Red Wine Quality Evaluation

Qi Wang[1], Author Name[2]
Department of Statistics, University of California, Santa Cruz[1]
Second Author Affiliation[2]

Abstract

DDDDDDDDDDD

KEY WORDS: Logistic regression, Classification, Red wine quality

1. Introduction

1.1 Background Information

Wine, maybe the oldest drink that its secret recipes, has been passed to us through centuries by our ancient. Like everything else, wineries are looking to evolve the way they are making wine and apply technology and innovations to this most popular drink in the world. To study more about this subject, the wine data collect from the north-west region, named Minho, of Portugal, and this data set is available from the UCI machine learning repository (UCI, 2015). It has been proposed for both, regression and classification, by Cortez et al. (2009). Cortez et al proposed a data mining approach to predict human wine taste preferences. Three regression techniques were applied, under a computationally efficient procedure that performs simultaneous variable and model selection(Cortez et al. 2009). Such model is useful to support the oenologist wine tasting evaluations and improve wine production. Later, Agyemang presents an analysis to extend what Cortez et al accomplished by using two logistic regression approaches to predict human wine taste preferences with the goal of better predictions(Agyemang 2010). Nebot et al used hybrid fuzzy logic techniques to predict human wine test preferences based on physicochemical properties from wine analyses (Nebot, Mugica, and Escobet 2015). The fuzzy technique result presents a better performance rather than other data mining techniques previously applied to the same data set, such are neural networks, support vector machines and multiple regression. Recently, Angus try to find out if it is possible to predict what score a wine would be given based on its chemical properties and wine testers' opinion on the wine quality (Angus, n.d.). The result opens up the possibility of assigning wine score without the use of wine testers.

1.2 Data Source and Description

In this paper, we are going to talk about how to make an evaluation of red wine based on several indexes in the data set. We are using R to make some descriptive statistics and analyzing with logistic regression methods to evaluate the most important index affecting the equality of wine. The data we used is from UCI machine learning repository, and originally from Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez.

There are 11 covariates and one categorical response variable. Fix acidity is the most acids involved with wine or fixed or nonvolatile (do not evaporate readily). Volatile acidity is the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste. Citric acid is the most important organic acid produced in tonnage by fermentation, with a taste of sour like lemons (Soccol et al. 2006). Residual sugar is the amount of sugar left in a wine, to some extent, it measures the sweetness of a wine. Chlorides is a key role in the salty taste of a wine, which will make customers feel uncomfortable. Sulfur dioxide (SO_2) is important in the winemaking process as it aids in preventing microbial growth and the oxidation of wine (Monro et al. 2012). The difference between free sulfur dioxide and total sulfur dioxide is the way of measuring them. Gaseous SO2 is released from the sample by addition of acid and swept into the ICP by an argon stream. The intensity of the sulfur atomic emission lines is measured in the vacuum UV region. Determination of total SO_2 is performed after hydrolysis of bound forms with sodium hydroxide ($NaOH$)(Čmelík et al. 2005). For sulfates, many experts believe that higher sulfurous content causes a duller taste in wine, and that high potency of sulfite ions presents a health risk and speeds up the wine's fermentation process. The other covariates including alcohol, pH and density are basically simple indexes of a red wine. Our response variable is an ordered categorical variable indicating the quality of red wine, from 0 to 10.

2. Data Cleaning and Variable Properties

2.1 Variable Transformation

First, renaming the variables into shorter words for further convenience to analysis by selecting several letters

Table 1: Name Transformation

Original	Transformed
fixed.acidity	fac
volatile.acidity	vac
citric.acid	cac
residual.sugar	res
chlorides	cho
free.sulfur.dioxide	fsu
total.sulfur.dioxide	tsu
density	den
pH	pH
sulphates	sul
alcohol	alc
quality	Q

from each word in the phrases. In table 1, there are two columns which indicates the original variable names and its transformed names.

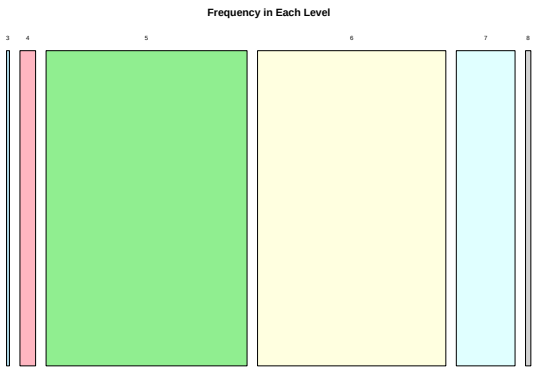

Figure 1: Mosaic Plot of Frequency

Then, the most important step is to find a proper method to combine 10 groups into 2 groups.Since the quality is ordinal, and the frequency of data in each group is as follows in figure 1 and table 2. From the table and the plot, most of the wines fall in quality level 5 and quality level 6. We will do three models to make better interpretation of the important parameters that affect the level of the wine. If we divide the data from the middle, that will be helpful for interpret the common criteria for red wine. However, if someone wants to distinguish those red wines with very perfect quality or very poor quality, I will also divide data in other ways, which I will talk more about it later. But the basic models we are talking about is the first case, in which I make the number of red wine in each group almost the same. So it is a good way to divide them into 2 groups by this criteria:

$$Y_{i,new} = 0, \ if \ \ Y_{i,data} \leq k$$

Table 2: Frequency in Each Level

Level	Count
3	5
4	35
5	513
6	464
7	170
8	13

and

$$Y_{i,new} = 1, \ if \ \ Y_{i,data} > k$$

In this way, we transformed the data from six levels into two levels so that we can make further logistic regression. Here, by setting different k values, we get different ways to separate the data, and we have three kinds of dividing data set. The first one is just set $k = 5$, which means we separate them in the middle and we want to know how to overall evaluate the quality of the red wine. The other ways are setting $k = 4$ or $k = 6$, which means we want to specify poor quality wines and excellent quality wines. For EDA parts we are using the case that $k = 5$ to make some basic conclusions since the length of the report is restricted, but later in the model part, all the three ways of division are included. And we are using first 1200 observations to fit the model and the rest 316 observations to make predictions.

2.2 Variable Properties Exploration

From the definition of independent variables mentioned above, there could be some inner relationship among them for example the different kinds of acidity and different kinds of sulfur dioxide. To begin with, here is a correlated plot and pairs plot.

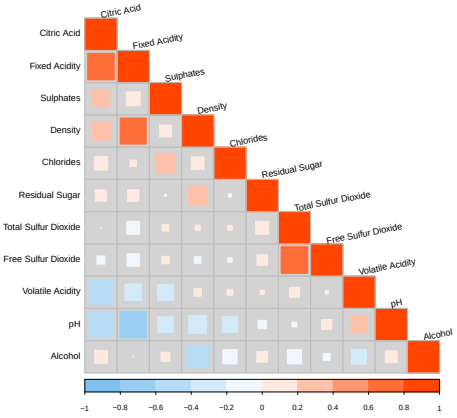

Figure 2: Correlation Plot of Indepedent Variables

From the chart 2, we can see there are some strong inner relationship between the independent variables. To be-

gin with, citric acid amount is strongly positively correlated with fixed acidity. And fixed acidity is also strongly positively correlated with the density of the wine. Researches have shown that the citric acid has an effect on the acidity of the liquid, and a more concentration of citric acid means a stronger acidity (Lustig et al. 2017). Also, the amount of free sulfur dioxide is positively correlated with the amount of total sulfur dioxide. However, there are still some negative correlations among variables. For example, the pH of the red wine is negatively correlated with the fixed acidity and the amount of citric acid in the wine. From common sense, we know a stronger acidity means a lower pH, that's why the pH is lower for those wine with more concentration of fixed acidity and citric acid. Also, larger concentration of alcohol gives a smaller density. As we know, the density of alcohol is smaller than water, so if more alcohol is included in the wine, the density must be lower than those without that much alcohol. Therefore, when we are doing the exploratory data analysis, we can care just several specific dependent variables. Citric acid, sulphates, density, residual sugar, total sulfur dioxide, and alcohol. Later, after constructing a base model, further model selection will be conducted by adding or subtracting variables to or from the base model.

3. Exploratory Data Analysis

3.1 One variable boxplot

In this subsection, we are going to show some basic one variable box plot for the selected variables I mentioned above. Then the relationship between the quality of red wine and the selected variables will be more obvious. From figure 3, not all the variables seem to have a very obvious difference between them. For example, the density, residual sugar, and total sulfur dioxide seem almost the same for the wine with high quality and poor quality. However, the citric acid, sulphates and alcohol concentration seems more significant. Those red wine with higher quality seem to have greater concentration on citric acid, sulphates and alcohol. However, there seems to be several outliers for total sulfur dioxide in the graph, and for residual sugar, the data set seems so sparse that the height of each box is too small, which makes it hard to observe the further information. In all, the difference between the groups is not so significant, but we can specify some information from the box plots.

3.2 t-test for Significance

From section 3.1, we have seen some differences of indexes between different wine quality groups. However, we cannot have a statistical significance measurement of each variable. Here, we are carrying out a t test for checking the significance. Here we only compare one variable for one time and don't consider the colinearity between variables.

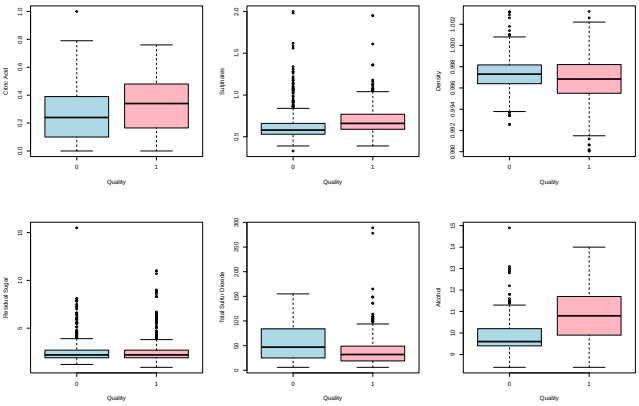

Figure 3: Boxplot of Variables

Table 3: t Test among Different Groups

	.95 CI[L]	.95 CI[U]	p-value
Citric Acid	0.046	0.090	0.00
Sulphates	0.047	0.087	0.00
Density	-0.001	0.000	0.00
Residual Sugar	-0.133	0.154	0.89
Total Sulfur Dioxide	-21.810	-14.193	0.00
Alcohol	0.847	1.063	0.00

From the table 3, we can see that citric acid, sulphates total sulfur dioxide, density and alcohol are significantly different among the two groups. However, the residual sugar seems not that significant. In other words, better red wines seem to have more citric acid, alcohol and sulphates concentration. Also, since the density of alcohol is less than water, the density is smaller according to our guess, which has been verified here. Another result worth discussing is the total sulfur dioxide, it is a harmful content for us, which seems to be less in red wine with better qualities. However, sulphates are also harmful to human's health, why does higher quality red wines have more sulphates? As we can see from the table, the difference is not so much, although it is significant. As the saying goes, it is impossible to discuss the poison without considering the dosage. We will discuss more in the model selection and regression part.

3.3 Outliers Detecting

Then, I will check whether there is outliers among these variables, I will use the strip chart to show an overview of the variable distribution.

After checking the data for these outliers in figure 4, since

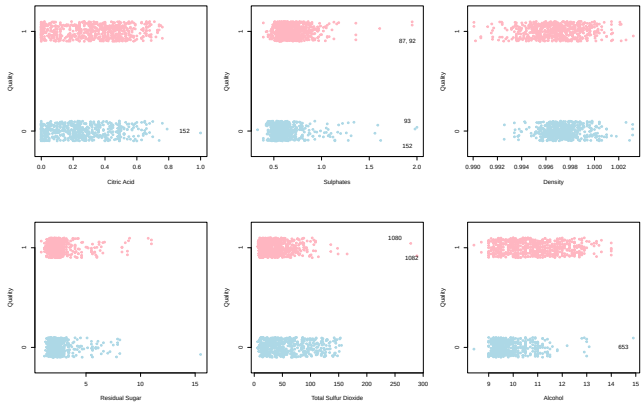

Figure 4: Check for Outlier

the measurement method is correct, there is no reason to delete these data. Those values are higher than the others but they are still possible values and is not caused by the measurement error. But we need to focus more on the data point 152 since it is both outliers for citric acid and sulphates.

4. Modeling

4.1 Modeling with Equal Points in Each Group

4.1.1 Base Model

I will use all the variables mentioned above in the data set as a base model. And here I am going to use this base model to check whether there are significant relationships between the red wine quality and them. Since the response variable is a categorical variable with two levels, I will use logistic regression to fit for the model. The model can be expressed as:

$$Logit(P_i) = \beta_0 + \beta_{fac}X_{i,fac} + \beta_{vac}X_{i,vac} + \beta_{cac}X_{i,cac}$$
$$+\beta_{cho}X_{i,cho} + \beta_{sul}X_{i,sul} + \beta_{den}X_{i,den} + \beta_{res}X_{i,res}$$
$$+\beta_{tsu}X_{i,tsu} + \beta_{fsu}X_{i,fsu} + \beta_{alc}X_{i,alc} + \beta_{pH}X_{i,pH}$$

There seems to be significant relationship between the volatile acid, chlorides, free sulfur dioxide, citric acid, sulphates, total sulfur dioxide and alcohol according to table 4. The indexes that positively affect the quality of red wine is sulphates, alcohol and free sulfur dioxide. And citric acid, total sulfur dioxide, chlorides and volatile acidity will have negative effects on the red wine quality.Some of our previous guess can be verified that more sulfur dioxide released from the wine, the better will the wine be. And chlorides and volatile acidity can lower the quality by the taste according to our materials. However, there are still some counter-intuitive results since citric acid is beneficial to our body but sulphates

Table 4: Coefficients for Base Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	84.293	93.418	0.902	0.367
fac	0.191	0.118	1.615	0.106
vac	-3.656	0.575	-6.361	0.000
cac	-1.681	0.670	-2.509	0.012
res	0.040	0.071	0.567	0.570
cho	-3.379	1.767	-1.913	0.056
fsu	0.025	0.010	2.436	0.015
tsu	-0.019	0.003	-5.465	0.000
den	-92.533	95.425	-0.970	0.332
pH	-0.205	0.835	-0.246	0.806
sul	2.359	0.484	4.872	0.000
alc	0.848	0.119	7.149	0.000

Table 5: Coefficients for Selected Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.720	1.068	-8.169	0.000
fac	0.142	0.061	2.336	0.019
vac	-3.821	0.566	-6.748	0.000
cho	-3.247	1.695	-1.916	0.055
fsu	0.024	0.010	2.366	0.018
tsu	-0.019	0.003	-5.496	0.000
cac	-1.685	0.670	-2.513	0.012
alc	0.916	0.085	10.774	0.000
sul	2.286	0.470	4.865	0.000

are harmful for us. Overall, this is just a base model and for the case that the data is divided in the middle, which means both two groups have almost the same quantities of data points. After the model selection, I will use other ways to divide data and detect which criteria is important to classify red wines.

4.1.2 Model Selection

A nice criteria for model selection is AIC for logistic regression. Similar to linear regression, we are also using add one or drop one variable continuously to get the final model. And the remained variables are volatile acidity, alcohol, sulphates and chlorides. So the final model in this case should be:

$$Logit(P_i) = \beta_0 + \beta_{fac}X_{i,fac} + \beta_{vac}X_{i,vac} + \beta_{cac}X_{i,cac}$$
$$+\beta_{cho}X_{i,cho} + \beta_{sul}X_{i,sul}$$
$$+\beta_{tsu}X_{i,tsu} + \beta_{fsu}X_{i,fsu} + \beta_{alc}X_{i,alc}$$

And the regression estimated coefficients can be derived in table 5.We can see that fixed acidity, alcohol, sul-

Table 6: Coefficients for Lower Biased Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-468.620	217.213	-2.157	0.031
fac	-0.779	0.252	-3.090	0.002
vac	-4.751	0.835	-5.688	0.000
res	-0.220	0.186	-1.183	0.237
cho	-8.263	2.825	-2.925	0.003
tsu	0.010	0.006	1.588	0.112
den	498.262	221.403	2.250	0.024
pH	-6.688	2.003	-3.339	0.001
alc	0.764	0.291	2.628	0.009

Table 7: Coefficients for Upper Biased Model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	304.942	101.696	2.999	0.003
fac	0.323	0.092	3.525	0.000
vac	-2.280	0.721	-3.164	0.002
res	0.316	0.081	3.914	0.000
cho	-7.031	3.376	-2.083	0.037
tsu	-0.013	0.004	-3.833	0.000
den	-319.862	102.215	-3.129	0.002
sul	3.329	0.561	5.938	0.000
alc	0.768	0.121	6.341	0.000

phates and free sulfur dioxide are still positively affecting the quality of the red wine. And the volatile acidity, concentration of citric acidity, chlorides and total sulfur dioxide are still having negative effects on the quality of wine. And here is the case for making a fair division, which means that the weight of the two groups (poor and good quality) are the same. Cases can also be that some people want to distinguish only the wine with very poor quality between others, or distinguish only the top wines. So here, I am going to use different methods to divide the data then rebuild the model.

4.2 Modeling with Biased Preference

4.2.1 Distinguish Poorly Qualified Red Wine

In this section, we prefer to distinguish the very poorly qualified red wine from others. Therefore, I reset the levels of quality as follows:

$$Y_{i,new} = 0,\ if\ \ Y_{i,data} \leq 4$$

and

$$Y_{i,new} = 1,\ if\ \ Y_{i,data} > 4$$

The base model is still the one we mentioned above, and to save space, we will only report the parameters of the selected model and make some inference. The selected model is:

$$Logit(P_i) = \beta_0 + \beta_{fac}X_{i,fac} + \beta_{vac}X_{i,vac}$$
$$+\beta_{cho}X_{i,cho} + \beta_{den}X_{i,den} + \beta_{res}X_{i,res}$$
$$+\beta_{tsu}X_{i,tsu} + \beta_{alc}X_{i,alc} + \beta_{pH}X_{i,pH}$$

To our surprise, the criteria changed a lot according to the table 6. First, there are new variables after the AIC selection and some important variables in the first case has been dropped. Fixed acidity, volatile acidity, residuals, chlorides and pH are negatively affecting the quality of the red wine. And alcohol, density and total sulfur dioxide are positively affecting it. So for those people who wants to distinguish only the poor quality with others, they should care more on the variables included in this model.

4.2.2 Distinguish Perfectly Qualified Red Wine

In this case, we divide the data into different groups by the following criteria:

$$Y_{i,new} = 0,\ if\ \ Y_{i,data} \leq 6$$

and

$$Y_{i,new} = 1,\ if\ \ Y_{i,data} > 6$$

Then, the selected model is:

$$Logit(P_i) = \beta_0 + \beta_{fac}X_{i,fac} + \beta_{vac}X_{i,vac}$$
$$+\beta_{cho}X_{i,cho} + \beta_{den}X_{i,den} + \beta_{res}X_{i,res}$$
$$+\beta_{tsu}X_{i,tsu} + \beta_{alc}X_{i,alc} + \beta_{pH}X_{i,sul}$$

Here, we can see that the variables in table 7 remained are similar to the last case. However, pH never makes significant effect on the quality. Fixed acidity's coefficient became to positive from negative, sulphates is added into the model and it positively affects the quality of the red wine. Volatile acidity, chlorides still negatively affect the qualification. Concentration of alcohol is still positively affecting the quality.

5. Conclusion

5.1 Overall Conclusion

From the coefficient value table above for all the three of them, we can find that there are some covariates that seems to be "robust" no matter which data they are using. Those variables are volatile acidity, chlorides, alcohol. These variables are significantly affecting all the

three cases and the absolute value of estimated parameter is large compared with others, which means these indexes have a strong effect on the quality of the red wine. Volatile acidity, as mentioned in the introduction part, is negatively affecting the quality of the red wine according to the result of the logistic regression, which verified our guess that volatile acidity affects the taste of the red wine. Chlorides, similarly, is negatively affecting the quality of the red wine as we imagined that it will give a salty taste to red wine which makes people feel unpleasant. Alcohol is positively affecting the quality of the red wine, which means that the red wine with greater concentration of alcohol, the larger the probability of the red wine is qualified "Good."

5.2 Case Specified Conclusion

There are more variables need to be cared about in certain cases. When we are distinguishing the "very poor" or "very good" quality red wine, density is also an important criteria. For the case that we distinguish the "very poor" quality red wine, the density is positively affecting the red wine quality, which means that wine with greater density does not tend to be classified in to "very poor" quality. However, when we want to distinguish "very good" quality red wine, the estimate of the coefficient for density became negative, which means that wines with too large density do not tend to be classified into "very good" group. The other similar variable is residual sugar, this index is not significant if we just want to make a "fair" division, but significant when we want to distinguish the very poor quality and very good quality red wine. Similarly, the estimate of the coefficients have different signs, which means it positively affect the quality when we distinguish the very poor quality red wine, but negatively affect the quality when we distinguish the very good ones. The other result needs to be specified is that pH is only significant when we distinguish the very poor quality, and greater pH (meaning that the red wine is more alkaline) will cause the red wine more tended to be classified as "very poor." Furthermore, sulphates is an important index when we distinguish the "very good" quality red wine and it positively affects the quality of the red wine, which is surprising after knowing that the sulphates is harmful to human's health.

6. Residual Analysis and Prediction Accuracy

6.1 Residual Analysis

Residual Analysis for three data set.

6.2 Prediction

Prediction using test data and make ROC curve and calculate prediction accuracy.

References

Agyemang, Perpetual. 2010. "Modeling the Preference of Wine Quality Using Logistic Regression Techniques Based on Physicochemical Properties." PhD thesis.

Angus, Dale. n.d. "Modeling Wine Quality from Physicochemical Properties." *Red* 895 (384): 320.

Čmelík, Jiří, Jiří Machát, Eva Niedobová, Vítězslav Otruba, and Viktor Kanickỳ. 2005. "Determination of Free and Total Sulfur Dioxide in Wine Samples by Vapour-Generation Inductively Coupled Plasma–Optical-Emission Spectrometry." *Analytical and Bioanalytical Chemistry* 383 (3): 483–88.

Cortez, Paulo, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. "Modeling Wine Preferences by Data Mining from Physicochemical Properties." *Decision Support Systems* 47 (4): 547–53.

Lustig, William P, Soumya Mukherjee, Nathan D Rudd, Aamod V Desai, Jing Li, and Sujit K Ghosh. 2017. "Metal–Organic Frameworks: Functional Luminescent and Photonic Materials for Sensing Applications." *Chemical Society Reviews* 46 (11): 3242–85.

Monro, Tanya M, Rachel L Moore, Mai-Chi Nguyen, Heike Ebendorff-Heidepriem, George K Skouroumounis, Gordon M Elsey, and Dennis K Taylor. 2012. "Sensing Free Sulfur Dioxide in Wine." *Sensors* 12 (8): 10759–73.

Nebot, Àngela, Francisco Mugica, and Antoni Escobet. 2015. "Modeling Wine Preferences from Physicochemical Properties Using Fuzzy Techniques." In *SIMULTECH*, 501–7.

Soccol, Carlos R, Luciana PS Vandenberghe, Cristine Rodrigues, and Ashok Pandey. 2006. "New Perspectives for Citric Acid Production and Application." *Food Technology & Biotechnology* 44 (2).