**Red Wines Logistic Regression**

In an effort to get a better feel for the data, I started out by using some simple commands to get a better look at the data file. The headers read as follows: “fixed acidity”, “volatile acidity”, “residual sugar”, “chlorides”, “total sulfur dioxide”, “density”, “sulphates” and “alcohol”. These are all the variables that are thought to be the most important ingredients that make a good wine. The final column reads “good”, which denotes whether the wine in question scored between a 90 and a 100 (‘yes”) or not (“no”) in Wine Enthusiast Magazine’s review.

Moving forward, I thought it would be worthwhile to get a look at the first six rows of the data file. Here is the output from this command:

fixed.acidity volatile.acidity residual.sugar chlorides

1 7.4 0.70 1.9 0.076

2 7.8 0.88 2.6 0.098

3 7.8 0.76 2.3 0.092

4 11.2 0.28 1.9 0.075

5 7.4 0.70 1.9 0.076

6 7.4 0.66 1.8 0.075

total.sulfur.dioxide density sulphates alcohol good

1 34 0.9978 0.56 9.4 No

2 67 0.9968 0.68 9.8 No

3 54 0.9970 0.65 9.8 No

4 60 0.9980 0.58 9.8 No

5 34 0.9978 0.56 9.4 No

6 40 0.9978 0.56 9.4 No

Looking more intently into the scales and central tendencies of these values, I used a command that revealed a great deal about the statistical information. Here is the output:

A screenshot of a computer

Description automatically generated

Finally, as a last step before using the backwards stepwise regression to truly reveal the most important variables, I decided to create a matrix outling all the correlations between all the variables. This command produced the following matrix:

A screenshot of a computer

Description automatically generatedA white background with black text

Description automatically generated  
A screenshot of a computer code

Description automatically generated

This matrix definitely produced some interesting results. The first thing that I noticed is that the only variable that has any correlation with the “good” column at all was the “residual sugar” column, and even this correlation was very weak (.05). However, this matrix did help us discover that there may be some collinearity among some of our independent variables! For instance, there seems to be a rather high correlational relationship between the following variables: ‘sulphates~ residual sugar’ and ‘volatile acidity~ residual sugar’. For these variables we will have to be aware of potential collinearity when it comes to building the final model.

The first model I built informed me that the categories with the highest level of significance were alcohol, sulphates and total sulfur dioxide (all with p values below .001). Here is the output from that command:

A screenshot of a computer

Description automatically generated

Yet, there is more we can do to be even more accurate when teasing out which variables are most valuable in engineering the best wine. The model depicted above only shows the significance of each variable when all variables are run simultaneously in the same model. Next, we will input a command that will run several models back-to-back, progressively removing the weakest variables as it goes. The following image depicts the final model the system generated using this method:

A screenshot of a computer

Description automatically generated

As you can see here, the system deduced that the following variables are most important when it comes to building a good wine: fixed acidity, volatile acidity, chlorides, total sulfur dioxide, sulphates, and alcohol. However, this doesn’t necessarily guarantee that all of these variables are high in significance. To ensure that all variables are truly significant, we must run further testing.

This matrix displays the results of a Chi squared anova, which tells us which of the variables were most important when manufacturing a wine that people rate highly.

A screenshot of a computer

Description automatically generated

In this instance, we want our deviance residuals to be as high as possible. While none of the coefficients are *very* high here, we can see that chlorides are not very important and alcohol, fixed acidity and sulphates all seem to be the most significant variables. Thus, moving forward with my analysis, I am quite confident that alcohol, fixed acidity and sulphates are the most important variables when engineering a good-tasting wine.

In order to make my final recommendations, I decided to separate all the wines that received a score of ‘good’ from the Wine Enthusiast Magazine. Then, I decided to further separate the data by excluding all the variables that we found to be weak predictors of a good wine (all but alcohol, fixed acidity and sulphates). In doing so, I was able to deduce a range of for each of these values that one should stay within if you hope to engineer a wine that is worthy of a 90+ in the Wine Enthusiast Magazine. For the fixed acidity, one should stay between 5 and 15.6 while primarily aiming for anywhere between 8.7 and 11.6. For sulphates, it is advisable to stay between .47 and 1.36 while striving to land anywhere between .69 and .85. Finally, for alcohol content, the data suggest that a good winemaker would do well to keep the alcohol content between 9.2 and 14 while attempting to finish somewhere between 10.5 and 12. Here is the matrix I used to make these suggestions:

A screenshot of a computer screen

Description automatically generated