Building a recommendation model for red wine from physiochemical properties

John Li and Pavan Kalyan

5/10/2022

1 Introduction

Wine has had a long history with human culture and archaeologists have discovered that wine was being produced in 6000 BC. The process of turning grape juice to wine depends on many factors and the resulting wine can vary in various qualities which can affect it's taste and appeal to consumers. Among wine producers, Portugal is a top ten wine exporting country with 3.17% of the market share in 2005. From the northwestern region of Portugal, a particular wine known as *vinho verde* has drawn consumer interest in recent years.

Within the industry the certification and quality assessment of wines are key in the wine making process. Certification assures quality and prevents illegal adulteration harmful to human health, while professional taste assessments are important as well in assuring quality and in identifying different price points for wines.

We wish to determine if we can use the physical and chemical properties of wine to accurately predict how professional wine raters would respond. With large datasets we can use statistical software such as R to analyze the key predictors associated with high quality ratings. We can also build a model to predict and classify new data into high and low quality ratings. For consumers, knowing what wines are rated highly by expert wine tasters can save time and hassle when shopping. Therefore, our goal is to build a multiple logistic regression model to predict between the classes of recommended wines or not recommended wines. Furthermore, our results will hopefully help us better understand the qualities of wine that contribute to good taste.

2 Data Description

The dataset used can be found online, hosted by UC Irvine Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/wine+quality]. The dataset is large and split into measurements taken on red *vinho verde* or white *vinho verde*. The data was collected from May 2004 to February 2007 using only protected designation in origin samples that were tested at the official certification entity (CVRVV). Because we are mainly interested in building a predictive model for red wine (and since taste differs from red and white wine), we choose to only utilize the red wine samples for analysis. Each wine sample was given a sensory score (reflected as "quality" in the dataset) which was the median score of three separate blind tastes. The other variables in the data reflect the measurements of common physiochemical tests. See Cortez et al. (2009) for more information about the data measurement process or background about *vinho verde*.

The dataset contains 1599 samples of red *vinho verde* wines with 11 predictor variables based on physiochemical measurements and the response variable which is the grade of the wine in a scale from 0 (very bad) to 10 (excellent). Figure 1 plots the histograms of the 12 target variables.

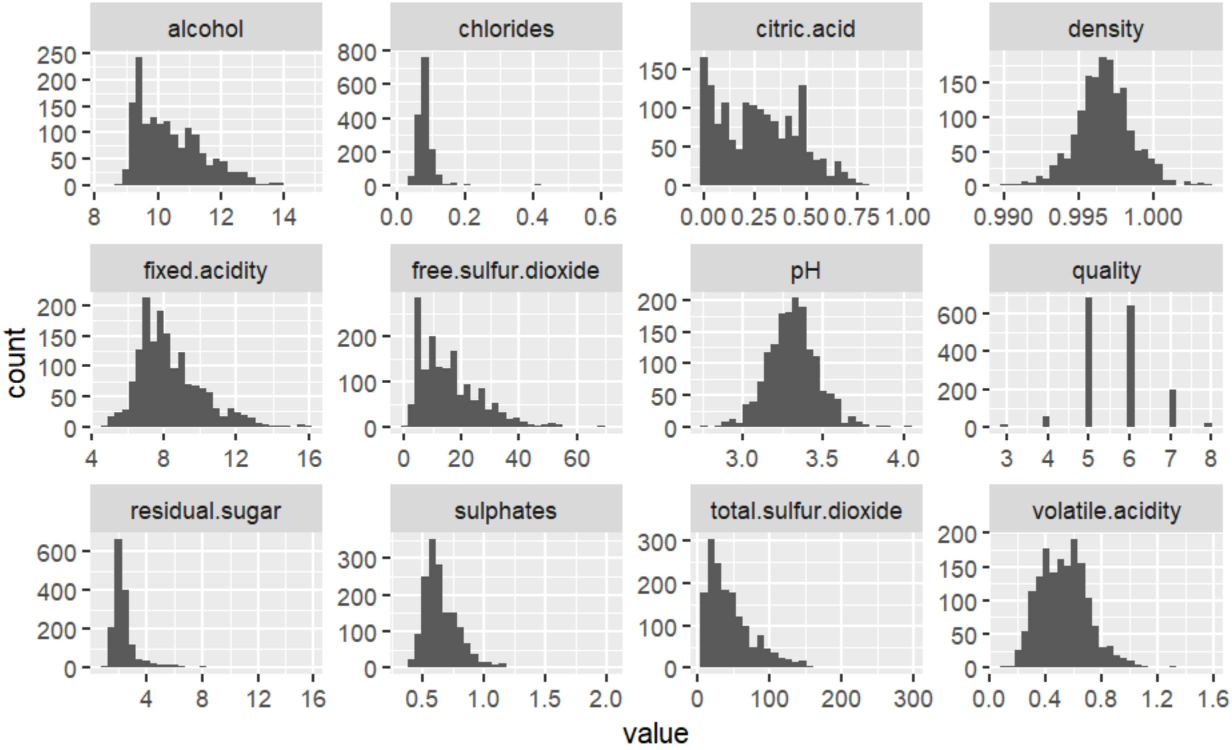
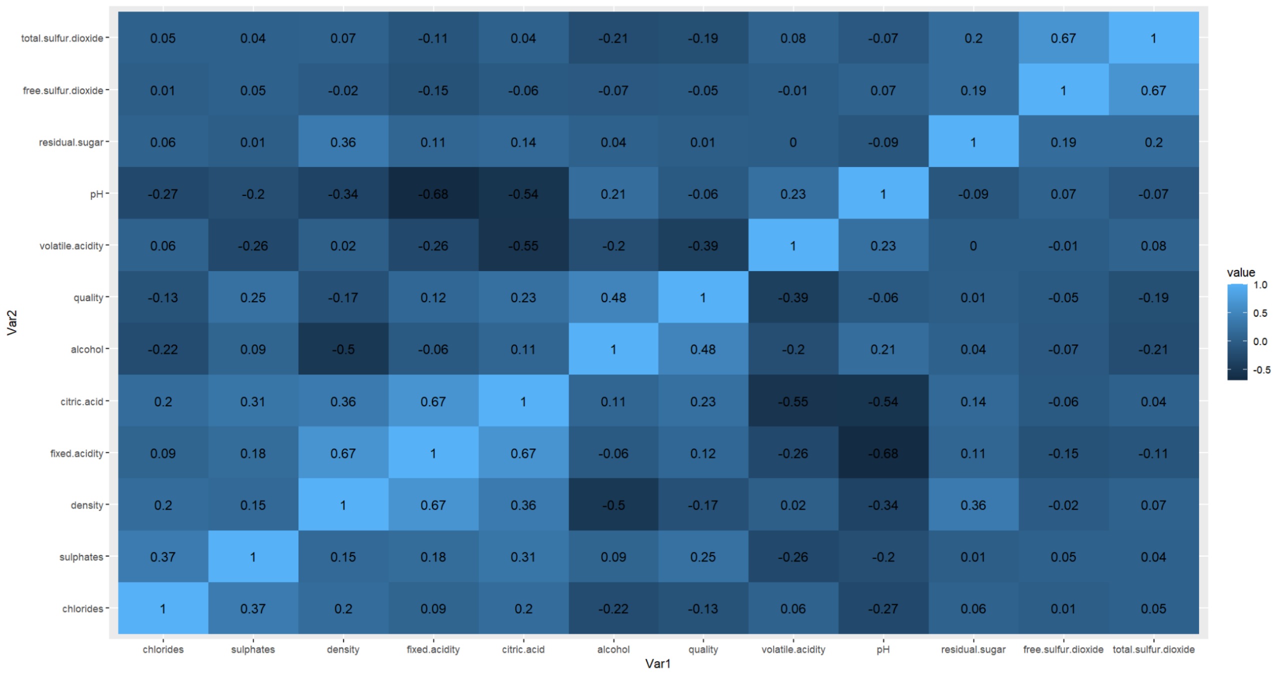


Figure 1: Distribution of target variables

We can see that the distribution of quality has a lower range of 3 and a upper range of 8. Also we observe that the majority of samples have a quality of either 5 or 6, while the extreme values are increasingly rare. Looking at the other 11 variables we can notice that these measurements are continuous variables, many of which have skewed or other non-normal distributions. For this reason, creating a logistic classification model allows us to make an accurate model without requiring the assumption of normally distributed variables.

We may be concerned with issues of multicollinearity, especially since several predictors like pH, citric acid, and fixed acidity are all various measurements of acidity. Figure 2. shows the correlation matrix, a 12 by 12 grid which gives a numerical measurement between -1 and 1 on the strength of correlation between any two variables.

As expected there appears to be moderate to high correlation between pH and fixed acidity (-0.68), ph and citric acid (-0.54). Alcohol and density also have a moderate negative correlation (-0.5). Because highly correlated predictors reduce the accuracy of estimated regression coefficients we will choose



	chlorides	sulphates	density	fixed.acidity	citric.acid	alcohol	quality	volatile.acidity	pH	residual.sugar	free.sulfur.dioxide	total.sulfur.dioxide
total.sulfur.dioxide	0.05	0.04	0.07	-0.11	0.04	-0.21	-0.19	0.08	-0.07	0.2	0.67	1
free.sulfur.dioxide	0.01	0.05	-0.02	-0.15	-0.06	-0.07	-0.05	-0.01	0.07	0.19	1	0.67
residual.sugar	0.06	0.01	0.36	0.11	0.14	0.04	0.01	0	-0.09	1	0.19	0.2
pH	-0.27	-0.2	-0.34	-0.68	-0.54	0.21	-0.06	0.23	1	-0.09	0.07	-0.07
volatile.acidity	0.06	-0.26	0.02	-0.26	-0.55	-0.2	-0.39	1	0.23	0	-0.01	0.08
quality	-0.13	0.25	-0.17	0.12	0.23	0.48	1	-0.39	-0.06	0.01	-0.05	-0.19
alcohol	-0.22	0.09	-0.5	-0.06	0.11	1	0.48	-0.2	0.21	0.04	-0.07	-0.21
citric.acid	0.2	0.31	0.36	0.67	1	0.11	0.23	-0.55	-0.54	0.14	-0.06	0.04
fixed.acidity	0.09	0.18	0.67	1	0.67	-0.06	0.12	-0.26	-0.68	0.11	-0.15	-0.11
density	0.2	0.15	1	0.67	0.36	-0.5	-0.17	0.02	-0.34	0.36	-0.02	0.07
sulphates	0.37	1	0.15	0.18	0.31	0.09	0.25	-0.26	-0.2	0.01	0.05	0.04
chlorides	1	0.37	0.2	0.09	0.2	-0.22	-0.13	0.06	-0.27	0.06	0.01	0.05

Figure 2: Correlation matrix of target variables

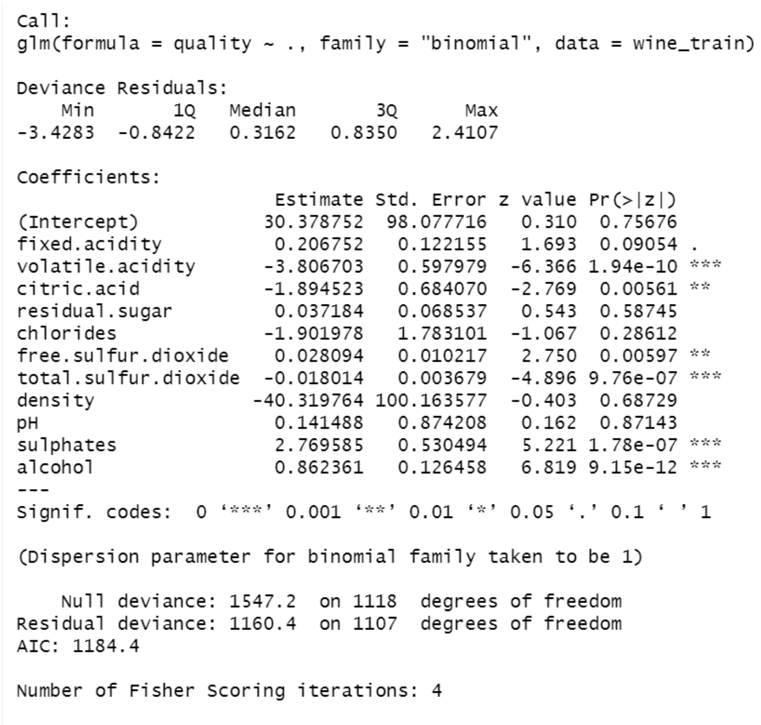
a stepwise approach to variable selection based on AIC. We will see that multicollinearity is much reduced in the final model after variable selection has been applied.

3 Methods and Results

After considering several alternatives we decided to adopt a multiple logistic regression approach. We considered other model building approaches such as classification trees or random forests but ultimately decided on MLR for our purposes. The reason for doing this is we wish to build recommendations for wine based on their qualities, targeted towards more practical use. For experts, they are able to differentiate and assess quality on a discrete scale of 0 to 10. However, consumers may be more interested in knowing wines which meet a minimum rating by experts and choose the one which fits their budget or tastes from among the recommended wines.

For the reasons explained, we modified quality into a binary variable with levels corresponding to 0 for all scores 5 or lower, and 1 for scores higher than 5. We are left with 744 observations for 0 and 855 for 1. There were no other transformations applied to the other variables.

We used the holdout validation approach to estimate the general capability and accuracy of our model. The red *vinho verde* dataset was randomly split into training and test sets; 70% was used for training the model, 30% was withheld for testing afterwards. With our training data we used R software to fit a multiple logistic regression model using our transformed quality as the response. Starting with all possible predictors, we can see the summary of the model in Figure 3 below.



```
Call:
glm(formula = quality ~ ., family = "binomial", data = wine_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4283  -0.8422   0.3162   0.8350   2.4107

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           30.378752  98.077716   0.310  0.75676
fixed.acidity          0.206752   0.122155   1.693  0.09054 .
volatile.acidity      -3.806703   0.597979  -6.366 1.94e-10 ***
citric.acid           -1.894523   0.684070  -2.769  0.00561 **
residual.sugar         0.037184   0.068537   0.543  0.58745
chlorides             -1.901978   1.783101  -1.067  0.28612
free.sulfur.dioxide    0.028094   0.010217   2.750  0.00597 **
total.sulfur.dioxide  -0.018014   0.003679  -4.896 9.76e-07 ***
density              -40.319764 100.163577  -0.403  0.68729
pH                     0.141488   0.874208   0.162  0.87143
sulphates              2.769585   0.530494   5.221 1.78e-07 ***
alcohol                0.862361   0.126458   6.819 9.15e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1547.2  on 1118  degrees of freedom
Residual deviance: 1160.4  on 1107  degrees of freedom
AIC: 1184.4

Number of Fisher Scoring iterations: 4
```

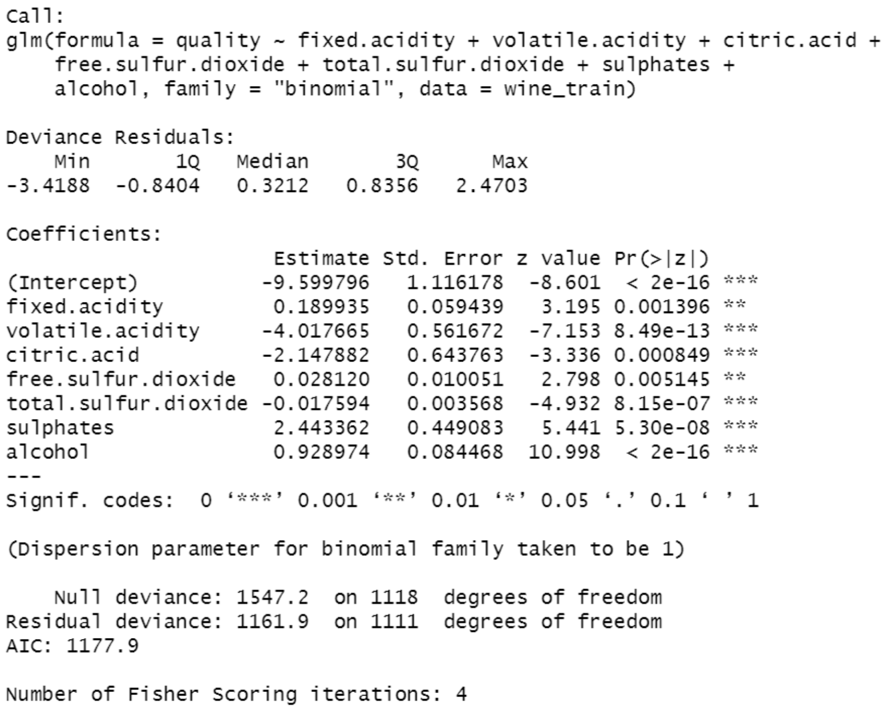
Figure 3: Summary of MLR using 11 predictors

Based on the summary output there are several predictors which are not statistically significant, while others are highly significant. We will use the AIC approach for variable selection using R. In this approach variables are removed one by one to lower the overall AIC of the model. We find that the reduced model has lower AIC compared to the full model from 1184.4 to 1177.9. The number of predictors has been reduced from 11 to 7.

The summary of the reduced model gives us the estimated coefficients of the equation for the MLR model. The coefficients indicate that Fixed acidity, free sulfur dioxide, sulphates, and alcohol are associated with higher log-odds of being a “recommend”, while volatile acidity, citric acid, total sulfur dioxide are associated with higher log-odds of being “not recommend”.

We can express the predicted probability of a sample of wine being lowly or highly rated with the following equation:

$logit(\hat{p}) = -9.60 + 0.19(fixed.acidity) - 4.02(volatile.acidity) - 2.15(citric.acid) + 0.028(free.sulfur.dioxide) - 0.018(total.sulfur.dioxide) + 2.44(sulphates) + 0.93(alcohol)$

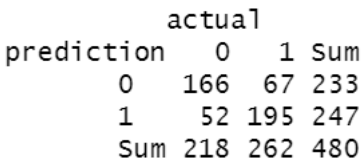


```
Call:
glm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
    free.sulfur.dioxide + total.sulfur.dioxide + sulphates +
    alcohol, family = "binomial", data = wine_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4188  -0.8404   0.3212   0.8356   2.4703

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -9.599796   1.116178  -8.601  < 2e-16 ***
fixed.acidity         0.189935   0.059439   3.195 0.001396 **
volatile.acidity     -4.017665   0.561672  -7.153 8.49e-13 ***
citric.acid          -2.147882   0.643763  -3.336 0.000849 ***
free.sulfur.dioxide   0.028120   0.010051   2.798 0.005145 **
total.sulfur.dioxide -0.017594   0.003568  -4.932 8.15e-07 ***
sulphates             2.443362   0.449083   5.441 5.30e-08 ***
alcohol               0.928974   0.084468  10.998  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1547.2  on 1118  degrees of freedom
Residual deviance: 1161.9  on 1111  degrees of freedom
AIC: 1177.9

Number of Fisher Scoring iterations: 4
```

Figure 4: Summary of model after stepwise variable selection

3.1 Cross validation and model diagnostics

Once we have built our model and identified the significant predictors we used the withheld 30% of the data to make predictions. We choose to use a 0.5 probability threshold to make predictions, since the two classes had approximately equal distribution. Then, we compared the predictions of our model versus the actual values (once transformed to a binary variable) on the quality of each sample of wine. The confusion matrix is given below.



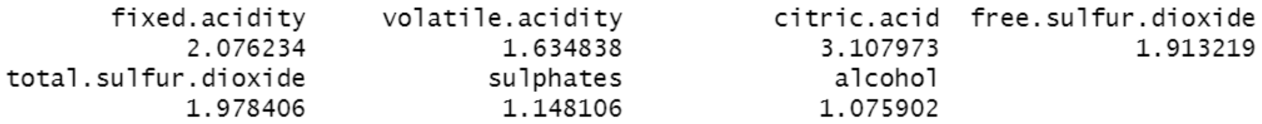
```
          actual
prediction   0   1 Sum
       0   166  67 233
       1    52 195 247
       Sum 218 262 480
```

Figure 5: Confusion matrix for predictions on test data

We can measure the predictive performance of the MLR model by calculating the accuracy. We found that a predictive accuracy of 0.752. In addition, the true negative rate or specificity was 0.761, the true postive rate or sensitivity was 0.744. For reference, the null model had an accuracy of 0.55, specificity of 0, and sensitivity of 1. Comparing our model with the null model we observed that not only was the accuracy higher, but the specificity and sensitivity were both indications that the model makes predictions equally well for both binary outputs.

One popularly used test for model performance is the AUC or the area under the ROC curve. Values of AUC close to 1 indicate better predictive performance of a model, while AUC closer to 0.5 indicates worse performance. For our reduced model the AUC was 0.818.

Back during the exploratory analysis stage we had some concerns regarding multicollinearity in our model. Using the vif() function from the *car* package in R, we can calculate the variable inflation factor (VIF) which measures collinearity amoung variables. A summary of the results is shown below.



```
      fixed.acidity    volatile.acidity         citric.acid free.sulfur.dioxide 
           2.076234            1.634838            3.107973            1.913219 
total.sulfur.dioxide           sulphates             alcohol 
           1.978406            1.148106            1.075902 
```

Figure 6: VIF values for predictor variables in the reduced model

The VIF values for the predictors indicate some level of correlation, with the highest being citric acid with a VIF of 3.10. For our purposes we could accept this level of correlation since the VIF values were within acceptable range (VIF less than 5).

We were interested in finding any outliers in our data so we checked the graph of standardized residuals vs. fitted. Often this is used to find points which have high absolute standard residuals since they may be outliers. The figure below shows the graph of standardized residuals vs. fitted values. Looking at the graph there is one point with a standardized residual greater than 3. This point varied greatly from what the model predicted so we may look deeper to see why this is so.

We can graph the estimated logit versus the value of the predictor variable for each of our predictor variables. This will show the probability changes for various values of each of our predictors. The figure below shows the graph of the estimated logit value of a specific point; the y-axis represents the value of the predictor variable of that specific point. A line of best fit was estimated using LOESS regression for each graph. On the left is the regression summary of the model.

We notice that the strongest associated variables with logit appear to be alcohol, volatile acidity, and sulphates. Samples with high alcohol content were associated with having higher odds of being highly rated, while volatile acidity was negatively associated with the same. For sulphates, there is a relatively weak but steady positive association. Using the varImp() function from the *caret* package in R we were able to see similar results with alcohol having the highest variable importance score of 11.00, followed by volatile acidity at 7.15, and sulphates with 5.44. A summary of the variable importance results are

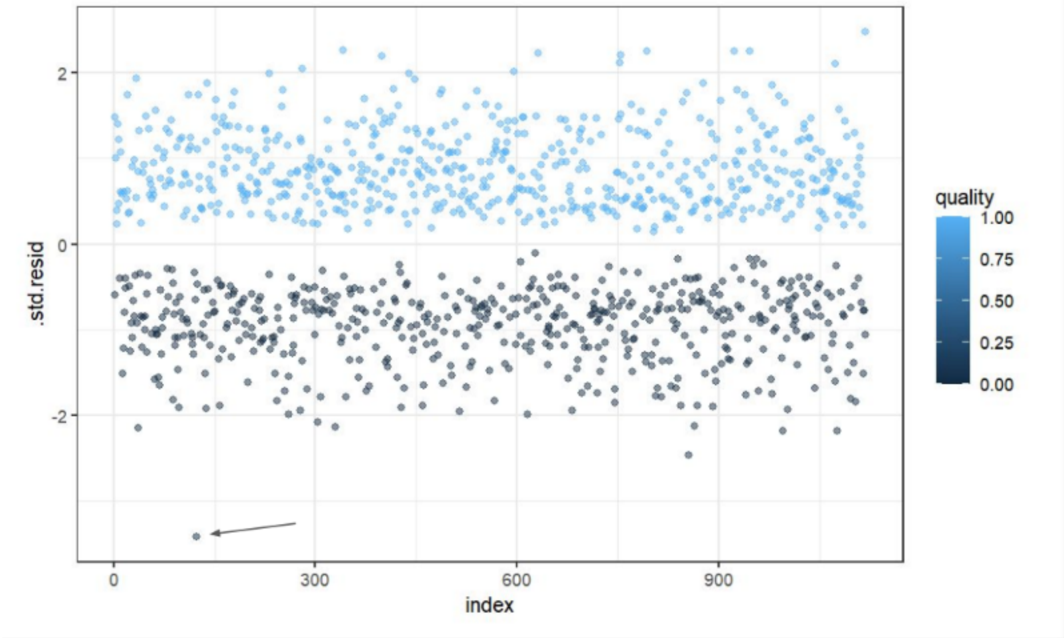
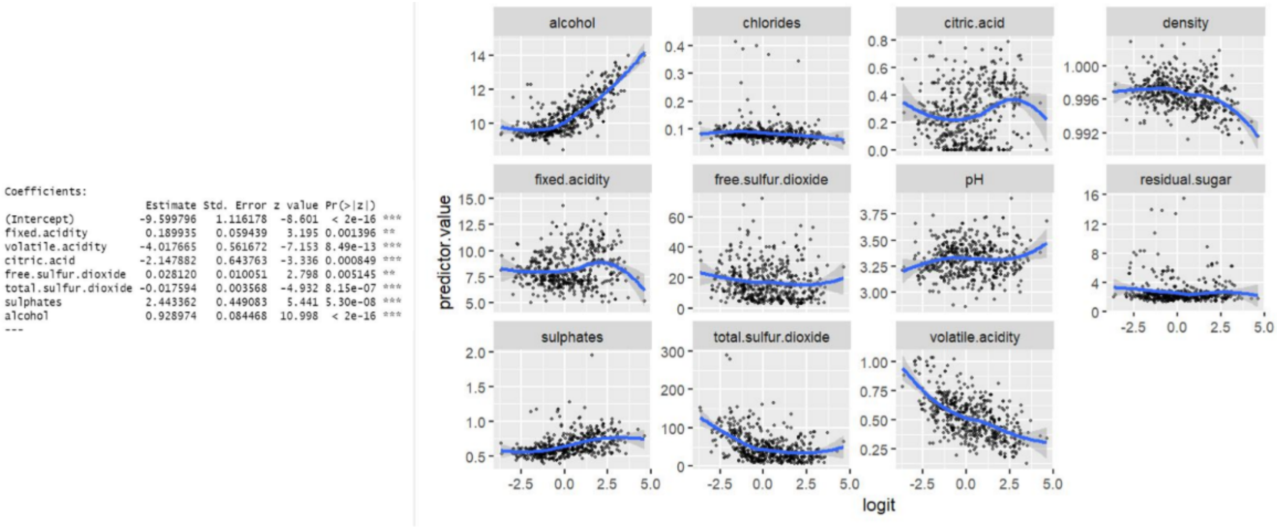


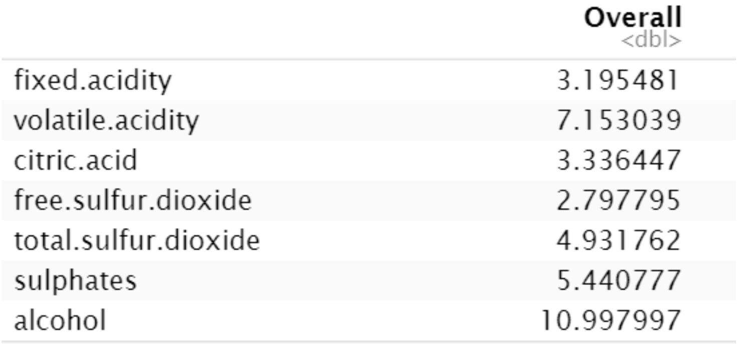
Figure 7: Standardized residuals vs. Fitted values



```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -9.599796   1.116178  -8.601  < 2e-16 ***
fixed.acidity        0.189935   0.059439   3.195 0.001396 ** 
volatile.acidity    -4.017665   0.561672  -7.153 8.49e-13 ***
citric.acid         -2.147882   0.643763  -3.336 0.000849 ***
free.sulfur.dioxide  0.028120   0.010051   2.798 0.005145 ** 
total.sulfur.dioxide -0.017594   0.003568  -4.932 8.15e-07 ***
sulphates            2.443362   0.449083   5.441 5.30e-08 ***
alcohol              0.928974   0.084468  10.998  < 2e-16 ***
---
```

Figure 8: Graph of estimated probabilities vs. predictor value

shown below.



	Overall <dbl>
fixed.acidity	3.195481
volatile.acidity	7.153039
citric.acid	3.336447
free.sulfur.dioxide	2.797795
total.sulfur.dioxide	4.931762
sulphates	5.440777
alcohol	10.997997

Figure 9: Summary of variable importance for predictor variables

4 Conclusion

We were interested in building a model which can classify wines based on experts' ratings of them. To this end, we built a logisitc regression model using all possible predictors and then performed stepwise variable selection to finally reduce the model to 7 predictors. These predictors were fixed acidity, volatile acidity, citric acid, free sulfur dioxide, total sulfur dioxide, sulphates, and alcohol. Furthermore given a new sample, we can estimate the log-odds or logit of the probability from the following equation.

$$logit(\hat{p}) = -9.60 + 0.19(fixed.acidity) - 4.02(volatile.acidity) - 2.15(citric.acid) + 0.028(free.sulfur.dioxide) - 0.018(total.sulfur.dioxide) + 2.44(sulphates) + 0.93(alcohol)$$

To see how well our model performed we built our model using a training set and made predictions on the withheld test data. We found our model did a decent job at correctly predicting high and low ratings with an accuracy of 0.752. The model correctly classified around 3/4 of the time and this was consistent whether it was for low or high scoring wines. We also saw that the three most important predictors of the model to be alcohol, volatile acidity, and sulphates in determining class prediction. While the other predictors were significant, their effect was less deciding than the three just mentioned.

We may be interested in finding ways to increase the predictors that positively contribute to higher ratings or minimizing those that negative affect good ratings. For example, the sugar content of grapes during their harvest affects the alcohol content of a wine. Grapes grown during warmer conditions tend to contain more sugar and thus will produce more alcohol upon fermentation. Volatile acidity does not affect wine quality at lower quantities but higher levels can give wine a sharp, vinegary taste. Sulphates minimize oxidation in wine and help to maintain freshness. Levels of sulphates are kept at a legally safe threshold but some wines contain more or less. All of these are considerations which are strictly monitored during the wine making process to ensure quality control.

While we were satisfied with our model, there are improvements that can be made. The accuracy was high enough to be useful but still had a large number of misclassifications. For our model we considered recommended wines as those with ratings greater than 5. In the future we may choose to use greater than 6 as being “recommended” and thus build recommendations with a lower minimum score. Another consideration would be to transform predictors to lessen the effect of colinearity. We may also compare different models to see how different methods perform.

Using our model as a framework we have shown that there is a relationship between the physiochemical properties of wine and how we experience it as taste. By analyzing a large dataset we can extract useful information and apply the results to our decision making. Consumer information greatly lacks compared to the producers’ so having a model capable of giving recommendations is an idea which others may want to consider for future work.

5 Code Appendix

The supplementary code can be found at:

https://github.com/Leonysus/STAT632_Spring2022_GroupProject/blob/main/STAT632_winequality_prediction.rmd