What Makes a Good Glass of Wine?

STA 210 - Summer 2022

Team 2 - Alicia Gong, Ashley Chen, Abdel Shehata, Claire Tan

6-16-2022

Introduction and Data

Introduction In 2020 alone, about 234 million hectoliters of wine were consumed worldwide, with the US making up approximately 14% of that consumption (Karlsson 2020). Both wine consumption and wine production are essential parts of many cultures and economies around the world. For some communities, wine production is their primary economic means to survival, and some wine regions are even recognized by UNESCO as World Hertiage sites. Thus, it is important to be able to assess the quality of wine based on its physicochemical components. Since wine composition and wine quality varies widely, it raises the question: what makes a good wine? More specifically, what chemical compositions of wine affect the variation of wine quality? To answer this question, we will analyze the wine quality dataset from Vinho Verde vineyard in Portugal, and more importantly try to narrow down our question to make it possible for it to be supported by evidence.

"Vinho Verde" wine is exclusively produced in the demarcated region of Vinho Verde in northwestern Portugal, which is one of the largest and oldest wine regions in the world and home to thousands of producers, generating a wealth of economic activity and jobs and strongly contributing to the development of the Vinho province and the country as a whole. This wine has its own unique system of production, as it is only produced from the indigenous grape varieties of the region. Additionally, the wine produced in Vinho Verde also enjoys high reputation worldwide, and is recurrently awarded in national and international competitions.

We are curious about what exactly makes a good glass of Vinho Verde wine, from a scientific standpoint. Wine-drinkers commonly relate the quality of a wine to its aroma or sweetness, but these standards are subjective to the person drinking the wine. Because the chemical factors of wine are what contribute to these varying characteristics, we predict that chemical components which contribute to wine's aroma, sweetness, ripeness and freshness will be influential to wine's quality measure points.

By investigating and analyzing the chemical factors attributable to the final rating of Vinho Verde wine, our research may shed light on future research and development directions for improving the quality of Vinho Verde wine, which may also contribute to the competitiveness of Portuguese wine industry. Thus, the overall goal of this project is to produce a classification model that best explains how different chemical compositions of the Portuguese “Vinho Verde” wine affects the variation of the wine quality.

Data Introduction The Wine Quality dataset was collected from Vinho Verde wine Samples, from northern Portugal. The data was originally donated in 2009 by Professor Cortez. The specific mechanism of the collection of the data was lab work done on different wines to measure their chemical attributes (such as acidity). The quality of the wine was obtained through the average rating of three wine experts. The dataset is divided into two: Red wine and White wine. Red Wine has 1599 observations, and white wine has 4898 observations (each observation being a specific wine). Key variables include:

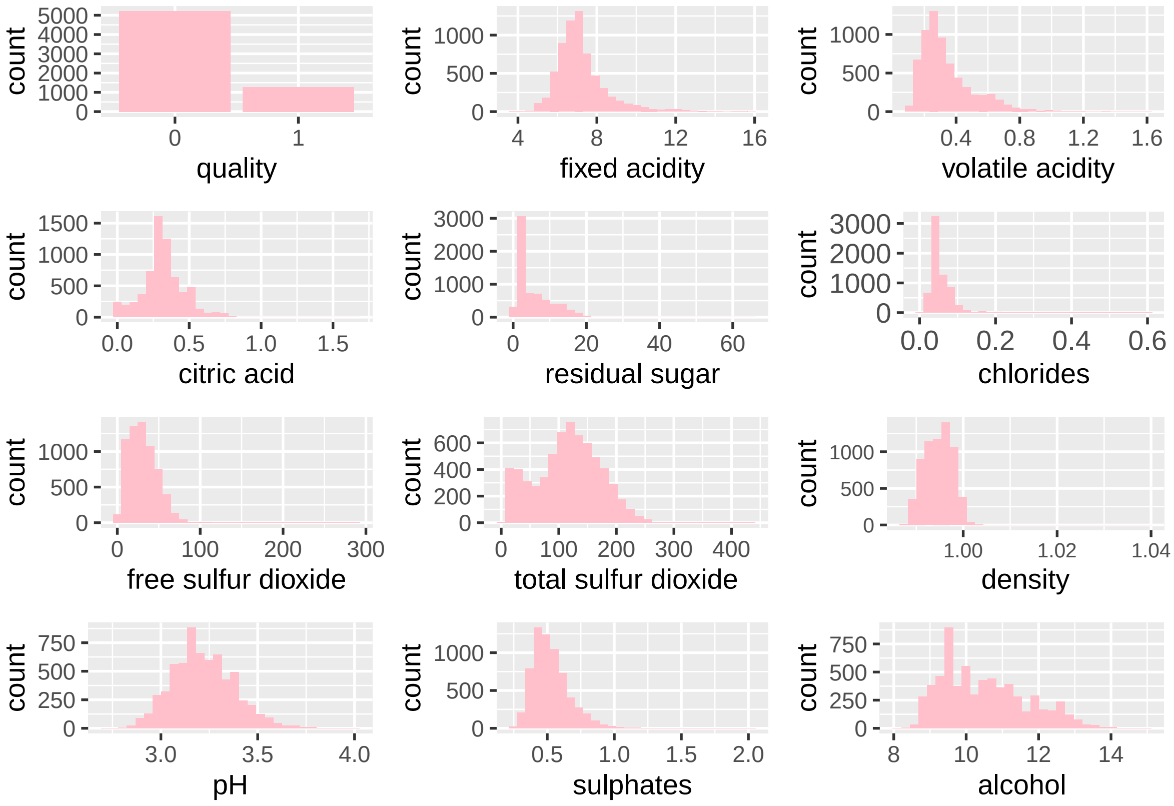
- `quality`: wine quality, rated on a scale from 0-9.
- `volatile.acidity`: amount of gaseous acid in wine
- `chlorides`: amount of salt in wine
- `total.sulfur.dioxide`: amount of free and bound forms of SO_2 in wine
- `pH`: acidity of wine
- `sulphates`: amount of sulphates in wine
- `alcohol`: percent alcohol content of the wine

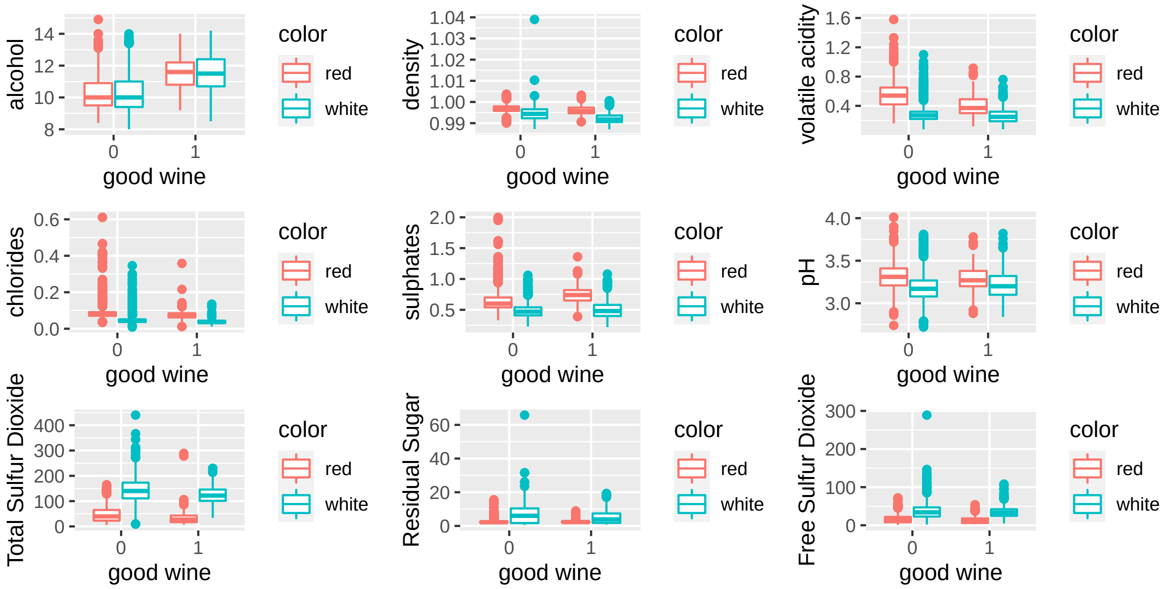
In our analysis, we split `quality` into “good wine” and “bad wine”, defining “good wine” as wine that receives a quality score higher than 7. We based this on the system of the wine magazine ‘Wine Spectator’, in which a score of at least ‘80’ (on a 100-point scale) is defined as ‘Good’. In our data, the highest score a wine can get is 9, and 0.8 * 9 = 7.2, so we chose 7 as our cut-off point, and created a new variable `good_wine` that is 1 for good wine, and 0 otherwise,

Exploratory Data Analysis We first visualized the distributions of potential predictor variables through histograms. Then, we visualized the relationships between key predictor variables and the response variable `good_wine` through boxplots, and also investigated possible interaction effects between these predictor variables and the color of the wine.

From this, we can see that most variables are normally distributed, but residual sugar, chlorides, and sulphates are slightly skewed to the right, while free sulfur dioxide and alcohol have a strongly right-skewed distribution. Additionally, the distribution of good and bad wine is quite unbalanced, as the number of observations categorized as not good greatly outnumbers the amount categorized as good. From the boxplot, we can see that for almost all variables expect alcohol, there exists a significant difference between red and white wine. Therefore, we will consider adding the interaction between color and other predictor variables in our model.

```
Rows: 5
```


count
5000
4000
3000
2000
1000
0
0
1
quality
count
1000
500
0
4
8
12
16
fixed acidity
count
1000
500
0
0.4
0.8
1.2
1.6
volatile acidity
count
1500
1000
500
0
0.0
0.5
1.0
1.5
citric acid
count
3000
2000
1000
0
0
20
40
60
residual sugar
count
3000
2000
1000
0
0.0
0.2
0.4
0.6
chlorides
count
1000
500
0
0
100
200
300
free sulfur dioxide
count
600
400
200
0
0
100
200
300
400
total sulfur dioxide
count
1000
500
0
1.00
1.02
1.04
density
count
750
500
250
0
3.0
3.5
4.0
pH
count
1000
500
0
0.5
1.0
1.5
2.0
sulphates
count
750
500
250
0
8
10
12
14
alcohol


alcohol
14
12
10
8
0
1
good wine
color
red
white
density
1.04
1.03
1.02
1.01
1.00
0.99
0
1
good wine
color
red
white
volatile acidity
1.6
1.2
0.8
0.4
0
1
good wine
color
red
white
chlorides
0.6
0.4
0.2
0.0
0
1
good wine
color
red
white
sulphates
2.0
1.5
1.0
0.5
0
1
good wine
color
red
white
pH
4.0
3.5
3.0
0
1
good wine
color
red
white
Total Sulfur Dioxide
400
300
200
100
0
0
1
good wine
color
red
white
Residual Sugar
60
40
20
0
0
1
good wine
color
red
white
Free Sulfur Dioxide
300
200
100
0
0
1
good wine
color
red
white

```
Columns: 2
$ Variables      <chr> "fixed acidity & citric acid", "fixed acidity & density~
$ R_Coefficients <dbl> 0.3244357, 0.4589100, -0.3779813, 0.5525170, 0.7209341
```

Lastly, we calculated the correlation coefficients for some of our predictor variables to investigate multicollinearity, which is the ocurrence of high intercorrelations among two or more independent variables in a multiple regression model. This is a problem because it undermines the statistical significance of the predictor variables and makes it difficult to establish relationships. From our calculations, we can see that the correlation between free.sulfur.dioxide and total.sulfur.dioxide is 0.72, indicating a strong positive linear relationship, so multicollinearity exists between these two variables.

Methodology

Best AIC Model To select the most optimal model to fit our data, we decided to conduct a step-wise AIC model. We can operationalise this as the model with the lowest AIC value to supplement which variables to add to or omit from to choose the model with the smallest amount of error or lowest residual sum of squares. AIC is an estimator of in-sample prediction error and a lower AIC values can indicate a more parsimonious model.

In this step-wise AIC test, an algorithm uses specific procedures in which the AIC values of different models are calculated to determine what covariates are added to or removed from the model, and this process is repeated several times in both directions. In our EDA, we found multicollinearity in the strong positive linear relationship between free.sulfur.dioxide and total.sulfur.dioxide. We also know from the data description that total.sulfur dioxide includes free.sulfur.dioxide - whose concentration level could impact the smell and taste of the wine, thus impacting quality. So it is sufficient to include only one of these two variables in the model, and we choose free.sulfur.dioxide. After conducting the step-wise AIC test, we found that the combination of variables with the lowest AIC values were fixed acidity, volatile acidity, residual sugar, chlorides, free sulfur dioxide, density, pH, and sulphates.

We split the data into 25% testing set and 75% training set. The training set is used to develop models, and the test set is used for estimating a final, unbiased assessment of the model's performance.

Model output

term	estimate	std.error	statistic	p.value
(Intercept)	401.932	74.428	5.400	0.000
c_fixed.acidity	0.486	0.076	6.356	0.000
c_volatile.acidity	-4.081	0.449	-9.088	0.000
c_citric.acid	-0.379	0.406	-0.935	0.350

term	estimate	std.error	statistic	p.value
c_residual.sugar	0.213	0.030	7.142	0.000
c_chlorides	-7.295	2.766	-2.638	0.008
c_free.sulfur.dioxide	0.007	0.003	2.579	0.010
c_density	-421.804	75.433	-5.592	0.000
c_pH	2.273	0.423	5.376	0.000
c_sulphates	2.528	0.325	7.787	0.000
c_alcohol	0.476	0.094	5.082	0.000
colorwhite	-1.198	0.244	-4.902	0.000

In this model, red wine is picked as the baseline. For example, the odds of being classified as 'good wine' for white wine are expected to be exp(-0.826) = 0.30 times the odds for red wine, holding all else constant.

After conducting the step- wise AIC test, the following variables are selected:

- `fixed.acidity`: the amount of acid in wine that's not volatile (do not evaporate fast)
- `volatile.acidity`: the amount of acetic acid in wine
- `citric.acid`: found in small quantities and can add freshness and flavor to wines
- `residual.sugar`: amount of sugar left after fermentation
- `chlorides`: amount of salt in wine
- `free.sulfur.dioxide`: free amount of SO_2 exists in equilibrium between molecular SO_2 and bisulfite ion
- `density`: density of wine measured in g/ml
- `pH`: acidity of wine and hydrogen ion concentration
- `sulphates`: produced by yeast, protecting wine against oxidation
- `alcohol`: percent alcohol content of the wine

Model Evaluation

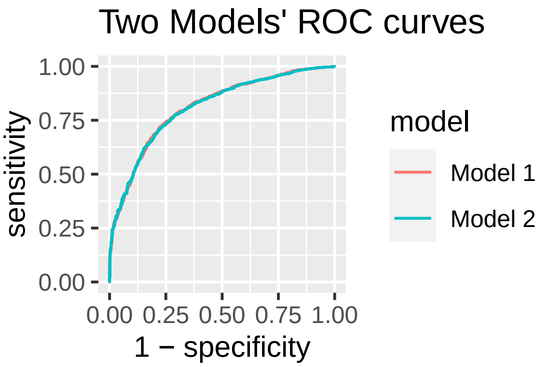
Full Model Then, considering that we have two types of wine–red wine and white wine, and we have seen potential interactive effects between the color of wine and the level of chemical components from the EDA, we decided to add interactive terms between color and other predictor variables. We compare the best AIC model with another logistic model that includes interactive terms. We will call the best AIC model the reduced model and the logistic model with interactive terms the full model.

term	estimate	std.error	statistic	p.value
(Intercept)	426.077	74.360	5.730	0.000
c_fixed.acidity	0.574	0.082	6.992	0.000
c_volatile.acidity	-3.002	0.684	-4.388	0.000
c_residual.sugar	0.222	0.030	7.344	0.000
c_chlorides	-8.897	2.811	-3.165	0.002
c_free.sulfur.dioxide	-0.017	0.010	-1.735	0.083
c_density	-447.221	75.380	-5.933	0.000
c_pH	2.303	0.423	5.439	0.000
c_sulphates	3.395	0.546	6.212	0.000
c_alcohol	0.436	0.092	4.731	0.000
color_white	1.277	0.856	1.492	0.136
c_fixed.acidity_x_color_white	-0.229	0.074	-3.085	0.002
c_volatile.acidity_x_color_white	-1.216	0.884	-1.376	0.169
c_free.sulfur.dioxide_x_color_white	0.026	0.010	2.546	0.011
c_sulphates_x_color_white	-1.310	0.657	-1.992	0.046

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 426.077 + 0.574 \times fixed\ acidity - 3.002 \times volatile\ acidity + 0.222 \times residual\ sugar - 8.897 \times chlorides + 0.017 \times free.sulfur.dioxide - 447.221 \times density + 2.303 \times pH + 3.395 \times sulphates + 0.436 \times alcohol + 1.277 \times colorwhite - 0.229 \times fixed\ acidity : colorwhite - 1.216 \times volatile\ acidity : colorwhite + 0.026 \times free\ sulfur\ dioxide : colorwhite - 1.310 \times sulphates : colorwhite$$

The interactive terms between predictor variables and colorwhite are used to assess the interactive effects between other predictor variables and wine's color. Red wine is used as baseline here. From the EDA, we can see that there is no significant difference between the alcohol concentration between red and white wine, so we did not consider the interactive effects between alcohol and colorwhite. We also removed citric acid, the interactive effect term of density:color_white, pH:color_white, residual.sugar:color_white, chlorides:color_white due to their high p-values.

Model Selection One measure to compare the two classification models above is to use the area under the ROC curve to assess their performance and determine which model performed more optimally. We fit both the full model with interactive terms and the initial Best AIC model without interactive terms to the training data.



The two ROC curves are very close, suggesting that the two models' performances are similar.

We also calculated the area under the ROC curve (AUC) to summarize the performance of each classifier into one single measure. The AUC value tells how well the model is capable of distinguishing between two binary classes. The higher the AUC, the better the model is predicting good wine quality and good wine quality and bad wine quality as bad wine quality.

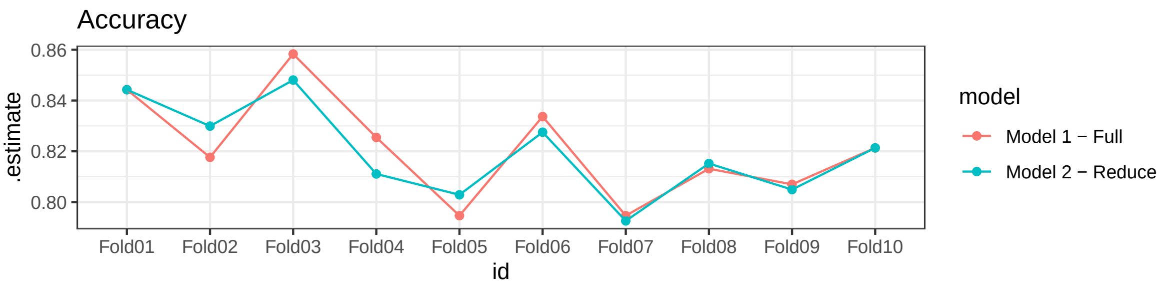
We found that the AUC value of the logistic model was 0.815 and the AUC value of the best AIC model was 0.811, suggesting that the two models have similar performances.

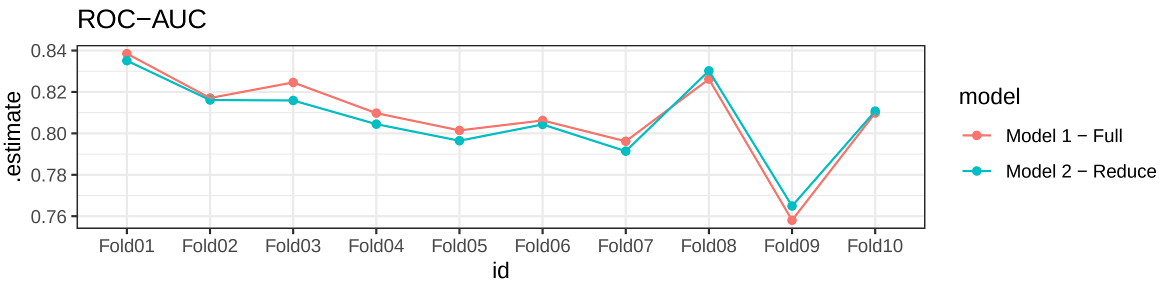
model	.metric	.estimator	.estimate
Logistic Model	roc_auc	binary	0.812
Best AIC Model	roc_auc	binary	0.811

After deciding that we are going to stick with the best AIC model, we decide to remove citric acid, due to its high p-value, to increase our model's accuracy.

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 407.674 + 0.472 \times fixed\ acidity - 3.397 \times volatile\ acidity + 0.214 \times residual\ sugar - 7.505 \times chlorides + 0.007 \times free\ sulfur\ dioxide - 427.514 \times density + 2.289 \times pH + 2.515 \times sulphates + 0.462 \times alcohol - 1.228 \times colorwhite$$

Cross Validation We finally conducted cross validation, a technique for assessing how the statistical analysis generalizes to the data set. It evaluates the regression model by training several models on subsets (folds) of the data set (we previously split into training and testing sets) and evaluating them on the complementary subset of the data. To perform the cross validation, we split the training data into 10 folds. After performing the cross validation, we graphed metrics of accuracy and ROC-AUC values of each individual fold for the two models on a corresponding graph.





The two models have similar accuracy and the ROC-AUC values for the 10 folds, so due to the principles of parsimony (i.e. model that provides simpler explanation and has fewer terms), we prefer the more reduced one or the best AIC model fit. Thus, the best AIC model is our final model.

Limitations When conducting a logistic regression, it is important to check the conditions are satisfied. When evaluating if the log-odds have a linear relationship with the predictors, we find that the conditions for linearity are not satisfied for all the variables. The linearity conditions are particularly not fulfilled for the variables free sulfur dioxide (the spread takes on a curved shape), total sulfur dioxide (the spread has similarly curved shape), and sulphates (the data points appear to be randomly dispersed).

The independence conditions, however, are satisfied. The Vinho Verde region is a vast region spreading 15500 hectares of vineyards in far-north Portugal allowing the observations to be collected independently from each other and the accumulated data are not collected across an extended period of time.

Results

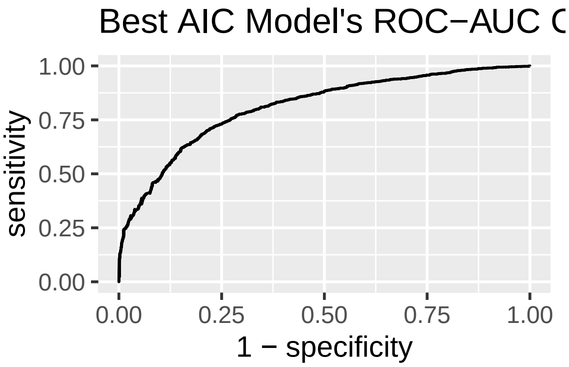
Our final model is the reduced Best AIC model. To achieve this conclusion, we first conducted a pairwise test to obtain a Best AIC Model with no interactive terms, which is our reduced model.

Then we include the interactive terms and build another logistic model, as our full model. We compared these two models' ROC curves, AIC, BIC. We also check the two models' accuracy and ROC-AUC through cross-validation. We found out that the two models have similar performances when predicting the quality of wine. Due to the principles of parsimonious, we prefer the simpler model. Therefore, we choose the reduced Best AIC Model as our best model.

$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = 407.674 + 0.472 \times fixed\ acidity - 3.397 \times volatile\ acidity + 0.214 \times residual\ sugar - 7.505 \times chlorides + 0.007 \times free\ sulfur\ dioxide - 427.514 \times density + 2.289 \times pH + 2.515 \times sulphates + 0.462 \times alcohol - 1.228 \times colorwhite$

From this model, we can see that some influential factors might be chlorides, pH, sulphates, and volatile acidity. For example, the coefficient for pH is 2.289, which means that for every 1-point increase in pH, we expect the odds of wine being good to increase by a factor of exp(2.289) = 9.87, holding all else constant.

To assess the model's predictive abilities, we fit it to test data. The following ROC curve plot shows that the Best AIC Model does fairly well in terms of predicting good wine vs bad wine.



Discussion & Conclusion

From our analysis, we found that the variables that contribute significantly to wine quality were fixed acidity, volatile acidity, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, and sulfates. Variables with biggest coefficients are chlorides, volatile acidity, pH and sulphates. It is worth noting that although the coefficient of density is very high, wine's density varies between approximate 0.99 to 1, therefore it is impossible for wine' density to jump 1 unit. So we did not include density as a influential factor.

Some of the variables that negatively affected wine quality were chlorides and volatile acidity (meaning when these variables increased, quality was less likely to be good). Scientific reasons for this could be that chlorides are a major contributor to saltiness, and volatile acidity refers to the amount of acetic acid in wine, which is a kind of acid which makes wine tastes like vinegar. Some variables that positively affected wine quality were pH and sulphates. This could be because wine with low pH tastes tart and crisp, while higher pH adds to wine's ripeness. Additionally, sulphates are a food preservative widely used in winemaking which protects wine against oxidation and maintains the flavor and freshness of wine.

The result shows that, under wine experts' standards, wine that tastes more salty and vinegary earns lower grades; on the other hand, wine taste riper, mellower, fresher receive higher grades. This results is coherent to our prediction. In the future, to improve the accuracy of the model, different stepwise algorithms (p-values for example) can be evaluated and adjusted. There are many different methods such as through feature engineering, viewing potential interaction terms, or other performance measurements such as machine learning algorithms to better predict our results. Furthermore, a limitation of our dataset is the unbalanced data, in which bad wine greatly outnumbers good wine. Thus, another potential step in future work could be assessing models that oversample good wine or downsample bad wine in order to address class unbalance, and compare these models to our Best AIC model.

Analytical methods in the past century have permitted researchers to gain detailed knowledge about the sugars and acids present in the wine-making operation. Analysis and control of oxidation through measuring both oxygen concentrations and sulfur dioxide concentrations in grape juice and wine have resulted in significant improvements in wine quality. Our results encourage further development in research to measuring sulfur dioxide and the pH of grapes.

Quality in any wine is a function of the potential quality in the grape and the conversion of grapes to wine. From our analysis, acidity is particularly important for clean, fruity, balanced tastes. Low acidity is also undesirable because it often leads to problems with fermentation and preservation of the wine which manifest themselves as undesirable tastes in the final wine.

Moreover analytical methods for sulfur dioxide determination can be improved by research for wineyards such as techniques for measuring phenols and enzymes or more precise methods of measuring acids of grapes and wines. Sulfur dioxide controls oxidation in wines, but too much gives the wine a very unpleasant flavor. Techniques for measuring phenols and enzymes are still being developed in enology laboratories, as are more precise methods of measuring residual sugars and the acids of grapes and wines.

Many analytical methods for sulfur dioxide determination and other factors in wine quality are quite outdated but our research can be used as an impetus to fund and significantly improve research in the specific characteristics of wine quality.

References

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal, 2009.