DATA621 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 data set 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 that we will remove so we have 14 predictor variables, 1 target and 12795 records.

Next we will preview datatypes in each of the columns

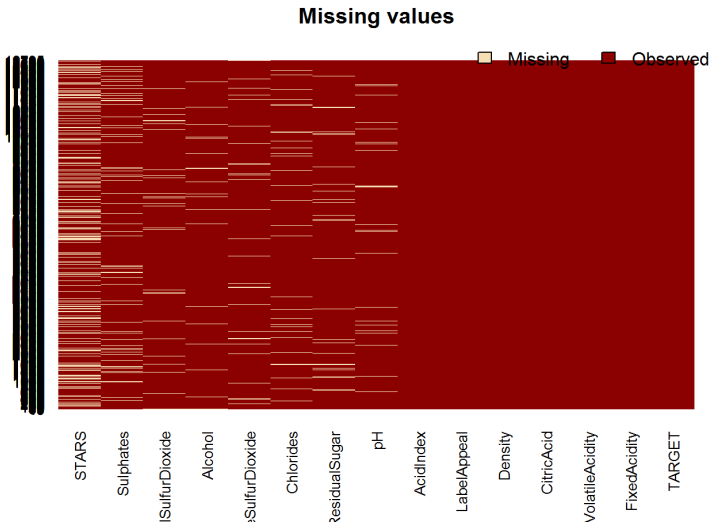
```
## 'data.frame':    12795 obs. of  15 variables:
##  $ TARGET           : int  3 3 5 3 4 0 0 4 3 6 ...
##  $ FixedAcidity     : num  3.2 4.5 7.1 5.7 8 11.3 7.7 6.5 14.8 5.5 ...
##  $ VolatileAcidity  : num  1.16 0.16 2.64 0.385 0.33 0.32 0.29 -1.22 0.27 -0.22 ...
##  $ CitricAcid       : num  -0.98 -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.

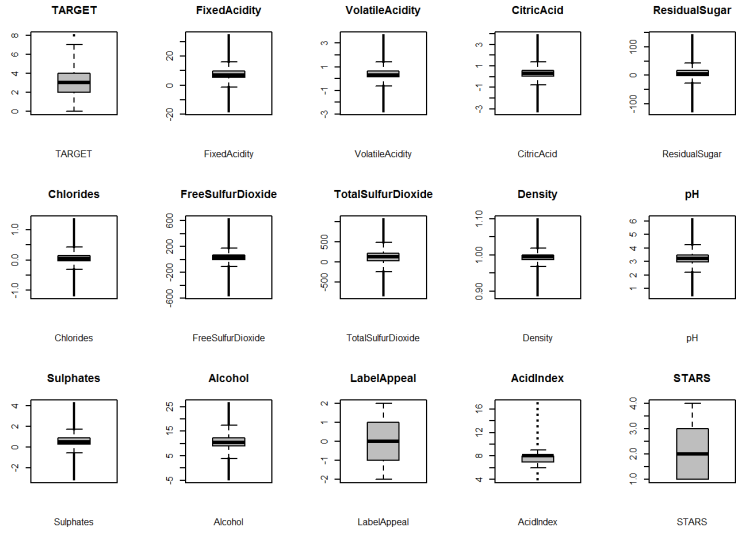


It appears there are 8 predictor variables that have some missing data and the amount of missing data ranges from 395 to 3359 in case of STARS. But overall we have a good data set and a good number of records to work with.

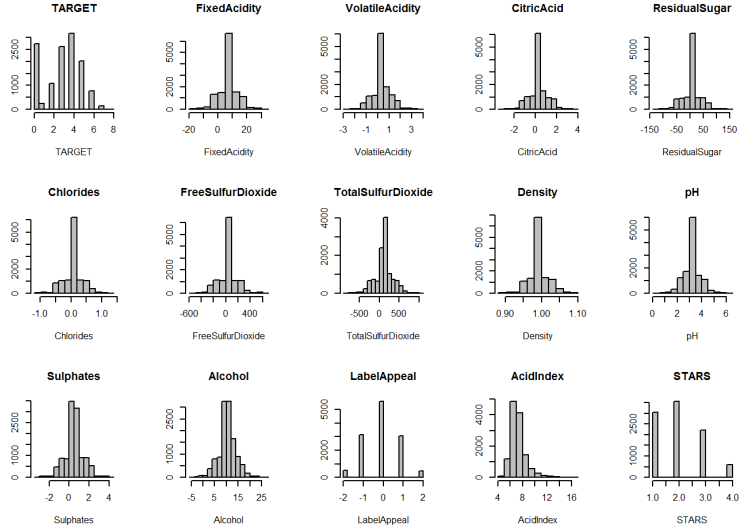
Next we will review the predictors given to us to better understand the data we are dealing with. The table gives us a basic overview of the data

	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 box plot 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 most of the data is well distributed and well centered and does not require much transformation. We do however see some negative values and they will need to be addressed. Also we need to impute some of the missing data.

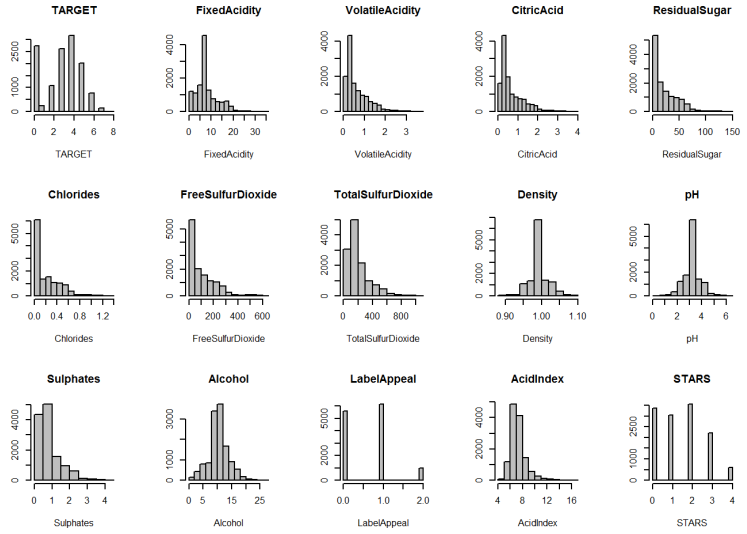
We will start with Stars and replace N/a with 0 with the assumption that the 0 stars were simply not rated and not given a rating

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 simply 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

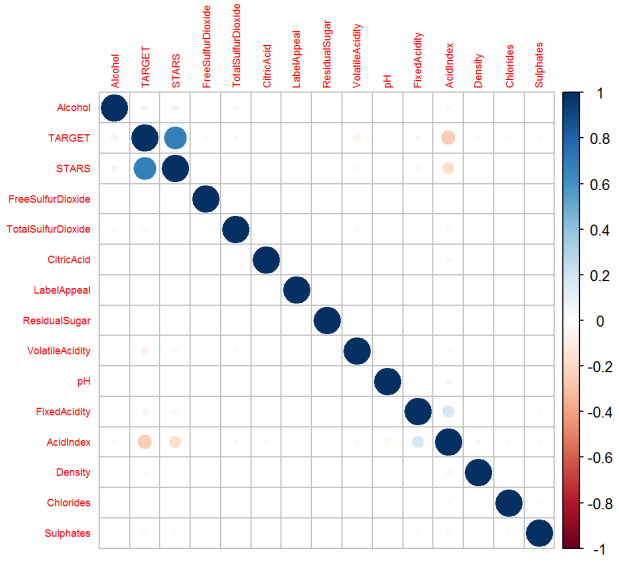
```
##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
##   missForest iteration 4 in progress...done!
```

At this point our data should be ready for the next 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 perform a box-cox transformation as it will take us one step further from the original raw data.

Finally we will review a Correlation plot to see any existing correlations within out data set.



We do not see any major correlations that need to be addressed, 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
## -2.7150  -0.8523   0.0215   0.5809   3.5494
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.622e+00  3.609e-01   4.495 6.96e-06 ***
## FixedAcidity       -8.940e-04  1.932e-03  -0.463  0.64353
## VolatileAcidity    -5.278e-02  1.719e-02  -3.069  0.00214 **
## CitricAcid         -2.970e-03  1.518e-02  -0.196  0.84489
## ResidualSugar       1.445e-04  3.780e-04   0.382  0.70216
## Chlorides          -6.967e-02  4.004e-02  -1.740  0.08184 .
## FreeSulfurDioxide   8.470e-05  8.697e-05   0.974  0.33012
## TotalSulfurDioxide  1.498e-04  5.848e-05   2.562  0.01041 *
## Density            -5.375e-01  3.540e-01  -1.518  0.12895
## pH                 -4.647e-04  1.360e-02  -0.034  0.97275
## Sulphates          -1.630e-04  1.498e-02  -0.011  0.99132
## Alcohol             1.567e-03  2.624e-03   0.597  0.55035
## LabelAppeal        -1.921e-02  1.499e-02  -1.281  0.20014
## AcidIndex          -7.427e-02  8.282e-03  -8.968  < 2e-16 ***
## STARS               3.447e-01  7.948e-03  43.365  < 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: 6839.2  on 3838  degrees of freedom
## Residual deviance: 4597.5  on 3824  degrees of freedom
## AIC: 14238
##
## Number of Fisher Scoring iterations: 5
```

Next we will use Step-wise backward method to eliminate the insignificant variables.

```
##
## Call:
## glm(formula = TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide +
##     Density + AcidIndex + STARS, family = "poisson", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7243  -0.8601   0.0266   0.5766   3.4987
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.642e+00  3.554e-01   4.621 3.82e-06 ***
## VolatileAcidity    -5.346e-02  1.718e-02  -3.112  0.00186 **
## Chlorides          -6.927e-02  4.001e-02  -1.731  0.08341 .
## TotalSulfurDioxide  1.499e-04  5.834e-05   2.569  0.01019 *
## Density            -5.419e-01  3.536e-01  -1.532  0.12542
## AcidIndex          -7.548e-02  8.153e-03  -9.259  < 2e-16 ***
## STARS               3.446e-01  7.896e-03  43.638  < 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: 6839.2  on 3838  degrees of freedom
## Residual deviance: 4600.8  on 3832  degrees of freedom
## AIC: 14226
##
## 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 = 42347.96844,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7149  -0.8523   0.0215   0.5809   3.5492
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.622e+00  3.610e-01   4.495 6.96e-06 ***
## FixedAcidity       -8.941e-04  1.932e-03  -0.463  0.64352
## VolatileAcidity    -5.278e-02  1.719e-02  -3.069  0.00215 **
## CitricAcid         -2.970e-03  1.518e-02  -0.196  0.84488
## ResidualSugar       1.445e-04  3.780e-04   0.382  0.70217
## Chlorides          -6.967e-02  4.004e-02  -1.740  0.08185 .
## FreeSulfurDioxide   8.470e-05  8.698e-05   0.974  0.33014
## TotalSulfurDioxide  1.498e-04  5.848e-05   2.562  0.01041 *
## Density            -5.375e-01  3.540e-01  -1.518  0.12897
## pH                 -4.653e-04  1.360e-02  -0.034  0.97271
## Sulphates          -1.635e-04  1.498e-02  -0.011  0.99129
## Alcohol             1.567e-03  2.624e-03   0.597  0.55039
## LabelAppeal        -1.921e-02  1.499e-02  -1.281  0.20009
## AcidIndex          -7.427e-02  8.282e-03  -8.968  < 2e-16 ***
## STARS               3.447e-01  7.948e-03  43.364  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(42347.97) family taken to be 1)
##
##     Null deviance: 6838.9  on 3838  degrees of freedom
## Residual deviance: 4597.3  on 3824  degrees of freedom
## AIC: 14240
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  42348
##           Std. Err.:  81792
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -14208.47
```

And use the Step-wise backwards method for the second model.

```
##
## Call:
## glm.nb(formula = TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide +
##     Density + AcidIndex + STARS, data = train, init.theta = 42462.94046,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7243  -0.8601   0.0266   0.5766   3.4986
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.642e+00  3.555e-01   4.621 3.83e-06 ***
## VolatileAcidity    -5.346e-02  1.718e-02  -3.112  0.00186 **
## Chlorides          -6.927e-02  4.001e-02  -1.731  0.08342 .
## TotalSulfurDioxide  1.499e-04  5.834e-05   2.569  0.01019 *
## Density            -5.419e-01  3.536e-01  -1.532  0.12543
## AcidIndex          -7.549e-02  8.153e-03  -9.259  < 2e-16 ***
## STARS               3.446e-01  7.897e-03  43.636  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(42462.94) family taken to be 1)
##
##     Null deviance: 6838.9  on 3838  degrees of freedom
## Residual deviance: 4600.6  on 3832  degrees of freedom
## AIC: 14228
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  42463
##           Std. Err.:  82004
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -14211.77
```

Linear Models

Next we will create the linear model

```
##
## Call:
## lm(formula = TARGET ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8519 -1.0090  0.1077  1.0235  6.5198
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.1939065  0.8731660   4.803 1.62e-06 ***
## FixedAcidity       -0.0028139  0.0046303  -0.608 0.543415
## VolatileAcidity    -0.1411451  0.0404683  -3.488 0.000493 ***
## CitricAcid          0.0064896  0.0374688   0.173 0.862504
## ResidualSugar       0.0003104  0.0009139   0.340 0.734153
## Chlorides          -0.1916039  0.0969316  -1.977 0.048148 *
## FreeSulfurDioxide   0.0001304  0.0002130   0.612 0.540479
## TotalSulfurDioxide  0.0004007  0.0001429   2.804 0.005074 **
## Density            -1.4980929  0.8596527  -1.743 0.081471 .
## pH                  0.0031002  0.0330558   0.094 0.925283
## Sulphates          -0.0021240  0.0357621  -0.059 0.952642
## Alcohol             0.0093395  0.0063804   1.464 0.143335
## LabelAppeal        -0.0278507  0.0362119  -0.769 0.441879
## AcidIndex          -0.1688439  0.0177955  -9.488  < 2e-16 ***
## STARS               1.0713644  0.0195234  54.876  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.395 on 3824 degrees of freedom
## Multiple R-squared:  0.4794, Adjusted R-squared:  0.4775
## F-statistic: 251.5 on 14 and 3824 DF,  p-value: < 2.2e-16
```

And the Linear model using Step-wise backwards method

```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + Chlorides + TotalSulfurDioxide +
##     Density + Alcohol + AcidIndex + STARS, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8133 -1.0065  0.1147  1.0228  6.5000
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.2045314  0.8651333   4.860 1.22e-06 ***
## VolatileAcidity    -0.1420528  0.0404189  -3.515 0.000446 ***
## Chlorides          -0.1917254  0.0967878  -1.981 0.047676 *
## TotalSulfurDioxide  0.0004019  0.0001425   2.821 0.004815 **
## Density            -1.4953237  0.8585692  -1.742 0.081651 .
## Alcohol             0.0093891  0.0063712   1.474 0.140648
## AcidIndex          -0.1715208  0.0174562  -9.826  < 2e-16 ***
## STARS               1.0715290  0.0194725  55.028  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.394 on 3831 degrees of freedom
## Multiple R-squared:  0.4792, Adjusted R-squared:  0.4782
## F-statistic: 503.5 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
## -1.94349 -0.47661  0.02657  0.49602  4.19253
##
## Count model coefficients (negbin with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         1.341e+00  3.746e-01   3.580 0.000344 ***
## FixedAcidity       -1.701e-04  1.999e-03  -0.085 0.932200
## VolatileAcidity    -2.035e-02  1.743e-02  -1.167 0.243030
## CitricAcid          1.339e-02  1.566e-02   0.855 0.392311
## ResidualSugar       1.832e-05  3.917e-04   0.047 0.962700
## Chlorides          -5.573e-02  4.111e-02  -1.356 0.175212
## FreeSulfurDioxide  -3.600e-05  8.876e-05  -0.406 0.685064
## TotalSulfurDioxide -1.057e-05  5.847e-05  -0.181 0.856533
## Density            -4.052e-01  3.661e-01  -1.107 0.268410
## pH                  1.788e-02  1.392e-02   1.285 0.198823
## Sulphates           1.334e-02  1.527e-02   0.874 0.382241
## Alcohol             6.785e-03  2.692e-03   2.521 0.011712 *
## LabelAppeal        -9.792e-03  1.538e-02  -0.637 0.524236
## AcidIndex          -7.024e-03  8.948e-03  -0.785 0.432521
## STARS               1.718e-01  9.004e-03  19.084  < 2e-16 ***
## Log(theta)          1.746e+01         NA      NA       NA
##
## Zero-inflation model coefficients (binomial with logit link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -4.2567587  2.3535169  -1.809  0.07050 .
## FixedAcidity        0.0085842  0.0123731   0.694  0.48782
## VolatileAcidity     0.2703979  0.1019352   2.653  0.00799 **
## CitricAcid          0.1156898  0.1001583   1.155  0.24806
## ResidualSugar      -0.0004668  0.0024043  -0.194  0.84607
## Chlorides          -0.0491714  0.2641476  -0.186  0.85233
## FreeSulfurDioxide  -0.0007161  0.0006102  -1.173  0.24060
## TotalSulfurDioxide -0.0015495  0.0003974  -3.899 9.67e-05 ***
## Density             0.4126272  2.3145126   0.178  0.85850
## pH                  0.1984785  0.0879720   2.256  0.02406 *
## Sulphates           0.1682712  0.0901609   1.866  0.06199 .
## Alcohol             0.0352118  0.0169522   2.077  0.03779 *
## LabelAppeal        -0.0961208  0.0949812  -1.012  0.31154
## AcidIndex           0.3959547  0.0433003   9.144  < 2e-16 ***
## STARS              -2.2092374  0.1058187 -20.878  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 38102338.9395
## Number of iterations in BFGS optimization: 63
## Log-likelihood: -6393 on 31 Df
```

And finally we will use zero inflation model combined with the predictor variables that were chosen by Step-wise backwards process.

```
##
## Call:
## zeroinfl(formula = TARGET ~ VolatileAcidity + FreeSulfurDioxide +
##     TotalSulfurDioxide + Sulphates + Alcohol + AcidIndex + STARS,
##     data = train, dist = "negbin")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -1.92094 -0.47543  0.03785  0.51248  4.09803
##
## Count model coefficients (negbin with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         9.954e-01  8.035e-02  12.388   <2e-16 ***
## VolatileAcidity    -2.046e-02  1.742e-02  -1.174   0.2402
## FreeSulfurDioxide  -3.910e-05  8.850e-05  -0.442   0.6586
## TotalSulfurDioxide -9.641e-06  5.833e-05  -0.165   0.8687
## Sulphates           1.334e-02  1.526e-02   0.874   0.3820
## Alcohol             6.990e-03  2.691e-03   2.597   0.0094 **
## AcidIndex          -8.644e-03  8.823e-03  -0.980   0.3272
## STARS               1.718e-01  8.982e-03  19.130   <2e-16 ***
## Log(theta)          1.733e+01  6.068e+00   2.856   0.0043 **
##
## Zero-inflation model coefficients (binomial with logit link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.1376709  0.4244290  -7.393 1.44e-13 ***
## VolatileAcidity     0.2748739  0.1015291   2.707  0.00678 **
## FreeSulfurDioxide  -0.0006922  0.0006049  -1.144  0.25249
## TotalSulfurDioxide -0.0015457  0.0003951  -3.912 9.16e-05 ***
## Sulphates           0.1631604  0.0894936   1.823  0.06828 .
## Alcohol             0.0341326  0.0169377   2.015  0.04389 *
## AcidIndex           0.3956618  0.0425036   9.309  < 2e-16 ***
## STARS              -2.2016924  0.1051672 -20.935  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 33512653.6271
## Number of iterations in BFGS optimization: 63
## Log-likelihood: -6399 on 17 Df
```

Select Models

Since we have different metrics for different methods, we 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 1.45535054020033
## 3 Model2 1.45367369223831
## 4 Model3 1.45535456506741
## 5 Model4 1.45367727240018
## 6 Model5 1.37095090236638
## 7 Model6  1.3708664003779
## 8 Model7 1.35759762870849
## 9 Model8  1.3552334336956
```

It appears that the Zero inflation model combined with step-wise 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 forecasts. 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 absolute 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 data set.

```
##   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!
```

FixedAcidity	VolatileAcidity	CitricAcid	ResidualSugar	Chlorides	FreeSulfurDioxide	TotalSulfurDioxide	Density	pH	Sulphates	Alcohol	LabelAppeal	Aci
5.4	0.860	0.27	10.700	0.092	23.000	398	0.98527	5.0200	0.6400	12.30	1	
12.4	0.385	0.76	19.700	1.169	37.000	68	0.99048	3.3700	1.0900	16.00	0	
7.2	1.750	0.17	33.000	0.065	9.000	76	1.04641	4.6100	0.6800	8.55	0	
6.2	0.100	1.80	1.000	0.179	104.000	89	0.98877	3.2000	2.1100	12.30	1	
11.4	0.210	0.28	1.200	0.038	70.000	53	1.02899	2.5400	0.0700	4.80	0	
17.6	0.040	1.15	1.400	0.535	250.000	140	0.95028	3.0600	0.0200	11.40	1	
15.5	0.530	0.53	4.600	1.263	10.000	17	1.00020	3.0700	0.7500	8.50	0	
15.9	1.190	1.14	31.900	0.299	115.000	381	1.03416	2.9900	0.3100	11.40	1	
11.6	0.320	0.55	50.900	0.076	35.000	83	1.00020	3.3200	2.1800	0.50	0	
3.8	0.220	0.31	7.700	0.039	40.000	129	0.90610	4.7200	0.6400	10.90	0	
6.8	1.680	0.44	13.300	0.046	132.785	583	1.00833	2.9888	1.6400	12.60	0	
9.0	0.210	0.04	51.400	0.237	213.000	527	0.99516	3.1600	0.7000	14.70	1	
24.6	0.030	1.20	1.300	0.035	241.000	297	0.99232	2.2200	0.5000	9.80	0	
13.0	0.210	0.32	3.200	0.263	111.000	141	0.95918	3.2000	0.7711	4.20	0	
17.9	0.420	0.91	7.100	0.045	177.000	169	0.95307	3.1700	1.1200	13.20	1	
10.0	0.200	1.27	30.900	0.050	19.000	152	0.99400	3.1758	0.4200	13.80	1	
7.4	0.290	0.50	8.500	0.480	178.000	647	0.97275	3.4500	0.5000	10.20	1	
11.7	1.180	0.94	62.000	0.675	7.000	393	0.99974	3.9600	0.6900	5.20	1	
9.7	0.410	1.00	19.718	0.235	24.000	113	0.99772	3.4400	0.5300	9.80	0	
5.2	0.980	0.08	6.400	0.046	180.000	166	0.99400	3.3000	2.1800	9.90	1	
7.3	0.190	0.49	35.550	0.380	75.000	343	0.99980	3.4200	0.3600	13.30	2	
6.7	0.780	0.31	16.400	0.030	20.000	16	0.99834	3.1100	0.5700	9.10	0	
4.1	0.540	0.30	2.200	0.088	9.000	427	0.99725	3.9800	2.0600	11.40	2	
6.5	1.530	0.84	14.400	0.047	54.000	184	0.94610	3.1700	0.5600	9.20	1	
11.8	0.270	0.46	11.750	0.316	61.000	794	1.01959	2.2000	0.0700	5.40	0	
5.0	0.160	0.78	1.400	0.898	10.000	67	0.94608	2.5000	0.4100	9.40	2	
8.3	0.260	1.93	12.700	0.042	41.000	479	0.99869	3.1000	1.6600	12.00	1	
3.3	0.230	1.39	1.800	0.043	183.000	69	0.99330	3.3800	0.3100	10.80	1	
8.7	0.300	0.30	4.500	0.053	48.500	314	0.94736	2.7300	2.0000	10.30	0	
1.2	1.010	0.32	34.379	0.038	345.000	277	0.99074	4.2900	1.0200	9.00	1	

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 (https://github.com/jelikish/Cuny1/tree/master/Spring2018/621/hw5)