Wine sales prediction

Sri Seshadri

11/11/2017

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.

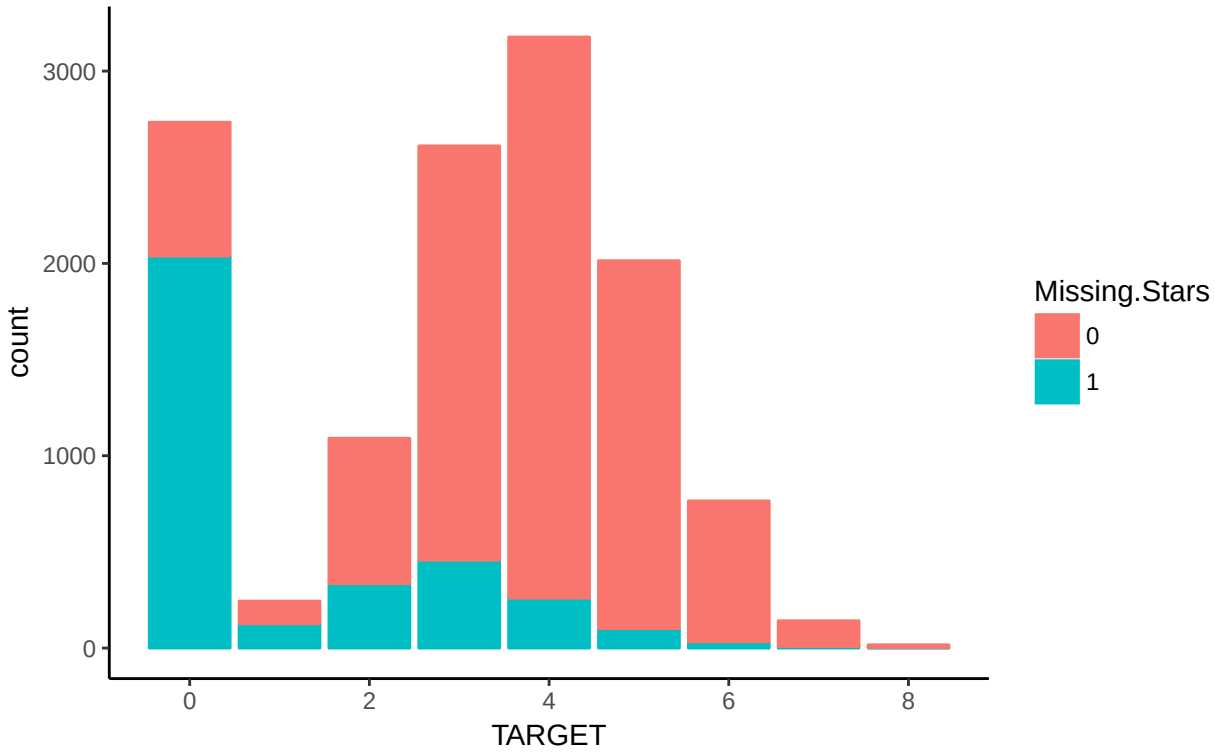

Figure 1: Missing STARS values' association with Number of cases sold

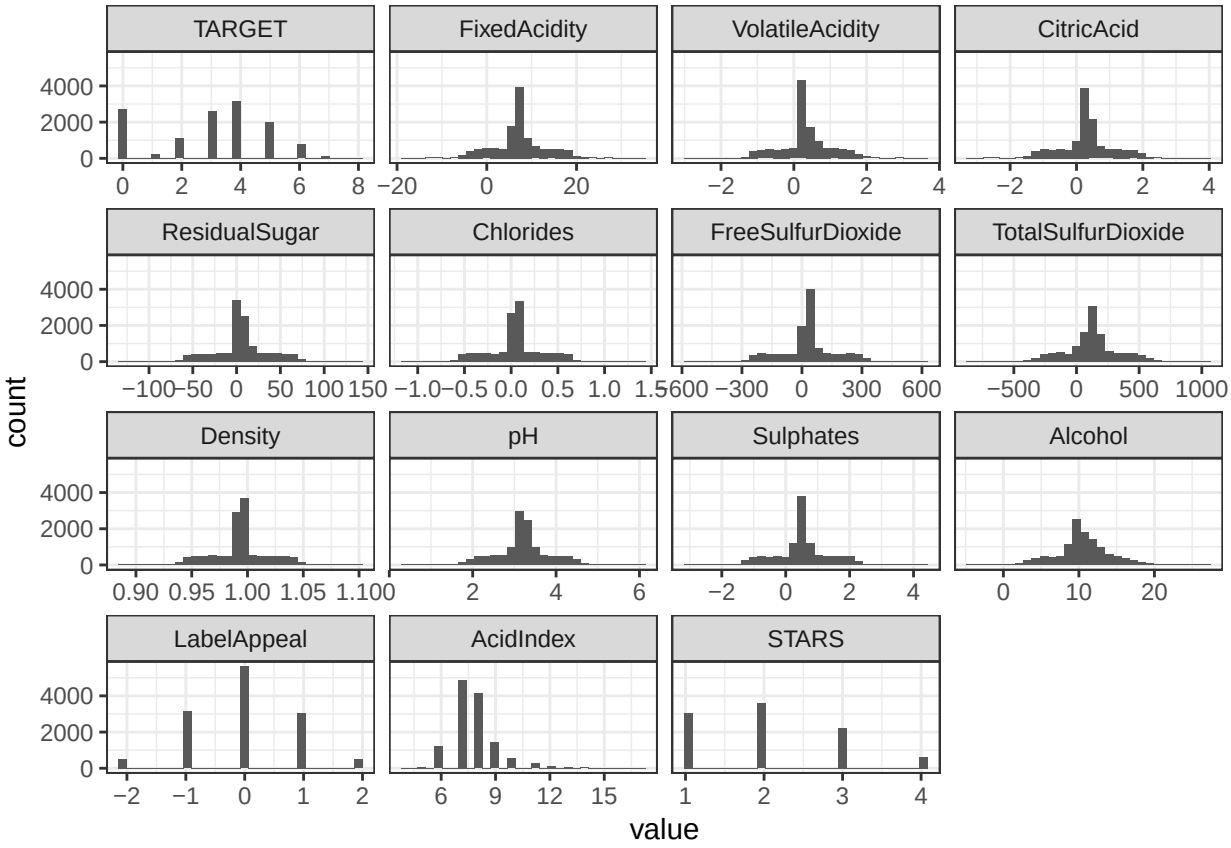

Figure 2: Histograms of features

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.

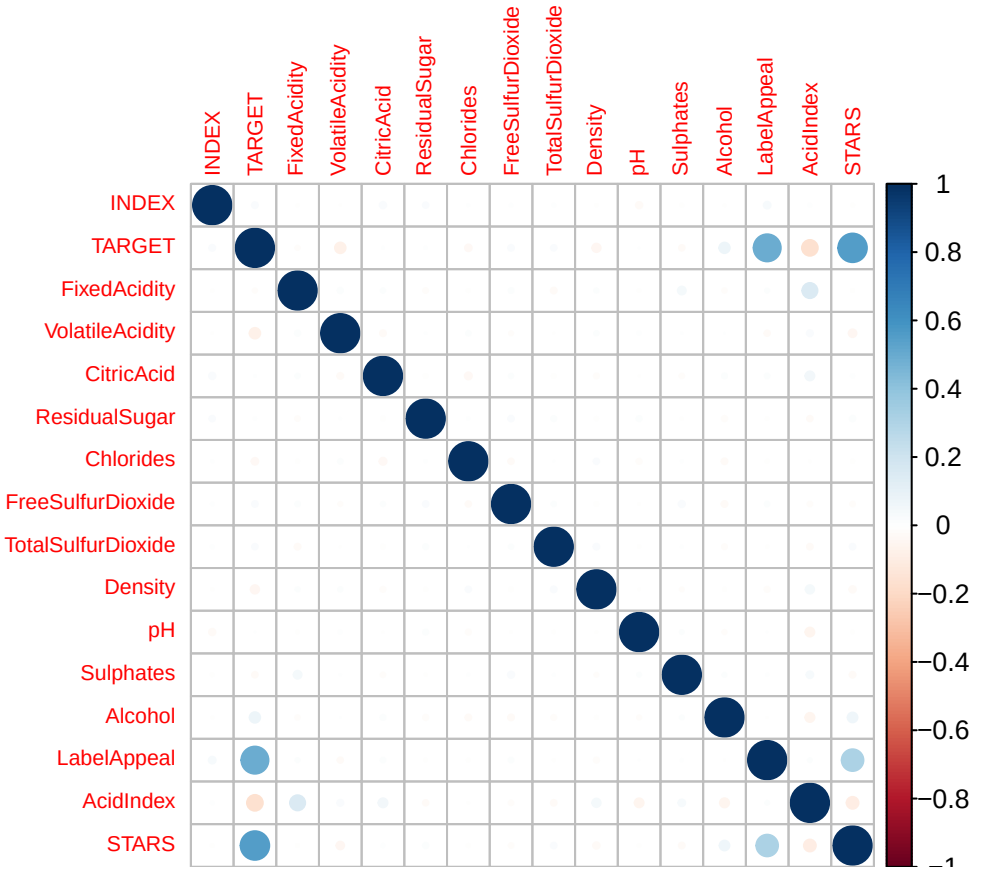

Figure 3: Correlation plot

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

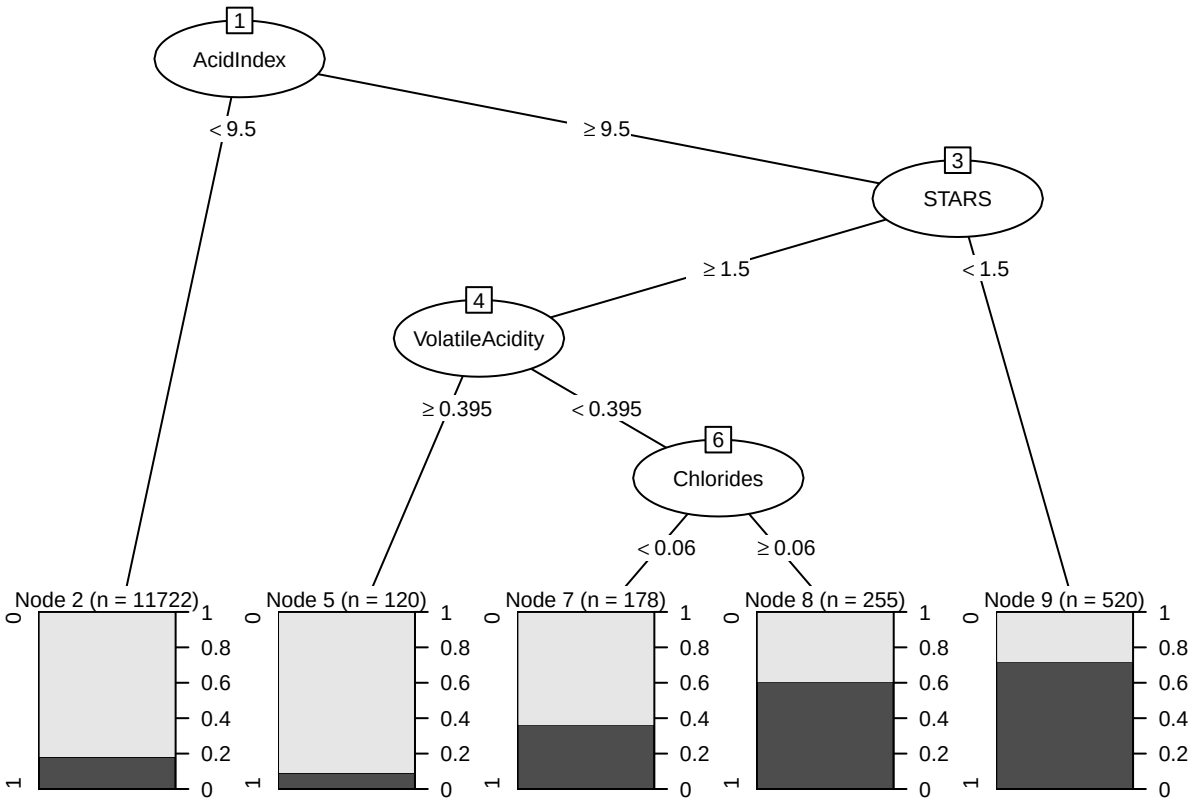

Figure 4: Decision tree - Sale (0) or no Sale (1)

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

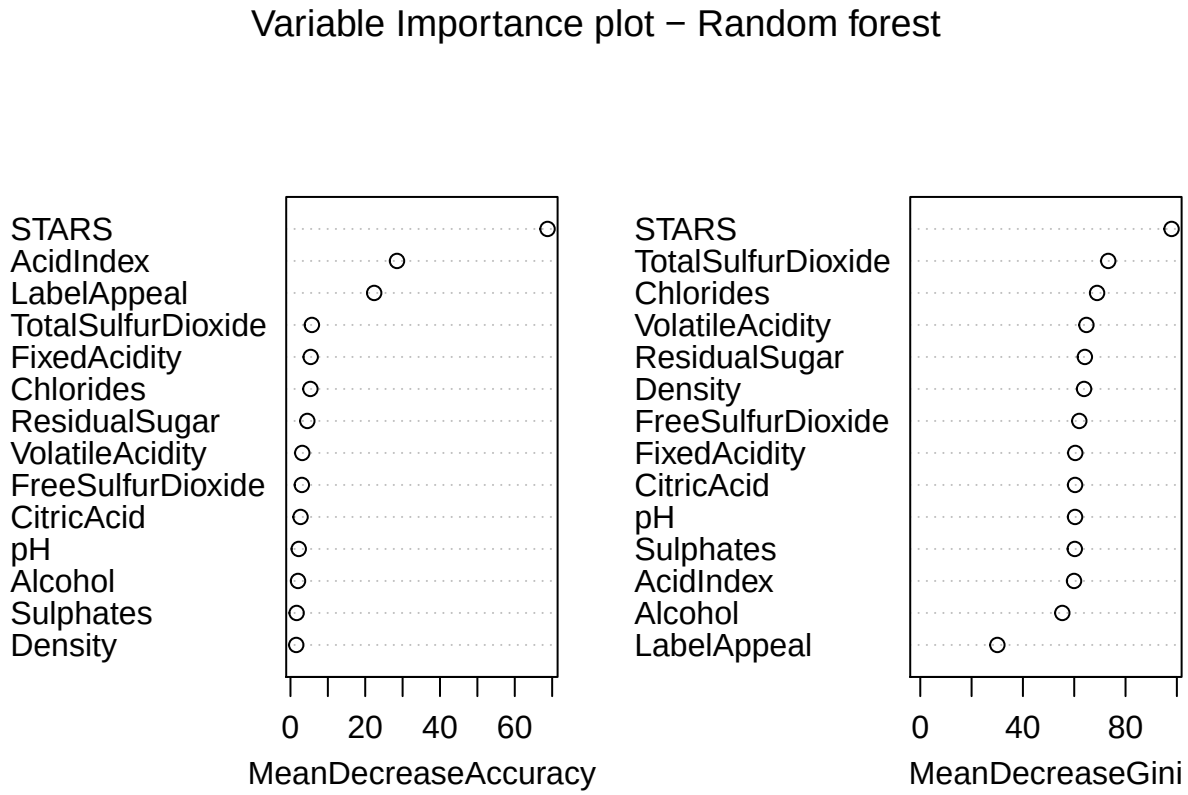

Figure 5: Variable Importance - Randomforest

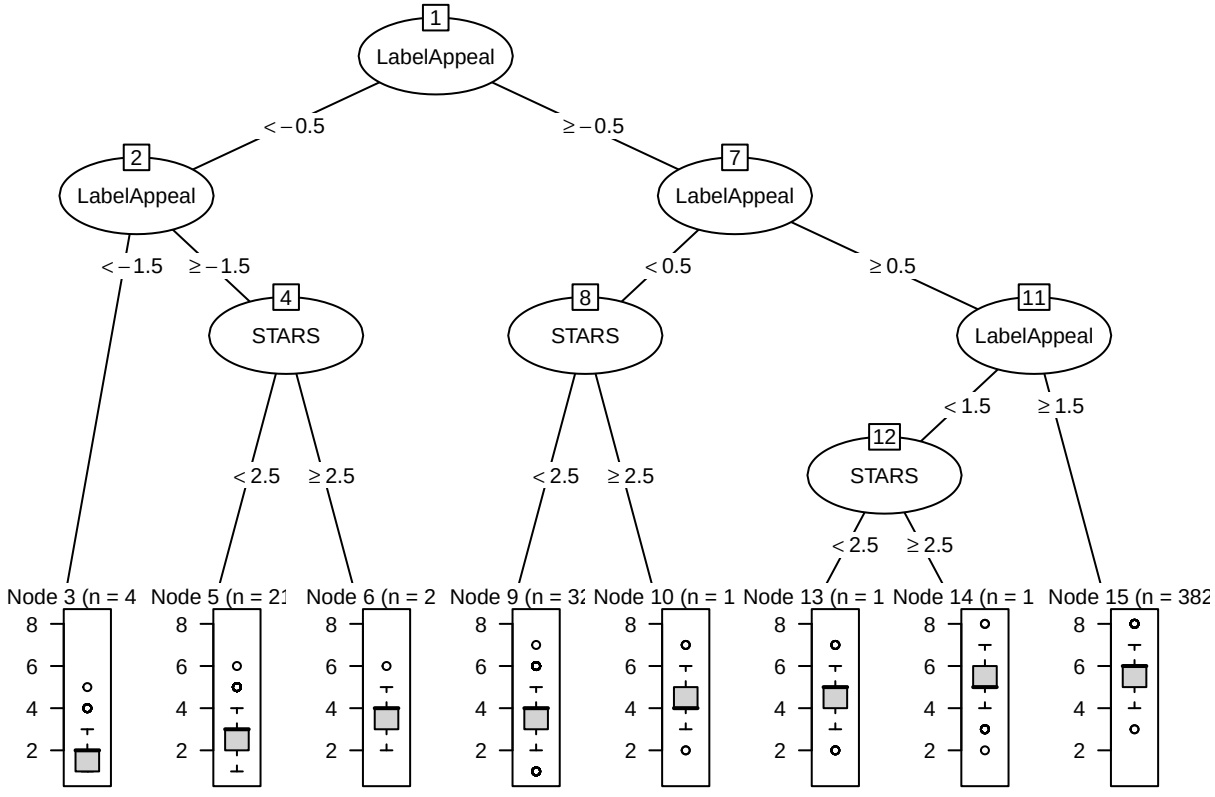

Figure 6: Simple tree for TARGET

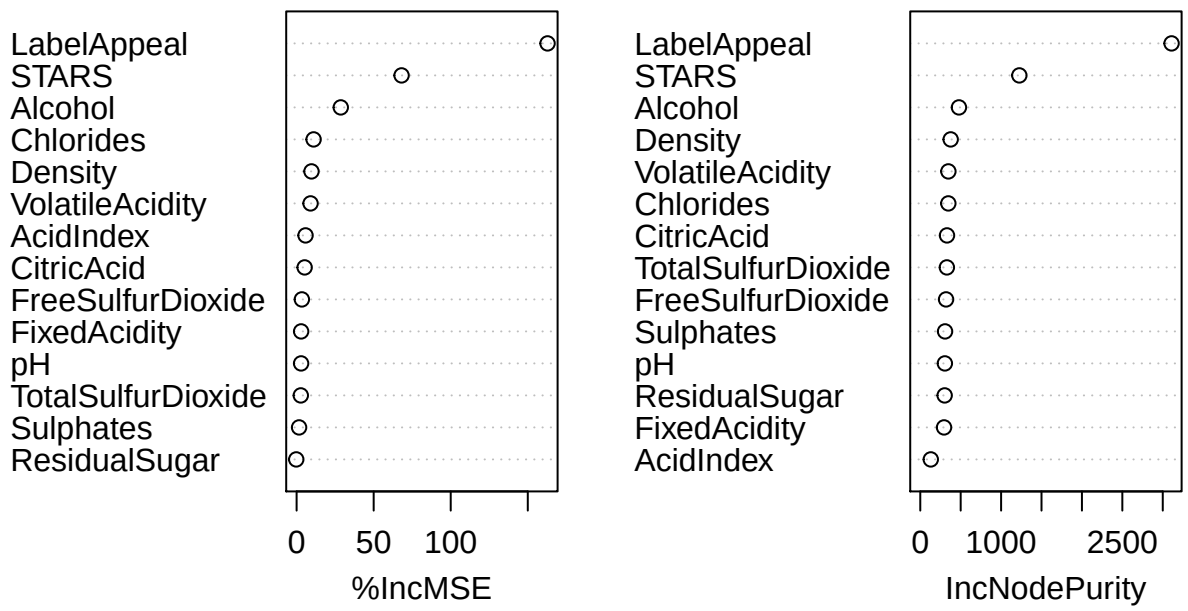

Figure 7: Variable importance plot - TARGET

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.

5. Modeling

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

- 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.

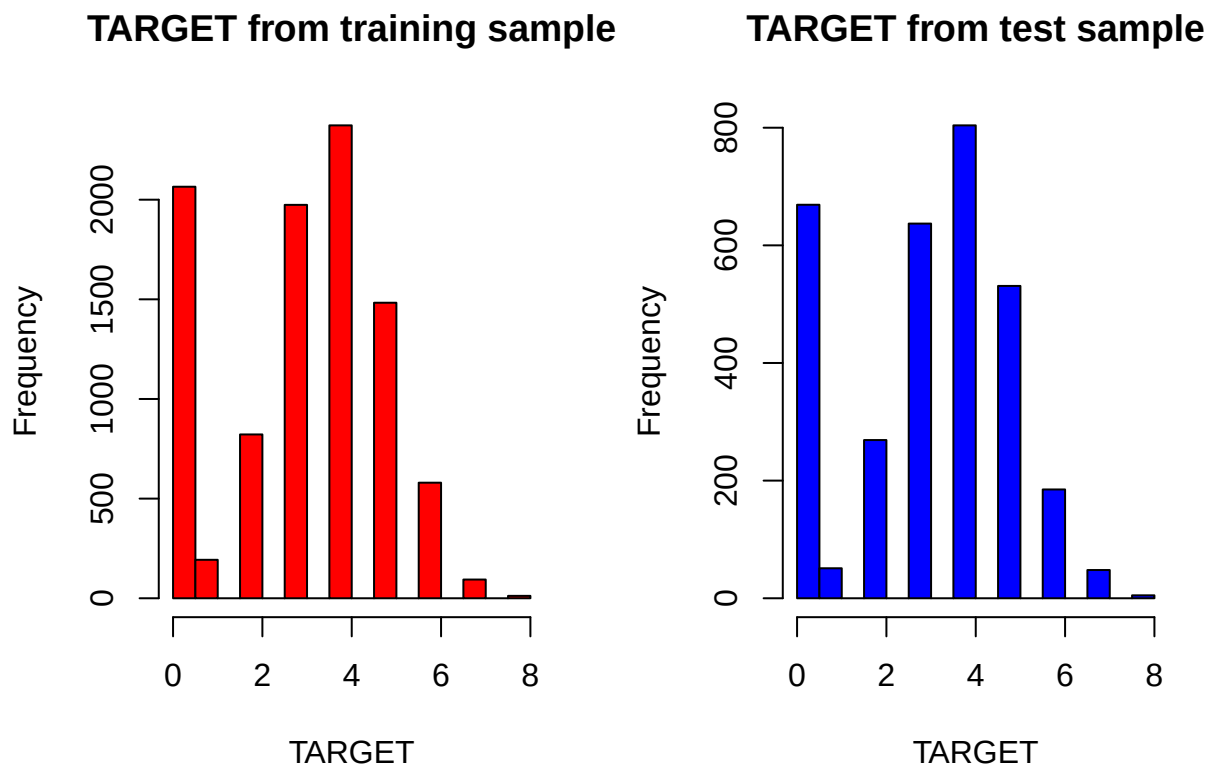

Figure 8: TARGET distribution for training and validation samples

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	1.39

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.

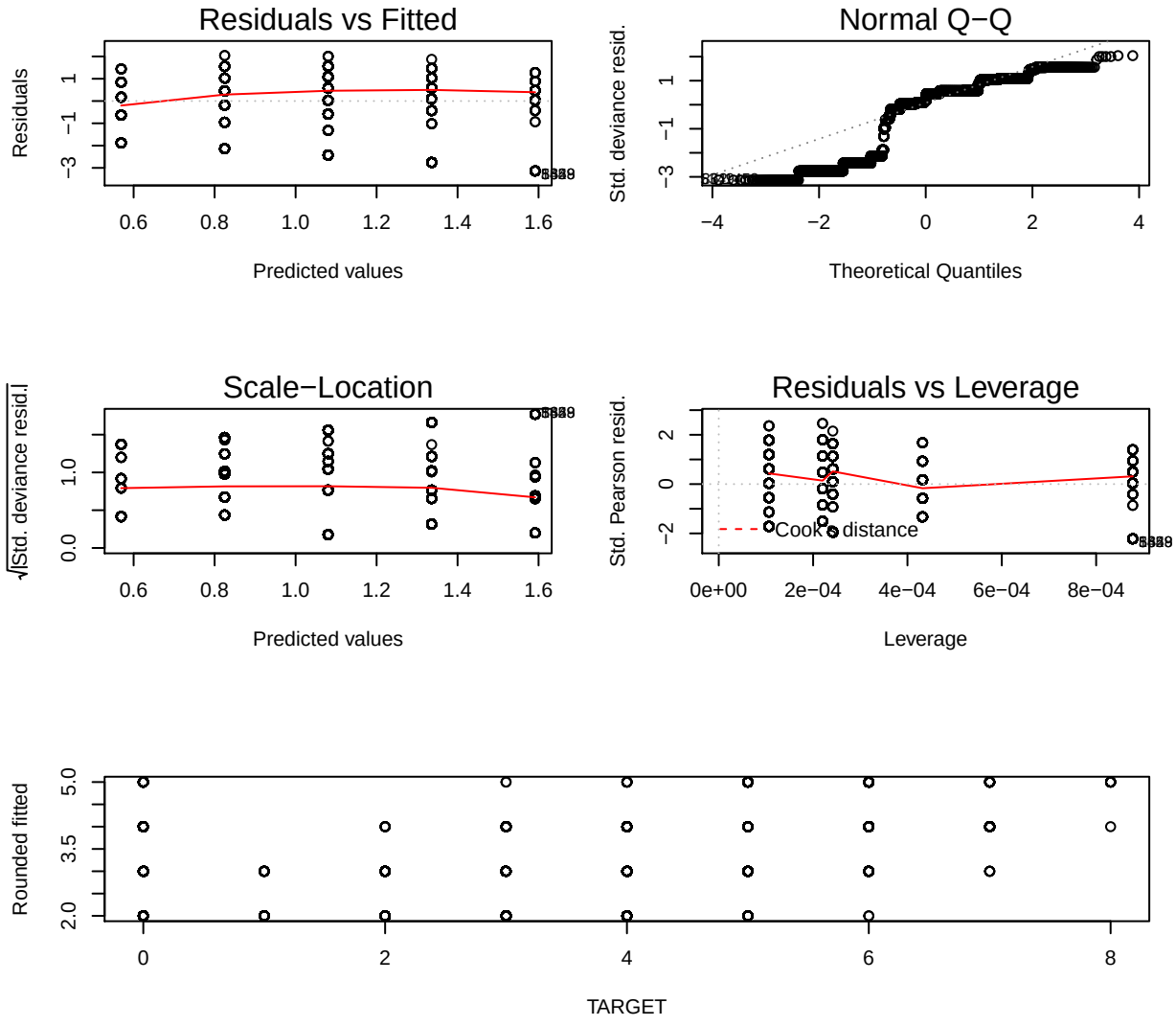

Figure 9: Simple poisson regression diagnostics