hw5

Joseph Elikishvili

May 15, 2018

### Overview

In this assignment, we 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. Your objective is to build a count regression model to predict the number of cases of wine that will be sold given certain properties of the wine. HINT: Sometimes, the fact that a variable is missing is actually predictive of the target. You can only use the variables given to you (or variables that you derive from the variables provided).

We will get started by loading the data and exploring the dimensions of the dataset and getting to know the variables

## [1] 12795 16

It appears we have over 12000 records and 16 variables, one is our target variable and one of the predictor variables is just an index whi we will remove so we have 14 predictor variables, 1 target and 12795 records.

Next we will preview datatypes in each of the columns

## 'data.frame': 12795 obs. of 15 variables:  
## $ TARGET : int 3 3 5 3 4 0 0 4 3 6 ...  
## $ FixedAcidity : num 3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...  
## $ VolatileAcidity : num 1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...  
## $ CitricAcid : num -0.98 -0.81 -0.88 0.04 -1.26 0.59 -0.4 0.34 1.05 0.39 ...  
## $ ResidualSugar : num 54.2 26.1 14.8 18.8 9.4 ...  
## $ Chlorides : num -0.567 -0.425 0.037 -0.425 NA 0.556 0.06 0.04 -0.007 -0.277 ...  
## $ FreeSulfurDioxide : num NA 15 214 22 -167 -37 287 523 -213 62 ...  
## $ TotalSulfurDioxide: num 268 -327 142 115 108 15 156 551 NA 180 ...  
## $ Density : num 0.993 1.028 0.995 0.996 0.995 ...  
## $ pH : num 3.33 3.38 3.12 2.24 3.12 3.2 3.49 3.2 4.93 3.09 ...  
## $ Sulphates : num -0.59 0.7 0.48 1.83 1.77 1.29 1.21 NA 0.26 0.75 ...  
## $ Alcohol : num 9.9 NA 22 6.2 13.7 15.4 10.3 11.6 15 12.6 ...  
## $ LabelAppeal : int 0 -1 -1 -1 0 0 0 1 0 0 ...  
## $ AcidIndex : int 8 7 8 6 9 11 8 7 6 8 ...  
## $ STARS : int 2 3 3 1 2 NA NA 3 NA 4 ...

It appears that most of the variables are either in the num or int format which is good for us and will require less data transformation.

Now our data is clean and we can move to the next step and check data for any missing values and develop a strategy to deal with those if we find any

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

Lets also visually check to see if there are any missing values, we will use Amelia library to do that and review results.

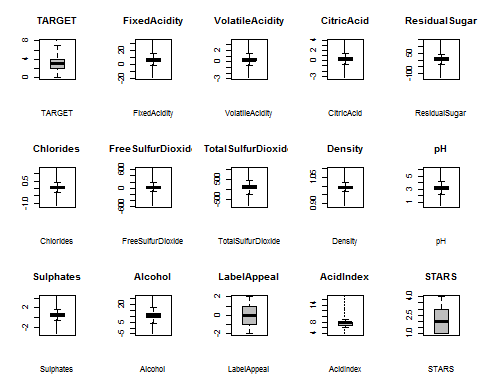


It appears there are 8 predictor variables that have some missing data and the amount of missing data ranges from 395 to 3359 in case of STARS. But overal we have a good dataset and a good number of records to work with.

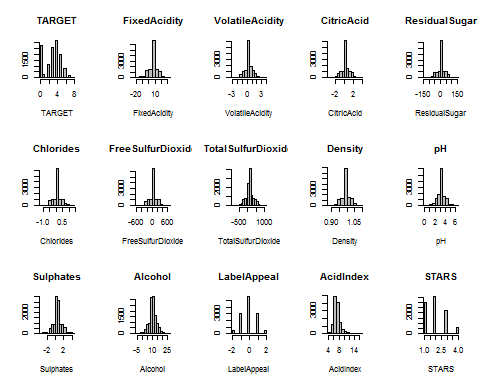
Next we will review the predictors given to us to better understand the data we are dealing with. The table gives us a basic overview of the data

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | vars | n | mean | sd | median | trimmed | mad | min | max | range | skew | kurtosis | se |
| TARGET | 1 | 12795 | 3.03 | 1.93 | 3.00 | 3.05 | 1.48 | 0.00 | 8.00 | 8.00 | -0.33 | -0.88 | 0.02 |
| FixedAcidity | 2 | 12795 | 7.08 | 6.32 | 6.90 | 7.07 | 3.26 | -18.10 | 34.40 | 52.50 | -0.02 | 1.67 | 0.06 |
| VolatileAcidity | 3 | 12795 | 0.32 | 0.78 | 0.28 | 0.32 | 0.43 | -2.79 | 3.68 | 6.47 | 0.02 | 1.83 | 0.01 |
| CitricAcid | 4 | 12795 | 0.31 | 0.86 | 0.31 | 0.31 | 0.42 | -3.24 | 3.86 | 7.10 | -0.05 | 1.84 | 0.01 |
| ResidualSugar | 5 | 12179 | 5.42 | 33.75 | 3.90 | 5.58 | 15.72 | -127.80 | 141.15 | 268.95 | -0.05 | 1.88 | 0.31 |
| Chlorides | 6 | 12157 | 0.05 | 0.32 | 0.05 | 0.05 | 0.13 | -1.17 | 1.35 | 2.52 | 0.03 | 1.79 | 0.00 |
| FreeSulfurDioxide | 7 | 12148 | 30.85 | 148.71 | 30.00 | 30.93 | 56.34 | -555.00 | 623.00 | 1178.00 | 0.01 | 1.84 | 1.35 |
| TotalSulfurDioxide | 8 | 12113 | 120.71 | 231.91 | 123.00 | 120.89 | 134.92 | -823.00 | 1057.00 | 1880.00 | -0.01 | 1.67 | 2.11 |
| Density | 9 | 12795 | 0.99 | 0.03 | 0.99 | 0.99 | 0.01 | 0.89 | 1.10 | 0.21 | -0.02 | 1.90 | 0.00 |
| pH | 10 | 12400 | 3.21 | 0.68 | 3.20 | 3.21 | 0.39 | 0.48 | 6.13 | 5.65 | 0.04 | 1.65 | 0.01 |
| Sulphates | 11 | 11585 | 0.53 | 0.93 | 0.50 | 0.53 | 0.44 | -3.13 | 4.24 | 7.37 | 0.01 | 1.75 | 0.01 |
| Alcohol | 12 | 12142 | 10.49 | 3.73 | 10.40 | 10.50 | 2.37 | -4.70 | 26.50 | 31.20 | -0.03 | 1.54 | 0.03 |
| LabelAppeal | 13 | 12795 | -0.01 | 0.89 | 0.00 | -0.01 | 1.48 | -2.00 | 2.00 | 4.00 | 0.01 | -0.26 | 0.01 |
| AcidIndex | 14 | 12795 | 7.77 | 1.32 | 8.00 | 7.64 | 1.48 | 4.00 | 17.00 | 13.00 | 1.65 | 5.19 | 0.01 |
| STARS | 15 | 9436 | 2.04 | 0.90 | 2.00 | 1.97 | 1.48 | 1.00 | 4.00 | 3.00 | 0.45 | -0.69 | 0.01 |

Next we will create a boxplot of each of the predictors to visualize the variability of the data



And finally we will review the histograms of the predictors to see the distribution and the skew.



It appears moxt of the data is well deistributed and well centered and does not require much transformation. We do however see some negative values and they will need to be adressed. Also we need to impute some of the missing data.

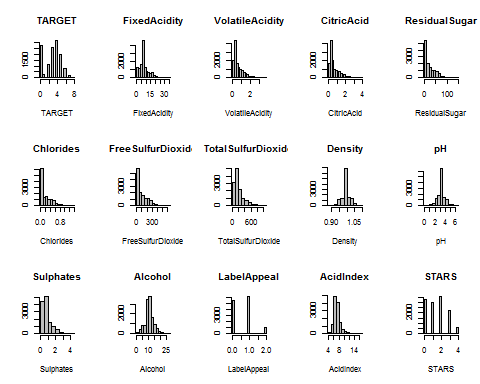
We will start with Stars and replace N/a with 0 with the assumption that the 0 stars were simply not rated and not given a rating

Next we will get rid of all the negative values in the data set since most of the predictors cannot have a negative value, we are making an assumption that it was simly a mistake and we will use an absolute value to remove any negative signs and retain the data we prefer this method over simply replacing the negative values by 0 since that would potentially alter data and have a major impact.

Next we will impute the missing data in the rest of the predictor variables. we will use missforest library, which takes a while to run but does a great job with imputation

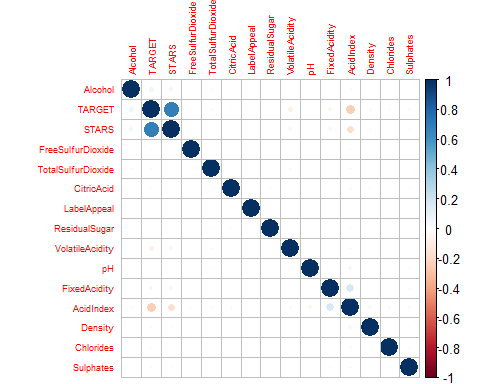
## missForest iteration 1 in progress...done!  
## missForest iteration 2 in progress...done!  
## missForest iteration 3 in progress...done!

At this point our data shoould be ready for the nect step, Lets review histograms after imputation



It appears that the data is now skewed to the right, but since we performed a removal of the negative sign, we will go ahead and leave the data unchanged and will not perofrm a box-cox transformation as it will take us one step further from the original raw data.

Finally we will review a Corelation plot to see any existing correlations within out dataset.



We do not see any major correlations that need to be adressed, so we are ready to start building our models

## Build Models

Before we proceed, We will create a holdout data set for validation, we will use 70% for training and 30% for validation purposes.

## [1] 12795 15

## [1] 8956 15

## [1] 3839 15

### Poison Models

We will start with poison model and will start first with the model that takes all predictors.

##   
## Call:  
## glm(formula = TARGET ~ ., family = "poisson", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6924 -0.8773 0.0428 0.5852 3.5278   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.244e+00 3.518e-01 3.537 0.000405 \*\*\*  
## FixedAcidity 5.011e-04 1.884e-03 0.266 0.790270   
## VolatileAcidity -3.426e-02 1.703e-02 -2.012 0.044243 \*   
## CitricAcid 1.370e-02 1.543e-02 0.888 0.374451   
## ResidualSugar -3.927e-04 3.902e-04 -1.007 0.314121   
## Chlorides -5.404e-02 4.112e-02 -1.314 0.188796   
## FreeSulfurDioxide 4.876e-05 8.608e-05 0.566 0.571133   
## TotalSulfurDioxide 7.861e-05 5.866e-05 1.340 0.180168   
## Density -1.707e-01 3.429e-01 -0.498 0.618673   
## pH 3.006e-03 1.371e-02 0.219 0.826444   
## Sulphates -2.419e-02 1.534e-02 -1.578 0.114677   
## Alcohol 5.837e-03 2.619e-03 2.229 0.025833 \*   
## LabelAppeal -8.228e-03 1.501e-02 -0.548 0.583605   
## AcidIndex -7.749e-02 8.188e-03 -9.464 < 2e-16 \*\*\*  
## STARS 3.373e-01 7.819e-03 43.139 < 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: 6874.7 on 3838 degrees of freedom  
## Residual deviance: 4654.1 on 3824 degrees of freedom  
## AIC: 14281  
##   
## Number of Fisher Scoring iterations: 5

Next we will use Stepwise backward method to eliminate the insignificant variables.

##   
## Call:  
## glm(formula = TARGET ~ VolatileAcidity + Sulphates + Alcohol +   
## AcidIndex + STARS, family = "poisson", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6756 -0.8884 0.0459 0.5795 3.5095   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.098722 0.074133 14.821 <2e-16 \*\*\*  
## VolatileAcidity -0.036078 0.016966 -2.127 0.0335 \*   
## Sulphates -0.025268 0.015306 -1.651 0.0988 .   
## Alcohol 0.005679 0.002615 2.171 0.0299 \*   
## AcidIndex -0.077725 0.008086 -9.613 <2e-16 \*\*\*  
## STARS 0.337466 0.007797 43.283 <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: 6874.7 on 3838 degrees of freedom  
## Residual deviance: 4660.3 on 3833 degrees of freedom  
## AIC: 14269  
##   
## Number of Fisher Scoring iterations: 5

### Negative Binomial Models

We will do the same for the binomial model and use all predictors first for a baseline model

##   
## Call:  
## glm.nb(formula = TARGET ~ ., data = train, init.theta = 42108.80059,   
## link = log)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6924 -0.8773 0.0428 0.5852 3.5277   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.244e+00 3.518e-01 3.537 0.000405 \*\*\*  
## FixedAcidity 5.012e-04 1.884e-03 0.266 0.790258   
## VolatileAcidity -3.426e-02 1.703e-02 -2.012 0.044246 \*   
## CitricAcid 1.370e-02 1.543e-02 0.888 0.374450   
## ResidualSugar -3.928e-04 3.902e-04 -1.007 0.314129   
## Chlorides -5.404e-02 4.112e-02 -1.314 0.188811   
## FreeSulfurDioxide 4.876e-05 8.609e-05 0.566 0.571145   
## TotalSulfurDioxide 7.862e-05 5.866e-05 1.340 0.180160   
## Density -1.707e-01 3.429e-01 -0.498 0.618673   
## pH 3.006e-03 1.371e-02 0.219 0.826483   
## Sulphates -2.419e-02 1.534e-02 -1.578 0.114668   
## Alcohol 5.837e-03 2.619e-03 2.229 0.025838 \*   
## LabelAppeal -8.231e-03 1.501e-02 -0.548 0.583507   
## AcidIndex -7.750e-02 8.188e-03 -9.464 < 2e-16 \*\*\*  
## STARS 3.373e-01 7.820e-03 43.137 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Negative Binomial(42108.8) family taken to be 1)  
##   
## Null deviance: 6874.4 on 3838 degrees of freedom  
## Residual deviance: 4653.9 on 3824 degrees of freedom  
## AIC: 14283  
##   
## Number of Fisher Scoring iterations: 1  
##   
##   
## Theta: 42109   
## Std. Err.: 82789   
## Warning while fitting theta: iteration limit reached   
##   
## 2 x log-likelihood: -14250.84

And use the Stepwise backwards method for the second model.

##   
## Call:  
## glm.nb(formula = TARGET ~ VolatileAcidity + Sulphates + Alcohol +   
## AcidIndex + STARS, data = train, init.theta = 42098.9714,   
## link = log)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6755 -0.8884 0.0459 0.5795 3.5094   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.098735 0.074136 14.821 <2e-16 \*\*\*  
## VolatileAcidity -0.036079 0.016967 -2.126 0.0335 \*   
## Sulphates -0.025270 0.015306 -1.651 0.0988 .   
## Alcohol 0.005679 0.002615 2.171 0.0299 \*   
## AcidIndex -0.077727 0.008086 -9.613 <2e-16 \*\*\*  
## STARS 0.337471 0.007797 43.282 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Negative Binomial(42098.97) family taken to be 1)  
##   
## Null deviance: 6874.4 on 3838 degrees of freedom  
## Residual deviance: 4660.1 on 3833 degrees of freedom  
## AIC: 14271  
##   
## Number of Fisher Scoring iterations: 1  
##   
##   
## Theta: 42099   
## Std. Err.: 82745   
## Warning while fitting theta: iteration limit reached   
##   
## 2 x log-likelihood: -14257.09

### Linear Models

Next we will create the linear model

##   
## Call:  
## lm(formula = TARGET ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8450 -1.0285 0.1141 0.9963 6.5197   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.397e+00 8.587e-01 3.956 7.77e-05 \*\*\*  
## FixedAcidity 3.137e-03 4.606e-03 0.681 0.49581   
## VolatileAcidity -1.003e-01 4.080e-02 -2.459 0.01398 \*   
## CitricAcid 2.861e-02 3.747e-02 0.764 0.44516   
## ResidualSugar -1.259e-03 9.329e-04 -1.349 0.17733   
## Chlorides -1.407e-01 9.949e-02 -1.415 0.15729   
## FreeSulfurDioxide 8.321e-05 2.112e-04 0.394 0.69355   
## TotalSulfurDioxide 2.207e-04 1.446e-04 1.527 0.12693   
## Density -6.779e-01 8.414e-01 -0.806 0.42048   
## pH 2.052e-02 3.335e-02 0.615 0.53851   
## Sulphates -7.316e-02 3.670e-02 -1.993 0.04629 \*   
## Alcohol 2.071e-02 6.382e-03 3.245 0.00118 \*\*   
## LabelAppeal 9.743e-03 3.648e-02 0.267 0.78945   
## AcidIndex -1.891e-01 1.783e-02 -10.607 < 2e-16 \*\*\*  
## STARS 1.054e+00 1.928e-02 54.682 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.398 on 3824 degrees of freedom  
## Multiple R-squared: 0.4774, Adjusted R-squared: 0.4755   
## F-statistic: 249.5 on 14 and 3824 DF, p-value: < 2.2e-16

And the Lineaer model using Stepwise backwards method

##   
## Call:  
## lm(formula = TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide +   
## Sulphates + Alcohol + AcidIndex + STARS, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.8177 -1.0370 0.1255 1.0040 6.4872   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.8014179 0.1700942 16.470 < 2e-16 \*\*\*  
## VolatileAcidity -0.1012756 0.0407175 -2.487 0.01291 \*   
## Chlorides -0.1441167 0.0993554 -1.451 0.14700   
## TotalSulfurDioxide 0.0002179 0.0001443 1.510 0.13107   
## Sulphates -0.0726253 0.0366546 -1.981 0.04762 \*   
## Alcohol 0.0206227 0.0063758 3.235 0.00123 \*\*   
## AcidIndex -0.1865397 0.0175233 -10.645 < 2e-16 \*\*\*  
## STARS 1.0551340 0.0192483 54.817 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.398 on 3831 degrees of freedom  
## Multiple R-squared: 0.4768, Adjusted R-squared: 0.4759   
## F-statistic: 498.8 on 7 and 3831 DF, p-value: < 2.2e-16

### Zero Inflated Poison

Finally we will use Zero inflated model which is used when we have an excess count of 0 values and it could be a suitable for our case. We will start with a base model

##   
## Call:  
## zeroinfl(formula = TARGET ~ ., data = train, dist = "negbin")  
##   
## Pearson residuals:  
## Min 1Q Median 3Q Max   
## -1.94619 -0.47663 0.01609 0.46840 4.34607   
##   
## Count model coefficients (negbin with log link):  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.424e+00 3.655e-01 3.895 9.82e-05 \*\*\*  
## FixedAcidity -1.850e-04 1.942e-03 -0.095 0.92411   
## VolatileAcidity -2.066e-02 1.734e-02 -1.192 0.23337   
## CitricAcid -2.113e-02 1.612e-02 -1.311 0.19002   
## ResidualSugar -3.671e-04 4.056e-04 -0.905 0.36549   
## Chlorides -4.104e-02 4.210e-02 -0.975 0.32965   
## FreeSulfurDioxide 1.489e-05 8.748e-05 0.170 0.86484   
## TotalSulfurDioxide -1.536e-05 5.879e-05 -0.261 0.79385   
## Density -4.620e-01 3.557e-01 -1.299 0.19401   
## pH 1.911e-02 1.417e-02 1.349 0.17737   
## Sulphates 8.713e-03 1.556e-02 0.560 0.57561   
## Alcohol 7.392e-03 2.659e-03 2.780 0.00544 \*\*   
## LabelAppeal -1.279e-02 1.543e-02 -0.829 0.40723   
## AcidIndex -6.201e-03 8.912e-03 -0.696 0.48655   
## STARS 1.662e-01 8.858e-03 18.763 < 2e-16 \*\*\*  
## Log(theta) 1.741e+01 NA NA NA   
##   
## Zero-inflation model coefficients (binomial with logit link):  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.3762561 2.3241659 -0.592 0.55375   
## FixedAcidity -0.0210265 0.0129121 -1.628 0.10343   
## VolatileAcidity 0.1629739 0.1069293 1.524 0.12748   
## CitricAcid -0.3009691 0.1047175 -2.874 0.00405 \*\*   
## ResidualSugar 0.0009155 0.0024854 0.368 0.71260   
## Chlorides -0.0666691 0.2740362 -0.243 0.80778   
## FreeSulfurDioxide 0.0000910 0.0005988 0.152 0.87921   
## TotalSulfurDioxide -0.0009132 0.0003964 -2.304 0.02125 \*   
## Density -2.4841518 2.2949956 -1.082 0.27907   
## pH 0.1236274 0.0868740 1.423 0.15472   
## Sulphates 0.3783487 0.0946208 3.999 6.37e-05 \*\*\*  
## Alcohol -0.0018011 0.0167326 -0.108 0.91428   
## LabelAppeal -0.2309362 0.0978638 -2.360 0.01829 \*   
## AcidIndex 0.5095160 0.0476341 10.696 < 2e-16 \*\*\*  
## STARS -2.2242679 0.1078714 -20.620 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1   
##   
## Theta = 36318237.0178   
## Number of iterations in BFGS optimization: 64   
## Log-likelihood: -6374 on 31 Df

And finally we will use zero inflation model combined with Stepwise backward selection process

##   
## Call:  
## zeroinfl(formula = TARGET ~ VolatileAcidity + CitricAcid + TotalSulfurDioxide +   
## Sulphates + Alcohol + LabelAppeal + AcidIndex + STARS, data = train,   
## dist = "negbin")  
##   
## Pearson residuals:  
## Min 1Q Median 3Q Max   
## -1.94185 -0.47774 0.02866 0.48081 4.37248   
##   
## Count model coefficients (negbin with log link):  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.020e+00 8.135e-02 12.539 < 2e-16 \*\*\*  
## VolatileAcidity -1.940e-02 1.730e-02 -1.122 0.26205   
## CitricAcid -2.125e-02 1.612e-02 -1.319 0.18723   
## TotalSulfurDioxide -2.111e-05 5.872e-05 -0.360 0.71917   
## Sulphates 8.593e-03 1.555e-02 0.553 0.58048   
## Alcohol 7.348e-03 2.657e-03 2.765 0.00569 \*\*   
## LabelAppeal -1.176e-02 1.543e-02 -0.762 0.44599   
## AcidIndex -7.734e-03 8.832e-03 -0.876 0.38119   
## STARS 1.667e-01 8.850e-03 18.835 < 2e-16 \*\*\*  
## Log(theta) 1.437e+01 1.304e+01 1.102 0.27050   
##   
## Zero-inflation model coefficients (binomial with logit link):  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.4256701 0.4333805 -7.905 2.69e-15 \*\*\*  
## VolatileAcidity 0.1814250 0.1056469 1.717 0.08593 .   
## CitricAcid -0.2937939 0.1041143 -2.822 0.00477 \*\*   
## TotalSulfurDioxide -0.0009205 0.0003954 -2.328 0.01992 \*   
## Sulphates 0.3758818 0.0943514 3.984 6.78e-05 \*\*\*  
## Alcohol -0.0022979 0.0166879 -0.138 0.89048   
## LabelAppeal -0.2298365 0.0976745 -2.353 0.01862 \*   
## AcidIndex 0.4858237 0.0461263 10.532 < 2e-16 \*\*\*  
## STARS -2.2113882 0.1068870 -20.689 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1   
##   
## Theta = 1740431.18   
## Number of iterations in BFGS optimization: 40   
## Log-likelihood: -6380 on 19 Df

## Select Models

Since we have diffirerent metrics for different methods, wre will need a unified method of measure the performance, so we will go ahead and use the validation set we set aside and then calculate the RMSE and compare the results.

## Model RMSE  
## 2 Model1 1.44151230539469  
## 3 Model2 1.44143166297745  
## 4 Model3 1.44151546869377  
## 5 Model4 1.44143463320284  
## 6 Model5 1.36983910739305  
## 7 Model6 1.36935012338059  
## 8 Model7 1.36571444796401  
## 9 Model8 1.36398042743749

It appears that the Zero inflation model combined with stepwise backwards selection is the most accurate model, so we will use that model to make our predictions.

## Predicting on testdata

We will go ahead and use model8 to make our forcasts. But before we do that we will need to do the same data imputation and transformation as with the train set. So we will need to do the following: 1. transform STARS predictor 2. Remove index 3. Use absilute values 4. Impute the missing values 5. Predict the target and round the result so that we have round number of cases.

Once we are done we will preview first 30 records of our prediction dataset.

## removed variable(s) 1 due to the missingness of all entries  
## missForest iteration 1 in progress...done!  
## missForest iteration 2 in progress...done!  
## missForest iteration 3 in progress...done!  
## missForest iteration 4 in progress...done!

## Appendix A

R markdown file with code along with full predictions csv file available at: <https://github.com/jelikish/Cuny1/tree/master/Spring2018/621/hw5>