NA's   :647     NA's   :682
##      Density          pH          Sulphates      Alcohol
##  Min.   :0.8881  Min.   :0.480  Min.   :-3.1300  Min.   :-4.70
##  1st Qu.:0.9877  1st Qu.:2.960  1st Qu.: 0.2800  1st Qu.: 9.00
##  Median :0.9945  Median :3.200  Median : 0.5000  Median :10.40
##  Mean   :0.9942  Mean   :3.208  Mean   : 0.5271  Mean   :10.49
##  3rd Qu.:1.0005  3rd Qu.:3.470  3rd Qu.: 0.8600  3rd Qu.:12.40
##  Max.   :1.0992  Max.   :6.130  Max.   : 4.2400  Max.   :26.50
##  NA's   :395       NA's   :1210    NA's   :653
##      LabelAppeal      AcidIndex      STARS          INDEX
##  Min.   :-2.000000  Min.   : 4.000  Min.   :1.000  Min.   : 1
##  1st Qu.: -1.000000  1st Qu.:  7.000  1st Qu.:1.000  1st Qu.: 4038
##  Median :  0.000000  Median :  8.000  Median :2.000  Median : 8110
##  Mean   : -0.009066  Mean   :  7.773  Mean   :2.042  Mean   : 8070
##  3rd Qu.:  1.000000  3rd Qu.:  8.000  3rd Qu.:3.000  3rd Qu.:12106
##  Max.   :  2.000000  Max.   :17.000  Max.   :4.000  Max.   :16129
##  NA's   :3359

```

Missing Data Check

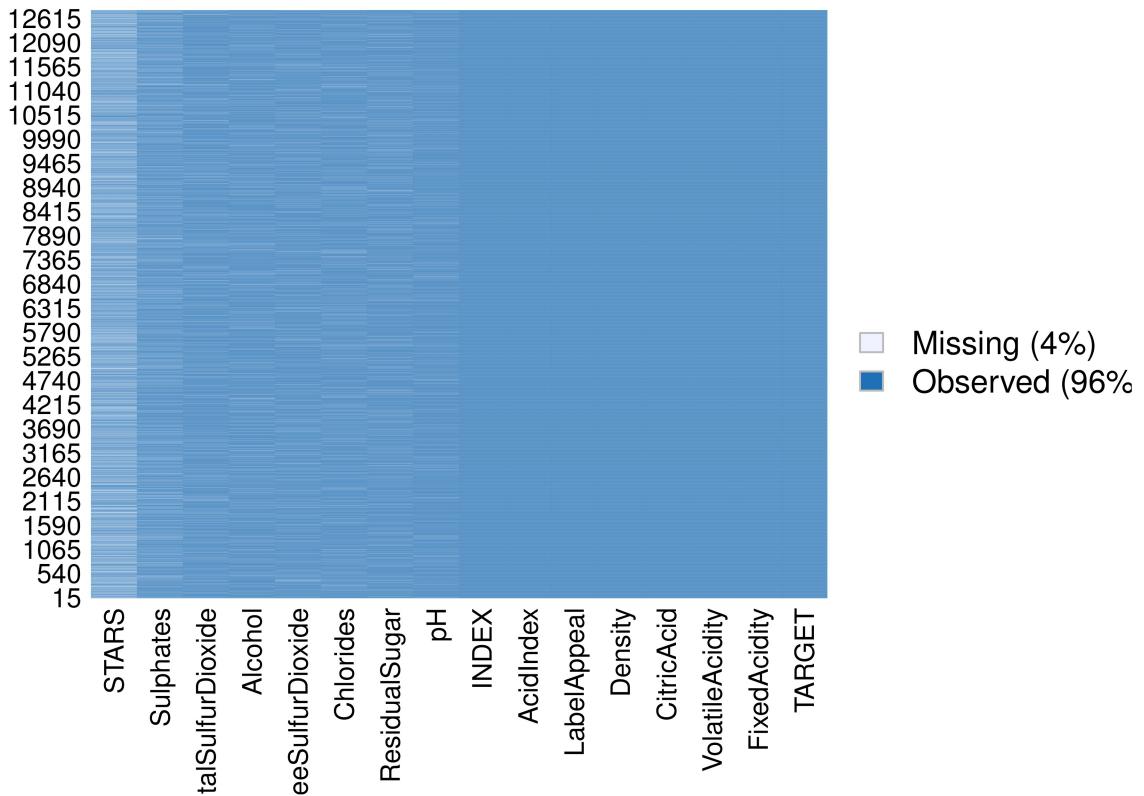
```

##      TARGET      FixedAcidity      VolatileAcidity      CitricAcid
##  0          0          0          0          0
##      ResidualSugar      Chlorides      FreeSulfurDioxide TotalSulfurDioxide
##  616        638        647        682
##      Density          pH          Sulphates      Alcohol
##  0          395        1210        653
##      LabelAppeal      AcidIndex      STARS          INDEX
##  0          0          3359        0

```

INDEX	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	To
3	NA	5.4	-0.860	0.27	-10.7	0.092	23	
9	NA	12.4	0.385	-0.76	-19.7	1.169	-37	
10	NA	7.2	1.750	0.17	-33.0	0.065	9	
18	NA	6.2	0.100	1.80	1.0	-0.179	104	
21	NA	11.4	0.210	0.28	1.2	0.038	70	
30	NA	17.6	0.040	-1.15	1.4	0.535	-250	

Missing Values



Wine Evaluation Data

Sample

Input Dataset Summaries

```

##      INDEX      TARGET      FixedAcidity      VolatileAcidity
##  Min.   : 3  Mode:logical  Min.   :-18.200  Min.   :-2.8300
##  1st Qu.: 4018 NA's:3335  1st Qu.: 5.200   1st Qu.: 0.0800
##  Median : 7906                           Median : 6.900   Median : 0.2800
##  Mean   : 8048                           Mean   : 6.864   Mean   : 0.3103
##  3rd Qu.:12061                          3rd Qu.: 9.000   3rd Qu.: 0.6300
##  Max.   :16130                          Max.   : 33.500  Max.   : 3.6100
##
##      CitricAcid      ResidualSugar      Chlorides      FreeSulfurDioxide

```

```

## Min.   :-3.1200  Min.   :-128.300  Min.   :-1.15000  Min.   :-563.00
## 1st Qu.: 0.0000  1st Qu.: -2.600  1st Qu.: 0.01600  1st Qu.: 3.00
## Median : 0.3100  Median : 3.600  Median : 0.04700  Median : 30.00
## Mean   : 0.3124  Mean   : 5.319  Mean   : 0.06143  Mean   : 34.95
## 3rd Qu.: 0.6050  3rd Qu.: 17.200  3rd Qu.: 0.17100  3rd Qu.: 79.25
## Max.   : 3.7600  Max.   : 145.400  Max.   : 1.26300  Max.   : 617.00
## NA's    :168      NA's    :138      NA's    :152
## TotalSulfurDioxide Density          pH                 Sulphates
## Min.   :-769.00   Min.   :0.8898  Min.   :0.600  Min.   :-3.0700
## 1st Qu.: 27.25    1st Qu.:0.9883  1st Qu.:2.980  1st Qu.: 0.3300
## Median : 124.00   Median :0.9946  Median :3.210  Median : 0.5000
## Mean   : 123.41   Mean   :0.9947  Mean   :3.237  Mean   : 0.5346
## 3rd Qu.: 210.00   3rd Qu.:1.0005  3rd Qu.:3.490  3rd Qu.: 0.8200
## Max.   :1004.00   Max.   :1.0998  Max.   :6.210  Max.   : 4.1800
## NA's    :157      NA's    :104      NA's    :310
## Alcohol     LabelAppeal       AcidIndex        STARS
## Min.   :-4.20    Min.   :-2.00000  Min.   : 5.000  Min.   :1.00
## 1st Qu.: 9.00    1st Qu.: -1.00000 1st Qu.: 7.000  1st Qu.:1.00
## Median :10.40    Median : 0.00000  Median : 8.000  Median :2.00
## Mean   :10.58    Mean   : 0.01349  Mean   : 7.748  Mean   :2.04
## 3rd Qu.:12.50    3rd Qu.: 1.00000  3rd Qu.: 8.000  3rd Qu.:3.00
## Max.   :25.60    Max.   : 2.00000  Max.   :17.000  Max.   :4.00
## NA's    :185      NA's    :841

```

Missing Data Check

	INDEX	TARGET	FixedAcidity	VolatileAcidity
##	0	3335	0	0
##	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide
##	0	168	138	152
##	TotalSulfurDioxide	Density	pH	Sulphates
##	157	0	104	310
##	Alcohol	LabelAppeal	AcidIndex	STARS
##	185	0	0	841



Findings

The findings from Data Exploration on Training and Evaluation dataset are below.

1. Imputation needs to be done for the missing values.

We will perform all of these exercises in the Data Preparation step.

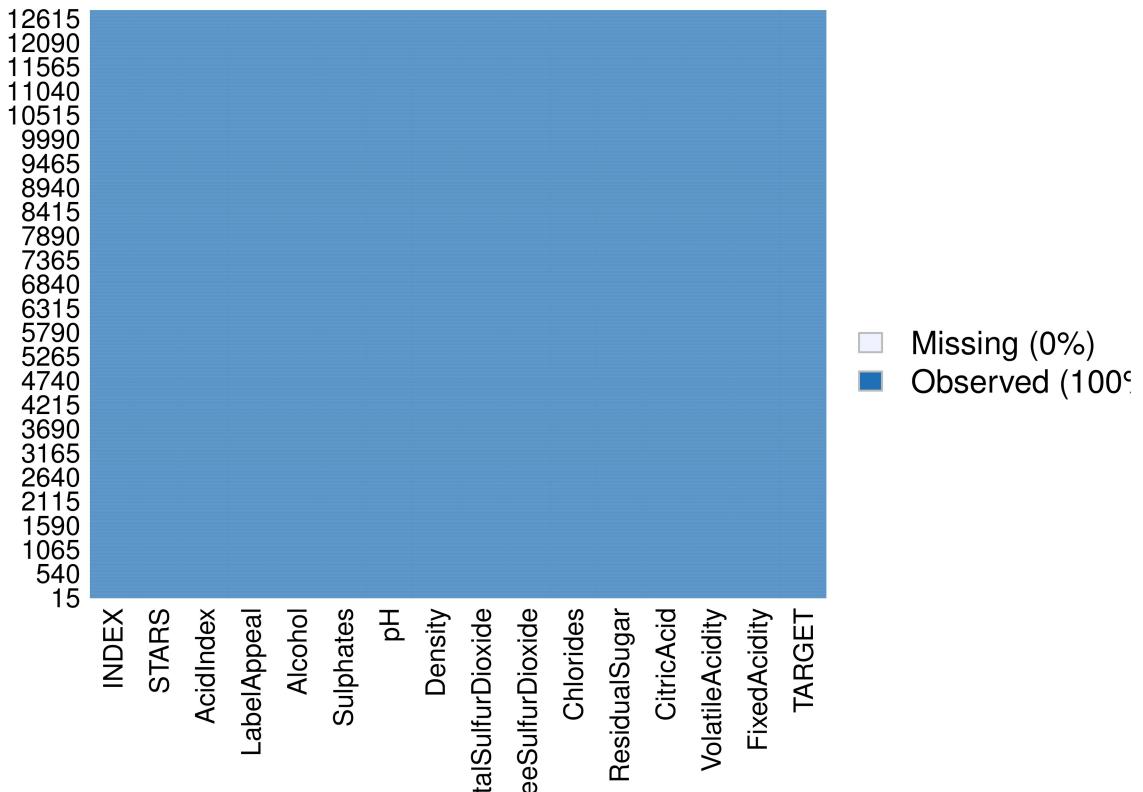
Data Preparation

Training Data - Missing Data Re-test

```
##          TARGET      FixedAcidity      VolatileAcidity      CitricAcid
##          0                  0                  0                  0
## ResidualSugar      Chlorides      FreeSulfurDioxide TotalSulfurDioxide
##          0                  0                  0                  0
##          Density      pH      Sulphates      Alcohol
##          0                  0                  0                  0
## LabelAppeal      AcidIndex      STARS      INDEX
##          0                  0                  0                  0
```

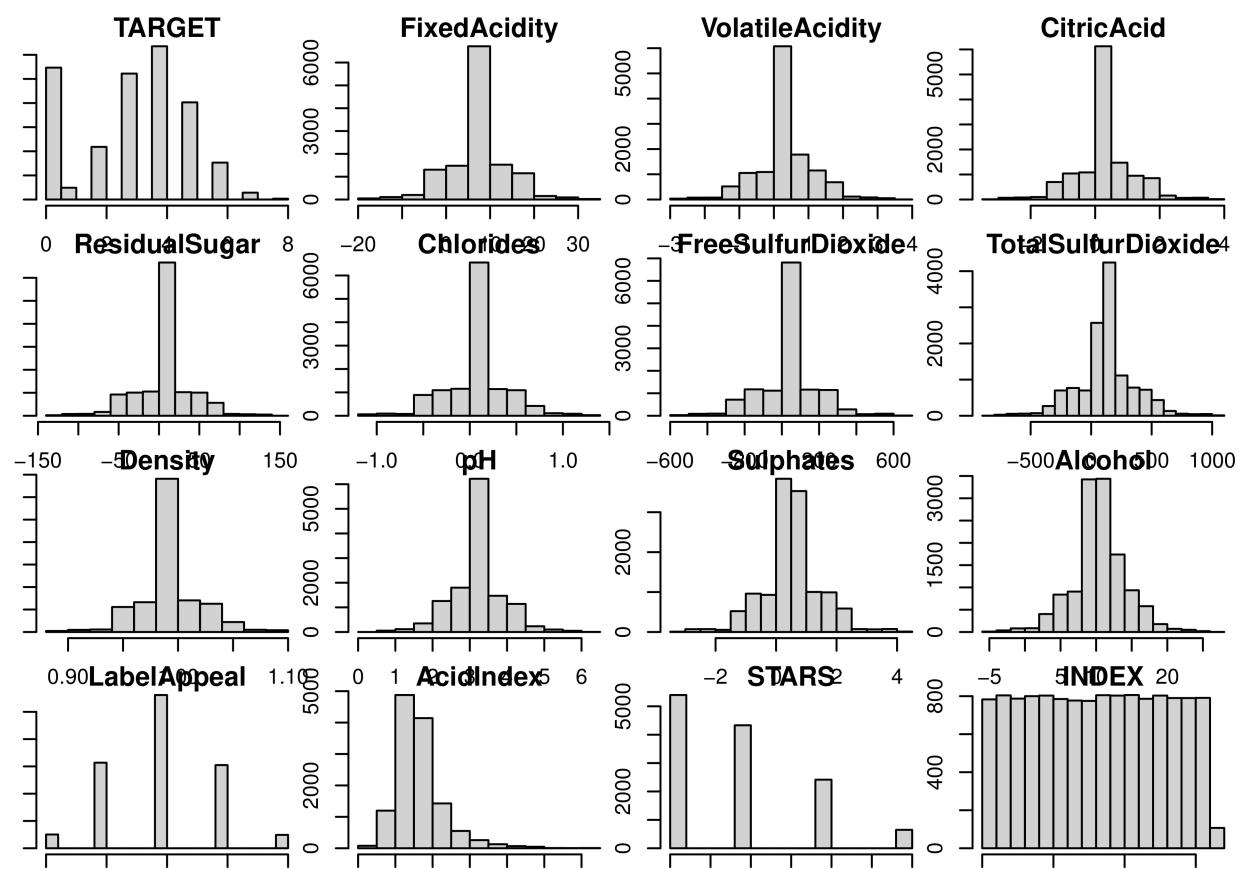
TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulf
Min. :0.000	Min. :-18.100	Min. :-2.7900	Min. :-3.2400	Min. :-127.800	Min. :-1.17100	Min. :-5
1st Qu.:2.000	1st Qu.: 5.200	1st Qu.: 0.1300	1st Qu.: 0.0300	1st Qu.: -2.800	1st Qu.:-0.03100	1st Qu.:
Median :3.000	Median : 6.900	Median : 0.2800	Median : 0.3100	Median : 3.750	Median : 0.04600	Median :
Mean :3.029	Mean : 7.076	Mean : 0.3241	Mean : 0.3084	Mean : 5.175	Mean : 0.05496	Mean :
3rd Qu.:4.000	3rd Qu.: 9.500	3rd Qu.: 0.6400	3rd Qu.: 0.5800	3rd Qu.: 15.600	3rd Qu.: 0.15200	3rd Qu.:
Max. :8.000	Max. : 34.400	Max. : 3.6800	Max. : 3.8600	Max. : 141.150	Max. : 1.35100	Max. : 6

Missing Values

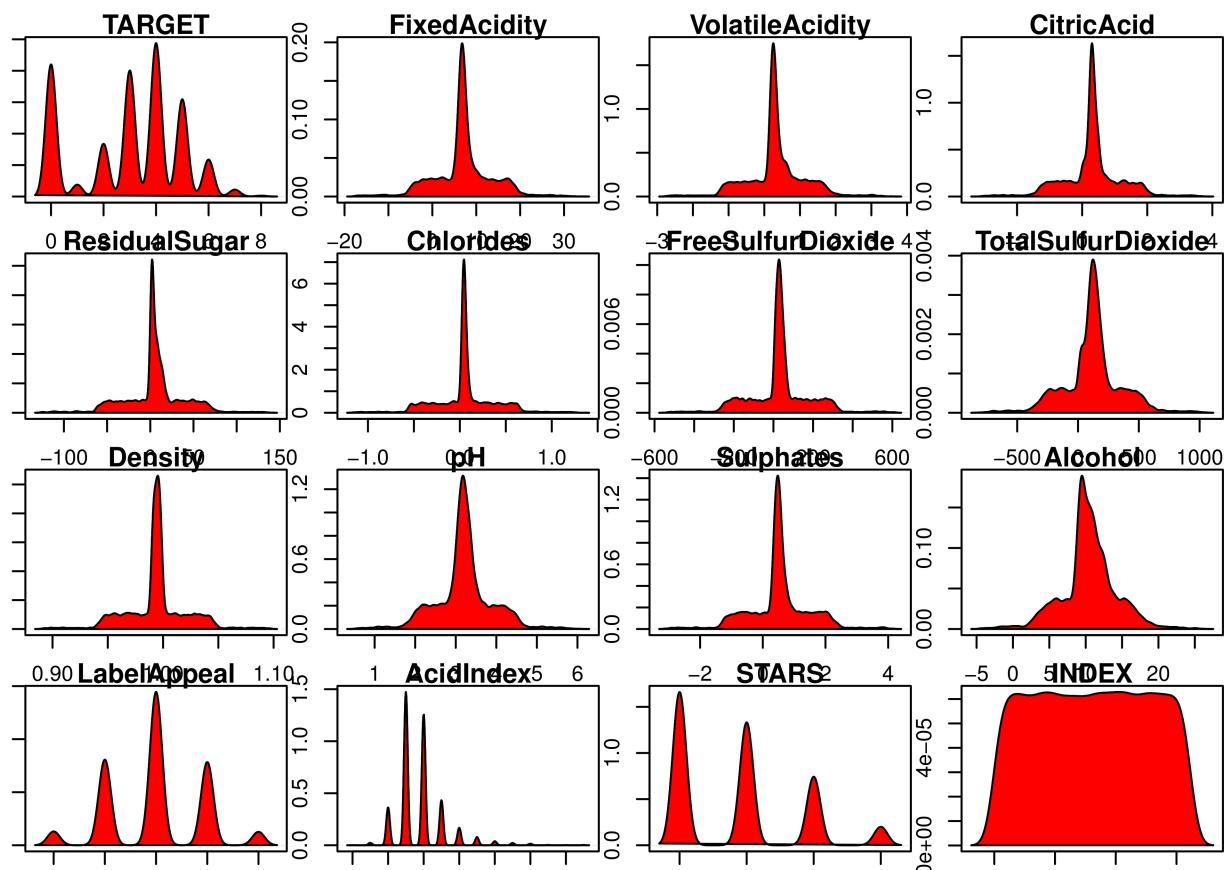


Training Data - Summary

Training Data - Histograms



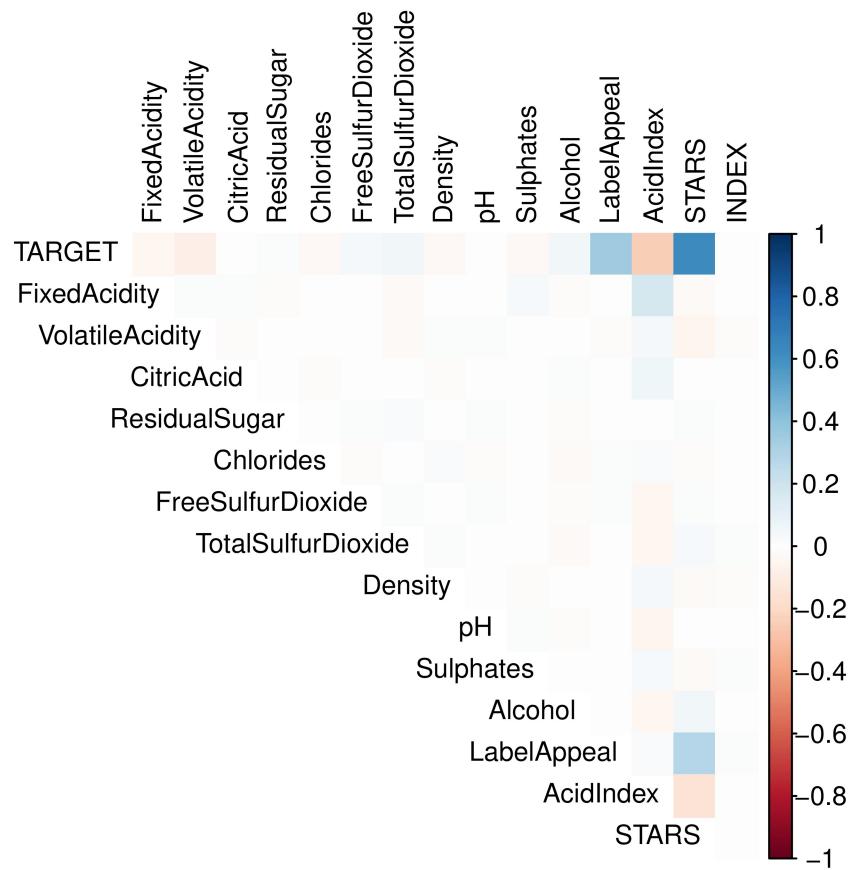
Training Data - Box Plots



Training Data - Skewness Report

```
##          TARGET      FixedAcidity    VolatileAcidity    CitricAcid
## -0.326301039 -0.022585961  0.020379965 -0.050307040
## ResidualSugar   Chlorides    FreeSulfurDioxide TotalSulfurDioxide
## -0.055094009  0.031981791  0.014569446 -0.009289989
##          Density       pH        Sulphates      Alcohol
## -0.018693764  0.037127896 -0.001408689 -0.036942156
## LabelAppeal     AcidIndex      STARS        INDEX
##  0.008429457  1.648495945  0.688688833 -0.003249620
```

Training Data - Correlation Report

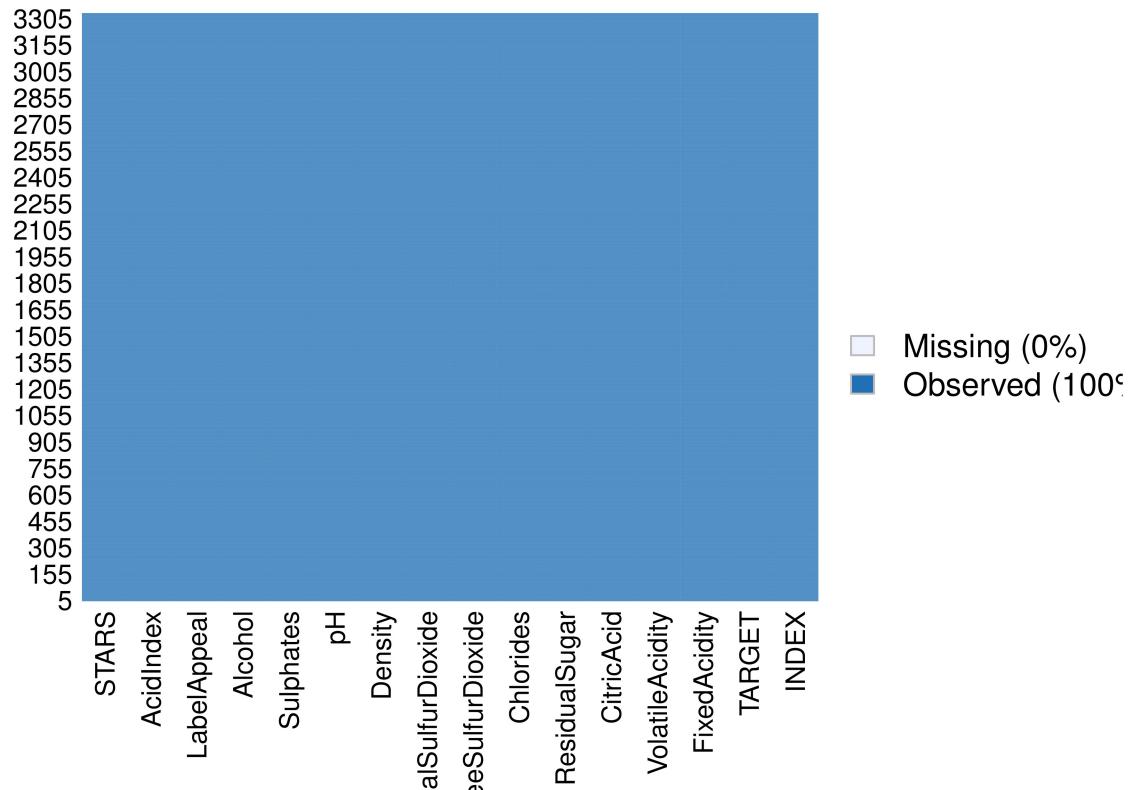


Evaluation Data - Missing Data Re-test

```
##          INDEX      TARGET  FixedAcidity  VolatileAcidity
##          0           0        0             0
## CitricAcid  ResidualSugar  Chlorides  FreeSulfurDioxide
##          0           0        0             0
## TotalSulfurDioxide  Density  pH           Sulphates
##          0           0        0             0
##          Alcohol  LabelAppeal  AcidIndex  STARS
##          0           0        0             0
```

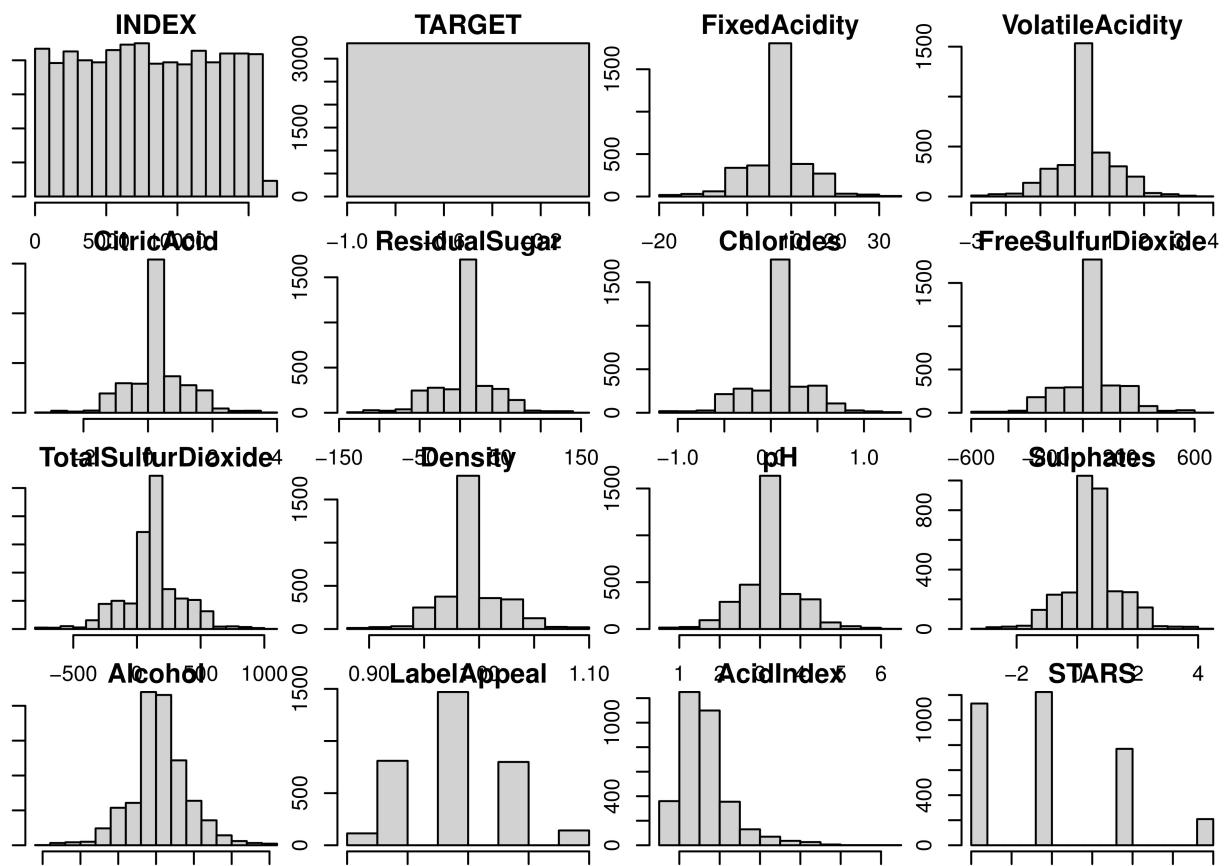
INDEX	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides
Min. : 3	Min. :0	Min. :-18.200	Min. :-2.8300	Min. :-3.1200	Min. :-128.30	Min. :-1.15000
1st Qu.: 4018	1st Qu.:0	1st Qu.: 5.200	1st Qu.: 0.0800	1st Qu.: 0.0000	1st Qu.: -3.15	1st Qu.: 0.01800
Median : 7906	Median :0	Median : 6.900	Median : 0.2800	Median : 0.3100	Median : 3.70	Median : 0.0470
Mean : 8048	Mean :0	Mean : 6.864	Mean : 0.3103	Mean : 0.3124	Mean : 5.19	Mean : 0.06097
3rd Qu.:12061	3rd Qu.:0	3rd Qu.: 9.000	3rd Qu.: 0.6300	3rd Qu.: 0.6050	3rd Qu.: 17.20	3rd Qu.: 0.16750
Max. :16130	Max. :0	Max. : 33.500	Max. : 3.6100	Max. : 3.7600	Max. : 145.40	Max. : 1.26300

Missing Values

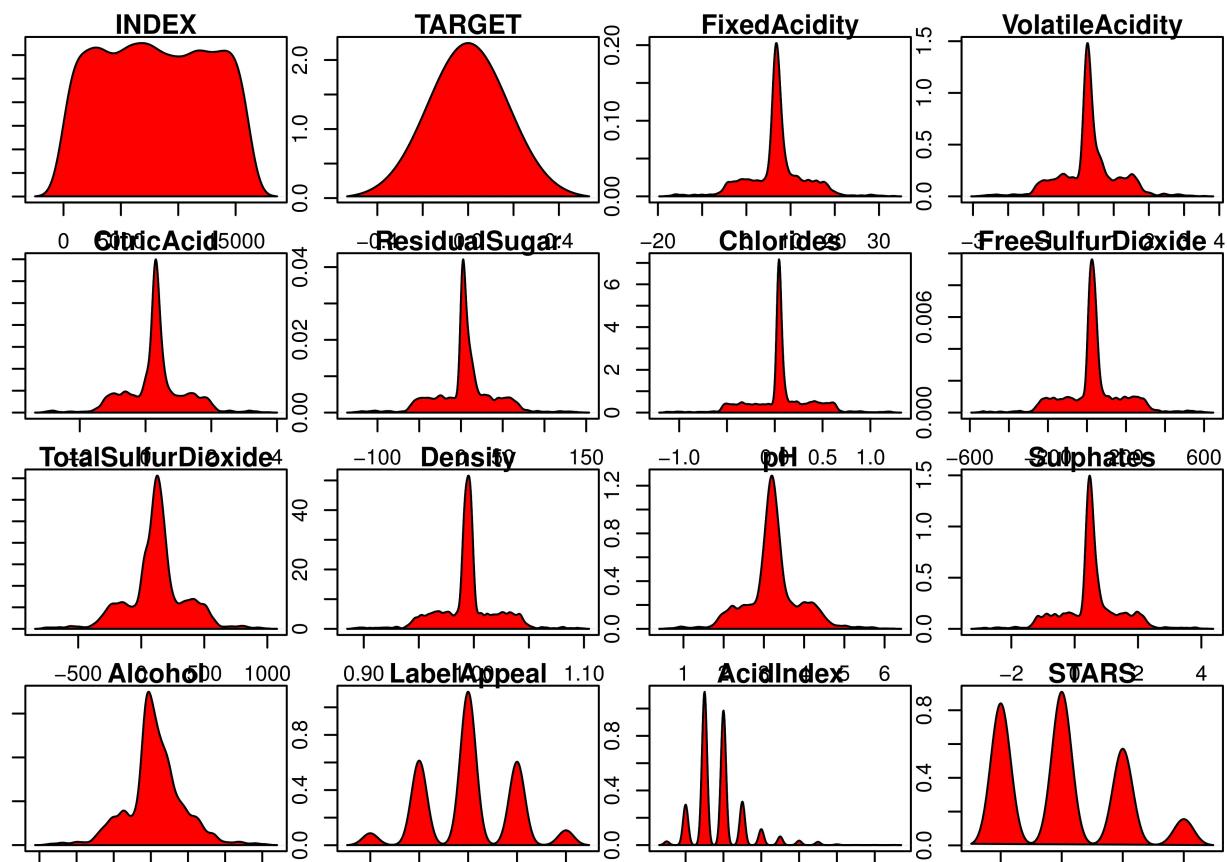


Evaluation Data - Summary

Evaluation Data - Histograms



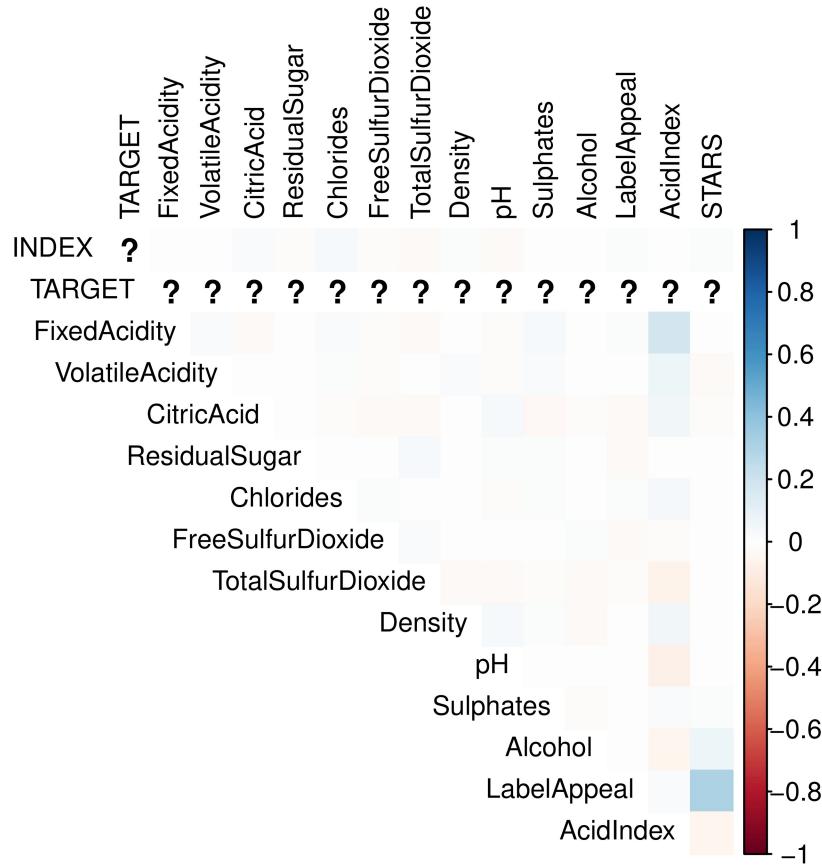
Evaluation Data - Box Plots



Evaluation Data - Skewness Report

##	INDEX	TARGET	FixedAcidity	VolatileAcidity
##	0.01246970	NaN	-0.11724599	-0.04373012
##	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide
##	-0.02848982	-0.04551615	-0.04334931	0.09591835
##	TotalSulfurDioxide	Density	pH	Sulphates
##	-0.08759696	-0.02965927	0.13813546	-0.01884956
##	Alcohol	LabelAppeal	AcidIndex	STARS
##	0.04003629	0.04548870	1.50665887	0.47249020

Evaluation Data - Correlation Report



Data Models

Model Preparation

The Training Insurance data is chosen and the train test split is created with 80% as factor. After the dataset split the plan is to create following models and predict evaluation dataset using the best model.

1. Poisson Regression - > TARGET and other variables
2. Zero Inflated Poisson - > TARGET and other variables
3. Negative Binomial - > TARGET and other variables
4. Linear Regression - > TARGET and other variables
5. Linear Regression - > TARGET and STARS
6. Step Wise Regression (Backward) -> TARGET and STARS
7. Linear Regression -> TARGET and Derived Variable

Poisson Regression Model

Poisson Regression models are best used for modeling events where the outcomes are counts. Or, more specifically, count data: discrete data with non-negative integer values that count something, like the number of times an event occurs during a given timeframe or the number of people in line at the grocery store.

##

```

## Call:
## glm(formula = TARGET ~ ., family = poisson, data = train2)
##
## Deviance Residuals:
##      Min     1Q   Median     3Q    Max 
## -3.2545 -0.6722  0.1238  0.6313  2.4180 
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 1.606e+00 2.187e-01 7.346 2.04e-13 ***
## FixedAcidity -1.672e-04 9.127e-04 -0.183 0.854634    
## VolatileAcidity -3.995e-02 7.232e-03 -5.524 3.32e-08 ***
## CitricAcid    1.432e-02 6.595e-03  2.171 0.029900 *  
## ResidualSugar -2.338e-06 1.690e-04 -0.014 0.988961    
## Chlorides     -5.118e-02 1.814e-02 -2.822 0.004773 ** 
## FreeSulfurDioxide 1.388e-04 3.853e-05  3.601 0.000317 *** 
## TotalSulfurDioxide 8.838e-05 2.462e-05  3.589 0.000332 *** 
## Density       -3.773e-01 2.145e-01 -1.759 0.078543 .  
## pH            -1.831e-02 8.365e-03 -2.189 0.028597 *  
## Sulphates     -1.249e-02 6.102e-03 -2.046 0.040746 *  
## Alcohol        2.150e-03 1.540e-03  1.396 0.162785    
## LabelAppeal    1.542e-01 6.744e-03 22.864 < 2e-16 ***
## AcidIndex     -1.016e-01 5.065e-03 -20.058 < 2e-16 *** 
## STARS          3.340e-01 6.267e-03 53.288 < 2e-16 *** 
## INDEX          -3.688e-07 1.221e-06 -0.302 0.762593    
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 18288  on 10237  degrees of freedom
## Residual deviance: 12745  on 10222  degrees of freedom
## AIC: 38332
##
## Number of Fisher Scoring iterations: 5

```

AIC of the Poisson Regression Model is 38388

Poisson Regression Model Prediction Metrics

Test dataset is used for predicting the output and the confusion matrix is used for comparing the output parameters.

```

## Confusion Matrix and Statistics
##
##           Reference
## Prediction 1 0
##           1 47 546
##           0 0 0
##
##           Accuracy : 0.0793
##                 95% CI : (0.0588, 0.104)
## No Information Rate : 0.9207

```

```

##      P-Value [Acc > NIR] : 1
##
##          Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##          Sensitivity : 1.00000
##          Specificity : 0.00000
##          Pos Pred Value : 0.07926
##          Neg Pred Value :      NaN
##          Prevalence : 0.07926
##          Detection Rate : 0.07926
##          Detection Prevalence : 1.00000
##          Balanced Accuracy : 0.50000
##
##          'Positive' Class : 1
##

```

Accuracy of the Model 1 is 7.9%

Zero Inflated Poisson

Zero-inflated poisson regression is used to model count data that has an excess of zero counts

```

##
## Call:
## zeroinfl(formula = TARGET ~ ., data = train2)
##
## Pearson residuals:
##      Min     1Q   Median     3Q    Max
## -2.28590 -0.45787  0.02647  0.43444  3.91450
##
## Count model coefficients (poisson with log link):
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.475e+00     NA     NA     NA
## FixedAcidity        4.293e-04     NA     NA     NA
## VolatileAcidity     -1.251e-02     NA     NA     NA
## CitricAcid         2.477e-04     NA     NA     NA
## ResidualSugar      -7.332e-05     NA     NA     NA
## Chlorides          -2.068e-02     NA     NA     NA
## FreeSulfurDioxide  2.787e-05     NA     NA     NA
## TotalSulfurDioxide -2.025e-05     NA     NA     NA
## Density            -3.643e-01     NA     NA     NA
## pH                 3.570e-03     NA     NA     NA
## Sulphates          -1.336e-03     NA     NA     NA
## Alcohol            7.181e-03     NA     NA     NA
## LabelAppeal        2.387e-01     NA     NA     NA
## AcidIndex          -2.005e-02     NA     NA     NA
## STARS              1.184e-01     NA     NA     NA
## INDEX              -2.885e-07     NA     NA     NA
##
## Zero-inflation model coefficients (binomial with logit link):
##                                         Estimate Std. Error z value Pr(>|z|)

```

```

## (Intercept) -2.940e+00 NA NA NA
## FixedAcidity 3.072e-03 NA NA NA
## VolatileAcidity 2.230e-01 NA NA NA
## CitricAcid -7.395e-02 NA NA NA
## ResidualSugar -2.558e-04 NA NA NA
## Chlorides 2.241e-01 NA NA NA
## FreeSulfurDioxide -9.411e-04 NA NA NA
## TotalSulfurDioxide -8.266e-04 NA NA NA
## Density 1.050e+00 NA NA NA
## pH 1.792e-01 NA NA NA
## Sulphates 1.017e-01 NA NA NA
## Alcohol 2.856e-02 NA NA NA
## LabelAppeal 6.487e-01 NA NA NA
## AcidIndex 4.652e-01 NA NA NA
## STARS -3.051e+00 NA NA NA
## INDEX 9.136e-06 NA NA NA
##
## Number of iterations in BFGS optimization: 39
## Log-likelihood: -1.681e+04 on 32 Df

```

AIC of the Zero Inflated Poisson is 38388

Vuong Test

The Vuong non-nested test is based on a comparison of the predicted probabilities of two models that do not nest. Examples include comparisons of zero-inflated count models with their non-zero-inflated analogs (e.g., zero-inflated Poisson versus ordinary Poisson, or zero-inflated negative-binomial versus ordinary negative-binomial).

```

## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
## null that the models are indistinguishable)
## -----
## Vuong z-statistic H_A p-value
## Raw -39.69360 model2 > model1 < 2.22e-16
## AIC-corrected -39.42270 model2 > model1 < 2.22e-16
## BIC-corrected -38.44288 model2 > model1 < 2.22e-16

```

As a result of Vuong test , Model 2 performs better

Zero Inflated Poisson Prediction Metrics

Test dataset is used for predicting the output and the confusion matrix is used for comparing the output parameters.

```

## Confusion Matrix and Statistics
##
## Reference
## Prediction 1 0
## 1 47 498
## 0 0 48
##
```

```

##          Accuracy : 0.1602
## 95% CI : (0.1316, 0.1922)
## No Information Rate : 0.9207
## P-Value [Acc > NIR] : 1
##
##          Kappa : 0.015
##
## McNemar's Test P-Value : <2e-16
##
##          Sensitivity : 1.00000
##          Specificity : 0.08791
## Pos Pred Value : 0.08624
## Neg Pred Value : 1.00000
##          Prevalence : 0.07926
## Detection Rate : 0.07926
## Detection Prevalence : 0.91906
## Balanced Accuracy : 0.54396
##
## 'Positive' Class : 1
##

```

Accuracy of the Model 2 is 15%

Negative Binomial

Negative binomial regression is for modeling count variables, usually for over-dispersed count outcome variables.

```

## 
## Call:
## glm.nb(formula = TARGET ~ ., data = train2, init.theta = 49164.47871,
##         link = log)
## 
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -3.2544   -0.6721    0.1238    0.6313    2.4179
## 
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)             1.606e+00  2.187e-01   7.346 2.05e-13 ***
## FixedAcidity           -1.672e-04  9.127e-04  -0.183 0.854624
## VolatileAcidity        -3.995e-02  7.232e-03  -5.524 3.32e-08 ***
## CitricAcid              1.432e-02  6.595e-03   2.171 0.029905 *
## ResidualSugar           -2.334e-06  1.690e-04  -0.014 0.988979
## Chlorides               -5.118e-02  1.814e-02  -2.822 0.004774 **
## FreeSulfurDioxide      1.388e-04  3.853e-05   3.601 0.000317 ***
## TotalSulfurDioxide     8.839e-05  2.463e-05   3.589 0.000332 ***
## Density                -3.773e-01  2.145e-01  -1.759 0.078548 .
## pH                     -1.831e-02  8.366e-03  -2.189 0.028597 *
## Sulphates              -1.249e-02  6.103e-03  -2.046 0.040748 *
## Alcohol                 2.150e-03  1.540e-03   1.396 0.162802
## LabelAppeal             1.542e-01  6.744e-03  22.864 < 2e-16 ***
## AcidIndex              -1.016e-01  5.065e-03 -20.057 < 2e-16 ***

```

```

## STARS           3.340e-01  6.268e-03  53.287  < 2e-16 ***
## INDEX          -3.689e-07  1.221e-06  -0.302  0.762576
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(49164.48) family taken to be 1)
##
## Null deviance: 18287  on 10237  degrees of freedom
## Residual deviance: 12745  on 10222  degrees of freedom
## AIC: 38334
##
## Number of Fisher Scoring iterations: 1
##
##
##             Theta:  49164
##             Std. Err.: 63187
## Warning while fitting theta: iteration limit reached
##
## 2 x log-likelihood: -38300.13

```

AIC of the Model 3 is 38390

Negative Binomial Prediction Metrics

Test dataset is used for predicting the output and the confusion matrix is used for comparing the output parameters.

```

## Confusion Matrix and Statistics
##
##             Reference
## Prediction   1   0
##             1 47 546
##             0   0   0
##
##             Accuracy : 0.0793
##                 95% CI : (0.0588, 0.104)
## No Information Rate : 0.9207
## P-Value [Acc > NIR] : 1
##
##             Kappa : 0
##
## Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.00000
##             Specificity : 0.00000
## Pos Pred Value : 0.07926
## Neg Pred Value :      NaN
## Prevalence : 0.07926
## Detection Rate : 0.07926
## Detection Prevalence : 1.00000
## Balanced Accuracy : 0.50000
##
## 'Positive' Class : 1
##

```

Linear Regression Model (All Variables)

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. ... A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable.

```
##  
## Call:  
## lm(formula = TARGET ~ ., data = train2)  
##  
## Residuals:  
##    Min     1Q Median     3Q    Max  
## -4.9112 -0.9987  0.1620  1.0255  4.0231  
##  
## Coefficients:  
##                               Estimate Std. Error t value Pr(>|t|)  
## (Intercept)            4.010e+00 5.333e-01 7.519 5.99e-14 ***  
## FixedAcidity          -9.414e-05 2.232e-03 -0.042 0.966362  
## VolatileAcidity      -1.198e-01 1.773e-02 -6.757 1.49e-11 ***  
## CitricAcid           4.091e-02 1.617e-02  2.531 0.011397 *  
## ResidualSugar        -9.005e-06 4.128e-04 -0.022 0.982595  
## Chlorides            -1.554e-01 4.404e-02 -3.530 0.000418 ***  
## FreeSulfurDioxide   3.989e-04 9.459e-05  4.218 2.49e-05 ***  
## TotalSulfurDioxide  2.414e-04 6.001e-05  4.023 5.79e-05 ***  
## Density             -1.094e+00 5.244e-01 -2.087 0.036946 *  
## pH                  -4.435e-02 2.055e-02 -2.158 0.030964 *  
## Sulphates           -3.309e-02 1.495e-02 -2.213 0.026905 *  
## Alcohol             1.069e-02 3.757e-03  2.845 0.004451 **  
## LabelAppeal         4.706e-01 1.631e-02 28.848 < 2e-16 ***  
## AcidIndex           -2.539e-01 1.098e-02 -23.114 < 2e-16 ***  
## STARS              1.144e+00 1.660e-02 68.948 < 2e-16 ***  
## INDEX              -1.865e-06 2.995e-06 -0.623 0.533538  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 1.41 on 10222 degrees of freedom  
## Multiple R-squared:  0.4645, Adjusted R-squared:  0.4637  
## F-statistic:  591 on 15 and 10222 DF,  p-value: < 2.2e-16
```

Linear Regression (All Variables) Prediction Metrics

Test dataset is used for predicting the output and the confusion matrix is used for comparing the output parameters.

```
## Confusion Matrix and Statistics  
##  
##             Reference  
## Prediction  1    0  
##           1  46 531  
##           0   1  15  
##  
##                 Accuracy : 0.1029  
##                 95% CI : (0.0796, 0.1302)
```

```

##      No Information Rate : 0.9207
##      P-Value [Acc > NIR] : 1
##
##              Kappa : 0.001
##
## McNemar's Test P-Value : <2e-16
##
##      Sensitivity : 0.97872
##      Specificity : 0.02747
##      Pos Pred Value : 0.07972
##      Neg Pred Value : 0.93750
##      Prevalence : 0.07926
##      Detection Rate : 0.07757
##      Detection Prevalence : 0.97302
##      Balanced Accuracy : 0.50310
##
##      'Positive' Class : 1
##

```

Linear Regression Model (STARS)

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. ... A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable.

```

##
## Call:
## lm(formula = TARGET ~ STARS, data = train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2040 -1.5506  0.1425  1.1425  4.1425
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.51090   0.03461   14.76  <2e-16 ***
## STARS       1.34657   0.01671   80.56  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.507 on 10236 degrees of freedom
## Multiple R-squared:  0.388, Adjusted R-squared:  0.388
## F-statistic: 6491 on 1 and 10236 DF, p-value: < 2.2e-16

```

Linear Regression (STARS) Prediction Metrics

Test dataset is used for predicting the output and the confusion matrix is used for comparing the output parameters.

```

## Confusion Matrix and Statistics
##
##      Reference

```

```

## Prediction 1 0
##      1 47 546
##      0 0 0
##
##          Accuracy : 0.0793
##      95% CI : (0.0588, 0.104)
##      No Information Rate : 0.9207
##      P-Value [Acc > NIR] : 1
##
##          Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##          Sensitivity : 1.00000
##          Specificity : 0.00000
##          Pos Pred Value : 0.07926
##          Neg Pred Value :      NaN
##          Prevalence : 0.07926
##          Detection Rate : 0.07926
##          Detection Prevalence : 1.00000
##          Balanced Accuracy : 0.50000
##
##          'Positive' Class : 1
##

```

Step Wise Linear Regression Model

The stepwise regression takes the predictors and adds/removes based on the significance of the predictors. At first the model is run with 0 predictors and the predictors are added in sequence based on its significance. Since the model chooses the predictors by itself all predictors (explanator variables) are considered for model against target variable.

Adding to the stepwise regression we are also considering the transformed dataset with new variables derived from the existing variables.

```

## Start: AIC=8392.95
## TARGET ~ STARS
##
##          Df Sum of Sq   RSS   AIC
## <none>            23231  8393
## - STARS 1     14731 37962 13419

##
## Call:
## lm(formula = TARGET ~ STARS, data = train2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.2040 -1.5506  0.1425  1.1425  4.1425
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.51090   0.03461   14.76   <2e-16 ***

```

```

## STARS      1.34657   0.01671   80.56   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.507 on 10236 degrees of freedom
## Multiple R-squared:  0.388, Adjusted R-squared:  0.388
## F-statistic:  6491 on 1 and 10236 DF, p-value: < 2.2e-16

```

Stepwise Linear Regression (STARS) Prediction Metrics

Test dataset is used for predicting the output and the confusion matrix is used for comparing the output parameters.

```

## Confusion Matrix and Statistics
##
##             Reference
## Prediction    1    0
##                 1   47 546
##                 0    0    0
##
##             Accuracy : 0.0793
##                 95% CI : (0.0588, 0.104)
## No Information Rate : 0.9207
## P-Value [Acc > NIR] : 1
##
##             Kappa : 0
##
## McNemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.00000
##             Specificity : 0.00000
## Pos Pred Value : 0.07926
## Neg Pred Value :      NaN
## Prevalence : 0.07926
## Detection Rate : 0.07926
## Detection Prevalence : 1.00000
## Balanced Accuracy : 0.50000
##
## 'Positive' Class : 1
##

```

Linear Regression (Derived Variable)

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. ... A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable.

```

##
## Call:
## lm(formula = TARGET ~ totalAcid, data = train2)
##
## Residuals:

```

```

##      Min     1Q Median     3Q    Max
## -3.4628 -1.1005  0.1079  1.1379  5.1611
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.162836  0.029466 107.34 < 2e-16 ***
## totalAcid   -0.017546  0.002929  -5.99 2.18e-09 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.922 on 10236 degrees of freedom
## Multiple R-squared: 0.003493, Adjusted R-squared: 0.003395
## F-statistic: 35.87 on 1 and 10236 DF, p-value: 2.175e-09

```

Linear Regression (Derived Variables) Prediction Metrics

Test dataset is used for predicting the output and the confusion matrix is used for comparing the output parameters.

```

## Confusion Matrix and Statistics
##
##             Reference
## Prediction 1 0
##           1 47 546
##           0 0 0
##
##           Accuracy : 0.0793
##           95% CI : (0.0588, 0.104)
## No Information Rate : 0.9207
## P-Value [Acc > NIR] : 1
##
##           Kappa : 0
##
## McNemar's Test P-Value : <2e-16
##
##           Sensitivity : 1.00000
##           Specificity : 0.00000
## Pos Pred Value : 0.07926
## Neg Pred Value :      NaN
##           Prevalence : 0.07926
## Detection Rate : 0.07926
## Detection Prevalence : 1.00000
## Balanced Accuracy : 0.50000
##
## 'Positive' Class : 1
##
```

Accuracy of the Model 3 is 78.3%

Model Selection

While comparing all models based on AIC, Accuracy values we can safely say Model 2 performs better.

INDEX	TARGET	FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	To
3	4.350038	5.4	-0.860	0.27	-10.7	0.092	23	
9	3.198689	12.4	0.385	-0.76	-19.7	1.169	-37	
10	1.811528	7.2	1.750	0.17	-33.0	0.065	9	
18	1.811045	6.2	0.100	1.80	1.0	-0.179	104	
21	2.318379	11.4	0.210	0.28	1.2	0.038	70	
30	6.190712	17.6	0.040	-1.15	1.4	0.535	-250	

Evaluation Data Prediction

The evaluation dataset is used for prediction purposes.

Conclusion and Output

```
## NULL
```

Overall we found that Model 2 (Zero Inflated Poisson) performs better in predicting the TARGET value for the evaluation data set.