Wine Sales Analysis

Logan Strouse

6/4/2019

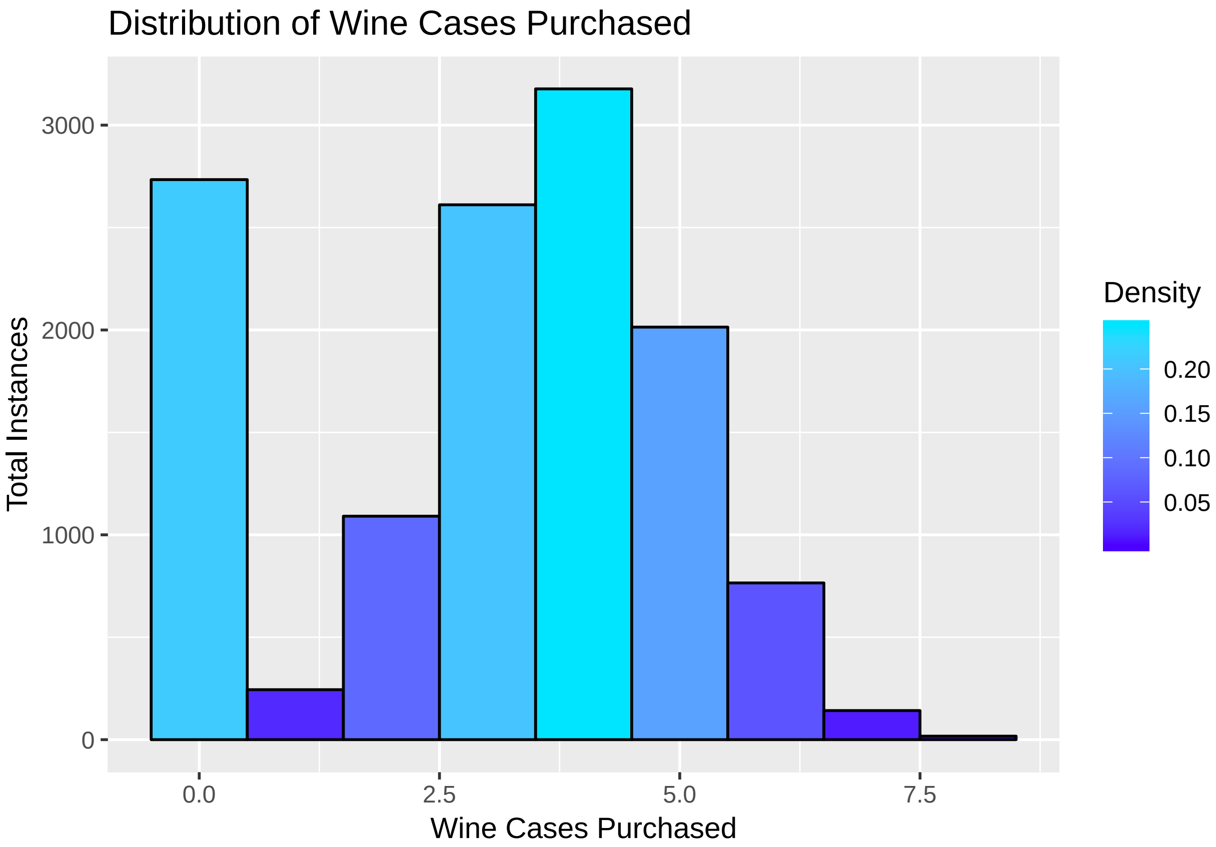
Bingo Bonus

. . . . I would like to be considered for a few bonus points for using a missForest package to impute missing variables for the first time. I had been using a different package called mice in prior assignments. I also would like to be considered for bonus points, in regards to my champion model. It was a hurdle model and took research to understand it's underworkings and to fine tune. ## Introduction

. . . .This assignment was tasking us to create a model that could successfully predict the amount of sample cases of wine a particular company would order after a sampling of the said wine. There was a multitude of different variables that were available to help with the prediction. They include: Acid Index, Alcohol, Chlorides, Citric Acid, Density, Fixed Acidity, Free Sulfur Dioxide, Label Appeal, Residual Sugar, Stars, Sulphates,Total Sulfur Dioxide,Volatile Acidity and pH. In order to most accurately predict the target variable, a group of different models and types will be created in order to choose a champion model that most accurately reflects the data available.

Data Exploration

. . . .The training data set contained 12,795 individual observations and 15 variables, not counting the index. In order to get an idea of how the target variable was distributed, I created the below histogram based on density and counts. Based on the below histogram, It appears that the variable is zero inflated. This is something to continue to be mindful of when building the models.The next step included investigating the data set further to see if there was any variables where the summary statistics appeared to be of interest. I used the describe function, as well as sapply to assess which variables had the most missing values. Below my histograms are the tables for these. After further inspection of these tables, Stars and Sulphates both appeared to have the most missing values. There is a possiblity that maybe a bunch of wines (3,359 were considered so poor of quality), that not many were ordered. I think the missing Sulphates values could be explained due to the lack of sulphate in red wine vs. white wine.



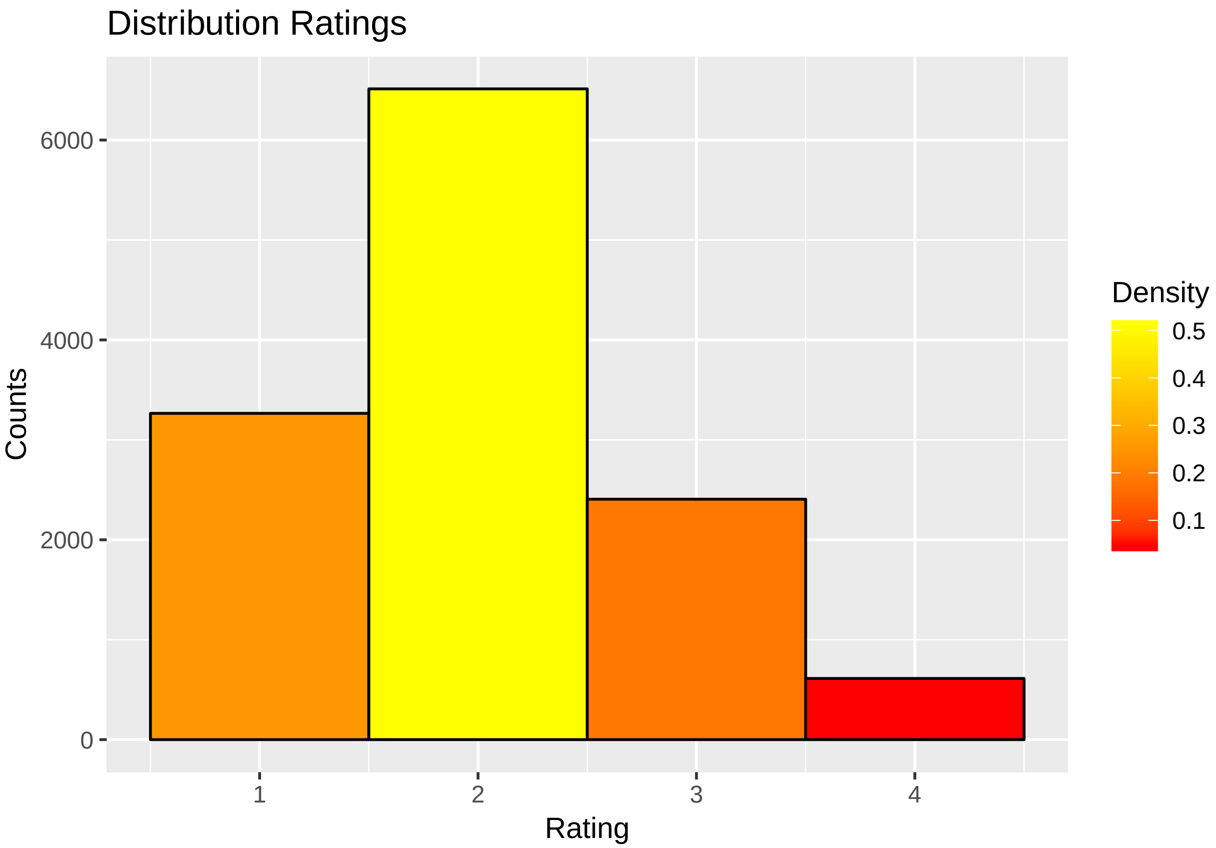
```
##                    vars     n    mean      sd  median trimmed     mad
## INDEX                 1 12795 8069.98 4656.91 8110.00 8071.03 5977.84
## TARGET                2 12795    3.03    1.93    3.00    3.05    1.48
## FixedAcidity          3 12795    7.08    6.32    6.90    7.07    3.26
## VolatileAcidity       4 12795    0.32    0.78    0.28    0.32    0.43
## CitricAcid            5 12795    0.31    0.86    0.31    0.31    0.42
## ResidualSugar         6 12179    5.42   33.75    3.90    5.58   15.72
## Chlorides             7 12157    0.05    0.32    0.05    0.05    0.13
## FreeSulfurDioxide     8 12148   30.85  148.71   30.00   30.93   56.34
## TotalSulfurDioxide    9 12113  120.71  231.91  123.00  120.89  134.92
## Density              10 12795    0.99    0.03    0.99    0.99    0.01
## pH                   11 12400    3.21    0.68    3.20    3.21    0.39
## Sulphates            12 11585    0.53    0.93    0.50    0.53    0.44
## Alcohol              13 12142   10.49    3.73   10.40   10.50    2.37
## LabelAppeal          14 12795   -0.01    0.89    0.00   -0.01    1.48
## AcidIndex            15 12795    7.77    1.32    8.00    7.64    1.48
## STARS                16  9436    2.04    0.90    2.00    1.97    1.48
##                        min      max    range  skew kurtosis    se
## INDEX                 1.00 16129.00 16128.00  0.00    -1.20 41.17
## TARGET                0.00     8.00     8.00 -0.33    -0.88  0.02
## FixedAcidity        -18.10    34.40    52.50 -0.02     1.67  0.06
## VolatileAcidity      -2.79     3.68     6.47  0.02     1.83  0.01
## CitricAcid           -3.24     3.86     7.10 -0.05     1.84  0.01
## ResidualSugar      -127.80   141.15   268.95 -0.05     1.88  0.31
## Chlorides            -1.17     1.35     2.52  0.03     1.79  0.00
## FreeSulfurDioxide  -555.00   623.00  1178.00  0.01     1.84  1.35
## TotalSulfurDioxide -823.00  1057.00  1880.00 -0.01     1.67  2.11
## Density             0.89     1.10     0.21 -0.02     1.90  0.00
## pH                  0.48     6.13     5.65  0.04     1.65  0.01
## Sulphates          -3.13     4.24     7.37  0.01     1.75  0.01
## Alcohol            -4.70    26.50    31.20 -0.03     1.54  0.03
## LabelAppeal        -2.00     2.00     4.00  0.01    -0.26  0.01
## AcidIndex           4.00    17.00    13.00  1.65     5.19  0.01
## STARS               1.00     4.00     3.00  0.45    -0.69  0.01
```

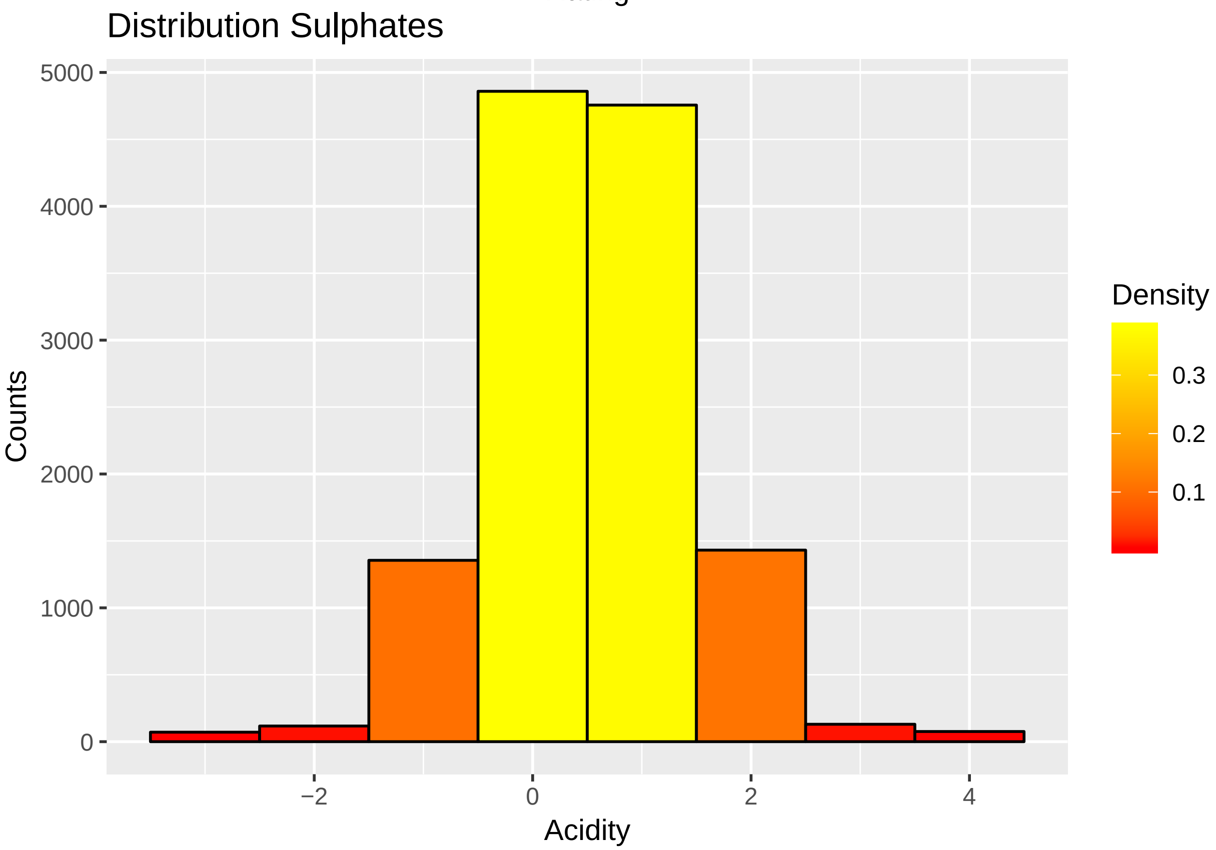
```
##              INDEX             TARGET       FixedAcidity
##                  0                  0                  0
##    VolatileAcidity         CitricAcid      ResidualSugar
##                  0                  0                616
##          Chlorides  FreeSulfurDioxide TotalSulfurDioxide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

Data Preparation

. . . . After assessing the above tables it was determined that the above variables would need to have their missing values imputed. It was at this point that I used the missForest package and it's algorithm to impute for the missing NA values. The sapply function was applied after doing this to ensure that all variables were dealt with in the process. It was at this point, that I also followed the same process to clean up the missing variables and NA values for the test data set so that it would be ready to apply my model to it. Once this was done,I decided to look further into some specific variables to see if they could be bucketed. I took the STARS varaible and ended up bucketing that variable into two and under stars. Everything from beyond two went into a higher rating bucket. I did this to see the difference from the highly concentrated two star rated wines and others. As the models were getting built, the binning of sulphates is what ended up having the biggest impact on the accuracy of the model. It helped to lower the AIC scores and helped to maintain the range of the predicted target.

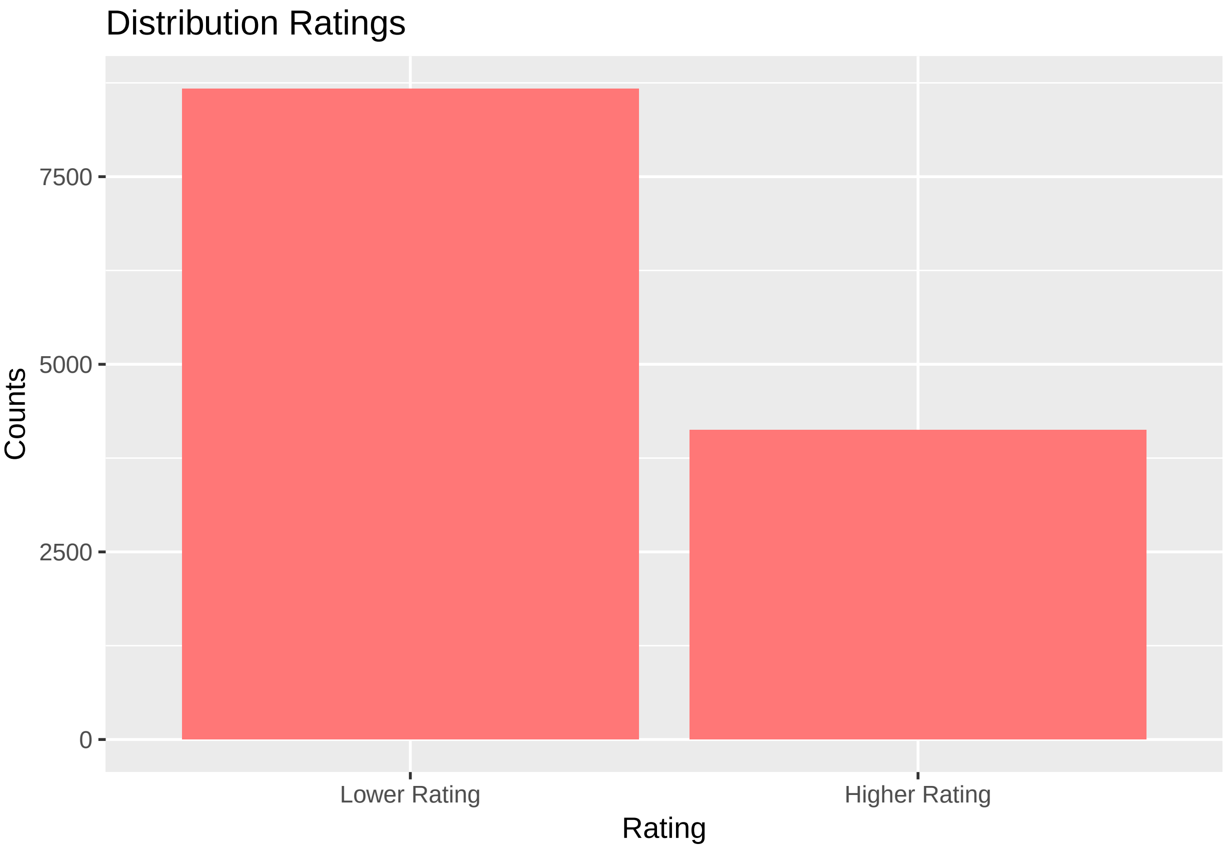
. . . . The below plot helps to illustrate the dispersion of the Star ratings. It shows a clear peak at around two with 50% of the density sitting there. I did the same type of ggplot for the sulphates as well. The distribution here is what suggested that binning would be a good possiblity with the density able to be separated into 4 distinct groupings, with a heavy concentration around 0 to 1.

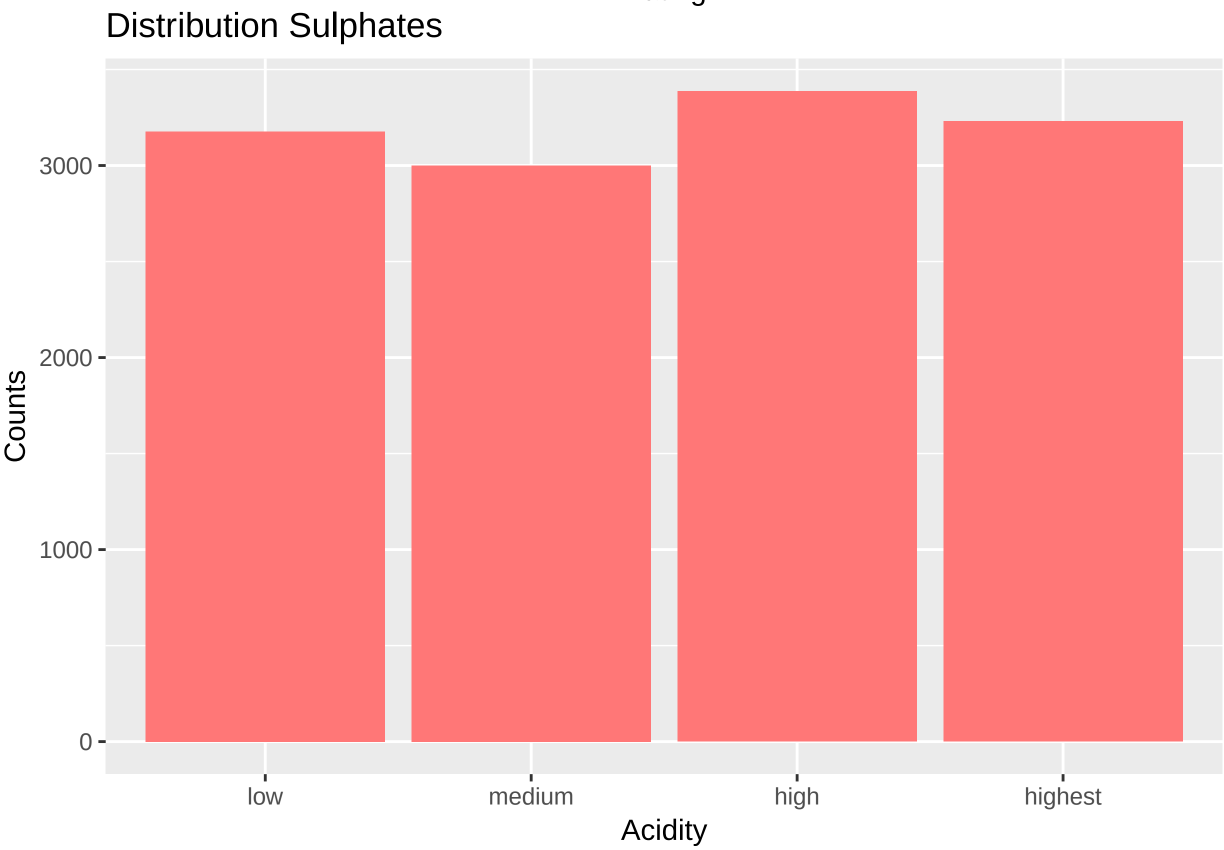




. . . . After the above graphs were binned, I replotted them to look at the new output and see the final results.

Below are examples of this.





Build Models

. . . .This was the point in the exercise, where I built a multitude of different models. I ended up using every variable avaiable in some form during my model building to make it robust. I did exclude some variables when I used the binned version of them that I had created. I will list out the 7 models below and some of their unique characteristics. As seen below, most of the coefficients from model to model kept the same sign for the most part. Citric Acid for example, is a positive coeffient in all models. The magnitudes of the actual coefficients did change though from model to model, especially when jumping from a standard linear regression model to something like a poisson model.

> The first model I ran was a standard linear model. It did not handle the zero inflated variables well.

```
##
## Call:
## lm(formula = TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid +
##     ResidualSugar + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     Density + pH + Sulphates_bin + Alcohol + LabelAppeal + AcidIndex +
##     STARS, data = train_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5550 -0.7418  0.3583  1.1170  4.2789
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.206e+00  5.544e-01   9.391  < 2e-16 ***
## FixedAcidity        -1.069e-03  2.320e-03  -0.461 0.644964
## VolatileAcidity     -1.509e-01  1.843e-02  -8.189 2.88e-16 ***
## CitricAcid           3.922e-02  1.677e-02   2.339 0.019361 *
## ResidualSugar        4.424e-04  4.379e-04   1.010 0.312395
## Chlorides           -1.939e-01  4.646e-02  -4.174 3.02e-05 ***
## FreeSulfurDioxide    4.310e-04  9.958e-05   4.328 1.52e-05 ***
## TotalSulfurDioxide   2.966e-04  6.397e-05   4.636 3.59e-06 ***
## Density             -1.160e+00  5.441e-01  -2.131 0.033082 *
## pH                  -5.813e-02  2.159e-02  -2.693 0.007100 **
## Sulphates_binmedium  5.965e-02  4.158e-02   1.435 0.151368
## Sulphates_binhigh   -1.871e-01  4.032e-02  -4.640 3.52e-06 ***
## Sulphates_binhighest -1.400e-01  4.081e-02  -3.431 0.000603 ***
## Alcohol              1.742e-02  3.981e-03   4.377 1.21e-05 ***
## LabelAppeal          5.413e-01  1.743e-02  31.055  < 2e-16 ***
## AcidIndex           -3.099e-01  1.128e-02 -27.476  < 2e-16 ***
## STARS                7.285e-01  1.971e-02  36.970  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.63 on 12778 degrees of freedom
## Multiple R-squared:  0.2846, Adjusted R-squared:  0.2837
## F-statistic: 317.8 on 16 and 12778 DF,  p-value: < 2.2e-16
```

```
##          (Intercept)         FixedAcidity      VolatileAcidity
##         5.2063616633        -0.0010691888        -0.1509019340
##           CitricAcid        ResidualSugar            Chlorides
##         0.0392218841         0.0004423803        -0.1939058809
##    FreeSulfurDioxide   TotalSulfurDioxide              Density
##         0.0004309837         0.0002965456        -1.1596453372
##                   pH  Sulphates_binmedium    Sulphates_binhigh
##        -0.0581250542         0.0596549628        -0.1870669994
## Sulphates_binhighest              Alcohol          LabelAppeal
##        -0.1400204797         0.0174240591         0.5412742884
##            AcidIndex                STARS
##        -0.3099027440         0.7285397015
```

```
## [1] "AIC IS:"
```

```
## [1] 48837.59
```

The second model I built was a stepwise model and it provided similar results to the first one, with slightly better AIC scores.

```
##
## Call:
## lm(formula = TARGET ~ VolatileAcidity + CitricAcid + Chlorides +
##     FreeSulfurDioxide + TotalSulfurDioxide + Density + pH + Sulphates_bin +
##     Alcohol + LabelAppeal + AcidIndex + STARS, data = train_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5578 -0.7441  0.3605  1.1142  4.2772
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.206e+00  5.544e-01   9.391  < 2e-16 ***
## VolatileAcidity     -1.510e-01  1.843e-02  -8.197 2.70e-16 ***
## CitricAcid           3.910e-02  1.677e-02   2.332 0.019730 *
## Chlorides           -1.940e-01  4.646e-02  -4.176 2.98e-05 ***
## FreeSulfurDioxide    4.323e-04  9.956e-05   4.342 1.42e-05 ***
## TotalSulfurDioxide   2.984e-04  6.394e-05   4.667 3.09e-06 ***
## Density             -1.157e+00  5.441e-01  -2.127 0.033421 *
## pH                  -5.788e-02  2.158e-02  -2.681 0.007339 **
## Sulphates_binmedium  5.928e-02  4.157e-02   1.426 0.153867
## Sulphates_binhigh   -1.876e-01  4.031e-02  -4.654 3.30e-06 ***
## Sulphates_binhighest -1.405e-01  4.079e-02  -3.443 0.000577 ***
## Alcohol              1.734e-02  3.980e-03   4.358 1.33e-05 ***
## LabelAppeal          5.413e-01  1.743e-02  31.058  < 2e-16 ***
## AcidIndex           -3.109e-01  1.110e-02 -27.995  < 2e-16 ***
## STARS                7.288e-01  1.970e-02  36.993  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.63 on 12780 degrees of freedom
## Multiple R-squared:  0.2846, Adjusted R-squared:  0.2838
## F-statistic: 363.1 on 14 and 12780 DF,  p-value: < 2.2e-16
```

```
##          (Intercept)      VolatileAcidity           CitricAcid
##         5.2060461539        -0.1510393067         0.0391013021
##            Chlorides    FreeSulfurDioxide   TotalSulfurDioxide
##        -0.1940060397         0.0004322914         0.0002984025
##              Density                   pH  Sulphates_binmedium
##        -1.1573661024        -0.0578797886         0.0592807489
##    Sulphates_binhigh Sulphates_binhighest              Alcohol
##        -0.1875843972        -0.1404521101         0.0173425922
##          LabelAppeal            AcidIndex                STARS
##         0.5412804423        -0.3108629857         0.7288261834
```

```
## [1] "AIC IS:"
```

```
## [1] 48834.84
```

The third model I built was a poisson model. Overall, I had a worse AIC score and I continued to refine the next couple models to gain better accuracy.

```
##
## Call:
## glm(formula = TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid +
##     ResidualSugar + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     Density + pH + Sulphates_bin + Alcohol + LabelAppeal + AcidIndex +
##     STARS, family = poisson(link = "log"), data = train_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6690  -0.5220   0.2010   0.6339   2.5171
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.978e+00  1.960e-01  10.092  < 2e-16 ***
## FixedAcidity         -4.046e-04  8.197e-04  -0.494 0.621602
## VolatileAcidity      -4.966e-02  6.494e-03  -7.648 2.04e-14 ***
## CitricAcid            1.341e-02  5.895e-03   2.275 0.022883 *
## ResidualSugar         1.339e-04  1.542e-04   0.868 0.385337
## Chlorides            -6.078e-02  1.643e-02  -3.700 0.000216 ***
## FreeSulfurDioxide     1.439e-04  3.510e-05   4.099 4.15e-05 ***
## TotalSulfurDioxide    1.036e-04  2.267e-05   4.571 4.86e-06 ***
## Density              -3.998e-01  1.921e-01  -2.082 0.037350 *
## pH                   -2.265e-02  7.635e-03  -2.966 0.003014 **
## Sulphates_binmedium   1.871e-02  1.426e-02   1.312 0.189502
## Sulphates_binhigh    -6.326e-02  1.430e-02  -4.424 9.67e-06 ***
## Sulphates_binhighest -4.932e-02  1.439e-02  -3.428 0.000607 ***
## Alcohol               4.899e-03  1.409e-03   3.476 0.000509 ***
## LabelAppeal           1.821e-01  6.199e-03  29.371  < 2e-16 ***
## AcidIndex            -1.179e-01  4.486e-03 -26.282  < 2e-16 ***
## STARS                 2.160e-01  6.553e-03  32.960  < 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: 22861  on 12794  degrees of freedom
## Residual deviance: 18444  on 12778  degrees of freedom
## AIC: 50420
##
## Number of Fisher Scoring iterations: 5
```

```
##          (Intercept)         FixedAcidity      VolatileAcidity
##         1.9784812018        -0.0004045715        -0.0496636463
##           CitricAcid        ResidualSugar            Chlorides
##         0.0134140536         0.0001338472        -0.0607822911
##    FreeSulfurDioxide   TotalSulfurDioxide              Density
##         0.0001438763         0.0001035973        -0.3998333305
##                   pH  Sulphates_binmedium    Sulphates_binhigh
##        -0.0226475637         0.0187135684        -0.0632576185
## Sulphates_binhighest              Alcohol          LabelAppeal
##        -0.0493172675         0.0048992621         0.1820578577
##            AcidIndex                STARS
##        -0.1178988480         0.2159994230

## [1] "AIC IS:"

## [1] 50419.77
```

The fourth model created was a negative binomial. It was very similar to the poisson model. The AIC was ever so slightly worse but the deviance was slightly less.

```
##
## Call:
## glm.nb(formula = TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid +
##     ResidualSugar + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     Density + pH + Sulphates_bin + Alcohol + LabelAppeal + AcidIndex +
##     STARS, data = train_clean, init.theta = 38064.88817, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6688  -0.5220   0.2010   0.6339   2.5170
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.979e+00  1.961e-01  10.091  < 2e-16 ***
## FixedAcidity         -4.046e-04  8.197e-04  -0.494 0.621603
## VolatileAcidity      -4.966e-02  6.494e-03  -7.648 2.05e-14 ***
## CitricAcid            1.341e-02  5.896e-03   2.275 0.022890 *
## ResidualSugar         1.339e-04  1.542e-04   0.868 0.385335
## Chlorides            -6.078e-02  1.643e-02  -3.700 0.000216 ***
## FreeSulfurDioxide     1.439e-04  3.510e-05   4.099 4.15e-05 ***
## TotalSulfurDioxide    1.036e-04  2.267e-05   4.571 4.86e-06 ***
## Density              -3.998e-01  1.921e-01  -2.082 0.037355 *
## pH                   -2.265e-02  7.635e-03  -2.966 0.003015 **
## Sulphates_binmedium   1.871e-02  1.426e-02   1.312 0.189506
## Sulphates_binhigh    -6.326e-02  1.430e-02  -4.424 9.68e-06 ***
## Sulphates_binhighest -4.932e-02  1.439e-02  -3.428 0.000607 ***
## Alcohol               4.899e-03  1.410e-03   3.476 0.000509 ***
## LabelAppeal           1.821e-01  6.199e-03  29.370  < 2e-16 ***
## AcidIndex            -1.179e-01  4.486e-03 -26.282  < 2e-16 ***
## STARS                 2.160e-01  6.554e-03  32.958  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(38064.89) family taken to be 1)
##
##     Null deviance: 22860  on 12794  degrees of freedom
## Residual deviance: 18443  on 12778  degrees of freedom
## AIC: 50422
##
## Number of Fisher Scoring iterations: 1
##
```

```
##
##               Theta:  38065
##           Std. Err.:  59764
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -50385.89
```

```
##          (Intercept)         FixedAcidity      VolatileAcidity
##         1.9785048910        -0.0004045889        -0.0496643874
##           CitricAcid        ResidualSugar            Chlorides
##         0.0134140243         0.0001338539        -0.0607833633
##    FreeSulfurDioxide   TotalSulfurDioxide              Density
##         0.0001438784         0.0001035992        -0.3998405104
##                   pH  Sulphates_binmedium    Sulphates_binhigh
##        -0.0226481597         0.0187142900        -0.0632582651
## Sulphates_binhighest              Alcohol          LabelAppeal
##        -0.0493182271         0.0048992825         0.1820591788
##            AcidIndex                STARS
##        -0.1179006415         0.2159989729
```

```
## [1] "AIC IS:"
```

```
## [1] 50421.89
```

The fifth model was a zero inflated poisson model. This model is where I really started to notice better performance results. By taking into account the zero inflated target correctly, the AIC dropped significantly and became more competitive. I ended up using this as my second best model, after the champion model that was chosen later.

```
##
## Call:
## zeroinfl(formula = TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid +
##     ResidualSugar + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     Density + pH + Sulphates_bin + Alcohol + LabelAppeal + AcidIndex +
##     STARS, data = train_clean)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.1657 -0.3708  0.1571  0.5020  4.3479
##
## Count model coefficients (poisson with log link):
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.389e+00  2.065e-01   6.729 1.71e-11 ***
## FixedAcidity          2.841e-04  8.570e-04   0.332  0.74025
## VolatileAcidity      -1.285e-02  6.862e-03  -1.873  0.06111 .
## CitricAcid            1.293e-05  6.139e-03   0.002  0.99832
## ResidualSugar        -9.858e-05  1.621e-04  -0.608  0.54304
## Chlorides            -1.675e-02  1.726e-02  -0.970  0.33189
## FreeSulfurDioxide     3.208e-05  3.597e-05   0.892  0.37245
## TotalSulfurDioxide   -3.414e-05  2.295e-05  -1.488  0.13679
## Density              -3.123e-01  2.023e-01  -1.544  0.12255
## pH                    7.410e-03  8.020e-03   0.924  0.35550
## Sulphates_binmedium  -1.196e-02  1.489e-02  -0.803  0.42215
## Sulphates_binhigh     3.276e-04  1.494e-02   0.022  0.98251
## Sulphates_binhighest  2.928e-03  1.502e-02   0.195  0.84539
## Alcohol               7.375e-03  1.465e-03   5.035 4.77e-07 ***
## LabelAppeal           2.414e-01  6.432e-03  37.526  < 2e-16 ***
## AcidIndex            -1.539e-02  4.977e-03  -3.092  0.00199 **
## STARS                 1.107e-01  6.479e-03  17.092  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -5.5603616  1.0458714  -5.316 1.06e-07 ***
## FixedAcidity          0.0035321  0.0043415   0.814 0.415894
## VolatileAcidity       0.2341668  0.0347531   6.738 1.61e-11 ***
## CitricAcid           -0.0793824  0.0316174  -2.511 0.012049 *
## ResidualSugar        -0.0013881  0.0008298  -1.673 0.094386 .
## Chlorides             0.2783787  0.0874446   3.183 0.001455 **
## FreeSulfurDioxide    -0.0006998  0.0001858  -3.766 0.000166 ***
## TotalSulfurDioxide   -0.0008527  0.0001196  -7.128 1.02e-12 ***
## Density               0.7794441  1.0253201   0.760 0.447138
## pH                    0.1876147  0.0407818   4.600 4.22e-06 ***
## Sulphates_binmedium  -0.2502130  0.0882285  -2.836 0.004569 **
## Sulphates_binhigh     0.3857611  0.0747402   5.161 2.45e-07 ***
## Sulphates_binhighest  0.3309373  0.0759275   4.359 1.31e-05 ***
## Alcohol               0.0113693  0.0074877   1.518 0.128914
## LabelAppeal           0.3638272  0.0334552  10.875  < 2e-16 ***
## AcidIndex             0.4586704  0.0197836  23.184  < 2e-16 ***
## STARS                -0.6550474  0.0383586 -17.077  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 41
## Log-likelihood: -2.247e+04 on 34 Df
```

```
##          count_(Intercept)         count_FixedAcidity
##               1.389427e+00               2.841031e-04
##      count_VolatileAcidity           count_CitricAcid
##              -1.284984e-02               1.293367e-05
##        count_ResidualSugar            count_Chlorides
##              -9.857828e-05              -1.674668e-02
##    count_FreeSulfurDioxide   count_TotalSulfurDioxide
##               3.207772e-05              -3.414291e-05
##              count_Density                   count_pH
##              -3.123489e-01               7.409966e-03
##  count_Sulphates_binmedium    count_Sulphates_binhigh
##              -1.195599e-02               3.275742e-04
## count_Sulphates_binhighest              count_Alcohol
##               2.928088e-03               7.374901e-03
##          count_LabelAppeal            count_AcidIndex
##               2.413575e-01              -1.538872e-02
##                count_STARS           zero_(Intercept)
##               1.107449e-01              -5.560362e+00
##          zero_FixedAcidity       zero_VolatileAcidity
##               3.532075e-03               2.341668e-01
##            zero_CitricAcid         zero_ResidualSugar
##              -7.938239e-02              -1.388084e-03
##             zero_Chlorides     zero_FreeSulfurDioxide
##               2.783787e-01              -6.998288e-04
##    zero_TotalSulfurDioxide               zero_Density
##              -8.527275e-04               7.794441e-01
##                    zero_pH   zero_Sulphates_binmedium
##               1.876147e-01              -2.502130e-01
##     zero_Sulphates_binhigh  zero_Sulphates_binhighest
##               3.857611e-01               3.309373e-01
##               zero_Alcohol           zero_LabelAppeal
##               1.136930e-02               3.638272e-01
##             zero_AcidIndex                 zero_STARS
##               4.586704e-01              -6.550474e-01
```

```
## [1] "AIC IS:"
```

```
## [1] 45000.91
```

The sixth model was a zero inflated negative binomial. The results for this model were very similar to the prior ZIP model, but just a slighly worse in regards to the AIC and a few other statistics.

```
##
## Call:
## zeroinfl(formula = TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid +
##     ResidualSugar + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     Density + pH + Sulphates_bin + Alcohol + LabelAppeal + AcidIndex +
##     STARS, data = train_clean, dist = "negbin", EM = TRUE)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.1657 -0.3708  0.1571  0.5020  4.3481
##
## Count model coefficients (negbin with log link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.389e+00  2.065e-01   6.729 1.71e-11 ***
## FixedAcidity          2.841e-04  8.570e-04   0.332  0.74023
## VolatileAcidity      -1.285e-02  6.862e-03  -1.873  0.06111 .
## CitricAcid            1.285e-05  6.139e-03   0.002  0.99833
## ResidualSugar        -9.858e-05  1.621e-04  -0.608  0.54302
## Chlorides            -1.675e-02  1.726e-02  -0.970  0.33189
## FreeSulfurDioxide     3.208e-05  3.597e-05   0.892  0.37245
## TotalSulfurDioxide   -3.414e-05  2.295e-05  -1.488  0.13681
## Density              -3.123e-01  2.023e-01  -1.544  0.12257
## pH                    7.410e-03  8.020e-03   0.924  0.35550
## Sulphates_binmedium  -1.196e-02  1.489e-02  -0.803  0.42216
## Sulphates_binhigh     3.280e-04  1.494e-02   0.022  0.98249
## Sulphates_binhighest  2.928e-03  1.502e-02   0.195  0.84537
## Alcohol               7.375e-03  1.465e-03   5.035 4.77e-07 ***
## LabelAppeal           2.414e-01  6.432e-03  37.526  < 2e-16 ***
## AcidIndex            -1.539e-02  4.977e-03  -3.092  0.00199 **
## STARS                 1.107e-01  6.479e-03  17.092  < 2e-16 ***
## Log(theta)            1.222e+01  3.787e+00   3.228  0.00125 **
##
## Zero-inflation model coefficients (binomial with logit link):
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -5.5610296  1.0458824  -5.317 1.05e-07 ***
## FixedAcidity          0.0035322  0.0043415   0.814 0.415877
## VolatileAcidity       0.2341645  0.0347534   6.738 1.61e-11 ***
## CitricAcid           -0.0793820  0.0316177  -2.511 0.012050 *
## ResidualSugar        -0.0013882  0.0008299  -1.673 0.094369 .  
## Chlorides             0.2783754  0.0874454   3.183 0.001455 ** 
## FreeSulfurDioxide    -0.0006998  0.0001858  -3.766 0.000166 ***
## TotalSulfurDioxide   -0.0008527  0.0001196  -7.128 1.02e-12 ***
## Density               0.7800490  1.0253295   0.761 0.446789    
## pH                    0.1876180  0.0407822   4.600 4.21e-06 ***
## Sulphates_binmedium  -0.2502156  0.0882296  -2.836 0.004569 ** 
## Sulphates_binhigh     0.3857652  0.0747408   5.161 2.45e-07 ***
## Sulphates_binhighest  0.3309397  0.0759282   4.359 1.31e-05 ***
## Alcohol               0.0113697  0.0074878   1.518 0.128904    
## LabelAppeal           0.3638383  0.0334555  10.875  < 2e-16 ***
## AcidIndex             0.4586752  0.0197838  23.184  < 2e-16 ***
## STARS                -0.6550487  0.0383589 -17.077  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Theta = 203119.2883 
## Number of iterations in BFGS optimization: 1 
## Log-likelihood: -2.247e+04 on 35 Df
```

```
##          count_(Intercept)         count_FixedAcidity 
##               1.389412e+00               2.841277e-04 
##      count_VolatileAcidity           count_CitricAcid 
##              -1.284999e-02               1.284678e-05 
##        count_ResidualSugar            count_Chlorides 
##              -9.858322e-05              -1.674670e-02 
##    count_FreeSulfurDioxide   count_TotalSulfurDioxide 
##               3.207776e-05              -3.414190e-05 
##              count_Density                   count_pH 
##              -3.123390e-01               7.410019e-03 
##  count_Sulphates_binmedium    count_Sulphates_binhigh 
##              -1.195586e-02               3.280278e-04 
## count_Sulphates_binhighest              count_Alcohol 
##               2.928469e-03               7.374953e-03 
##          count_LabelAppeal            count_AcidIndex 
##               2.413590e-01              -1.538851e-02 
##                count_STARS           zero_(Intercept) 
##               1.107454e-01              -5.561030e+00 
##          zero_FixedAcidity       zero_VolatileAcidity 
##               3.532237e-03               2.341645e-01 
##            zero_CitricAcid         zero_ResidualSugar 
##              -7.938196e-02              -1.388169e-03 
##             zero_Chlorides     zero_FreeSulfurDioxide 
##               2.783754e-01              -6.998311e-04 
##    zero_TotalSulfurDioxide               zero_Density 
##              -8.527284e-04               7.800490e-01 
##                    zero_pH   zero_Sulphates_binmedium 
##               1.876180e-01              -2.502156e-01 
##     zero_Sulphates_binhigh  zero_Sulphates_binhighest 
##               3.857652e-01               3.309397e-01 
##               zero_Alcohol           zero_LabelAppeal 
##               1.136969e-02               3.638383e-01 
##             zero_AcidIndex                 zero_STARS 
##               4.586752e-01              -6.550487e-01
## [1] "AIC IS:"
```

```
## [1] 45003.05
```

The seventh model I built is a hurdle model. I researched this model on line and found out that it is a good two way model that does automated truncating as part of it's algorithm, amoungst other advanced features. It had the best AIC score of any of the prior models! Below are the results.

```
##
## Call:
## hurdle(formula = TARGET ~ FixedAcidity + VolatileAcidity + CitricAcid +
##     ResidualSugar + Chlorides + FreeSulfurDioxide + TotalSulfurDioxide +
##     Density + pH + Sulphates_bin + Alcohol + LabelAppeal + AcidIndex +
##     STARS, data = train_clean, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.1575 -0.3854  0.1579  0.5037  3.1642
##
## Count model coefficients (truncated negbin with log link):
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.401e+00  2.076e-01   6.750 1.48e-11 ***
## FixedAcidity          2.363e-04  8.635e-04   0.274 0.784327
## VolatileAcidity      -1.209e-02  6.910e-03  -1.749 0.080231 .
## CitricAcid            6.573e-04  6.188e-03   0.106 0.915411
## ResidualSugar        -9.584e-05  1.629e-04  -0.588 0.556390
## Chlorides            -1.934e-02  1.737e-02  -1.113 0.265496
## FreeSulfurDioxide     2.967e-05  3.637e-05   0.816 0.414581
## TotalSulfurDioxide   -3.295e-05  2.323e-05  -1.418 0.156088
## Density              -3.146e-01  2.034e-01  -1.547 0.121850
## pH                    8.122e-03  8.073e-03   1.006 0.314354
## Sulphates_binmedium  -1.158e-02  1.501e-02  -0.771 0.440560
## Sulphates_binhigh     4.874e-04  1.505e-02   0.032 0.974168
## Sulphates_binhighest  3.968e-03  1.512e-02   0.262 0.792976
## Alcohol               7.567e-03  1.480e-03   5.112 3.19e-07 ***
## LabelAppeal           2.458e-01  6.558e-03  37.483  < 2e-16 ***
## AcidIndex            -1.672e-02  4.997e-03  -3.347 0.000817 ***
## STARS                 1.082e-01  6.617e-03  16.349  < 2e-16 ***
## Log(theta)            1.707e+01  5.672e+00   3.010 0.002614 **
## Zero hurdle model coefficients (binomial with logit link):
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           4.8612309  0.8835986   5.502 3.76e-08 ***
## FixedAcidity         -0.0026282  0.0036946  -0.711  0.47687
## VolatileAcidity      -0.2114657  0.0295120  -7.165 7.76e-13 ***
## CitricAcid            0.0600035  0.0268348   2.236  0.02535 *
## ResidualSugar         0.0011085  0.0007021   1.579  0.11436
## Chlorides            -0.2345231  0.0741220  -3.164  0.00156 **
## FreeSulfurDioxide     0.0006351  0.0001591   3.991 6.57e-05 ***
## TotalSulfurDioxide    0.0006948  0.0001021   6.807 9.94e-12 ***
## Density              -0.8249037  0.8660836  -0.952  0.34087
## pH                   -0.1583360  0.0344928  -4.590 4.42e-06 ***
## Sulphates_binmedium   0.1966597  0.0715097   2.750  0.00596 **
## Sulphates_binhigh    -0.3246949  0.0637146  -5.096 3.47e-07 ***
## Sulphates_binhighest -0.2854874  0.0646995  -4.413 1.02e-05 ***
## Alcohol              -0.0063641  0.0063648  -1.000  0.31736    
## LabelAppeal          -0.1896530  0.0275841  -6.875 6.18e-12 ***
## AcidIndex            -0.4096906  0.0172447 -23.757  < 2e-16 ***
## STARS                 0.6386047  0.0348282  18.336  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Theta: count = 25966149.6945
## Number of iterations in BFGS optimization: 65 
## Log-likelihood: -2.247e+04 on 35 Df
```

```
##          count_(Intercept)         count_FixedAcidity 
##               1.401025e+00               2.363359e-04 
##      count_VolatileAcidity           count_CitricAcid 
##              -1.208764e-02               6.572752e-04 
##        count_ResidualSugar            count_Chlorides 
##              -9.584048e-05              -1.933993e-02 
##    count_FreeSulfurDioxide   count_TotalSulfurDioxide 
##               2.967389e-05              -3.295353e-05 
##              count_Density                   count_pH 
##              -3.145991e-01               8.122335e-03 
##  count_Sulphates_binmedium    count_Sulphates_binhigh 
##              -1.158015e-02               4.874143e-04 
## count_Sulphates_binhighest              count_Alcohol 
##               3.968439e-03               7.567384e-03 
##          count_LabelAppeal            count_AcidIndex 
##               2.458029e-01              -1.672441e-02 
##                count_STARS           zero_(Intercept) 
##               1.081766e-01               4.861231e+00 
##          zero_FixedAcidity       zero_VolatileAcidity 
##              -2.628194e-03              -2.114657e-01 
##            zero_CitricAcid         zero_ResidualSugar 
##               6.000353e-02               1.108498e-03 
##             zero_Chlorides     zero_FreeSulfurDioxide 
##              -2.345231e-01               6.350713e-04 
##    zero_TotalSulfurDioxide               zero_Density 
##               6.947958e-04              -8.249037e-01 
##                    zero_pH   zero_Sulphates_binmedium 
##              -1.583360e-01               1.966597e-01 
##     zero_Sulphates_binhigh  zero_Sulphates_binhighest 
##              -3.246949e-01              -2.854874e-01 
##               zero_Alcohol           zero_LabelAppeal 
##              -6.364111e-03              -1.896530e-01 
##             zero_AcidIndex                 zero_STARS 
##              -4.096906e-01               6.386047e-01
```

```
## [1] "AIC IS:"
```

```
## [1] 45000.24
```

Select Models

.... Overall, I ended up selecting the hurdle model. It had the best AIC score and will be the easiest to explain due to it’s two part structure. It includes a count model, along with a zero hurdle model coeffient. Based on my research, I read that this type of model is good for when there is only one source for why a zero would happen. In this case, it was simply a customer deciding to not by a case of wine. I believe that this was a good scenerio to deploy this type of model.

Stand Alone Data Step and Scores

These were both done and submitted separately.

Conclusion

. . . . This was an excellent assignment. It allowed me to put together everthing I had learned in the quarter to build a couple good models. I was able to discover a new way to impute variables with the missForest package, as well as use a new/different modeling package called Hurdle. I also was able to build a few nice looking ggplot charts as well to illustrate the data. My final champion model had a good distribution of the target variable. The mean ended up being close to around 3 and the range was from about 0 to just under 8. Overall, this project was a success and if we had unaltered data from the beginning it might have been possible to further tighten the model. Thanks for a great quarter Professor!