Scott Jue
MSDS 410
2/26/2023

Modeling Assignment #8: Modeling Dichotomous Responses

**Introduction**

In this analysis I will be modeling data obtained from a dataset containing approximately 12,000 commercially available wines. Each record of the data set is an observation for a different type of wine. In addition to the characteristics of the wine, the dataset contains a Purchase variable that reflects whether or not a purchase was made of that wine. As such, a large wine manufacturer is interested in being able to predict if a bottle of wine will be purchased based on the wine's characteristics. Therefore, the Purchase variable will be the response variable for our model and since this is a dichotomous response variable (0 = no purchase and 1 = purchased), a logistic regression model will be used to model the predicted probability of purchase. In this analysis, I will fit a logistic regression model using this data and then assess the model fit and also determine which variables provide the most information in predicting if a customer purchases the bottle of wine or not remove any that are not useful in the model. Additionally, for this analysis the sample population is commercially available wines that are available to the manufacturer.

**Tasks**

1. Before fitting the logistic regression model, I first performed exploratory data analysis to get a better understanding of the dataset. There is a total of 12,795 observations and 16 total variables, including the response variable. Of these 12,795 wines in the dataset, 10,061 were purchased which is approximately 79% of the wines. Additionally, due to the large amount of observations, we should be concerned with having "too much statistical power", so I will keep this mind throughout the analysis when looking at statistical significances. Furthermore, due to the large sample size I will split this data into training and testing datasets using a 70/30 split for model testing and validation. Next, I created a table of summary statistics for each of the variables which can be seen below.

| key | < 0 Count | Avg | Max | Med | Med > 0 | Min | NA Cnt | Std |
|---|---|---|---|---|---|---|---|---|
| AcidIndex | 0 | 7.77 | 17.00 | 8.00 | 8.00 | 4.00 | 0 | 1.32 |
| Alcohol | 118 | 10.49 | 26.50 | 10.40 | 10.40 | -4.70 | 653 | 3.73 |
| Cases | 0 | 3.03 | 8.00 | 3.00 | 3.00 | 0.00 | 0 | 1.93 |
| Chlorides | 3197 | 0.05 | 1.35 | 0.05 | 0.06 | -1.17 | 638 | 0.32 |
| CitricAcid | 2966 | 0.31 | 3.86 | 0.31 | 0.39 | -3.24 | 0 | 0.86 |
| Density | 0 | 0.99 | 1.10 | 0.99 | 0.99 | 0.89 | 0 | 0.03 |
| FixedAcidity | 1621 | 7.08 | 34.40 | 6.90 | 7.20 | -18.10 | 0 | 6.32 |
| FreeSulfurDioxide | 3036 | 30.85 | 623.00 | 30.00 | 43.00 | -555.00 | 647 | 148.71 |
| LabelAppeal | 3640 | -0.01 | 2.00 | 0.00 | 0.00 | -2.00 | 0 | 0.89 |
| pH | 0 | 3.21 | 6.13 | 3.20 | 3.20 | 0.48 | 395 | 0.68 |
| Purchase | 0 | 0.79 | 1.00 | 1.00 | 1.00 | 0.00 | 0 | 0.41 |
| ResidualSugar | 3136 | 5.42 | 141.15 | 3.90 | 8.70 | -127.80 | 616 | 33.75 |
| STARS | 0 | 2.04 | 4.00 | 2.00 | 2.00 | 1.00 | 3359 | 0.90 |
| Sulphates | 2361 | 0.53 | 4.24 | 0.50 | 0.58 | -3.13 | 1210 | 0.93 |
| TotalSulfurDioxide | 2504 | 120.71 | 1057.00 | 123.00 | 150.00 | -823.00 | 682 | 231.91 |
| VolatileAcidity | 2827 | 0.32 | 3.68 | 0.28 | 0.36 | -2.79 | 0 | 0.78 |

From this table, we can see that several variables have negative values. After researching these variables, I've determined that it is not possible for these variables to have negative values. Therefore, these values could be negative due to data entry errors. As such, I have to flipped their sign so that they are positive values. Next, I noticed that there are several variables that have a high amount of NA values. Due to the high amount of observations with NA values, I did not want to remove them since I wanted to retain as much of the original dataset as possible. Therefore, I replaced all the NA values for the continuous variables with the median value of the variable. This allowed me to preserve as many observations as possible, otherwise a significant portion of the dataset would have to be removed.

Next, I looked at the discrete variable STARS which is reflects the quality rating assigned by a panel of experts. Approximately, 26% of the observations lack a STARS rating. In the below table, we can see a summary of differences in mean purchases for each rating and for the wines with no rating.

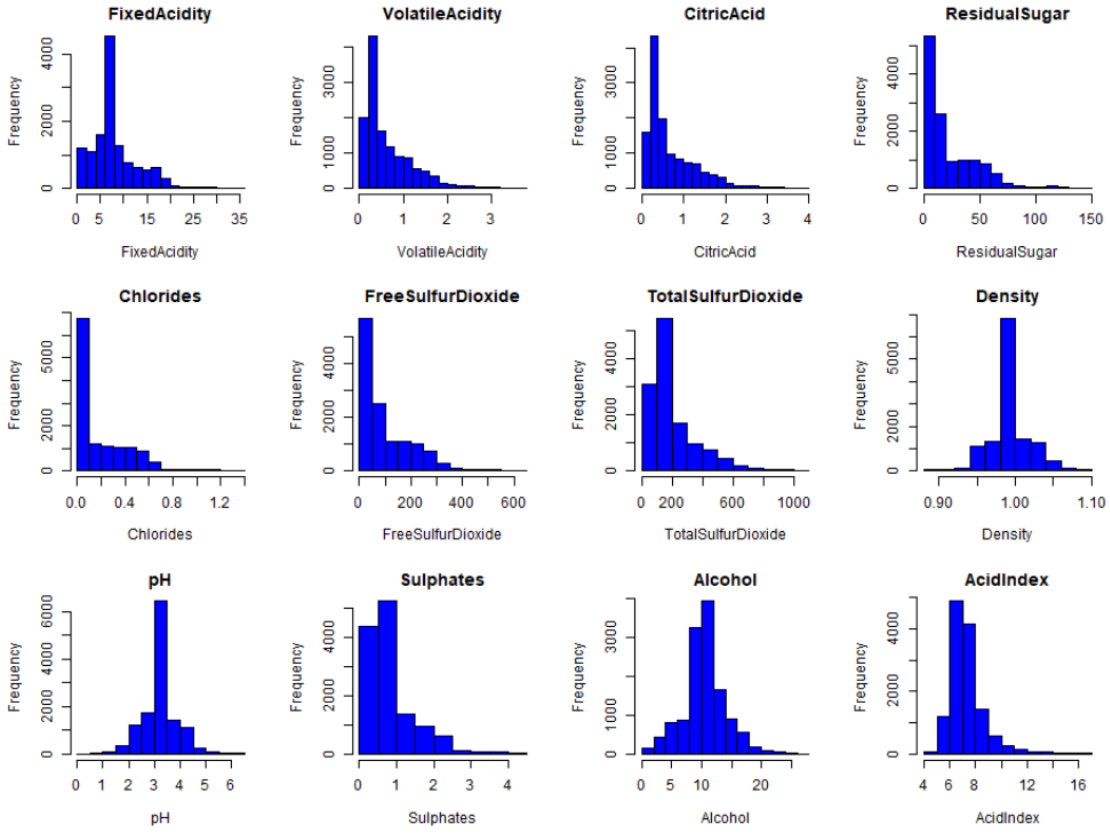
| STARS | Mean |
|---|---|
| 1 | 0.800 |
| 2 | 0.975 |
| 3 | 1.000 |
| 4 | 1.000 |
| NA | 0.393 |

From this table, we can also see that there is not much difference in the ratings 1-4 as even 80% of the lowest rated wines were purchased. However, there is a significantly large reduction in the amount purchases for wines that do not have a STARS rating. Therefore, I have determined that having a STARS rating is an important determinant to be considered for the model. As such, I have removed the original STARS variable and created a dummy coded categorical variable (“rated”) that is equal to 1 if the wine has a STARS rating and equal to 0 if there is no STARS rating. I have also determined that the Cases variable is not useful for the model as it is simply the amount of wine purchased and does not provide any information about the wine’s characteristics. Therefore, I have dropped this variable from the dataset.

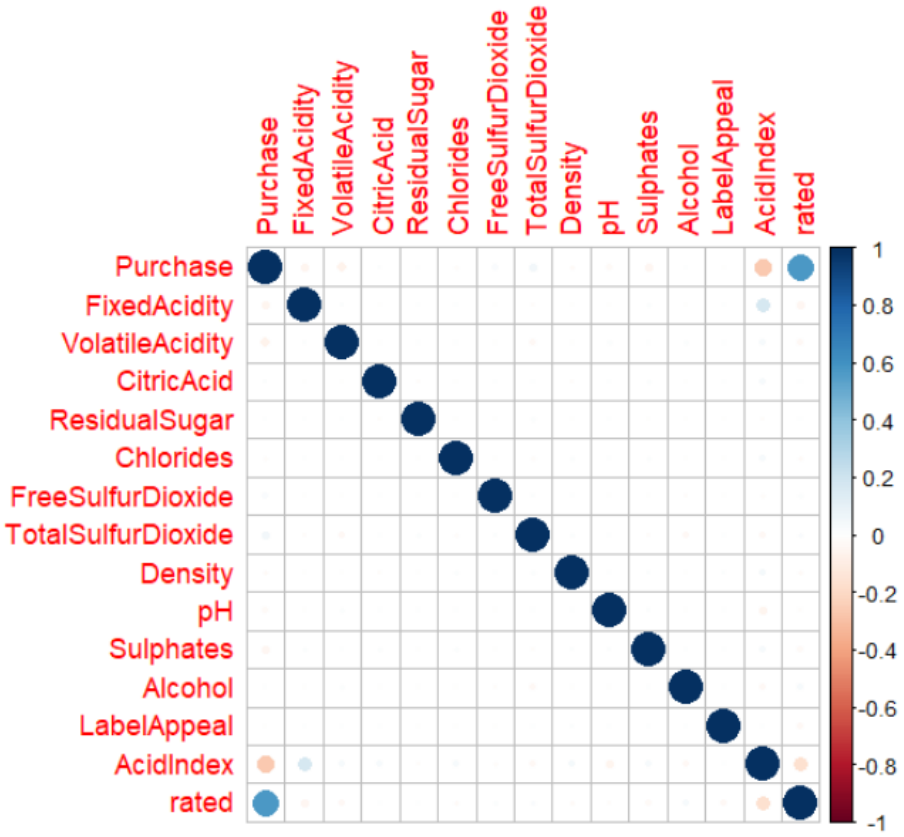
I have reproduced the summary statics table with the cleaned data and final variables that will be considered in the feature selection process for the logistic regression model below. We can see from this updated table that there are no longer any missing or negative values.

| key | < 0 Count | Avg | Max | Med | Med > 0 | Min | NA Cnt | Std |
|---|---|---|---|---|---|---|---|---|
| AcidIndex | 0 | 7.77 | 17.00 | 8.00 | 8.00 | 4.00 | 0 | 1.32 |
| Alcohol | 0 | 10.52 | 26.50 | 10.40 | 10.40 | 0.00 | 0 | 3.54 |
| Chlorides | 0 | 0.22 | 1.35 | 0.10 | 0.10 | 0.00 | 0 | 0.23 |
| CitricAcid | 0 | 0.69 | 3.86 | 0.44 | 0.44 | 0.00 | 0 | 0.61 |
| Density | 0 | 0.99 | 1.10 | 0.99 | 0.99 | 0.89 | 0 | 0.03 |
| FixedAcidity | 0 | 8.06 | 34.40 | 7.00 | 7.00 | 0.00 | 0 | 5.00 |
| FreeSulfurDioxide | 0 | 104.12 | 623.00 | 56.00 | 56.00 | 0.00 | 0 | 105.92 |
| LabelAppeal | 0 | 0.64 | 2.00 | 1.00 | 1.00 | 0.00 | 0 | 0.62 |
| pH | 0 | 3.21 | 6.13 | 3.20 | 3.20 | 0.48 | 0 | 0.67 |
| Purchase | 0 | 0.79 | 1.00 | 1.00 | 1.00 | 0.00 | 0 | 0.41 |
| rated | 0 | 0.74 | 1.00 | 1.00 | 1.00 | 0.00 | 0 | 0.44 |
| ResidualSugar | 0 | 22.86 | 141.15 | 12.90 | 12.90 | 0.00 | 0 | 24.44 |
| Sulphates | 0 | 0.82 | 4.24 | 0.59 | 0.59 | 0.00 | 0 | 0.63 |
| TotalSulfurDioxide | 0 | 201.64 | 1057.00 | 154.00 | 154.00 | 0.00 | 0 | 159.11 |
| VolatileAcidity | 0 | 0.64 | 3.68 | 0.41 | 0.41 | 0.00 | 0 | 0.56 |

Next, I created histograms of the continuous variables which can be seen below. We can see that some variables appear to close to normally distributed while several others are right-skewed. The right-skewed variables could cause issues for logistic regression modeling, so I will reference these histograms when selecting the final variables for the model.



Next, I created a correlation matrix of the variables which can be seen below. The only variables that show noticeable correlations to Purchase are the AcidIndex and rated variables. Additionally, there appears to be very little to no correlation between the explanatory variables. One of the assumptions of logistic regression models is that there is no multicollinearity among the explanatory variables, so seeing that we don't have collinearity between the explanatory variables this assumption has not been violated.




2. After concluding my exploratory data analysis, I then proceeded to perform a train/test split of the dataset using a 70/30 split. 70% of the data will be used to train the model and then I will test and validate the model on the remaining 30% of the unseen data. Below is a table that shows the number of observations in the training and test datasets.

| DataFrame | ObsCounts | PercentOfObs |
|---|---|---|
| Training Data | 8990 | 0.703 |
| Validation Data | 3805 | 0.297 |

Once I obtained a training data set, I then used automated feature selection as a starting point for the logistic regression model. To do this I utilized the forward, backward, and stepwise variable selection methods using the R function stepAIC() from the MASS library. First, I created 3 models to be used for this process. The first is a full model that uses all explanatory variables. The next model is the lower

model which is an intercept only model. The last model is a simple linear regression model that uses Alcohol as the explanatory variable in order to initialize a “dual” or “both” stepwise selection process. Then I passed these models to a stepwise variable selection process using forward, backward, and both directions techniques. The results of this variable selection process can be seen below along with the VIF value of each variable selected and a table that summarizes the AIC and BIC of each the models.

Forward Model:

```
AcidIndex             rated                pH TotalSulfurDioxide    VolatileAcidity  FreeSulfurDioxide
 1.021558          1.017783          1.007508          1.003235          1.002703          1.002317
Sulphates
 1.001795
```

Backward Model:

```
AcidIndex             rated                pH TotalSulfurDioxide    VolatileAcidity  FreeSulfurDioxide
 1.021558          1.017783          1.007508          1.003235          1.002703          1.002317
Sulphates
 1.001795
```

Stepwise (both) Model:

```
AcidIndex             rated                pH TotalSulfurDioxide    VolatileAcidity  FreeSulfurDioxide
 1.021558          1.017783          1.007508          1.003235          1.002703          1.002317
Sulphates
 1.001795
```

| Model | AIC | BIC |
|---|---|---|
| Forward | 6285.272 | 6342.103 |
| Backward | 6285.272 | 6342.103 |
| Stepwise | 6285.272 | 6342.103 |

We can see that each method selected the same variables. Additionally, all the VIF values are very close to 1. As such, there is little to no collinearity between the variables selected by the automated variable selection process. This is to be expected based on the correlation matrix from above. If any of the VIF were greater than 5, we would start to get concerned that there is collinearity between the explanatory variables and could possible violate one of the assumptions of logistic regression models. We can also see that the AIC and BIC for each model are the same due to the same variables being selected. I have selected the forward model to perform further analysis and model refining before validating the model on the test data. Below is a coefficient summary table and ANOVA table of this model.

*Model 1:*

```
Call:
glm(formula = Purchase ~ rated + AcidIndex + TotalSulfurDioxide +
    VolatileAcidity + pH + FreeSulfurDioxide + Sulphates, family = binomial,
    data = train_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7801   0.2567   0.3366   0.4196   2.4381

Coefficients:
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)        3.2905051  0.2621178  12.554 < 0.0000000000000002 ***
rated              2.9329438  0.0656432  44.680 < 0.0000000000000002 ***
AcidIndex         -0.4139367  0.0231933 -17.847 < 0.0000000000000002 ***
TotalSulfurDioxide 0.0008250  0.0002058   4.010            0.0000608 ***
VolatileAcidity   -0.1815570  0.0565257  -3.212              0.00132 **
pH                -0.1369542  0.0480368  -2.851              0.00436 **
FreeSulfurDioxide  0.0006091  0.0003124   1.950              0.05118 .
Sulphates         -0.0811881  0.0505360  -1.607              0.10816
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9338.3  on 8989  degrees of freedom
Residual deviance: 6269.3  on 8982  degrees of freedom
AIC: 6285.3

Number of Fisher Scoring iterations: 5


Analysis of Deviance Table

Model: binomial, link: logit

Response: Purchase

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev
NULL                                 8989     9338.3
rated              1  2679.96      8988     6658.4
AcidIndex          1   345.28      8987     6313.1
TotalSulfurDioxide 1    17.99      8986     6295.1
VolatileAcidity    1    10.85      8985     6284.2
pH                 1     8.56      8984     6275.7
FreeSulfurDioxide  1     3.85      8983     6271.8
Sulphates          1     2.56      8982     6269.3
```

Based on the above summary, we can see that the rated dummy coded variable has the largest impact on the model’s predicted log odds ratio. However, there are several variables that appear to have minimal influence on the model. This is likely to due to the fact that we have “too much statistical power” as mentioned previously. These variables might technically be statistically significant; however, they may not actually provide any meaningful predictive value to the model. Therefore, I proceeded to drop variables one by one from the model to see how the model changed in order to assess the impact of the variable being dropped. Additionally, by comparing the log-likelihoods I also performed a chi-square test to determine if the variable being dropped is contributing any information to the model.

For reference the critical chi-square value I will be using for the hypothesis tests is 3.841. First, I removed the Sulphates variable due to having a higher p-value and called this Model 2. This resulted in chi-square value of 2.65 meaning that we can fail to reject the null hypothesis can concluded that the Sulphates variable is not contributing and meaningful information to the model any can be removed. Below is the coefficient table for Model 2.

*Model 2:*

```
Call:
glm(formula = Purchase ~ rated + AcidIndex + TotalSulfurDioxide +
    VolatileAcidity + pH + FreeSulfurDioxide, family = binomial,
    data = train_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7740   0.2569   0.3360   0.4179   2.4223

Coefficients:
                   Estimate Std. Error z value             Pr(>|z|)
(Intercept)         3.2356072  0.2595998  12.464 < 0.0000000000000002 ***
rated               2.9329266  0.0656247  44.692 < 0.0000000000000002 ***
AcidIndex          -0.4149797  0.0231734 -17.908 < 0.0000000000000002 ***
TotalSulfurDioxide  0.0008309  0.0002058   4.038             0.000054 ***
VolatileAcidity    -0.1832379  0.0564558  -3.246              0.00117 **
pH                 -0.1384517  0.0480104  -2.884              0.00393 **
FreeSulfurDioxide   0.0006095  0.0003124   1.951              0.05102 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9338.3  on 8989  degrees of freedom
Residual deviance: 6271.8  on 8983  degrees of freedom
AIC: 6285.8

Number of Fisher Scoring iterations: 5
```

Next, I removed FreeSulfurDioxide since it had the next highest p-value and called this Model 3. This resulted in chi-square value of 3.852 meaning that we can reject the null hypothesis can concluded that the FreeSulfurDioxide variable is contributing to the model. I also checked the accuracy and AUC for Model 3 and these were relatively the same as the original model. Therefore, although FreeSulfurDioxide is statistically significant, removing the variable does have a material negative impact on the model’s predictive performance. Again, this seems to be attributable to having “too much statistical power”. As such, I have removed this variable to opt for parsimony and a more simplified model. Below is the coefficient table for Model 3.

*Model 3:*

```
Call:
glm(formula = Purchase ~ rated + AcidIndex + TotalSulfurDioxide +
    VolatileAcidity + pH, family = binomial, data = train_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7375   0.2576   0.3356   0.4179   2.4141

Coefficients:
                   Estimate Std. Error z value             Pr(>|z|)
(Intercept)       3.3135630  0.2566335  12.912 < 0.0000000000000002 ***
rated             2.9325455  0.0655961  44.706 < 0.0000000000000002 ***
AcidIndex        -0.4166408  0.0231823 -17.972 < 0.0000000000000002 ***
TotalSulfurDioxide 0.0008394  0.0002057   4.081            0.0000448 ***
VolatileAcidity  -0.1816842  0.0564738  -3.217              0.00129 **
pH               -0.1401789  0.0479601  -2.923              0.00347 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9338.3  on 8989  degrees of freedom
Residual deviance: 6275.7  on 8984  degrees of freedom
AIC: 6287.7

Number of Fisher Scoring iterations: 5
```

Next, I removed pH since it had the next highest p-value and called this Model 4. This resulted in chi-square value of 8.561 meaning that we can reject the null hypothesis can concluded that the pH variable is contributing to the model. I also checked the accuracy and AUC for Model 4 and these were relatively the same as the original model. Therefore, although pH is statistically significant based on the chi-square test, removing the variable does have a material negative impact on the model's predictive performance. Again, this seems to be attributable to having "too much statistical power". As such, I have removed this variable to opt for parsimony and a more simplified model. Below is the coefficient table for Model 4.

*Model 4:*

```
Call:
glm(formula = Purchase ~ rated + AcidIndex + TotalSulfurDioxide +
    VolatileAcidity, family = binomial, data = train_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7133   0.2600   0.3341   0.4137   2.4375

Coefficients:
                   Estimate Std. Error z value             Pr(>|z|)
(Intercept)       2.8315039  0.1954174  14.490 < 0.0000000000000002 ***
rated             2.9310982  0.0655226  44.734 < 0.0000000000000002 ***
AcidIndex        -0.4122158  0.0231183 -17.831 < 0.0000000000000002 ***
TotalSulfurDioxide 0.0008407  0.0002055   4.090            0.0000431 ***
VolatileAcidity  -0.1867921  0.0563693  -3.314             0.000921 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9338.3  on 8989  degrees of freedom
Residual deviance: 6284.2  on 8985  degrees of freedom
AIC: 6294.2

Number of Fisher Scoring iterations: 5
```

Next, I removed VolatileAcididty since it had the next highest p-value and called this Model 5. This resulted in chi-square value of 10.85 meaning that we can reject the null hypothesis can concluded that the VolatileAcidity variable is contributing to the model. I also checked the accuracy and AUC for Model 5 and these were relatively the same as the original model. Therefore, although VolatileAcidity is statistically significant based on the chi-square test, removing the variable does have a material negative impact on the model's predictive performance. Again, this seems to be attributable to having "too much statistical power". As such, I have removed this variable to opt for parsimony and a more simplified model. Below is the coefficient table for Model 5.

*Model 5:*

```
Call:
glm(formula = Purchase ~ rated + AcidIndex + TotalSulfurDioxide,
    family = binomial, data = train_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7952   0.2668   0.3350   0.4111   2.4651

Coefficients:
                    Estimate Std. Error z value             Pr(>|z|)
(Intercept)        2.7145662  0.1919133  14.145 < 0.0000000000000002 ***
rated              2.9318913  0.0654490  44.797 < 0.0000000000000002 ***
AcidIndex         -0.4136289  0.0231120 -17.897 < 0.0000000000000002 ***
TotalSulfurDioxide 0.0008599  0.0002054   4.186             0.0000284 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9338.3  on 8989  degrees of freedom
Residual deviance: 6295.1  on 8986  degrees of freedom
AIC: 6303.1

Number of Fisher Scoring iterations: 5
```

Next, I removed TotalSulfurDioxide since it had the next highest p-value and called this Model 6. This resulted in chi-square value of 17.99 meaning that we can reject the null hypothesis can concluded that the TotalSulfurDioxide variable is contributing to the model. I also checked the accuracy and AUC for Model 6 and these were relatively the same as the original model. Therefore, although TotalSulfurDioxide is statistically significant based on the chi-square test, removing the variable does have a material negative impact on the model's predictive performance. Again, this seems to be attributable to having "too much statistical power". As such, I have removed this variable to opt for parsimony and a more simplified model. Below is the coefficient table for Model 6.

*Model 6:*

```
Call:
glm(formula = Purchase ~ rated + AcidIndex, family = binomial,
    data = train_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7486   0.2641   0.3238   0.3961   2.4386

Coefficients:
            Estimate Std. Error z value            Pr(>|z|)
(Intercept)  2.90587    0.18666   15.57 <0.0000000000000002 ***
rated        2.92936    0.06530   44.86 <0.0000000000000002 ***
AcidIndex   -0.41619    0.02308  -18.03 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9338.3  on 8989  degrees of freedom
Residual deviance: 6313.1  on 8987  degrees of freedom
AIC: 6319.1

Number of Fisher Scoring iterations: 5
```

Next, I removed AcidIndex since it had the next highest p-value and called this Model 7. This resulted in chi-square value of 345.3 meaning that we can reject the null hypothesis can concluded that the AcidIndex variable is contributing to the model. I also checked the accuracy and AUC for Model 7 and the AUC decreased by .05. Therefore, removing the AcidIndex variable does have a material negative impact on the model's performance. So, I have chosen to keep this variable in the final model. Below is the coefficient table for Model 6.

*Model 7:*

```
Call:
glm(formula = Purchase ~ rated, family = binomial, data = train_df)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.286   0.390   0.390   0.390   1.363

Coefficients:
            Estimate Std. Error z value            Pr(>|z|)
(Intercept) -0.42615    0.04191  -10.17 <0.0000000000000002 ***
rated        2.96404    0.06314   46.95 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9338.3  on 8989  degrees of freedom
Residual deviance: 6658.4  on 8988  degrees of freedom
AIC: 6662.4

Number of Fisher Scoring iterations: 5
```

A summary of these model metrics can be seen below. We can see how in Models 1 through 6 the AIC increases slightly for each model which is to be expected when removing variables; however, for the most part they all have similar metrics and there is only a noticeable decrease in AUC when looking at Model 7 which only uses the rated variable as the explanatory variable. As such, in opting for a more simplified model over a more complex model I have selected Model 6 as the final model for testing even though the chi-square tests for models 3, 4, and 5, suggest that those variables are statistically significant to the model. As previously mentioned, this is likely to due to having "too much statistical power" from having a large number of observations in the dataset.

| Model | AIC | BIC | Chi_Square | AUC | Accuracy |
|---|---|---|---|---|---|
| Model 1 | 6285.272 | 6342.103 | NA | 0.857 | 0.843 |
| Model 2 | 6285.832 | 6335.559 | 2.56 | 0.857 | 0.843 |
| Model 3 | 6287.684 | 6330.307 | 3.852 | 0.857 | 0.843 |
| Model 4 | 6294.246 | 6329.765 | 8.561 | 0.856 | 0.844 |
| Model 5 | 6303.095 | 6331.510 | 10.849 | 0.856 | 0.841 |
| Model 6 | 6319.086 | 6340.397 | 17.991 | 0.853 | 0.840 |
| Model 7 | 6662.371 | 6676.578 | 345.285 | 0.808 | 0.841 |

Below is the coefficient summary table and ANOVA for the final model (Model 6).

```
Call:
glm(formula = Purchase ~ rated + AcidIndex, family = binomial,
    data = train_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7486   0.2641   0.3238   0.3961   2.4386

Coefficients:
            Estimate Std. Error z value             Pr(>|z|)
(Intercept)  2.90587    0.18666   15.57 <0.0000000000000002 ***
rated        2.92936    0.06530   44.86 <0.0000000000000002 ***
AcidIndex   -0.41619    0.02308  -18.03 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9338.3  on 8989  degrees of freedom
Residual deviance: 6313.1  on 8987  degrees of freedom
AIC: 6319.1

Number of Fisher Scoring iterations: 5

Analysis of Deviance Table

Model: binomial, link: logit

Response: Purchase

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev
NULL                       8989     9338.3
rated      1  2679.96       8988     6658.4
AcidIndex  1   345.28       8987     6313.1
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Purchase

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev
NULL                       8989     9338.3
rated      1  2679.96       8988     6658.4
AcidIndex  1   345.28       8987     6313.1
```

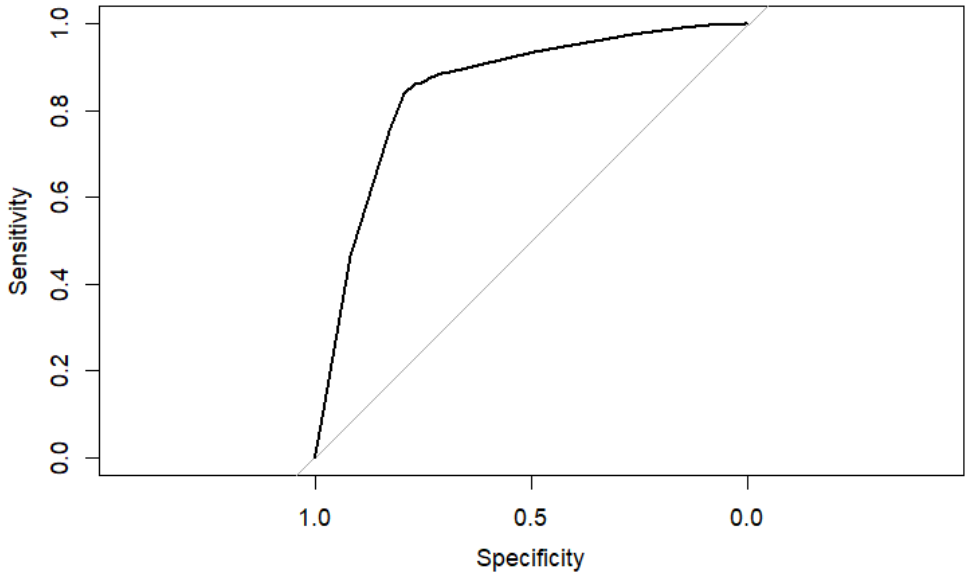
The intercept for the logistic regression model is 2.906. This means that log of odds of the wine being purchased is equal to 2.906 when AcidIndex is equal to 0 and the wine has no STARS rating. This intercept cannot be interpreted outside the context of this model since the minimum value of AcidIndex in this dataset is 4. So, having an AcidIndex equal to 0 may not be feasible. The coefficient for the rated variable is 2.293. This can be interpreted as a wine's log of odds for being purchased will increase by 2.293 is the wine has a STARS rating given everything else held constant. The coefficient for the AcidIndex is -0.416 which can be interepreted as the wine's log of odds for being purchased will decrease by 0.416 given everything else held constant. We can see that this coefficient is negative meaning that AcidIndex value is negatively correlated with a wine's probability of being purchased. This matches the correlation matrix from before. Logically, these coefficients make sense. Wines that had a STARS rating were more likely to be purchased based on observations in the dataset. Additionally, it also makes sense that more acidic wines were less likely to be purchased since they might not be as pleasant to drink compared to less acidic wines, therefore decreasing the likelihood of being purchased. Furthermore, based on the correlation matrix, these variables had the strongest linear relationship to the Purchase variable so it also makes sense that these variables are the explanatory variables used in the model to predict the probability of a purchase.

Prior to model testing, I also checked to see if adding an interaction variable between AcidIndex and rated provided any useful information to the model. Based on this model's coefficient summary table, the interaction variable has a p-value of 0.231 which is greater than the alpha value of .05. Additionally, adding this variable to the model resulted in a chi-square value of 1.427 which is less than the critical chi-square value meaning that we can fail to reject the null hypothesis can concluded that the interaction variable is not contributing any meaningful information to the model and can be removed. Therefore, the final model uses the rated variable and the AcidIndex variables as the explanatory variables to predict the probability of a wine being purchased by customers. From the histograms displayed in the EDA section, we can see that the AcidIndex is approximately normally distributed. It does have a slight right-skew but it is minimal, so I will move forward with including this variable in the final mode.

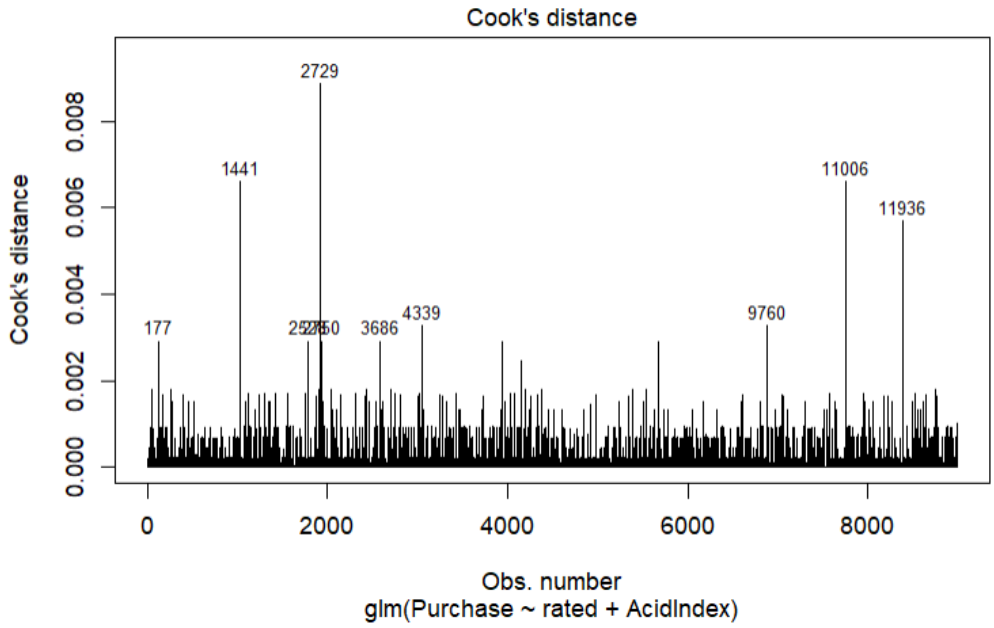
Next, I proceeded to validate the model by using the test dataset from the train/test split. Using a 0.5 probability threshold, I converted predicted probabilities that were greater than 0.5 to reflect a "purchase (1)" and less than 0.5 to reflect "no purchase (0)". Below is a table that shows the actual purchases vs the predicted purchases using the test dataset.

| | Predict | |
|---|---|---|
| Purchase | 0 | 1 |
| 0 | 567 | 242 |
| 1 | 341 | 2655 |

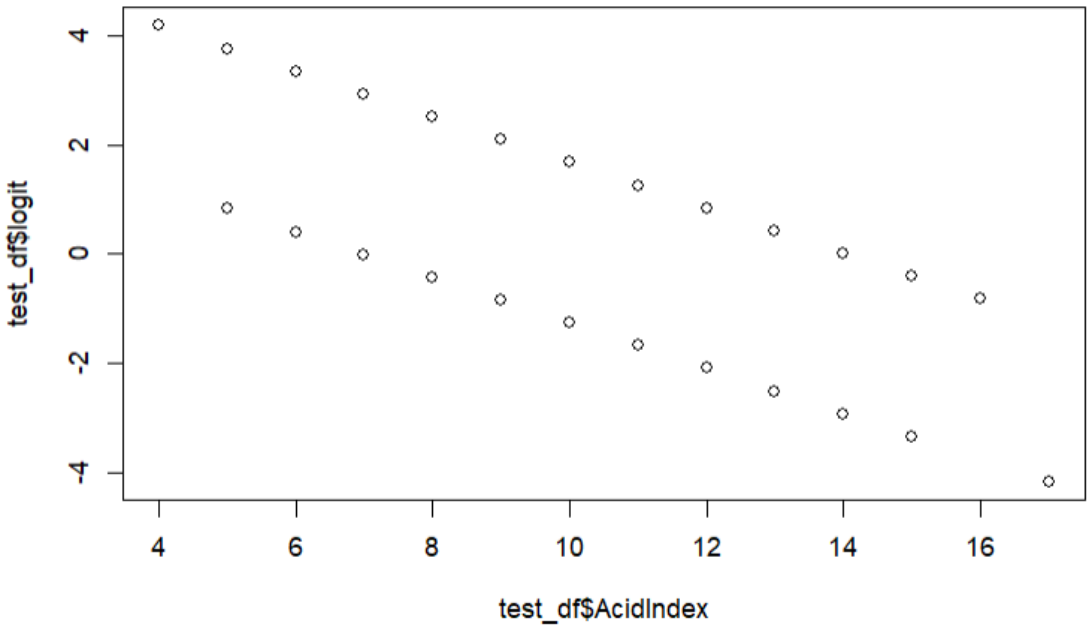
We can see that the model correctly predicted 3,222 observations which is equivalent to 84.7% accuracy. This is actually slightly better than the model's performance on the training dataset which had an accuracy of 84%. This is a good sign that shows that our model has not been overfit to the training data. Below is the ROC curve for the final model on the test dataset. This ROC curve results in an AUC of 0.8517 which is similar to the AUC observed using the training dataset.




Next, I created the below Cook's Distance Chart to visualize the influential observations of the model. This chart shows that there are several potential outliers based on the Cook's Distance relative to the rest of the data points. We should be concerned that these observations are outliers that may have a high influence on the model. We could potentially remove these observations to increase the model performance and fit. However, these observations could also be valid representations of the sample population. So, further analysis should be done on these influential observations before determining if they should be removed and refitting the model.




Lastly, I checked the logistic regression assumption that the continuous explanatory variables have linear relationship with the predicted logits. Below we can see a scatter plot that shows this linear relationship assumption has not been violated between the predicted logits and the AcidIndex variable. Additionally, I also used a Box-Tidwell test which resulted in a p-value of 0.114 for the AcidIndex variable which is greater than the alpha value of 0.05 which also suggests that we can accept the null hypothesis of linearity and conclude here is a linear relationship between the predicted logits and the AcidIndex values.



## 3. Conclusion

Overall, I am satisfied with the fit of the final logistic regression model. The final model has achieved a prediction accuracy on the unseen test data better than the training accuracy and also the baseline proportions in the original dataset. From my analysis, the major determinant if a wine is likely to purchased or not is if the wine has a STARS rating. The original variable STARS variable was transformed to a binary dummy coded variable due to a large portion of the observations not having a STARS rating. It would be interesting if all the wines in the data could have received a STARS rating and keeping the original discrete STARS variable, how the model would change to reflect each STARS rating (1-5). Through this modeling process, I have learned that in the wine world the STARS rating and the Acid Index level of the wine contribute most to a customer's purchasing decision. As such, my recommendation for a wine manufacture is that they should aim to sell wines with relatively lower Acid Index levels and that have a STARS rating or seek to have wines rated by a panel of experts in order to receive a STARS rating.

From this modeling process, my perspective has changed in that while it is good to have large amounts of data to look for patterns and draw insights for analytical purposes, we must also be cautious of having "too much statistical power". As seen from this analysis, several variables were considered statistically significant from the automated feature selection and the chi-square test. However, by removing these variables one-by-one and refitting the model, I was able to determine that these variables did not provide significant predictive information to the model as the model metrics remained relatively unchanged. As such, I removed these variables from the final model to obtain a more parsimonious model. I have also learned that data modeling is a rigorous and time-consuming process in terms of EDA, data cleaning and processing, variable and model selection and interpretation, re-assessing each variable through hypothesis tests and model impact, checking model assumptions and model diagnostics, and communicating findings and insights. With that said, I now have a greater appreciation for amount of effort needed to obtain a fully developed model that considers the possibility of having too much statistical power with large datasets.