Predicting Wine Quality based on Physicochemical Properties

Phillip Efthimion

Wine is a popular alcoholic beverage widely enjoyed by many. In recent years, there has been an explosion in the number of wineries throughout the world as more wine is being consumed yearly. Creating wine is a tightrope act of mixing ingredients to find a delicious formula, but what factors are most important for stimulating the users taste buds?

The data consists of 4,898 white wines sampled. The researchers collected the physicochemical and sensory properties through a computerized system (Cortez). The system controls the making of the wine. The samples in their data set were recorded from May 2004 until February 2007. There is no mention to how the seasons could have affected the ingredients of the wine. The wine is all from the same winery, Vinho Verde, in Portugal.

The objective of this analysis is to perceive any relationship between each of out explanatory variables and the response variable quality in order to see what of the physicochemical properties are important in the winemaking process.

The response variable, quality, was evaluated by 3 wine experts who rated each of the wines from a scale of 1 to 10 with a 10 rating being reserved for only the most excellent wine and 1 being a poor quality wine. The quality scores for each of the 3 experts were averaged together to give the quality rating. There is no information on each of the judge's individual ratings. The ability to be an expert on wine is a contentious subject. The are some that believe it is a "junk science" because very few judges are able to be consistent with there rankings (Hodgson 2012), yet it is still a widely respected field.

The first predictor refers to the fixed acidity of a wine, which for the referenced study was tartaric acid in grams per density meter cubed (g/dm^3) maintains the chemical stability of the wine. It is most significantly found in grapes. Tartaric acid also is largely responsible for the flavor of the wine (Boulton 1980).

The second predictor is volatile acidity, which here is acetic acid in (g/dm^3). It is responsible for any sour taste in a wine and is created during fermentation (Boulton 1980).

The third predictor is citric acid. The concentration of citric acid is commonly less than tartaric acid in wines and is mostly an additive (Boulton 1980).

The fourth predictor is residual sugar. Residual sugar is leftover from the fermentation process by design because it is responsible for the sweetness of a wine. In winemaking, sugar is responsible for breaking down yeast into alcohol (Robinson 1994).

The fifth predictor refers to chlorides. They are responsible for the salinity of a wine. They are a determinate of species of grape used. Chlorides also affect the fermentation process. Some countries put an artificial constraint on the concentration of chlorides in a wine, though that is not the case for the data set (Logothetis 2010).

The sixth and seventh predictors refer to free sulfur dioxide and total sulfur dioxide. Sulfur emits a categorically unpleasant smell so it is not used in large amounts. Sulfur dioxide acts as an anti-oxidant for the wine. In the wine making process though, only a portion of the sulfur dioxide will do this (robinson 1994). This part of the sulfur dioxide is referred to as the free sulfur dioxide. The rest of it is chemically bound into the wine. The total sulfur dioxide is the sum of the free sulfur dioxide and the bound sulfur dioxide. Sulfur dioxide serves 2 purposes in winemaking: being an anti-oxidant and preventing bacterial growth.

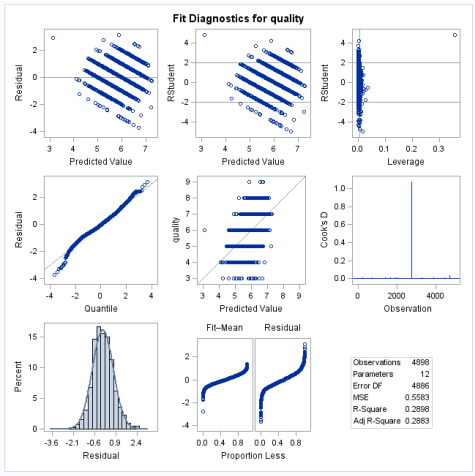
The eighth predictor consists of the density of the wine measured in g/dm^3. It is the density measurement of the liquid.

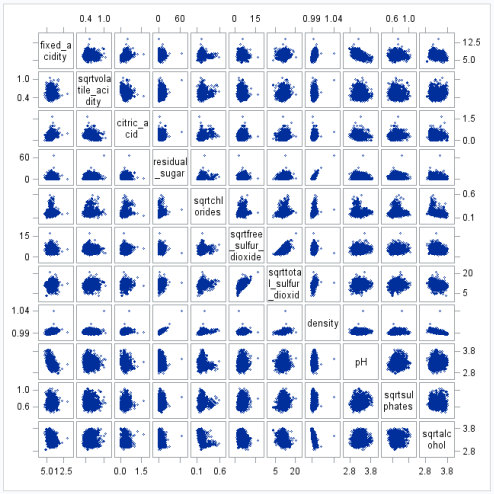
The ninth predictor refers to pH, a scale of how acidic or basic a solution is. It is measured on a scale from 0 to 14 with 0 being the most acidic solution and 14 being the most basic. Water has a pH of 7.

The tenth predictor of the data is sulphates, measure in g/dm^3. The data refers specifically to potassium sulphate. Sulphates are a binder for many bacterias, acids, and sugars (Robinson 1994).

The eleventh and final predictor is alcohol, by percentage of volume.

We will first created histograms of all eleven of the predictor variables in order to identify any outliers. Not all of the predictor variables appear to be normally distributed so they will be square root transformed. We are going to take the square root transformation of the predictor variables: volatile acidity, free sulfur dioxide, total sulfur dioxide, sulphates, and alcohol. With this transformation, each of the variables histograms will be better distributed. However, for chlorides we will be taking the square root transformation.





We start with our initial multiple regression model, as shown below. We can see that this model's R^2 and Adj. R^2 models are quite low at 0.2903 and 0.2887 respectively. Our first method of fitting will be the backward elimination method of model selection. We will decide which factor to remove by looking at the p-values of each variable. We will also be taking into account the variance inflation factor of each variable as it shows collinearity. We will begin with removing the predictive variable total sulfur dioxide, which we have taken the square root of.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	1115.00838	101.36440	181.68	<.0001
Error	4886	2725.98141	0.55792		
Lack of Fit	3949	2725.98141	0.69030	Infty	<.0001
Pure Error	937	0	0		
Corrected Total	4897	3840.98979			

Root MSE	0.74694	R-Square	0.2903
Dependent Mean	5.87791	Adj R-Sq	0.2887
Coeff Var	12.70755		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	149.52984	18.85591	7.93	<.0001	0
fixed_acidity	1	0.06820	0.02067	3.30	0.0010	2.67053
sqrtvolatile_acidity	1	-2.07595	0.12633	-16.43	<.0001	1.13208
citric_acid	1	0.01220	0.09453	0.13	0.8973	1.14872
residual_sugar	1	0.08011	0.00743	10.79	<.0001	12.45416
invchlorides	1	0.00349	0.00168	2.07	0.0383	1.48205
sqrtfree_sulfur_dioxide	1	0.07170	0.01017	7.05	<.0001	1.86905
sqrttotal_sulfur_dioxide	1	-0.00626	0.00887	-0.71	0.4799	2.33242
density	1	-151.56426	18.90805	-8.02	<.0001	28.07114
pH	1	0.65769	0.10425	6.31	<.0001	2.17526
sqrtsulphates	1	0.89717	0.14418	6.22	<.0001	1.14203
sqrtalcohol	1	1.22194	0.15758	7.75	<.0001	7.68916

Removing the predictive variable total sulfur dioxide incrementally improved the adjusted R squared of our model from 0.2887 to 0.2888. It makes sense that this variable could potentially be removed as it is the free sulfur dioxide not the total that is responsible for anti-oxidants. The next variable that will be removed is citric acid, which has a p-value of 0.9188, well above our alpha =0.5. Doing so only increases the adjusted R square value incrementally again to 0.2889. There is also an issue with multicollinearity. Though all of our p-values are below our alpha threshold of 0.5, we still have two high variance inflation factors.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	1114.72404	123.85823	222.07	<.0001
Error	4888	2726.26575	0.55775		
Lack of Fit	3945	2726.26575	0.69107	Infty	<.0001
Pure Error	943	0	0		
Corrected Total	4897	3840.98979			

Root MSE	0.74682	R-Square	0.2902
Dependent Mean	5.87791	Adj R-Sq	0.2889
Coeff Var	12.70561		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	151.84447	18.46933	8.22	<.0001	0
fixed_acidity	1	0.06894	0.02052	3.36	0.0008	2.63322
sqrtvolatile_acidity	1	-2.09744	0.12183	-17.22	<.0001	1.05331
residual_sugar	1	0.08076	0.00735	10.99	<.0001	12.19908
invchlorides	1	0.00360	0.00167	2.15	0.0317	1.46644
sqrtfree_sulfur_dioxide	1	0.06738	0.00803	8.39	<.0001	1.16535
density	1	-153.91908	18.50992	-8.32	<.0001	26.90965
pH	1	0.65782	0.10375	6.34	<.0001	2.15485
sqrtsulphates	1	0.89021	0.14362	6.20	<.0001	1.13352
sqrtalcohol	1	1.21763	0.15673	7.77	<.0001	7.60943

The next multiple linear regression model that was tried was done using forward selection for variable selection. One predictive variable was added at a time until adding variables no longer increased the R squared value. Our final model with this method is shown

below. It has slightly lower R squared and adjusted R squared values, however, there are no longer any signs of correlation.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1077.43451	134.67931	238.26	<.0001
Error	4889	2763.55528	0.56526		
Corrected Total	4897	3840.98979			

Root MSE	0.75184	R-Square	0.2805
Dependent Mean	5.87791	Adj R-Sq	0.2793
Coeff Var	12.79090		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-1.12740	0.29864	-3.78	0.0002	0
sqrtalcohol	1	2.33024	0.07723	30.17	<.0001	1.82322
fixed_acidity	1	-0.06035	0.01299	-4.65	<.0001	1.04099
sqrtvolatile_acidity	1	-2.16834	0.12498	-17.35	<.0001	1.09358
residual_sugar	1	0.02339	0.00250	9.34	<.0001	1.39669
invchlorides	1	0.00577	0.00167	3.46	0.0005	1.43743
sqrtfree_sulfur_dioxide	1	0.08420	0.01007	8.36	<.0001	1.80988
sqrtsulphates	1	0.62411	0.13838	4.51	<.0001	1.03828
sqrttotal_sulfur_dioxide	1	-0.01789	0.00868	-2.06	0.0394	2.20647

Our model from the forwards selection method looks to be more accurate as there are no high variance inflation factors suggesting collinearity.

The next selection methods used were the LASSO and LAR method. Neither of them had a much different effect. One thing that could be learned to better improve the model is by looking at the lack of fit test. As it has a p-value of <0.0001, which is less than our alpha 0.05, this model can be improved by adding higher power variables to the equation. After adding squared terms to the model, there was not a change in the R squared. This is something that would warrant further exploratory analysis in a further study.

Therefore, our final model is quality = -1.127 + 2.33 (sqrt(alcohol)) - 0.06 (fixed acidity) -2.168 (sqrt(volatile acidity)) + 0.023 (residual sugar) + 0.00577 (1/chlorides) + 0.084 (sqrt(free sulfur dioxide) + 0.624 (sqrt(sulphates)) - 0.018 (sqrt(total sulfur dioxide))

The final model tells us that the amount of fixed acidity, volatile acidity, and total sulfur dioxide have negative effects on the quality of the wine. The volatile acid is responsible for any sour taste in the wine, therefore our model is telling us that as the wine became more sour, it's quality level decreased. The strongest indicator for higher quality of wine is the variable alcohol,

by percent of volume. The more alcoholic the wine was, the higher the quality rating by the 3 judges as it's 1.22 coefficient is the highest of all of the coefficients in the model.

In further analysis, I would look into higher order terms to better fit the model. Though the final R squared and adjusted R squared terms are low, I believe this could be due to factors outside of the variables in this study, such as the ingredients used in the wines themselves, like grapes. Also, the model may not be accurate to the quality of wine because as cited in, Hodgson's 'An Examination of Judge Reliability at a major U.S. Wine Competition', wine quality experts are not always consistent, if not poor, judges of wine. To go with that, people have different preferences to what makes a wine 'good quality'. Further information that could improve this study to further reduce bias are more judges and wine from more than a single winery.

Bibliography

Boulton, R. 1980. The Relationships between Total Acidity, Titratable Acidity and pH in Wine. American Journal of Enology and Viticulture 31:76-80.

Cortez, Paulo, Antonio Cerdeira, Fernando Almeida, Telmo Matos, and Jose Reis. “Modeling Wine Preferences by Data Mining from Physicochemical Properties.” *Decision Support Systems* 47.4 (2009): 547-53. Web.

Hodgson, R.T. (2008) ‘An Examination of Judge Reliability at a major U.S. Wine Competition’, Journal of Wine Economics, 3(2), pp. 105–113. doi: 10.1017/S1931436100001152.

Logothetis, Stelios, and Graeme Walker. “Influence of Sodium Chloride on Wine Yeast Fermentation Performance.” *International Journal of Wine Research* (2010):35. Web.

Robinson, Jancis. *The Oxford Companion to Wine*. Oxford: Oxford UP, 1994. Web.

Appendix

SAS Code

```
data Wine;
infile "\\Client\C$\Users\Phillip\Downloads\WhiteWineData.csv" dlm=","
firstobs = 2;
input fixed_acidity volatile_acidity citric_acid residual_sugar chlorides
free_sulfur_dioxide total_sulfur_dioxide density pH sulphates alcohol
quality;
run;
*proc univariate data = Wine;
*     histogram;
*run;
data Wine_log;
set Wine;
logvolatile_acidity = log(volatile_acidity+1);
logcitric_acid = log(citric_acid+1);
logresidual_sugar = log(residual_sugar+1);
logchlorides = log(chlorides+1);
logfree_sulfur_dioxide = log(free_sulfur_dioxide+1);
logtotal_sulfur_dioxide = log(total_sulfur_dioxide+1);
logsulphates = log(sulphates+1);
logalcohol = log(alcohol+1);
run;
proc reg data = Wine_log;
model quality = fixed_acidity logvolatile_acidity logcitric_acid
logresidual_sugar logchlorides logfree_sulfur_dioxide logtotal_sulfur_dioxide
density pH logsulphates logalcohol;
run;
proc univariate data = Wine_log;
      var fixed_acidity logvolatile_acidity logcitric_acid logresidual_sugar
logchlorides logfree_sulfur_dioxide logtotal_sulfur_dioxide density pH
logsulphates logalcohol;
      histogram;
run;
proc reg data=Wine plots(unpack);
model quality = fixed_acidity volatile_acidity citric_acid residual_sugar
chlorides free_sulfur_dioxide total_sulfur_dioxide density pH sulphates
alcohol/VIF;
run;

data Wine_inv;
set Wine;
invvolatile_acidity = 1/(volatile_acidity);
invcitric_acid = 1/(citric_acid);
invresidual_sugar = 1/(residual_sugar);
invchlorides = 1/(chlorides);
invfree_sulfur_dioxide = 1/(free_sulfur_dioxide);
invtotal_sulfur_dioxide = 1/(total_sulfur_dioxide);
invsulphates = 1/(sulphates);
invalcohol = 1/(alcohol);
run;
proc univariate data = Wine_inv;
      var fixed_acidity invvolatile_acidity invcitric_acid invresidual_sugar
invchlorides invfree_sulfur_dioxide invtotal_sulfur_dioxide density pH
invsulphates invalcohol;
      histogram;
```

```
run;
proc reg data = Wine_inv;
model quality = fixed_acidity invvolatile_acidity invcitric_acid
invresidual_sugar invchlorides invfree_sulfur_dioxide invtotal_sulfur_dioxide
density pH invsulphates invalcohol;
run;

data Wine_sqrt;
set Wine;
sqrtvolatile_acidity = sqrt(volatile_acidity);
sqrtchlorides = sqrt(chlorides);
sqrtfree_sulfur_dioxide = sqrt(free_sulfur_dioxide);
sqrttotal_sulfur_dioxide = sqrt(total_sulfur_dioxide);
sqrtsulphates = sqrt(sulphates);
sqrtalcohol = sqrt(alcohol);
run;
proc univariate data = Wine_sqrt;
      var fixed_acidity sqrtvolatile_acidity citric_acid residual_sugar
sqrtchlorides sqrtfree_sulfur_dioxide sqrttotal_sulfur_dioxide density pH
sqrtsulphates sqrtalcohol;
      histogram;
run;
proc sgscatter data = Wine_sqrt;
matrix fixed_acidity sqrtvolatile_acidity citric_acid residual_sugar
sqrtchlorides sqrtfree_sulfur_dioxide sqrttotal_sulfur_dioxide density pH
sqrtsulphates sqrtalcohol;
run;
proc reg data = Wine_sqrt;
model quality = fixed_acidity sqrtvolatile_acidity citric_acid residual_sugar
sqrtchlorides sqrtfree_sulfur_dioxide sqrttotal_sulfur_dioxide density pH
sqrtsulphates sqrtalcohol/lackfit VIF;
run;
proc reg data = Wine_sqrt;
model quality = fixed_acidity sqrtvolatile_acidity citric_acid residual_sugar
sqrtchlorides sqrtfree_sulfur_dioxide sqrttotal_sulfur_dioxide pH
sqrtsulphates sqrtalcohol/lackfit VIF;
run;
proc reg data = Wine_sqrt;
model quality = fixed_acidity sqrtvolatile_acidity citric_acid sqrtchlorides
sqrtfree_sulfur_dioxide sqrttotal_sulfur_dioxide density pH sqrtsulphates
sqrtalcohol/lackfit VIF;
run;
data Wine2;
set Wine;
sqrtvolatile_acidity = sqrt(volatile_acidity);
invchlorides = 1/(chlorides);
sqrtfree_sulfur_dioxide = sqrt(free_sulfur_dioxide);
sqrttotal_sulfur_dioxide = sqrt(total_sulfur_dioxide);
sqrtsulphates = sqrt(sulphates);
sqrtalcohol = sqrt(alcohol);
run;
proc reg data = Wine2;
model quality = fixed_acidity sqrtvolatile_acidity citric_acid residual_sugar
invchlorides sqrtfree_sulfur_dioxide sqrttotal_sulfur_dioxide density pH
sqrtsulphates sqrtalcohol/lackfit VIF;
run;
quit;
proc reg data = Wine2;
```

```
model quality = fixed_acidity sqrtvolatile_acidity citric_acid residual_sugar
invchlorides sqrtfree_sulfur_dioxide density pH sqrtsulphates sqrtalcohol/
lackfit VIF;
run;
quit;
proc reg data = Wine2;
model quality = fixed_acidity sqrtvolatile_acidity residual_sugar
invchlorides sqrtfree_sulfur_dioxide density pH sqrtsulphates sqrtalcohol/
lackfit VIF;
run;
quit;
proc reg data = Wine2;
model quality = fixed_acidity sqrtvolatile_acidity residual_sugar
invchlorides sqrtfree_sulfur_dioxide density pH sqrtsulphates sqrtalcohol/
lackfit VIF;
run;
quit;
proc glmselect data=Wine2 plots=all;
model quality = fixed_acidity sqrtvolatile_acidity citric_acid residual_sugar
invchlorides sqrtfree_sulfur_dioxide sqrttotal_sulfur_dioxide density pH
sqrtsulphates sqrtalcohol/selection=lar;
run;
quit;

proc reg data = Wine2;
model quality = fixed_acidity sqrtvolatile_acidity residual_sugar
invchlorides sqrtfree_sulfur_dioxide density pH sqrtsulphates sqrtalcohol/
lackfit VIF;
run;
proc reg data = Wine2;
model quality = fixed_acidity sqrtvolatile_acidity residual_sugar
invchlorides sqrtfree_sulfur_dioxide pH sqrtsulphates sqrtalcohol/lackfit
VIF;
run;
proc reg data = Wine2;
model quality = fixed_acidity sqrtvolatile_acidity residual_sugar
invchlorides sqrtfree_sulfur_dioxide sqrtsulphates sqrtalcohol/lackfit VIF;
run;

proc reg data = Wine2;
model quality = sqrtalcohol fixed_acidity sqrtvolatile_acidity residual_sugar
invchlorides sqrtfree_sulfur_dioxide sqrtsulphates sqrttotal_sulfur_dioxide/
VIF;
run;
proc glmselect data = Wine2;
model quality = sqrtalcohol fixed_acidity sqrtvolatile_acidity residual_sugar
invchlorides sqrtfree_sulfur_dioxide sqrtsulphates sqrttotal_sulfur_dioxide/
selection=lasso;
run;
quit;

/*
data Wine_sq;
set Wine;
sq_volatile_acidity = (volatile_acidity)**2;
sq_citric_acid = (citric_acid)**2;
sq_residual_sugar = (residual_sugar)**2;
sq_chlorides = (chlorides)**2;
```

```
sq_free_sulfur_dioxide = (free_sulfur_dioxide)**2;
sq_total_sulfur_dioxide = (total_sulfur_dioxide)**2;
sq_sulphates = (sulphates)**2;
sq_alcohol = (alcohol)**2;
run;
proc univariate data = Wine_sq;
      var fixed_acidity sq_volatile_acidity sq_citric_acid sq_residual_sugar
sq_chlorides sq_free_sulfur_dioxide sq_total_sulfur_dioxide density pH
sq_sulphates sq_alcohol;
      histogram;
run;
```