Wine sales prediction

Sri Seshadri

11/11/2017

Contents

1. Introduction

A large wine manufacturer is studying data collected on 12,000 commercially available wines, with a goal to predict the number of cases ordered based upon the characteristics. The manufacturer intends to adjust the winde offerings based on the findings. The data collected is related to chemical properties of wine with response variable being the number of sample cases sold to distribution companies. the data also includes features like review stars provided by the tasters and label appeal.

2. Exploratory Data analysis

The below table shows the summary statistics of the data. It is seen that there are missing data. "Stars" has the most missing values. The missing values mostly correspond to NO (0) cases sold. Which is most likely due to lack of opportunity to sample because of none sold.

Table 1: Summary Stats and missing values

	min	Q1	median	Q3	max	mean	sd	n	missing
INDEX	1.00	4037.50	8110.00	12106.50	16129.00	8069.98	4656.91	12795	0
TARGET	0.00	2.00	3.00	4.00	8.00	3.03	1.93	12795	0
FixedAcidity	-18.10	5.20	6.90	9.50	34.40	7.08	6.32	12795	0
VolatileAcidity	-2.79	0.13	0.28	0.64	3.68	0.32	0.78	12795	0
CitricAcid	-3.24	0.03	0.31	0.58	3.86	0.31	0.86	12795	0
ResidualSugar	-127.80	-2.00	3.90	15.90	141.15	5.42	33.75	12179	616
Chlorides	-1.17	-0.03	0.05	0.15	1.35	0.05	0.32	12157	638
FreeSulfurDioxide	-555.00	0.00	30.00	70.00	623.00	30.85	148.71	12148	647
TotalSulfurDioxide	-823.00	27.00	123.00	208.00	1057.00	120.71	231.91	12113	682
Density	0.89	0.99	0.99	1.00	1.10	0.99	0.03	12795	0
pH	0.48	2.96	3.20	3.47	6.13	3.21	0.68	12400	395
Sulphates	-3.13	0.28	0.50	0.86	4.24	0.53	0.93	11585	1210
Alcohol	-4.70	9.00	10.40	12.40	26.50	10.49	3.73	12142	653
LabelAppeal	-2.00	-1.00	0.00	1.00	2.00	-0.01	0.89	12795	0
AcidIndex	4.00	7.00	8.00	8.00	17.00	7.77	1.32	12795	0
STARS	1.00	1.00	2.00	3.00	4.00	2.04	0.90	9436	3359

Figure 2 shows the historgram of features in the data. The chemical properties of wines appear to share an identical distribution with peaks closer to zero. This mau be likely to some standardization done to the day. The variable "TARGET" looks to be poisson or negative biniomially distributed with inflation at 0.

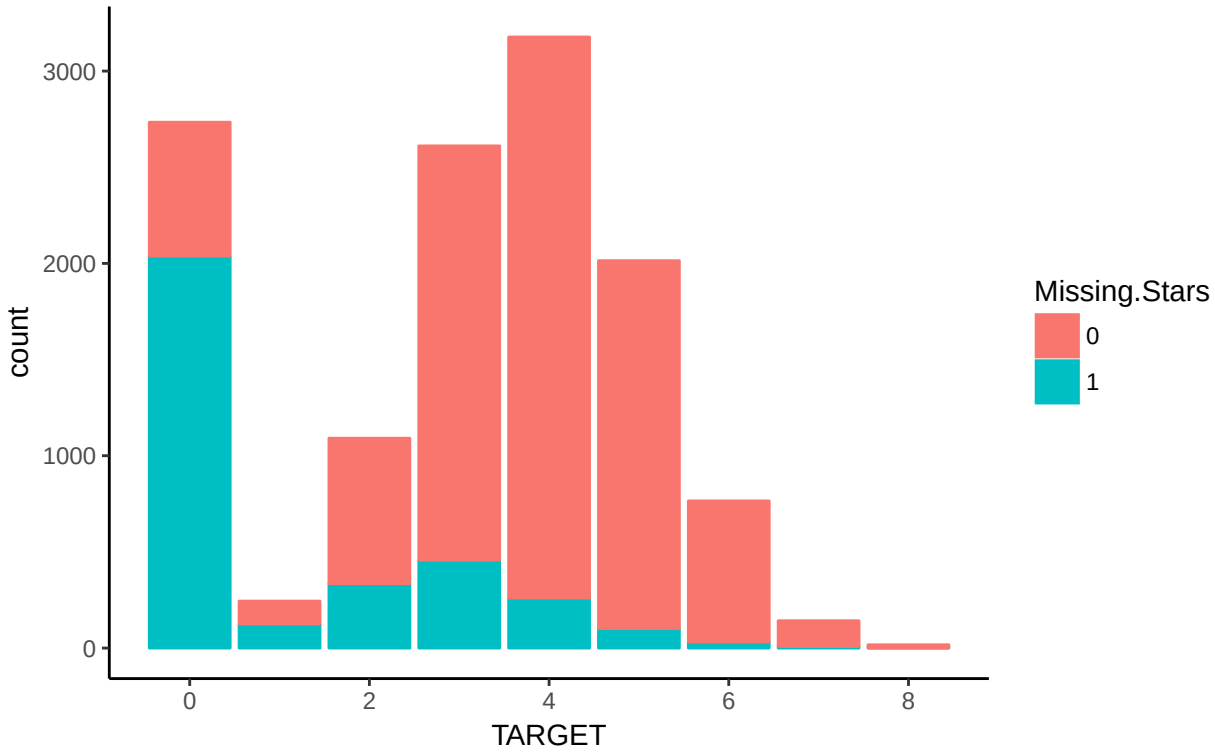


Figure 1: Missing STARS values' association with Number of cases sold

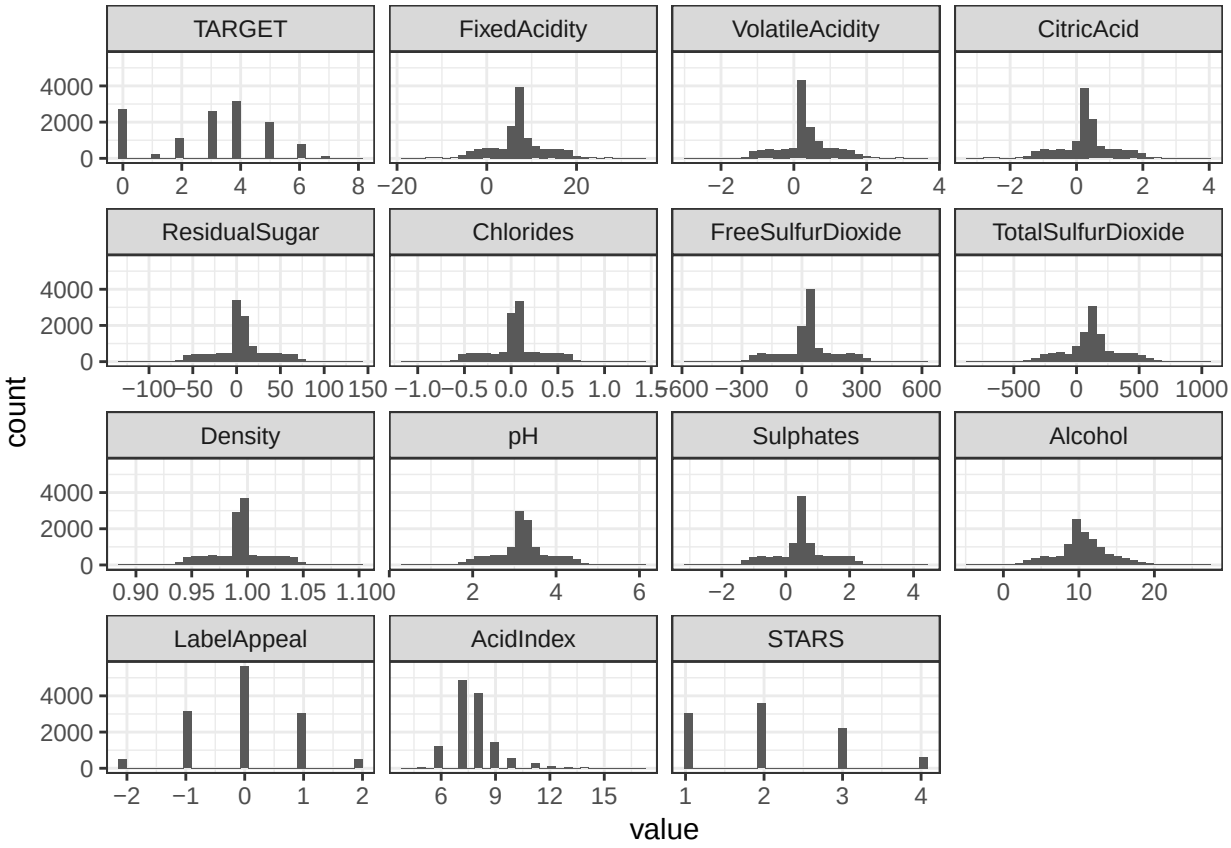


Figure 2: Histograms of features

3. Feature selection

In this section we'll attempt to select important features that explain the target variable. We'll explore the correlations that might exist in the data. There is positive correlations between TARGET and STARS and LabelAppeal.

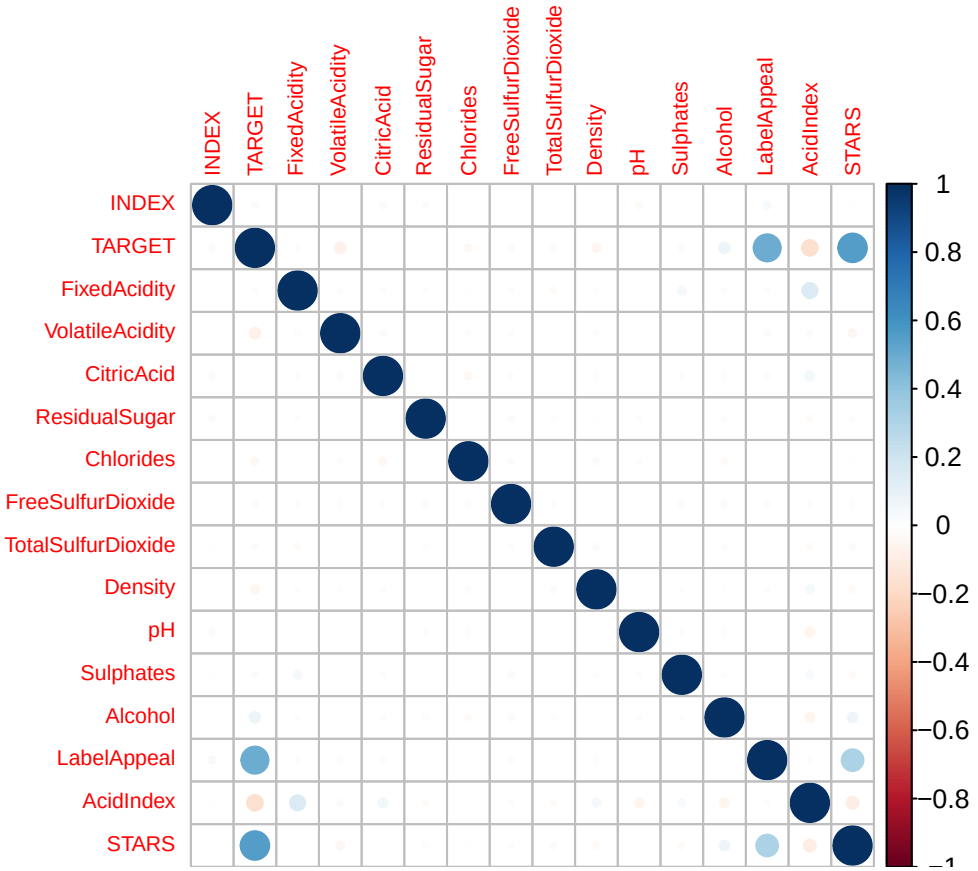


Figure 3: Correlation plot

It'll be useful to identify what contributes to the zero inflation in the TARGET variable. For which let's create an indicator variable "TARGET0" to equal 1 when TARGET is 0 and 0 otherwise.

3.1 Decision Trees

3.1.1 Predictors for zero cases sold (No sale)

Figure 4 shows the decision tree for TARGET0. Where 1 is no sale (cases sold = 0) and 0 is sale (cases sold > 0). Figure 5 shows the variable importance plot after a random forest bootstrap. While LabelAppeal did not contribute to node purity (Decrease in Gini index), it did affect accuracy on out of bag (OOB) samples.

3.1.2 Predictors for cases sold; when successfuly sold (cases > 0)

Figure 6 shows the decision tree of cases sold when they are greater than 0. It can be seen that the LabelAppeal and STARS are the top hitters. Figure 7 shows the variable importance plot when a random forest method is employed, where a random set of predictors are chosen at each iteration to fit a decision tree. Variables

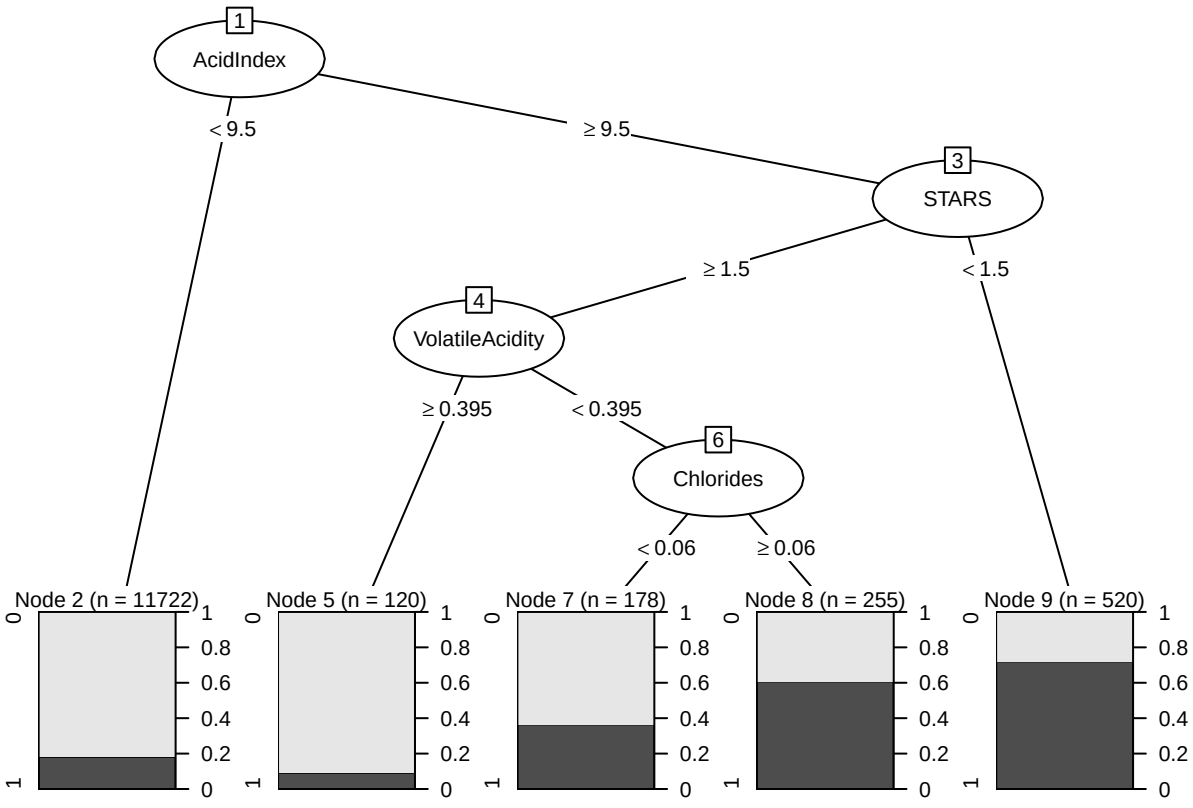


Figure 4: Decision tree - Sale (0) or no Sale (1)

whose exclusion contributes to higher Mean Squared Error (MSE) is deemed important. LabelApeal, STARS and Alcohol are top 3 variables that are important.

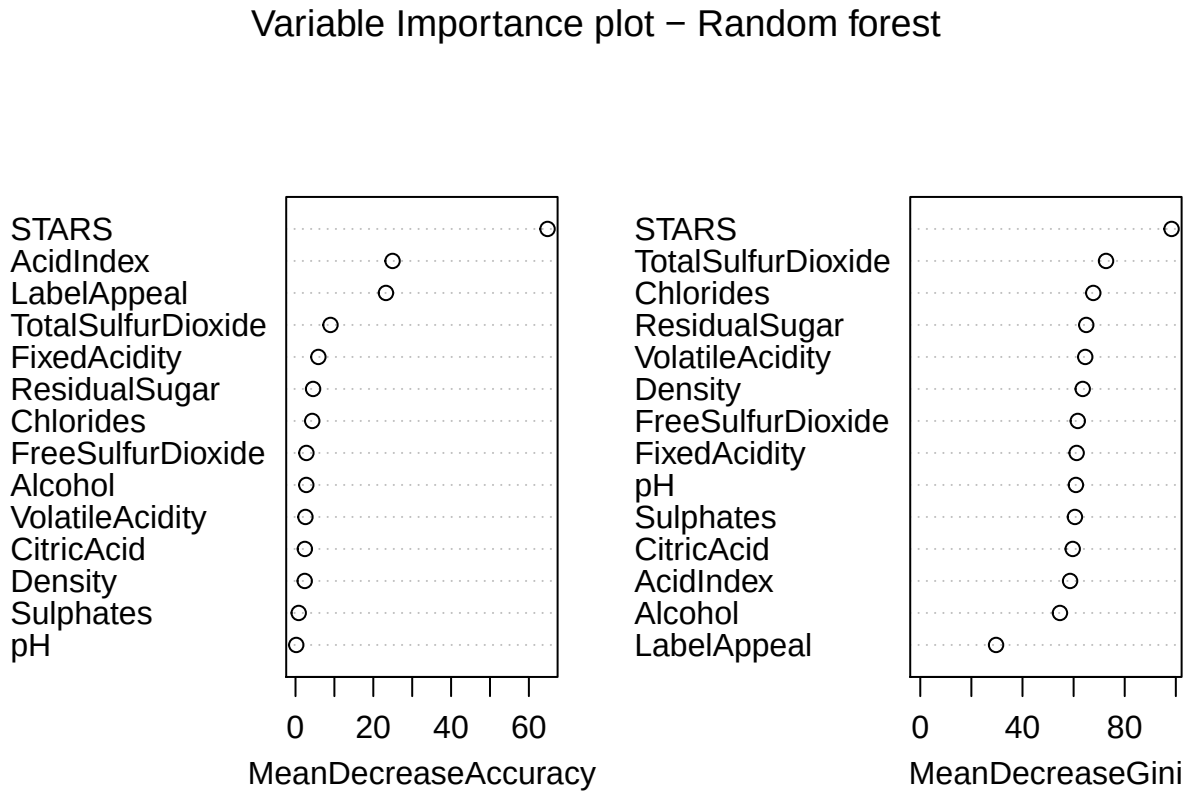


Figure 5: Variable Importance - Randomforest

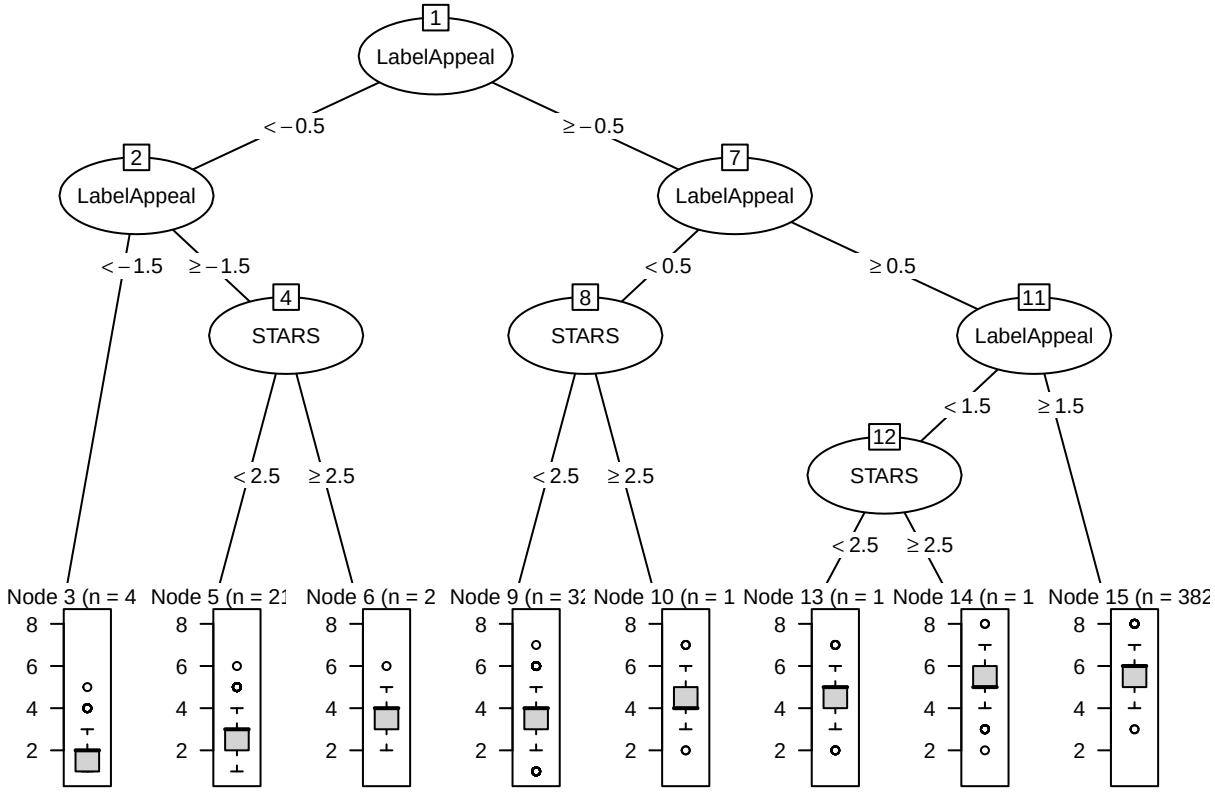


Figure 6: Simple tree for TARGET

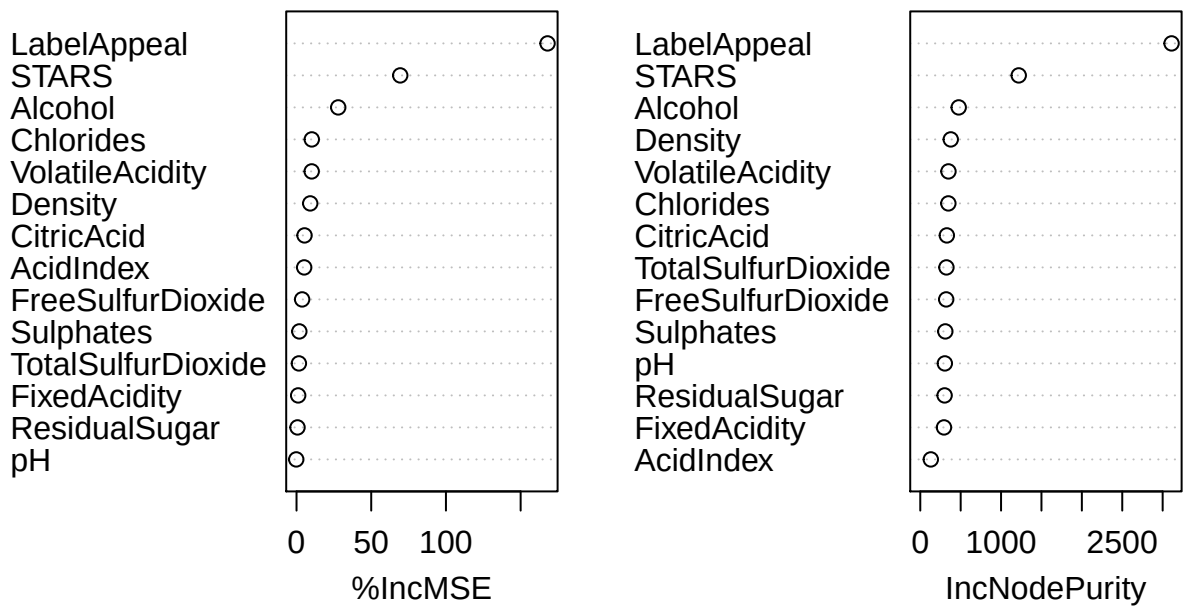


Figure 7: Variable importance plot - TARGET

4. Training and validation samples

The data is split into training and validation sample for evaluating models for final selection. The 75% of the data is sampled as training sample and the rest is set aside as validation set. Figure 8 shows that the TARGET variable is identically distributed for both training and validation samples.

4.1. The problem.

The TARGET variable from both training and validation (test) sample are associated with missing values of key predictors like STARS. GLM models in R ignores missing records for model fitting. In which case the zero inflation in the histograms in Figure 8 may not be representative in the training sample. Figure 9 illustrates the problem.

4.2 Problem Aleviation

From figure 1 it is inferred that most of the missing STARS are attributed to TARGET of 0. This implies that since no cases were distributed, it is likely that the wines did not have the opportunity to be tasted for review stars. Therefore we will impute zeros to STARS when cases distributed (TARGET) is zero.

5. Modeling

The following modeling approaches will be tried for predicting the number of cases sold.

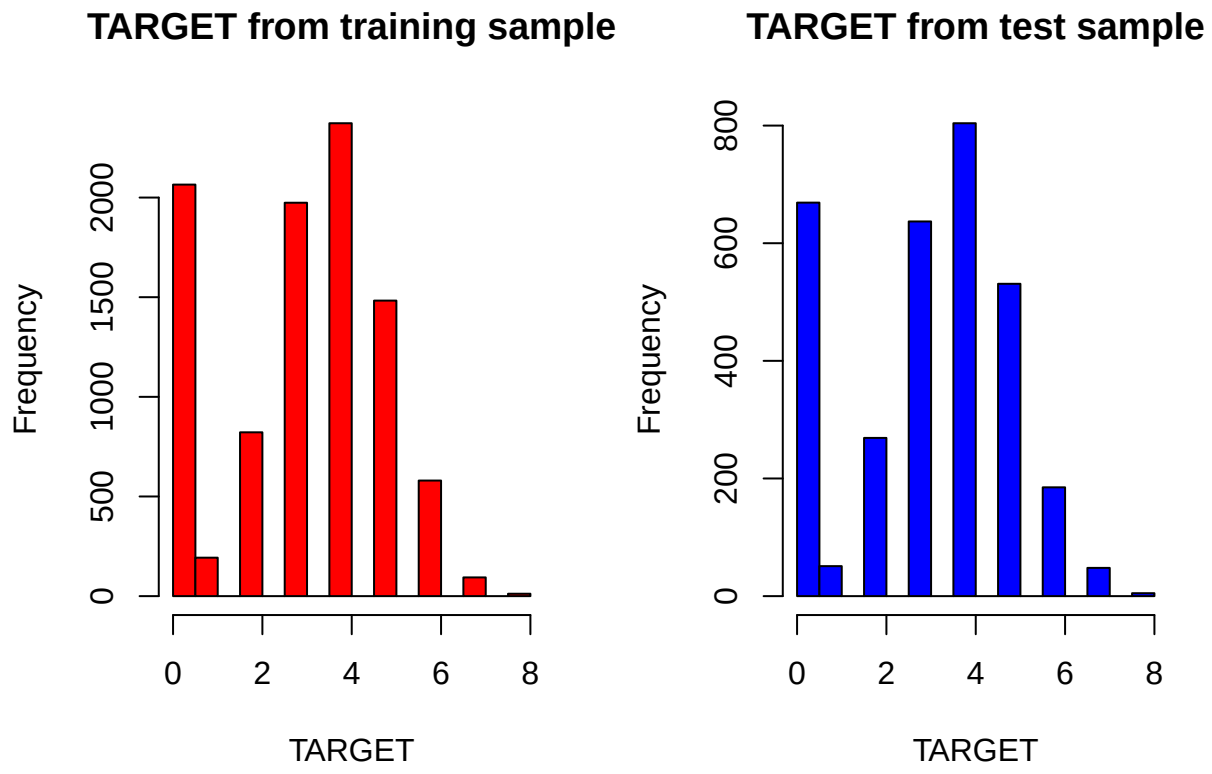


Figure 8: TARGET distribution for training and validation samples

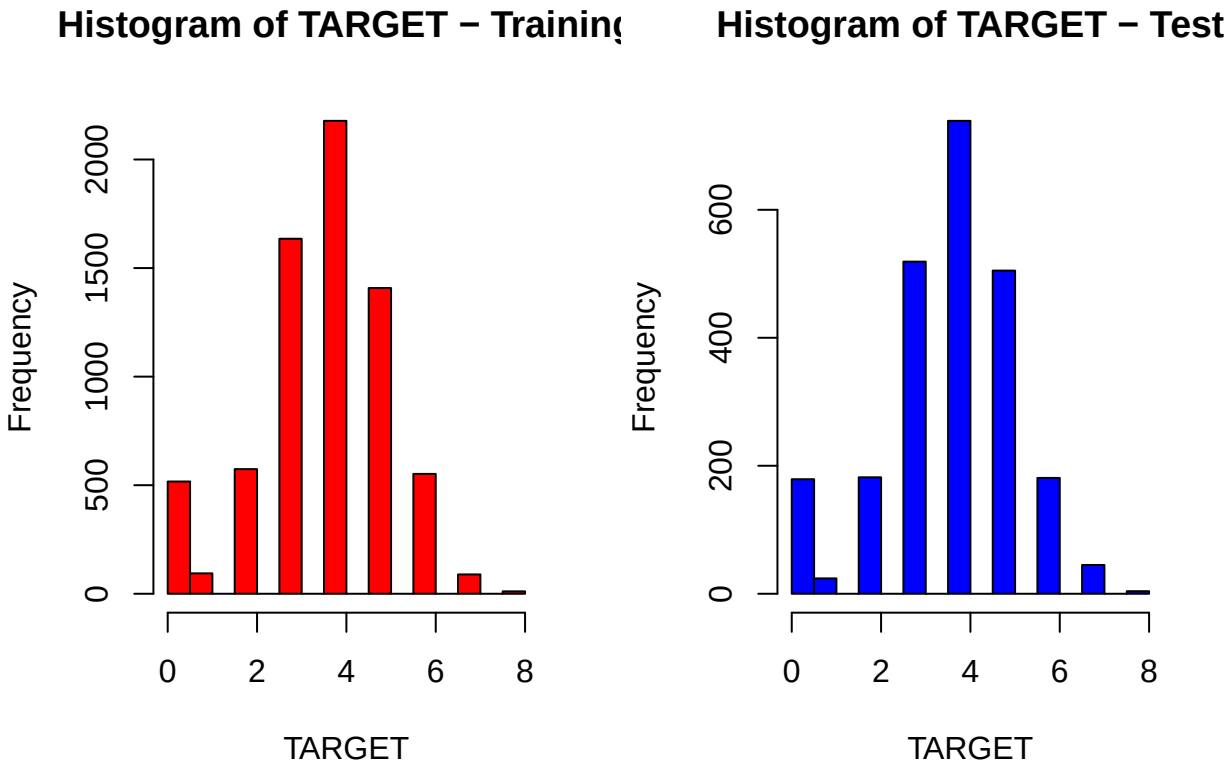


Figure 9: Histograms of TARGET when missing values of STARS is ignored

- Poisson regression
- Negative binomial regression
- Zero infalted Poisson regression
- Zero inflated Negative binomial regression
- OLS regression

5.1 Poisson regression.

As seen in table 1, the mean and standard deviation of TARGET are not too far apart. Poisson model is a good candidate for regression modeling.

5.1.1 Simple model with Label appeal as the predictor.

While we know that the TARGET variable is zero inflated and there are atleast more than one important predictor from the section above, we'll attempt to build the model ground up with LabelAppeal as the single predictor.

```
##
## Call:
## glm(formula = TARGET ~ LabelAppeal, family = poisson(link = "log"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1342  -0.4281   0.1717   0.5821   2.0412
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.080397   0.006012  179.70   <2e-16 ***
## LabelAppeal 0.255594   0.006623   38.59   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 17183  on 9595  degrees of freedom
## Residual deviance: 15689  on 9594  degrees of freedom
## AIC: 39571
##
## Number of Fisher Scoring iterations: 5
```

```
## Likelihood ratio test
##
## Model 1: TARGET ~ LabelAppeal
## Model 2: TARGET ~ 1
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   2 -19783
## 2   1 -20530 -1 1494.4  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 2: Simple Poisson model statistics

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	pseudoR.Sq	MAE
17182.9	9595	-19783.35	39570.7	39585.04	15688.51	9594	0.09	1.37

5.1.1.1 Model interpretation

It can be seen from the Likelihood ratio test ab0ve, that the slope of LabelAppeal is NOT zero and an unit increase in label appeal increases by 1. However the model is not an adequate fit with residual deviance over degrees of freedom being much higher than 1 (in comparison to a saturated model). This is also reflected in the Pseudo R Squared value.

The model does not predict 0s. As can be see from figure 9.

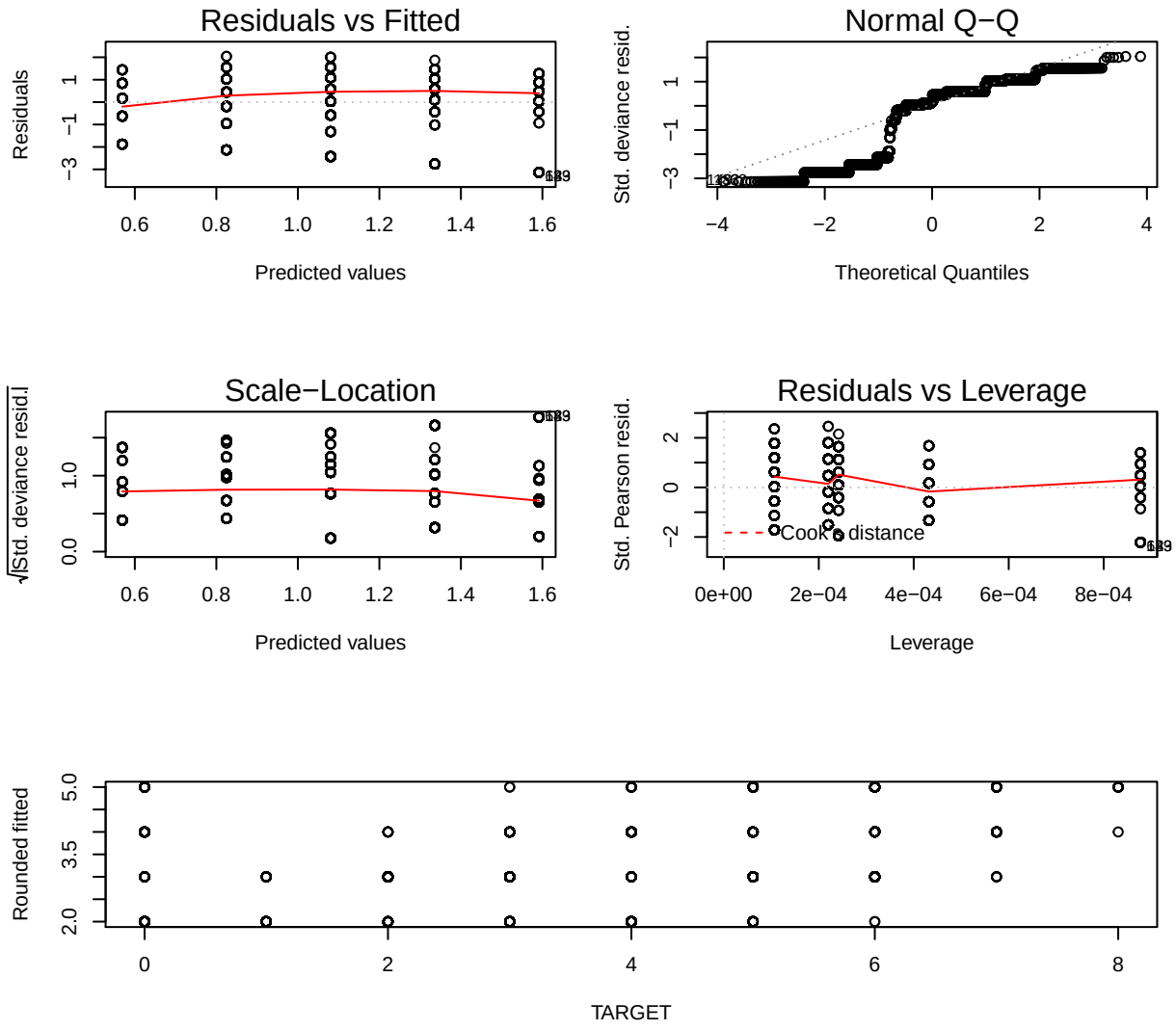


Figure 10: Simple poisson regression diagnostics

5.1.2 Poisson model with Label appeal & Stars as predictors

Now, we'll use Label appeal and stars as predictors and compare the model with the simple model in the above section.

```
##
## Call:
## glm(formula = TARGET ~ LabelAppeal + STARS, family = poisson(link = "log"),
##     data = train[!is.na(train$LabelAppeal), ])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.72926  -1.18544  -0.01392   0.67749   2.29938
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.253794   0.014112   17.98   <2e-16 ***
## LabelAppeal 0.104140   0.007474   13.93   <2e-16 ***
## STARS       0.426421   0.005828   73.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16678.9  on 8605  degrees of freedom
## Residual deviance:  9759.5  on 8603  degrees of freedom
##   (990 observations deleted due to missingness)
## AIC: 30760
##
## Number of Fisher Scoring iterations: 5
```

```
## Likelihood ratio test
##
## Model 1: TARGET ~ LabelAppeal + STARS
## Model 2: TARGET ~ 1
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   3 -15377
## 2   1 -18837 -2 6919.3  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 3: Poisson model statistics with STARS & LabelAppeal as predictors

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	pseudoR.Sq	MAE
16678.86	8605	-15377.01	30760.02	30781.2	9759.51	8603	0.41	1.16

5.1.2.1 Model interpretation.

It is seen that the regression model is statistically significant from the Likelihood Ratio Test (LRT) and the regression coefficents are also significant. The MAE is 1.5 which is a marginally higher than the simple model with LabelAppeal as predictor. However the ratio of residual deviance to its degrees of freedom is closer to 1, which can be interpreted as poor model fit to the data. This model also does not fit the zero counts.

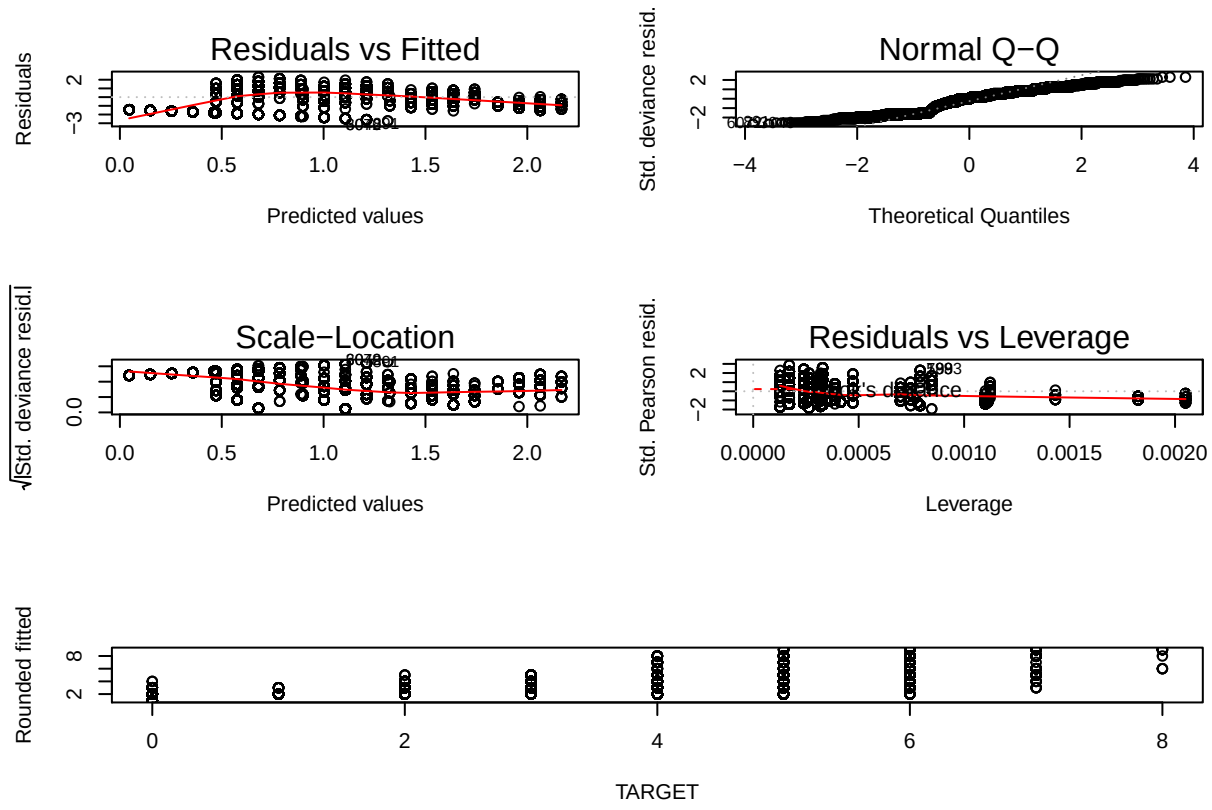


Figure 11: Model diagnostics TARGET ~ LabelAppeal + STARS

5.1.3 Dropping one predcitor from model.

It'll be interesting to see if there is value in dropping one of the regressors from the model. Below, LRT is performed by dropping one regressor at a time from the model.

```
## Single term deletions
##
## Model:
## TARGET ~ LabelAppeal + STARS
##             Df Deviance   AIC    LRT  Pr(>Chi)
## <none>           9759.5 30760
## LabelAppeal  1   9953.7 30952  194.2 < 2.2e-16 ***
## STARS        1  15390.3 36389 5630.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From LRT, both Label Appeal and STARS appear significant. Dropping Label Appeal (kepping STARS as the only predictor) provides a marginal increase in model fit compared to dropping STARS(keeping Label Appeal as the only predictor). This interpretation is onsistent with the decision tree interpretation in the section 3.1.2.

The LRT test for model comparison may be biased, since the models may have different samples sizes due to missing data. However, LRT can be used as a guidance for model selection. It is examined in the below section just to be sure

5.1.3.1 Two models - A. STARS as predictor B. LabelAppeal and Interpretation.

Verifying the results in the above section to make sure the drop1 test is providing us the expected results. It can be seen that when the data is kept contant between the models, the results are as expected. The result for model B is different from that of in section 5.1.1, this is because the data used to fit the model is different. Based on AIC and MAE comparison Label Appeal seems to be a better predcitor so far, when using a Poisson model.

```
##
## Call:
## glm(formula = TARGET ~ STARS, family = poisson(link = "log"),
##     data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.47661  -1.42755  -0.03828   0.51817   2.31937
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.216923   0.014001   15.49   <2e-16 ***
## STARS       0.451855   0.005584   80.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16678.9  on 8605  degrees of freedom
## Residual deviance:  9953.7  on 8604  degrees of freedom
##   (990 observations deleted due to missingness)
## AIC: 30952
##
## Number of Fisher Scoring iterations: 5
```

```
## Likelihood ratio test
##
## Model 1: TARGET ~ STARS
## Model 2: TARGET ~ 1
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   2 -15474
## 2   1 -18837 -1 6725.2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 4: Simple Poisson model statistics with STARTS as regressor

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	pseudoR.Sq	MAE
16678.86	8605	-15474.1	30952.2	30966.33	9953.7	8604	0.4	1.21

```
##
## Call:
## glm(formula = TARGET ~ LabelAppeal, family = poisson(link = "log"),
##     data = train[complete.cases(train[, c(14, 16)]), ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1113  -0.6184   0.4694   0.6115   2.0593
```

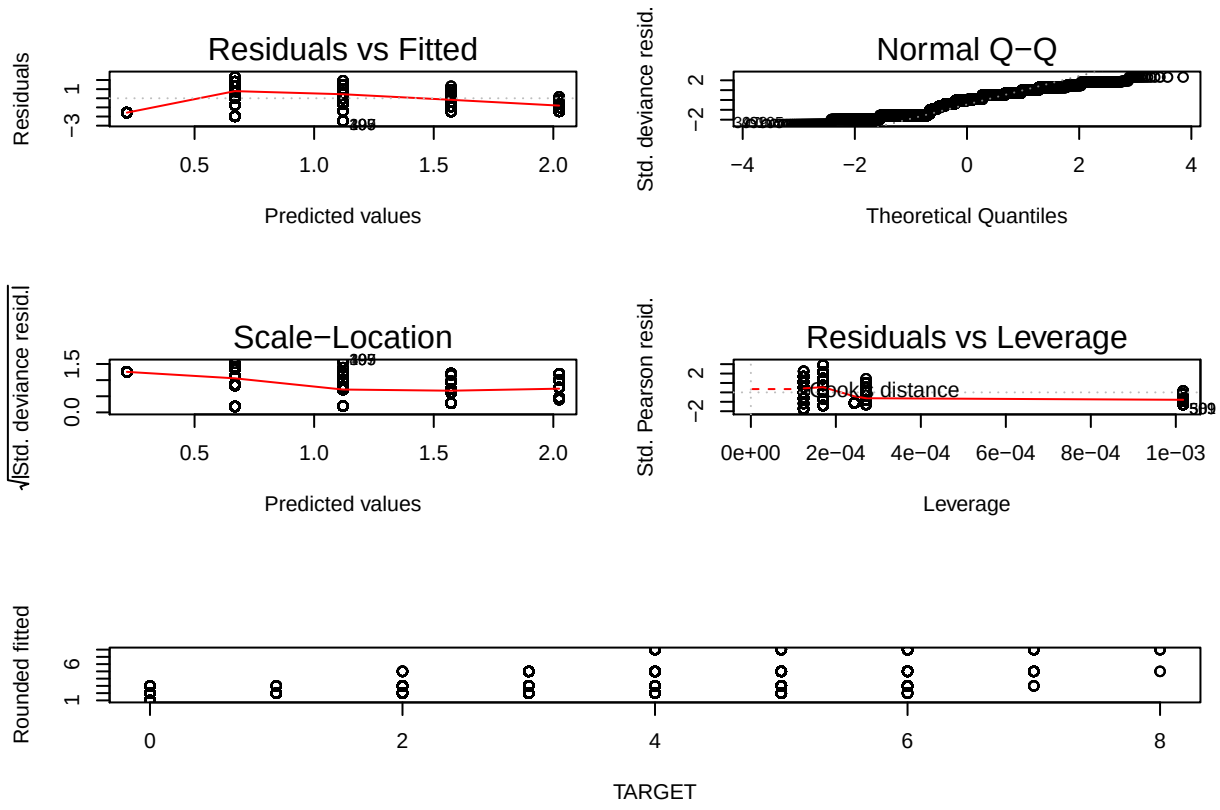


Figure 12: Model diagnostics TARGET ~ STARS

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.068859   0.006423  166.41   <2e-16 ***
## LabelAppeal 0.254055   0.007091   35.83   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16679  on 8605  degrees of freedom
## Residual deviance: 15390  on 8604  degrees of freedom
## AIC: 36389
##
## Number of Fisher Scoring iterations: 5
```

```
## Likelihood ratio test
##
## Model 1: TARGET ~ LabelAppeal
## Model 2: TARGET ~ 1
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   2 -18192
## 2   1 -18837 -1 1288.6  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 5: Simple Poisson model statistics with LabelAppeal as regressor; non missing cases for STARS and LabelAppeal

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	pseudoR.Sq	MAE
16678.86	8605	-18192.4	36388.8	36402.92	15390.29	8604	0.08	1.46

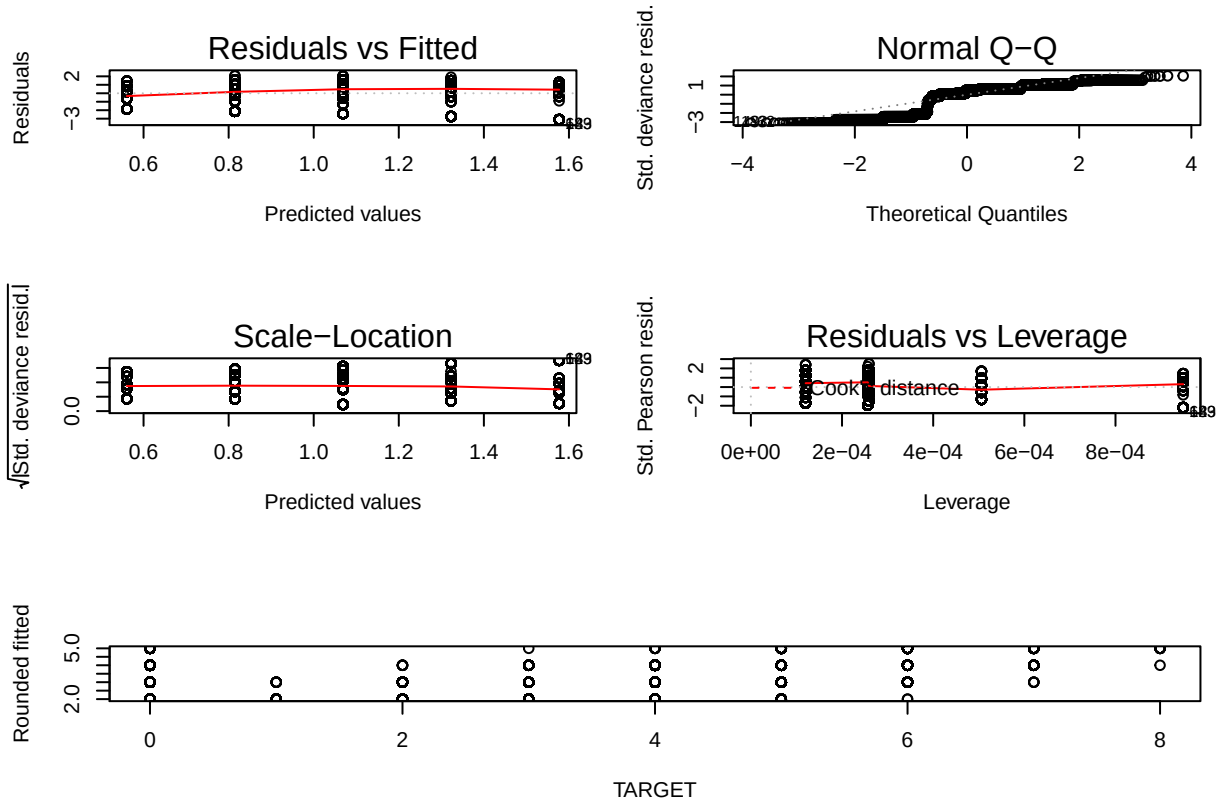


Figure 13: Model diagnostics TARGET ~ LABEL APPEAL non missing data for STARS and LabelAppeal

5.1.4 Poisson regression with LabelAppeal, STARS and Alcohol as predictors

In this section Alcohol is added as a preditor to the model in 5.1.2. Then LRT is performed by dropping one variable from the predictor. Though Alcohol is statistically significant, it does not add significant predictive power to the model. There is no significant reduction in MAE. Also we see that there is pattern in residuals based on the levels of the TARGET variable. It is to be noted that the predicted values are in log scale.

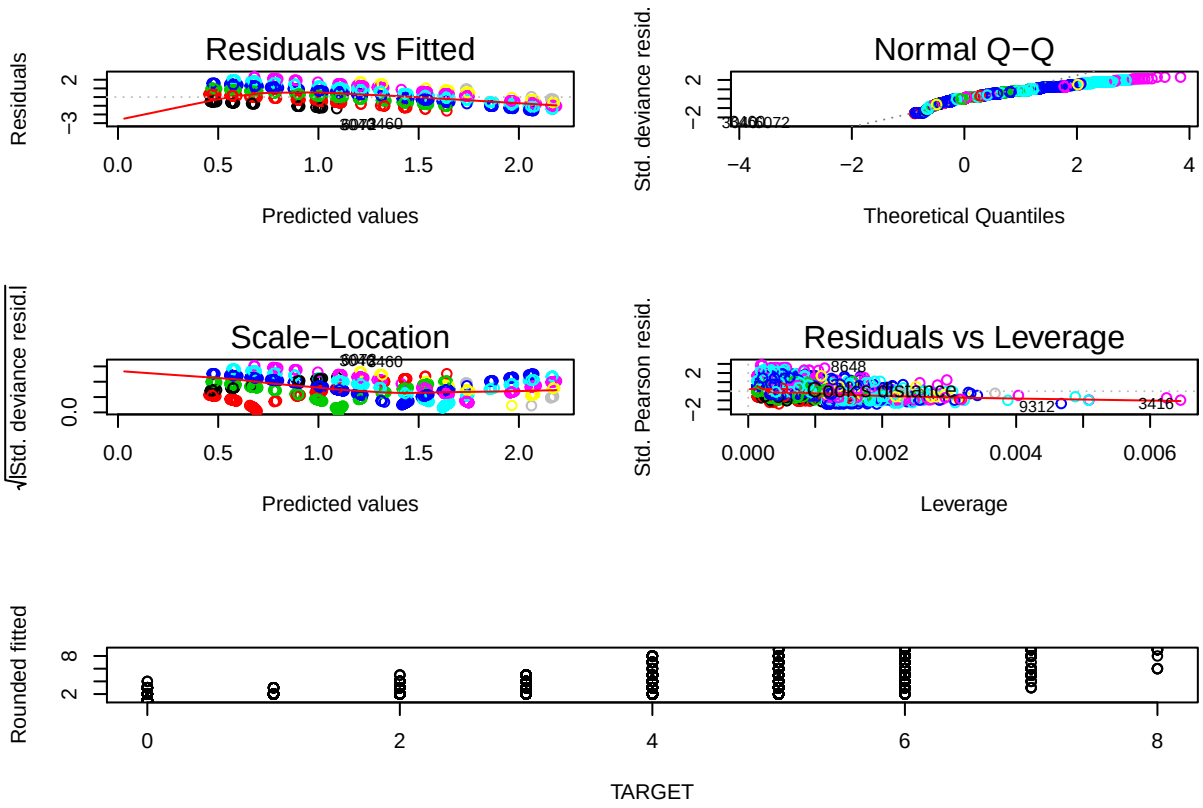
```
##
## Call:
## glm(formula = TARGET ~ LabelAppeal + STARS + Alcohol, family = poisson(link = "log"),
##     data = train[complete.cases(train[, c(13, 14, 16)]), ])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.73977  -1.18697   0.00272   0.67254   2.30094
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) 0.235588   0.022670  10.392   <2e-16 ***
## LabelAppeal 0.103426   0.007683  13.461   <2e-16 ***
## STARS       0.427821   0.006019  71.075   <2e-16 ***
## Alcohol     0.001434   0.001729   0.829    0.407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 15804.5  on 8162  degrees of freedom
## Residual deviance:  9241.3  on 8159  degrees of freedom
## AIC: 29168
##
## Number of Fisher Scoring iterations: 5
```

```
## Likelihood ratio test
##
## Model 1: TARGET ~ LabelAppeal + STARS + Alcohol
## Model 2: TARGET ~ 1
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   4 -14580
## 2   1 -17861 -3 6563.1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 6: Poisson model statistics with LabelAppeal, STARS, ALCOHOL as regressors

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	pseudoR.Sq	MAE
15804.46	8162	-14579.89	29167.79	29195.82	9241.35	8159	0.42	1.16



5.1.4.1 Model interpretation

```
## Single term deletions
##
## Model:
## TARGET ~ LabelAppeal + STARS + Alcohol
##             Df Deviance   AIC    LRT Pr(>Chi)
## <none>           9241.3 29168
## LabelAppeal  1   9422.6 29347  181.2   <2e-16 ***
## STARS        1  14563.8 34488 5322.5   <2e-16 ***
## Alcohol      1   9242.0 29166    0.7    0.407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From LRT by droping one predictor at every iteration, it is seen that the exclusion of Alcohol does not affect the fit considerably. Hence it is best to drop Alcohol from the model.

5.1.5 Addition of AcidIndex as predictor

One of the other predictors that was found key in section 3.1.2 is AcidIndex. It is seen below that though Acidindex's coefficient is statsitically significant, it does not improve prediction. Alcohol's coefficient is marginally significant.

```
##
## Call:
## glm(formula = TARGET ~ LabelAppeal + STARS + Alcohol + AcidIndex,
##     family = poisson(link = "log"), data = train[complete.cases(train[,
##         c(13, 14, 15, 16)]), ])
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -2.81466  -1.17100   0.03683   0.59784   2.76342
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  8.922e-01  5.139e-02  17.362   <2e-16 ***
## LabelAppeal  1.097e-01  7.701e-03  14.246   <2e-16 ***
## STARS        4.121e-01  6.139e-03  67.140   <2e-16 ***
## Alcohol     -2.463e-05  1.731e-03  -0.014    0.989
## AcidIndex   -7.986e-02  5.670e-03 -14.085   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 15804  on 8162  degrees of freedom
## Residual deviance:  9034  on 8158  degrees of freedom
## AIC: 28962
##
## Number of Fisher Scoring iterations: 5
```

```
## Likelihood ratio test
##
## Model 1: TARGET ~ LabelAppeal + STARS + Alcohol + AcidIndex
## Model 2: TARGET ~ 1
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   5 -14476
## 2   1 -17861 -4 6770.5  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 7: Poisson model statistics with LabelAppeal, STARS, ALCOHOL, AcidIndex as regressors

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	pseudoR.Sq	MAE
15804.46	8162	-14476.22	28962.45	28997.48	9034.01	8158	0.43	1.13

```
## Single term deletions
##
## Model:
## TARGET ~ LabelAppeal + STARS + Alcohol + AcidIndex
##             Df Deviance   AIC    LRT Pr(>Chi)
## <none>           9034.0 28962
## LabelAppeal  1   9237.0 29163  203.0   <2e-16 ***
## STARS        1  13735.0 33661 4701.0   <2e-16 ***
## Alcohol      1   9034.0 28960    0.0   0.9887
## AcidIndex    1   9241.3 29168  207.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

drop1 test shows that the exclusion of Alcohol provides the best AIC. In the next section we will fit a model with LabelAppeal, AcidIndex and STARS as predictors.

5.1.5.1 Removal of Alcohol and inclusion of AcidIndex

The model with LabelAppeal, STARS and AcidIndex is fit below and it is seen that all the predictor variables' coefficients are statistically significant. However the MAE is not improved. LRT when one predictor is left out of the model at each iteration shows that inclusion of AcidIndex does NOT impact the AIC considerably.

```
##
## Call:
## glm(formula = TARGET ~ LabelAppeal + STARS + AcidIndex, family = poisson(link = "log"),
##     data = train[complete.cases(train[, c(14, 15, 16)]), ])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.81673  -1.16342   0.03607   0.60401   2.75588
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.894561   0.046080   19.41   <2e-16 ***
## LabelAppeal  0.110305   0.007492   14.72   <2e-16 ***
## STARS        0.410285   0.005949   68.97   <2e-16 ***
## AcidIndex   -0.079677   0.005512  -14.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 16678.9  on 8605  degrees of freedom
## Residual deviance:  9541.1  on 8602  degrees of freedom
## AIC: 30544
##
## Number of Fisher Scoring iterations: 5
```

```
## Likelihood ratio test
##
## Model 1: TARGET ~ LabelAppeal + STARS + AcidIndex
## Model 2: TARGET ~ 1
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   4 -15268
## 2   1 -18837 -3 7137.7  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 8: Poisson model statistics with LabelAppeal, STARS, ALCOHOL, AcidIndex as regressors

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	pseudoR.Sq	MAE
16678.86	8605	-15267.82	30543.63	30571.88	9541.13	8602	0.43	1.14

5.1.5.1 Model interpretation

Droping one predictor and performing the LRT, it can be seen below that removing AcidIndex does NOT move the needle on AIC.

```
## Single term deletions
##
## Model:
```

```
## TARGET ~ LabelAppeal + STARS + AcidIndex
##             Df Deviance   AIC    LRT  Pr(>Chi)
## <none>           9541.1 30544
## LabelAppeal  1   9757.9 30758  216.8 < 2.2e-16 ***
## STARS        1  14493.9 35494 4952.8 < 2.2e-16 ***
## AcidIndex    1   9759.5 30760  218.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5.1.6 Poisson regression summary

```
##
##
## +--------------------------------+---------------+---------+--------+-------+
## |             model              | null.deviance | df.null | logLik |  AIC  |
## +================================+===============+=========+========+=======+
## |    TARGET = 1.08 + 0.256 *     |     17183     |  9595   | -19783 | 39571 |
## |          LabelAppeal           |               |         |        |       |
## +--------------------------------+---------------+---------+--------+-------+
## |    TARGET = 0.254 + 0.104 *    |     16679     |  8605   | -15377 | 30760 |
## |  LabelAppeal + 0.426 * STARS   |               |         |        |       |
## +--------------------------------+---------------+---------+--------+-------+
## | TARGET = 0.217 + 0.452 * STARS |     16679     |  8605   | -15474 | 30952 |
## +--------------------------------+---------------+---------+--------+-------+
## |    TARGET = 1.069 + 0.254 *    |     16679     |  8605   | -18192 | 36389 |
## |          LabelAppeal           |               |         |        |       |
## +--------------------------------+---------------+---------+--------+-------+
## |    TARGET = 0.236 + 0.103 *    |     15804     |  8162   | -14580 | 29168 |
## | LabelAppeal + 0.428 * STARS +  |               |         |        |       |
## |        0.001 * Alcohol         |               |         |        |       |
## +--------------------------------+---------------+---------+--------+-------+
## |    TARGET = 0.892 + 0.11 *     |     15804     |  8162   | -14476 | 28962 |
## | LabelAppeal + 0.412 * STARS -  |               |         |        |       |
## | 0 * Alcohol - 0.08 * AcidIndex |               |         |        |       |
## +--------------------------------+---------------+---------+--------+-------+
## |    TARGET = 0.895 + 0.11 *     |     16679     |  8605   | -15268 | 30544 |
## |  LabelAppeal + 0.41 * STARS -  |               |         |        |       |
## |        0.08 * AcidIndex        |               |         |        |       |
## +--------------------------------+---------------+---------+--------+-------+
##
## Table: Poisson fit summary (continued below)
##
##
##
## +-------+----------+-------------+------------+-----+
## |  BIC  | deviance | df.residual | pseudoR.Sq | MAE |
## +=======+==========+=============+============+=====+
## | 39585 |  15689   |    9594     |   0.087    | 1.4 |
## +-------+----------+-------------+------------+-----+
## | 30781 |   9760   |    8603     |    0.41    | 1.2 |
## +-------+----------+-------------+------------+-----+
## | 30966 |   9954   |    8604     |    0.4     | 1.2 |
## +-------+----------+-------------+------------+-----+
```

```
## | 36403 |  15390   |    8604     |   0.077    | 1.5 |
## +-------+----------+-------------+------------+-----+
## | 29196 |   9241   |    8159     |    0.42    | 1.2 |
## +-------+----------+-------------+------------+-----+
## | 28997 |   9034   |    8158     |    0.43    | 1.1 |
## +-------+----------+-------------+------------+-----+
## | 30572 |   9541   |    8602     |    0.43    | 1.1 |
## +-------+----------+-------------+------------+-----+
```

The poisson regression fits is summarized in the table above. The models

```
TARGET = 0.254 + 0.104 * LabelAppeal  + 0.426 *  STARS
TARGET = 0.217 + 0.452 * STARS
```

seems to be a reasonable model without running the risk of overfitting. While the model that includes AcidIndex to the above model seems marginally better, but requires another predictor for relatively similar MAE. However none of the models predict "no sales" well. Zero inflation poisson model would be useful and is explored in the next section.

5.2 Zero-Inflation Poisson regression (ZIP) model

The challenge we faced in the above section is to predict the zeros cases sold. In this section the predictors that were identified for zero cases sold in section 3.1.1 (figure 5) will be used as predictors to predict no sale and predictors that were identified in section 3.1.2 (figure 7) would be used to predict cases sold when there was a sale.

5.2.1 Zero Inflated Poisson (ZIP) model.

Figure 5 shows STARS has the maximum affect on accuracy of predicting a sale (TARGET > 0) or no sale (TARGET $= 0$). STARS is used as predictor for predicting zero sale and for a start,STARS and LabelAppeal would be used for predicting TARGET > 0.

```
##
## Call:
## pscl::zeroinfl(formula = TARGET ~ LabelAppeal + STARS | STARS, data = train)
##
## Pearson residuals:
##       Min        1Q    Median        3Q       Max
## -2.362621 -0.135533 -0.002754  0.214413  1.724116
##
## Count model coefficients (poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.076782   0.017741   60.70   <2e-16 ***
## LabelAppeal 0.205960   0.007764   26.53   <2e-16 ***
## STARS       0.115422   0.007517   15.35   <2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   11.496      8.271   1.390    0.165
## STARS        -13.131      8.271  -1.587    0.112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Number of iterations in BFGS optimization: 12
## Log-likelihood: -1.237e+04 on 5 Df
```

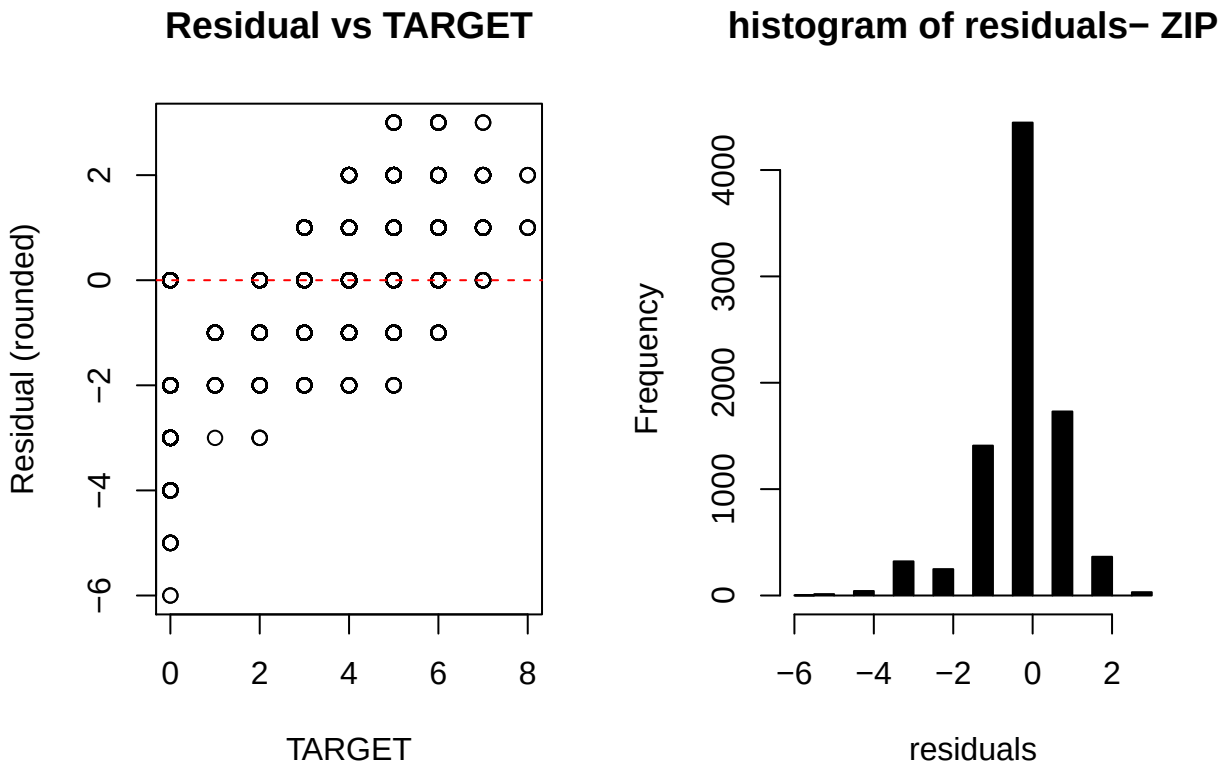


Figure 14: ZIP - rounded residuals vs TARGET

5.2.2 ZIP model's surprising result

The logistic component of the ZIP model shows that STARS is not a statistically significant contributor. However randomforest approach tells a different story. Also figure 14 shows that the residuals of the model increases as the TARGET value increases. In the next section we will attempt to verify this by modeling using the Logistic Hurdle model.

5.2.3 Logistic Hurdle model to verify ZIP

The same input to the model as provided in the ZIP model is supplied to the logistic regression. STAR is used as a predictor to predict when NO cases were sold. It is seen that STARS are good predictor of no sale. And most of the imputed data for stars was associated with TARGET = 0. The imputation has no effect on STARS being a good predictor. Here missing data itself is a signal more so than a noise.

Figure 15 shows the ROC curve of of logictic regression of both the training (black) and test data (red). The lines are on top of each other and both have the same AUC of 95%.

LabelAppeal and STARS are used for predicting the counts in poisson model.

```
##
## Call:
## glm(formula = TARGET0 ~ STARS, family = binomial(link = "logit"),
##     data = train, subset = !is.na(train$STARS))
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7515  -0.0887  -0.0887  -0.0011   3.3285
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.2956     0.1258   26.19   <2e-16 ***
## STARS        -4.4155     0.1287  -34.32   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9484.2  on 8605  degrees of freedom
## Residual deviance: 3164.2  on 8604  degrees of freedom
## AIC: 3168.2
##
## Number of Fisher Scoring iterations: 8

## Single term deletions
##
## Model:
## TARGET0 ~ STARS
##        Df Deviance    AIC  LRT  Pr(>Chi)
## <none>      3164.2 3168.2
## STARS   1   9484.2 9486.2 6320 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## glm(formula = TARGET ~ LabelAppeal + STARS, family = poisson(link = "log"),
##     data = train[train$TARGET0 == 0 & !is.na(train$STARS), ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7051  -0.2974  -0.0525   0.3032   1.4760
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.130586   0.017053   66.30   <2e-16 ***
## LabelAppeal 0.201647   0.007676   26.27   <2e-16 ***
## STARS       0.100198   0.007315   13.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2454.4  on 6540  degrees of freedom
## Residual deviance: 1143.5  on 6538  degrees of freedom
## AIC: 22144
##
## Number of Fisher Scoring iterations: 4

## Single term deletions
```

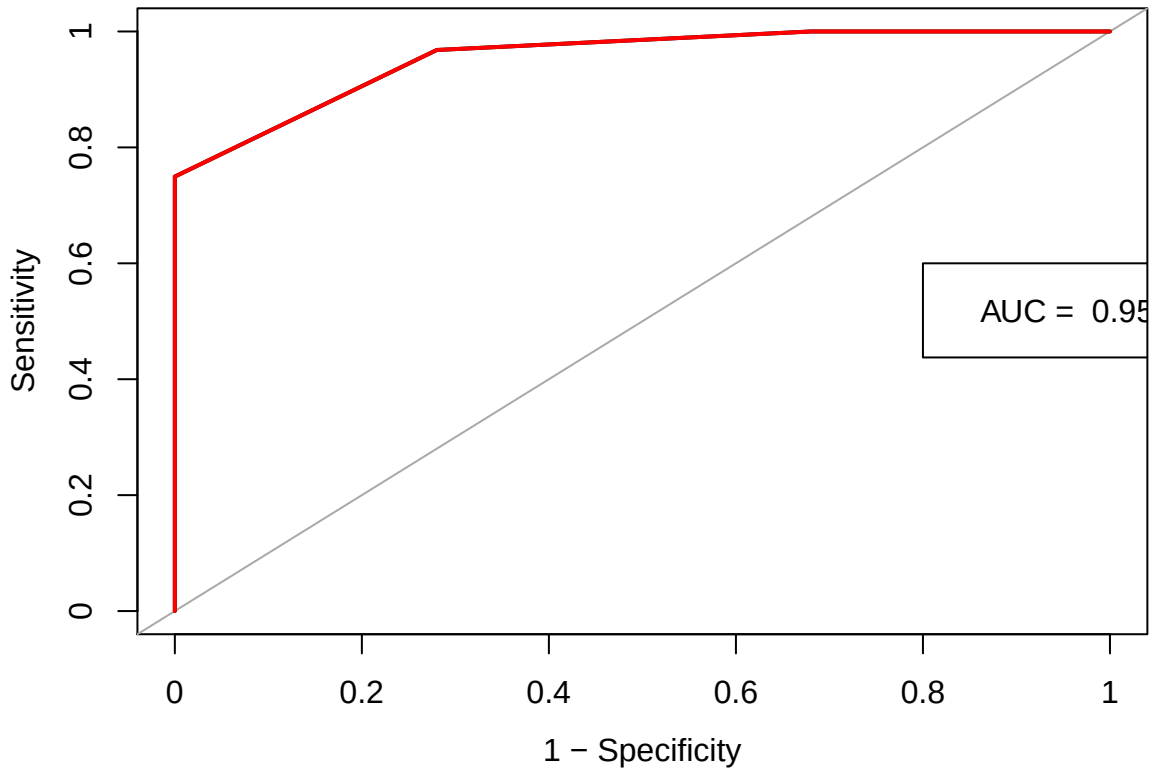


Figure 15: ROC curve for No sale prediction

```
##
## Model:
## TARGET ~ LabelAppeal + STARS
##             Df Deviance   AIC    LRT  Pr(>Chi)
## <none>           1143.5 22144
## LabelAppeal  1   1834.4 22833 690.94 < 2.2e-16 ***
## STARS        1   1330.0 22328 186.57 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 9: TARGET ~ LabelAppeal+ STARS|STARS

LogLik	AIC	BIC	MAE
-12370.4	24750.8	NA	0.6596561

It is seen that Figure 16 is identical to figure 14. ZIP model is verified. However I find that the ZIP model's logistic component not very intuitive to interpret. I am hoping that the professor Mickelson would be able to throw more light on this through email response of my question to him on 11/19/2017.

5.2.3.1 Model interpretation

From the drop1 test for logistic regression above, we see that dropping STARS would not increase AIC significantly. This warrants a fit with just LabelAppeal as predictor. Also it is seen that the residuals are skewed right, as the TARGET increases. There is bias in the model.

The model makes physical sense, one would expect the cases shipped to be low if the rating stars are low. The log odds of cases shipped < 0 increases with increase in STARS. The theory also holds for the poisson component of the regression model.

5.2.4 ZIP with LabelAppeal for count prediction and STARS for No sale prediction

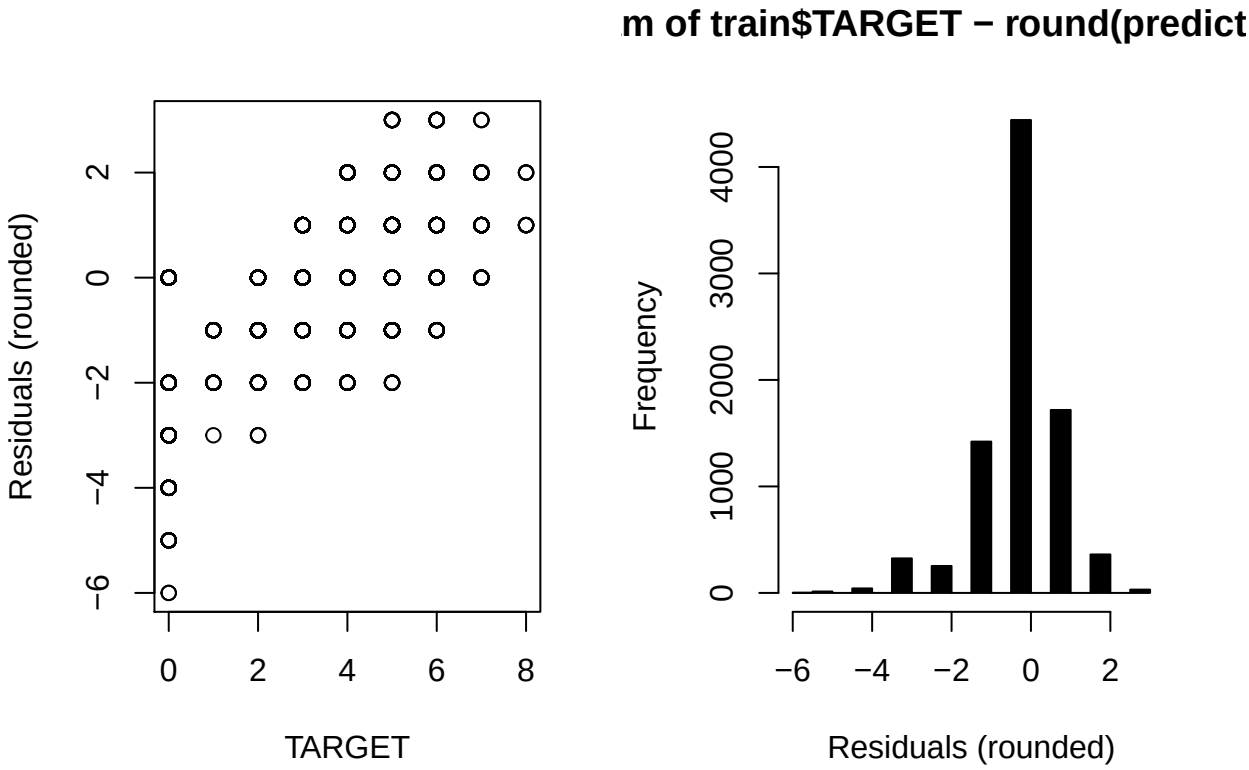


Figure 16: Verify ZIP with Logistic Hurdle model

```
## Number of iterations in BFGS optimization: 14
## Log-likelihood: -1.249e+04 on 4 Df
```

Table 10: TARGET ~ LabelAppeal|STARS

LogLik	AIC	BIC	MAE
-12488.22	24984.44	NA	0.71

5.2.4.1 Model interpretation

The Figure 17 shows that the residuals are left tailed. The model does not predict zeros well. For goodness of fit assessment, a model is fit for samples where counts of sale is greater than 0. The residula deviance over degrees of freedom is less than 1 indicating a good fit.

```
##
## Call:
## glm(formula = TARGET ~ LabelAppeal, family = poisson(link = "log"),
##     data = train[train$TARGET0 == 0 & !is.na(train$STARS), ])
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.72554  -0.40647   0.05914   0.08511   1.51180
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.343436   0.006427  209.04   <2e-16 ***
## LabelAppeal 0.239437   0.007153   33.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

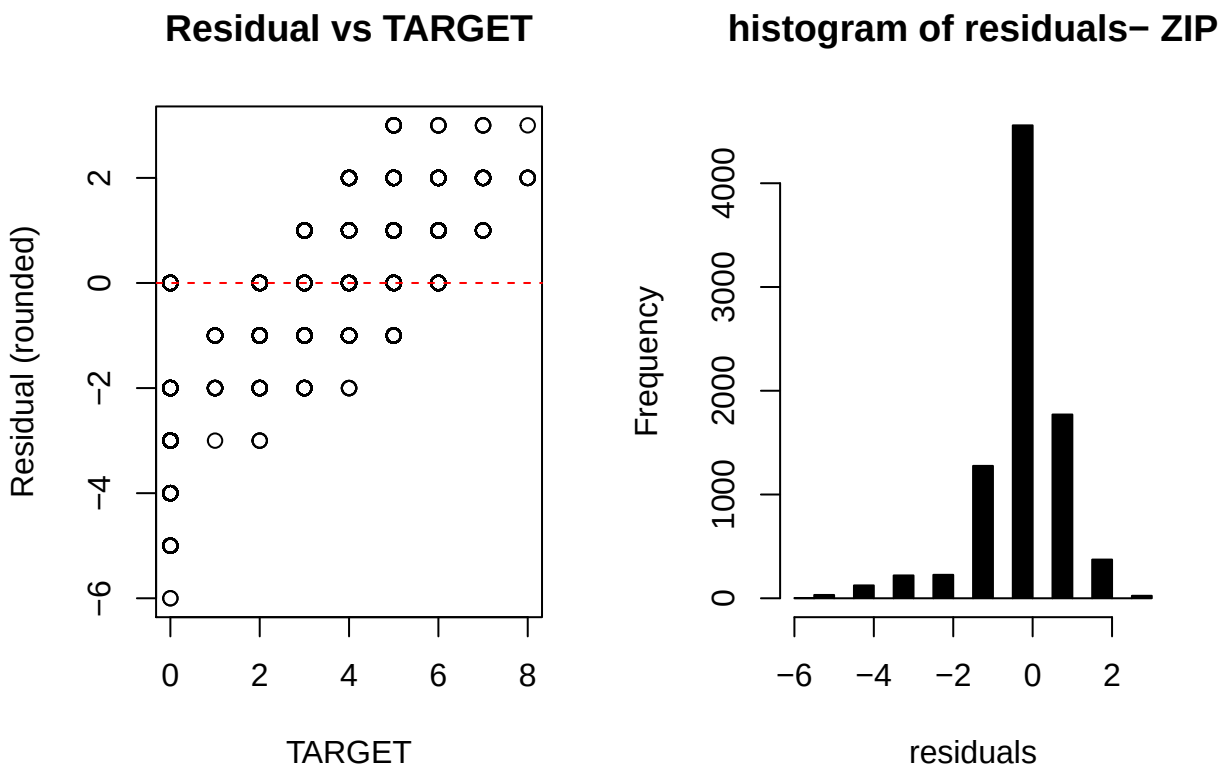


Figure 17: ZIP TARGET ~ LabelAppeal|STARS

```
## | LabelAppeal | TARGET = 22.15 - |        |       |     |      |
## |          23.7 * STARS          |        |       |     |      |
## +--------------------------------+--------+-------+-----+------+
##
## Table: ZIP fit summary - Note model equation before | is for count and after | is prob of No sale
```

5.3 Negative Binomial models.

In this section Negative binomial model fits are explored.

5.3.1 Negative binomial fit with LabelAppeal and STARS as predictors.

Below is the results for the negative binomial fit, the results are identical to that of poisson models with a slight increase in standard errors.

```
##
## Call:
## MASS::glm.nb(formula = TARGET ~ STARS + LabelAppeal, data = train,
##     init.theta = 46099.61566, link = log)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.72920  -1.18542  -0.01392   0.67748   2.29932
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.253774   0.014113   17.98   <2e-16 ***
## STARS       0.426431   0.005828   73.17   <2e-16 ***
## LabelAppeal 0.104139   0.007474   13.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(46099.68) family taken to be 1)
##
##     Null deviance: 16678.1  on 8605  degrees of freedom
## Residual deviance:  9759.1  on 8603  degrees of freedom
##   (990 observations deleted due to missingness)
## AIC: 30762
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  46100
##           Std. Err.:  58612
## Warning while fitting theta: alternation limit reached
##
##  2 x log-likelihood:  -30754.2
```

```
## Single term deletions
##
## Model:
## TARGET ~ STARS + LabelAppeal
##             Df Deviance   AIC    LRT  Pr(>Chi)
## <none>           9759.1 30760
## STARS        1  15389.6 36389 5630.5 < 2.2e-16 ***
## LabelAppeal  1   9953.3 30952  194.2 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

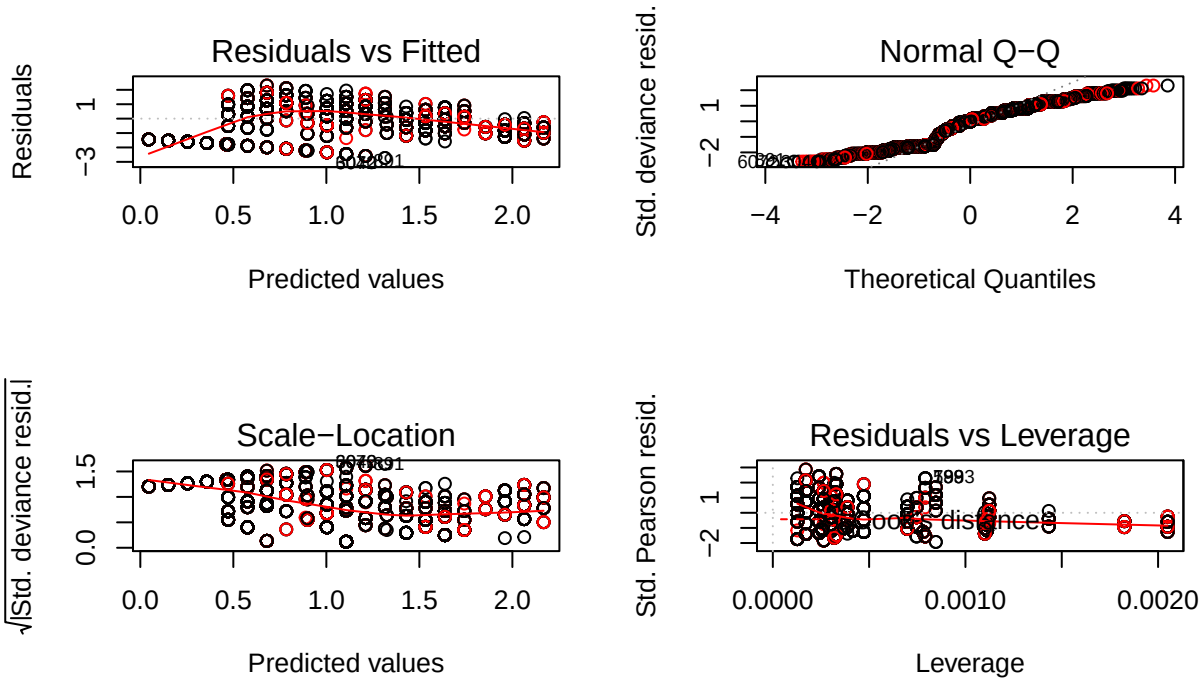


Figure 18: Negative binomial model

Table 11: negative binomial model

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	MAE
16678.12	8605	-15377.1	30762.2	30790.44	9759.13	8603	1.21

5.3.1.1 Model interpretation

The model is identical to poisson model with the same predictors. Label Appeal has a marginal effect compared to STARS.

5.4. Zero inflated Negative binomial model

In this section a Zero inflated negative binomial model is fit.

```
##
## Call:
## pscl::zeroinfl(formula = TARGET ~ LabelAppeal + STARS | STARS, data = train,
##     dist = "negbin", EM = T)
##
## Pearson residuals:
##        Min         1Q     Median         3Q        Max
## -2.363e+00 -1.353e-01 -1.804e-05  2.144e-01  1.724e+00
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.076746   0.017741   60.69   <2e-16 ***
## LabelAppeal  0.205959   0.007764   26.53   <2e-16 ***
## STARS        0.115435   0.007517   15.36   <2e-16 ***
## Log(theta)  17.339717   1.567470   11.06   <2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    21.55    1262.68   0.017    0.986
## STARS         -23.19    1262.68  -0.018    0.985
##
```

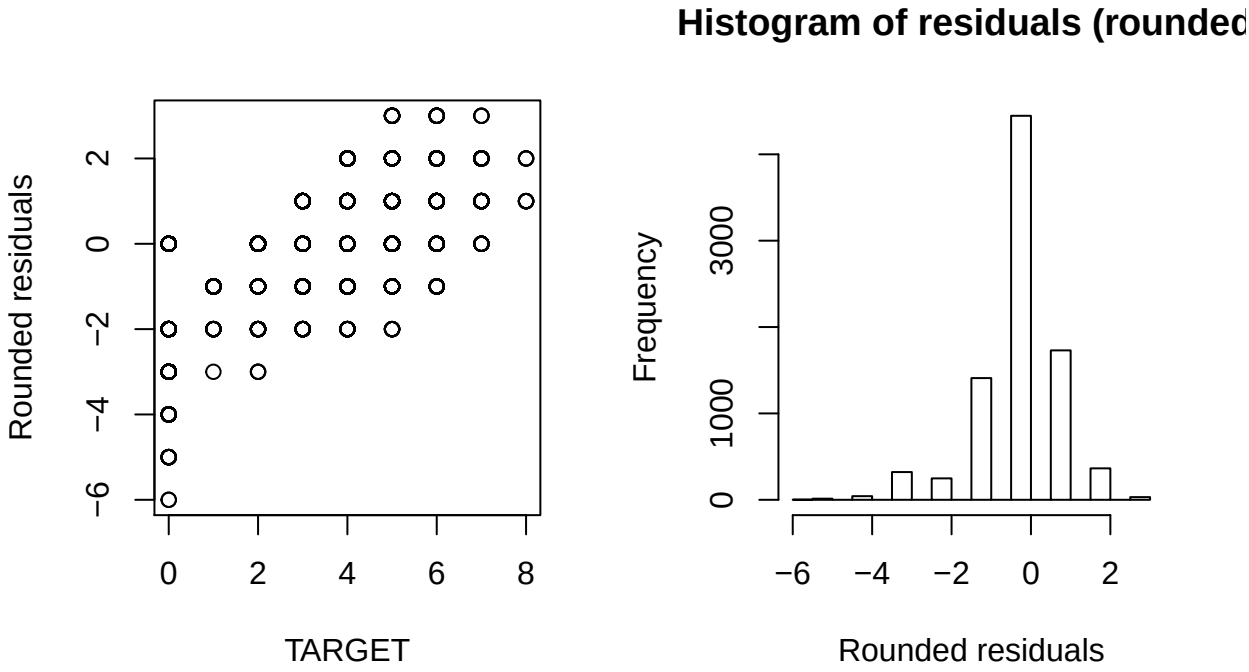


Figure 19: Zero inflated negative binomial model

```
## | TARGET = 21.55 - 23.19 * STARS |        |       |     |     |
## +--------------------------------+--------+-------+-----+-----+
##
## Table: Zero inflated negative binomial model
```

5.5 OLS regression

Lastly, an OLS regression is fit to the data to see how it performs.

```
##
## Call:
## lm(formula = TARGET ~ STARS + LabelAppeal, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1862 -0.8364 -0.0459  0.8688  3.8688
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.83643    0.02374   35.23   <2e-16 ***
## STARS        1.29476    0.01192  108.60   <2e-16 ***
## LabelAppeal  0.38010    0.01531   24.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.215 on 8603 degrees of freedom
##   (990 observations deleted due to missingness)
## Multiple R-squared:  0.6263, Adjusted R-squared:  0.6262
## F-statistic:  7209 on 2 and 8603 DF,  p-value: < 2.2e-16
```

```
## Single term deletions
##
## Model:
## TARGET ~ STARS + LabelAppeal
##             Df Sum of Sq   RSS     AIC  F value    Pr(>F)
## <none>                   12710  3362.0
## STARS        1   17424.2 30135 10789.2 11793.55 < 2.2e-16 ***
## LabelAppeal  1     910.3 13621  3955.3   616.11 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

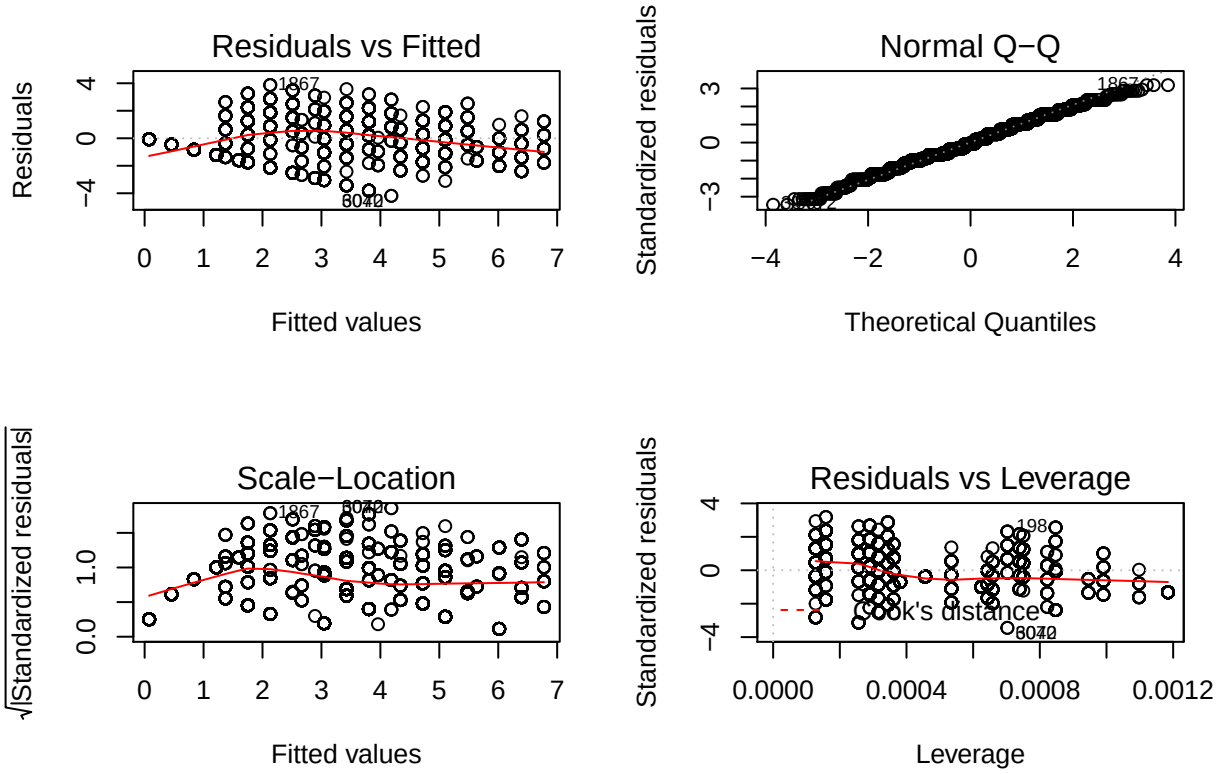


Figure 20: OLS regression diagostics

5.5.1 model interpretation and summary

The model surprisingly beats expectation. The coefficients make logical sense and can definitely be a candidate model. The summary table is shown below. The MAE is comprarble to ZIP model above. There are patterns in the residuals (Figure 20), making it not structly meeting OLS assumptions.

```
##
##
## +--------------------------------+---------------+-------+-------+--------+
## |             model              | adj.r.squared |  AIC  |  BIC  |  MAE   |
## +================================+===============+=======+=======+========+
## | TARGET = 0.84 + 1.29 * STARS + |    0.6262     | 27787 | 27815 | 0.9914 |
## |       0.38 * LabelAppeal       |               |       |       |        |
## +--------------------------------+---------------+-------+-------+--------+
##
## Table: OLS regression
```

6. Model selection

Comparing the model summaries in Section 5, the Zero inflated binomial model:

TARGET = 1.08 + 0.21 LabelAppeal + 0.12 * STARS | TARGET = 11.5 - 13.13 * STARS*

has the minimum MAE. While STARS alone may have been sufficient as predictors as in the model:

TARGET = 1.33 + 0.25 LabelAppeal | TARGET = 22.15 - 23.7 * STARS*

given that STARS could be missing in new data sets, Label Appeal may be a handy predictor. It is surprising that Alcohol did not come out as an important predictor. It was expected that too high or too high proof (measure of alcohol by volume) may not be a popular choice for customers.

7. Model deployment

Along with this report an addendum file is provided that contains the R code for model deployment.

8. Conclusion

More than any chemical properties Label appeal and review stars are dominant factors in wine sales. Care in label appeal is as much required as in production of wine. It is recommended tha the wine manufacturer market wine with a catchy label to receive as many review stars as possible.