DOI: 10.54254/2754-1169/6/20220140

Development and Comparison of Regression Models Predicting Wine Quality using Excel - From a Profit Maximization Perspective

Wanling Xie[1,a,*], Shining Hu[2,b], Jincheng Ji[3,c] and Qianqi Chen[4,d]

[1] Economics and Management School, Wuhan University, Wuhan, 430071, China
[2] Ningbo University of Technology, Ningbo, 315211, China
[3] College of Liberal Arts, University of Minnesota twin cities, Minneapolis,55455, United States
[4] School of Science, University of Rensselaer Polytechnic Institute, Troy, 12180, US
a. wanling.xie@foxmail.com, b. 240179107@qq.com,
c. jjcray1218@gmail.com, d. chenq11@rpi.edu
**corresponding author email: wanling.xie@foxmail.com*

Abstract: Wine reviewers and the general public usually determine the quality of the wine. There is a great deal of subjec-tivity in such reviews and differences in the reviewers' preferences, making it difficult for the winery to obtain favorable information. This paper will use Excel to build a related model of chemical substances in wine and wine quality. Both linear and logistic regression is used in this article to predict wine quality. In addition, differ-ent from past literature that indicates the quality of red wine based on chemical substances, this paper creatively constructs a profit model based on predicting wine quality. It thus helps wine sellers to make model selections. All these can help producers understand how to make good wine and get higher profits.

Keywords: linear regression; logistic regression; wine quality; profit maximization.

1. Introduction

Nowadays, the wine market is large and in high demand worldwide. However, a better price does not always represent better quality. On the one hand, higher prices may be considered better quality. In Tier's experiment [1], given two identical wines, one with a tag of $5 and the other with a title of 45$, the drinker will say that the wine with a label of $45 tastes better than the other one. On the other hand, better quality wine always comes with a higher price. Salespeople earn more by selling better quality wine and thus want to improve the quality of wine and make more profit. This paper uses linear and logistic regression to determine which chemical component better influences the wine quality. By changing certain chemical elements, the Salespeople can maximize the profit by maximizing wine quality.

The paper is organized as follows: Section 2 provides literature reviews from the past article. Section 3 listed the datasheet which is used during the research. Section 4 explains the method which is used during the study. The result and discussion are shown in Section 5. The conclusion is displaced in Section 6.

2. Literature Review

Past literature has already focused on the prediction of wine quality. Some made analyses based on external factors of quality: Ira Horowitz et al. [2] set price range, winery rating, vintage, size, and winery and region as independent variables to predict wine quality. Antonio Capurso [3] predicted wine quality by checking the wine's balance, intensity, clarity, complexity, and finish length. Kwak, Young-Sik [4]. used wine CI(Collective Intelligence) to predict wine quality, statistically showing the same association degree to price as wine guru Robert Parker's score. Others predicted wine quality based on internal factors, mainly chemical characteristics. Badole and Mayur [5] used machine learning and correlation to find the bonding and relationship between variables and quality. Yogesh Gupta [6] also used machine learning, specifically neural networks and support vector machines, to predict wine quality after using linear regression results to choose important variables. Niggl, Dennis [7] did a small amount of data cleaning, used the vitalization method to find the relationship between independent and dependent variables, and then used a random forest model to predict the quality of the wine. Tingwei, Zhou [8] used active and semi-supervised machine learning to predict wine quality through query strategy. Shaw, B., and Suman [9] compare different classification algorithms for wine quality analysis to know which algorithms give a more accurate result. Kothawade [10] used machine learning with three algorithms (SVM, NB, ANN) to identify wine quality with certain features. While some previous literature used the same chemical dataset as this paper to predict wine quality, this paper uses linear regression and logistic regression compared to the machine learning method that previous literature used. Further, while previous literature mainly focused on quality prediction, this paper also tries to maximize profit by maximizing quality. The previous literature was on the producer side of the market, while this paper is on the seller side.

3. Dataset

These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The total data are 1599 rows and 12 columns. The independent variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfate, alcohol, and quality.

4. Methods

4.1. Model 1: Linear Regression

To obtain the linear relationship between the chemical composition and the quality of wine, a linear equation with the quality as the dependent variable should be established from the data. A model with k variables should be expressed as (1)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k+1} x_{k+1} \tag{1}$$

These data need to be preprocessed to make the prediction more reasonable before implementing the linear regression model.

Visualization.

To see how the values of each variable are distributed, this paper first plots the original twelve variables, including histogram and box plot, by the R language. The details are shown in Figure 1.

These graphs show many problems with these data, such as they contain many outliers and the distribution of some variables is not normally distributed. The result of the experiment was that the distribution of alcohol, chlorides, citric acid, fixed acidity, free sulfur dioxide, residual sugar, sulfates, total sulfur dioxide, and volatile acidity variables has different degrees of right-skew.

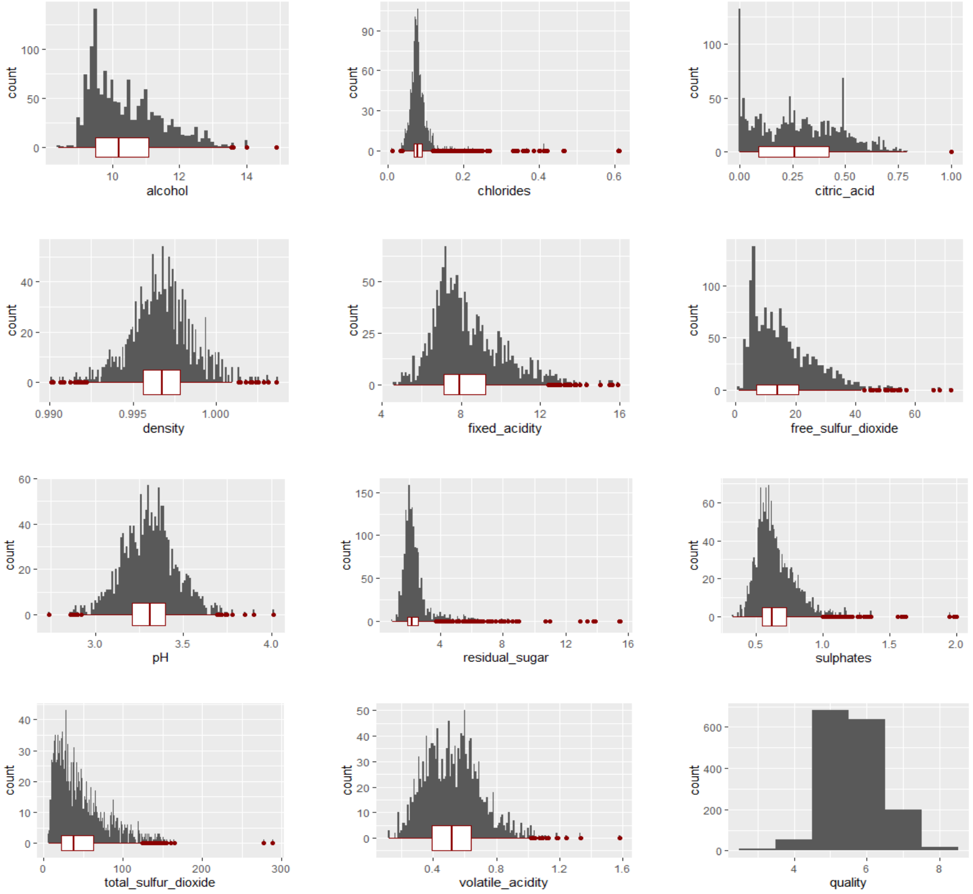

Figure 1: variables visualization.

Descriptive Statistics.

Descriptive statistics of twelve variables were obtained with the help of Excel. The results of the descriptive statistics are shown in *Table 1*.

Table 1: the results of the descriptive statistics.

	Mean	**Standard deviation**	**Median**	**Minimum**	**Maximum**
fixed acidity	8.32	0.04	7.90	4.60	15.90
volatile acidity	0.53	0.00	0.52	0.12	1.58
critic acid	0.27	0.00	0.26	0.00	1.00
residual sugar	2.54	0.04	2.20	0.90	15.50
chlorides	0.09	0.00	0.08	0.01	0.61
free sulfur dioxide	15.87	0.26	14.00	1.00	72.00
total sulfur dioxide	46.47	0.82	38.00	6.00	289.00
density	1.00	0.00	1.00	0.99	1.00
pH	3.31	0.00	3.31	2.74	4.01
sulfates	0.66	0.00	0.62	0.33	2.00
alcohol	10.42	0.03	10.20	8.40	14.90

Descriptive statistics provide essential criteria for outliers, such as each variable's minimum and maximum mean and standard deviation. From this information, this paper can develop criteria for outliers, such as three times the standard deviation centered on the mean.

Outliers.

The box plot has some lonely outliers, perhaps due to data entry errors or probably because there are wines with extreme conditions. The presence of these data can interfere with the fitness and should therefore be excluded from the data used. The original dataset of 1599 rows of data analyzed by combining images and formulas had 18 rows of outliers. They come from different rows of different variables, which are line 1300 of alcohol, line 1080, 1082 of total sulfur dioxide, line 152 of citric acid, line 481, 1435, 1436 of residual sugar, line 152, 1317, 1322 of PH, line 87, 92, 93, 152 of sulfates, line 152, 259 of chlorides, line 653 of alcohol, line 397, 401, 1245, 1559 of free sulfur dioxide.

Correlation.

To avoid multicollinearity between variables, conclusions should be drawn from correlation coefficients. Again, Excel's data analysis tool was used to obtain the correlation coefficient matrix of all variables. The result is shown in *Table 2:*

There are only five correlation coefficients above 0.5 and three in fixed acidity. However, the selected edge cannot be deleted directly because it might still have some explanatory power; only after the adjusted R² of both the deleted and undeleted regressions are compared can this paper tell whether to drop the seemingly highly correlated variables or not. Dropping the "fixed acidity" variable, which has a relatively high correlation with three other variables, makes the adjusted R² only fall from 0.364914 to 0.364878, while F increases from 83.531947 to 91.77107. In terms of the results, dropping the "fixed acidity" variable is suitable for the overall regression model, so it was removed from the model.

Table 2: correlation between all variables.

	fixed acidity	volatile acidity	critic acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulfates	alcohol
fixed acidity	1.0000										
volatile acidity	-0.2561	1.0000									
critic acid	0.6717	-0.5525	1.0000								
residual sugar	0.1148	0.0019	0.1436	1.0000							
chlorides	0.0937	0.0613	0.2038	0.0556	1.0000						
free sulfur dioxide	-0.1538	-0.0105	-0.0610	0.1870	0.0056	1.0000					
total sulfur dioxide	-0.1132	0.0765	0.0355	0.2030	0.0474	0.6677	1.0000				
density	0.6680	0.0220	0.3649	0.3553	0.2006	-0.0219	0.0713	1.0000			
pH	-0.6830	0.2349	-0.5419	-0.0857	-0.2650	0.0704	-0.0665	-0.3417	1.0000		
sulphates	0.1830	-0.2610	0.3128	0.0055	0.3713	0.0517	0.0429	0.1485	-0.1967	1.0000	
alcohol	-0.0617	-0.2023	0.1099	0.0421	-0.2211	-0.0694	-0.2057	-0.4962	0.2056	0.0936	1.0000

Log Transformation.

As mentioned in the visualization, the distribution of these variables has a right-skewed problem. Some transformation of variables can effectively solve or improve this problem. In this experiment, log transformation is used to adjust the variables. The variable fixed acidity, citric acid, residual sugar, free sulfur dioxide, and sulfates make the adjusted R^2 rise after log transformation. Still, variable alcohol and total sulfur dioxide do not have the same influence, even if their distribution is right-skewed. Therefore, this paper refuses to transfer them. Some data whose value is smaller than one will lose their meaning after the log, so they will be preceded after adding one.

P-value.

P-value is the value that reflects how much the null hypothesis is rejected. The test of regression coefficient P is the t-test. When P-value is lower than the value of α, the regression coefficient is significant, and the null hypothesis is rejected. The regression model test tests whether the model is suitable through the F test. When the P-value of the F test is lower than α, the model is significant, meaning the overall regression is insignificant. After the initial regression was established, the F-test result was obtained with the help of the Excel data analysis tool, and α was set as 0.05. This paper deletes the variables with a P-value greater than 0.05 because they were not significant enough for the dependent variables. The variable log(1+citric acid) and log(residual sugar) density should drop out from the dataset.

Now that the data has been processed, we have a linear regression equation for eight variables after using linear regression in the Excel data analysis tool.

4.2. Model 2: Logistic Regression

To ensure the model's accuracy, this logistic regression model continues to follow the treatment of data in the previous linear regression model, using the data with outliers and correlation variables removed and log-transformed.

Table 3: alterable cells with reset value.

b1	0
b2	0
b3	0
b4	0
b5	0
b6	0
b7	0
b8	0
b9	0
b10	0
b11	0

Step 1: *Create cells for regression coefficients.* This paper first calculates the mean of quality values, which is 5.623, then sets all the quality which is smaller than the mean quality to 0, otherwise to 1. As there are ten explanatory variables in the model: volatile_acidity, log(1+citric_acid), log(residual _sugar), chlorides, log(free_sulfur_dioxide), total_sulfur_dioxide, density, pH, log(sulfates), alcohol, this paper creates alterable cells for each explanatory variables and intercept. The cells are given a temporary value of 0, which will be optimized later (*Table 3*).

Step 2: *Create values for the logit.* Then use formula (2) to obtain values of the logit:

b1+b2*volatile_acidity+b3*log(1+citric_acid)+b4*log(residual_suger)+b5*chlorides+b6*log(free_sulfur_dioxide)+b7*total_sulfur_dioxide+b8*density+b9*ph+b10*log(sulphates)+b11*alcohol (2)

Step 3: *Create values for the e^(logit).* Use the exponential of the logit values to get e^(logit).

Step 4: *Create values for the probability.* Then use the values of e^(logit) divided by one plus e^(logit) to get the probability values.

Step 5: *Create values for the log-likelihood.* Next, use the formula (3) to get log-likelihood :

(quality*log(probability))+((1-quality)*log(1-probability)) (3)

Step 6: *Calculate the maximum of the sum of log-likelihood.* Finally, calculate the sum of log-likelihood using Solver (a statistic package from excel) to calculate the maximum likelihood by setting b1 to b11 as changing variables and setting the sum of log-likelihood to Max and get the value of cells(*Table 4*).

Table 4: value of cells in the final model.

b1	0.09527662
b2	-2.9591784
b3	-2.6857105
b4	0.13922955
b5	-4.4281472
b6	1.23262156
b7	-0.0221476
b8	-1.210272
b9	-1.5279756
b10	5.92019716
b11	0.89937146

5. Results and Discussion

5.1. Summary of Two Models

The original data set owns 12 variables. Since the quality variable is set as the variable to be predicted, there are 11 independent variables left. After the preparation process of multicollinearity exclusion and log transformation, models of better fitting can be obtained. This paper builds two different regression models: a linear regression model (model 1) and a logistic regression model (model 2).

Table 4 represents the summary of model 1. The top row shows R, R^2, adjusted R^2, and F. R represents the complex correlation coefficient used in regression analysis to describe the correlation and regression relationship between the dependent and independent variables, indicating how much the dependent variable is co-related to all the independent variables as a whole. R^2 equals 0.366361, which means that the predictors can explain quality at a level of about 36.64%. Adjusted R^2, which is 0.363541, implies approximately 36.35% of quality's dependency on all the independent variables more accurately than R^2. Adjusted R^2 is relatively low, indicating that these variables fail to predict the value of quality perfectly. The F-value is used to test the extent to which the sample results represent the overall results. In this model, F-value equals 137.1617, which shows a suitable fitting.

The value of coefficients, standard error, t stat, P-value, lower 95%, and upper 95% are also shown in table 1. These values indicate the differences between the degree of influence of independent variables on quality. Coefficients show that a 1% increase in the value of each of the independent variables may bring how many percent changes in quality. Since the predictors' P-values are all lower than 0.05, they are crucial predictors of quality. So, we can conclude the final equation (equation (4)) of model 1:

$$\begin{aligned} Quality &= 5.5389 - 0.8725 * volatile\ acidity - 1.9201 * chlorides + 0.2607 * \log(free\ sulfur\ dioxide) \\ &- 0.0044 * total\ sulfur\ dioxide - 0.5968 * pH + 1.9167 * \log(sulphates) + 0.2868 * alcohol \end{aligned} \tag{4}$$

From the model equation, to improve wine quality, free sulfur dioxide, sulfates, and alcohol are elements that need to be enhanced, and volatile acidity, chlorides, total sulfur dioxide, and pH are elements that need to be decreased.

Table 5: Summary of the linear regression model (model 1).

Multiple R = 0.605278	R²= 0.366361	Adjusted R²= 0.363541	F = 137.1617			
	Coefficients	Standard error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	5.538866	0.40058	13.826949	4.01875E-41	4.753129	6.324602
volatile acidity	-0.872454	0.101470	-8.598127	1.92469E-17	-1.071485	-0.673423
chlorides	-1.920067	0.418301	-4.590153	4.78120E-06	-2.740554	-1.099580
log(free sulfur dioxide)	0.260681	0.075655	3.445656	0.000585	0.112286	0.409076
Total sulfur dioxide	-0.004391	0.000721	-6.089460	1.42117E-09	-0.005805	-0.002977
pH	-0.596760	0.118160	-5.050431	4.92180E-07	-0.828528	-0.364992
log(sulphates)	1.916690	0.193344	9.913390	1.63830E-22	1.537452	2.295928
alcohol	0.286770	0.016867	17.002228	1.20431E-59	0.253686	0.319853

After the logistic regression operation performed in the previous section, the results of model 2 are obtained and shown as equation (5):

$$ln(P/(1-P)) = 0.0953 - 2.9592 * volatile\ acidity - 2.6857 * log(1+citric\ acid) + 0.1392 * log(residual\ sugar) - 4.4281 * chlorides + 1.2326 * log(free\ sulfur\ dioxide) - 0.0221 * total\ sulfur\ dioxide - 1.2103 * density - 1.5280 * pH + 5.9202 * log(sulphates) + 0.8994 * Alcohol \tag{5}$$

In this equation, the dependent variable is ln(P/(1-P)), where P represents the probability of quality equal to 1, and (1-P) represents the probability of quality equal to 0. Unlike model 1, there is no way to get a specific value for wine quality here, only the possibility of good and bad wine. From this equation, we can also tell how the predictors influence wine quality. With higher residual sugar, free sulfur dioxide, sulfates, and alcohol, there is a higher possibility of having good wine. While with higher volatile acidity, citric acid, chlorides, total sulfur dioxide, density, and pH, there is a lower possibility of getting a good wine. Since ln(P/(1-P)) and P are positively correlated, the variables with positive coefficients should be increased, and those with negative coefficients should be decreased to improve the wine quality.

5.2. Profit Comparison

In the previous section, for logistic regression, wines with quality greater than 5.6382 are categorized as good wines (noted as 1), and those less than 5.6382 are classified as bad wines (indicated as 0). To compare the linear and logistic regression models, quality data in model 1 also needs to be transferred into 0 or 1. Therefore, the equation in model 1 is used to evaluate the quality. After substituting the data into the equation in model 1 to calculate the predicted rate and classifying the expected quality in 0 or 1 according to the above rules, 737 bottles of good and 844 bottles of bad wines are obtained. Simple multiplication is needed to determine the projected profit when calculating the profit. Given that the yield of good wine is A and lousy wine is B, the formula for calculating profit is as equation (6):

$$Profit\ 1 = 737 \cdot A + 844 \cdot B \tag{6}$$

In model 2, probability values are used to perform probability calculations on prices. Since the probability of good wine is obtained, the operation is as follows: multiply the cost of good wine by the likelihood of getting good wine, multiply the cost of bad wine by the likelihood of getting sour wine, and then add the two together to get the expected cost of each glass of wine. Set as the probability of the nth data is good wine. The formula of the sum of the price is:

$$Profit\ 2 = \sum_{1}^{1581} [P_n \cdot A + (1 - P_n) \cdot B] = 846.0023 \cdot A + 734.9977 \cdot B \tag{7}$$

According to common sense, higher quality wine tends to be more popular and expensive, so A is more significant than B (both A and B are positive). On the basis that the sum of the coefficients in both profit models is equal to 1581, the coefficient of A in profit model 2 is higher, so the profit calculated in model 2 is more elevated. From the standpoint of sellers, they will seek to maximize the profitability of the sale. When comparing the two models, the one that predicts higher profits for the same batch of wine will be more favored by them. From this perspective, profit model 2 is better based on the logistic regression model.

6. Conclusion and Future Direction

Research on wine quality has been widespread. This paper predicts the quality of the wine based on chemical features in two ways: linear regression and logistic regression. The two models reached similar conclusions about which variables have a positive influence and which have a negative one on quality. Further, this paper creatively stands on the seller's perspective and compares the two models with the goal of profit maximization. The quality with multiple levels is divided into good and evil to facilitate pricing. On the basis that the price of good wine is greater than that of sour wine, model 2 can predict more good wine and bring more revenue, which is a better model.

This study, however, is not without limitations. Firstly, there are only 12 variables in the data, and after setting quality as the dependent variable, there are only 11 independent variables. Thus, not all good chemical properties are examined comprehensively. In addition, the price of red wine may be influenced by external factors, such as brand value, vintage, origin, and other features. Limited by the data set used, these were not taken into account in the model in this study. Secondly, this paper assumes that all the variables have a single effect on quality, meaning that the variables can only have a positive or negative impact on quality. But some products may be favorable and then change to negative after reaching a particular value. Thirdly, since model 2 is obtained by using the solver plug-in in excel instead of the regression analysis function, there is no way to get the t-values of each variable or F-value of the whole equation, so this paper fails to detect the significance of the equation and analyze the degree of fit from the statistical point of view. Past research can improve these limitations by refining the data, modifying the model, and using other software for assistance.

References

[1] Trei, Lisa, and Lisa Trei. Price Changes Way People Experience Wine, Study Finds. Stanford University, January 16, 2008. https://news.stanford.edu/news/2008/january16/wine-011608.html.

[2] *Horowitz I, Lockshin L. What Price Quality? An Investigation into the Prediction of Wine-quality Ratings. Journal of Wine Research. 2002; 13: 7-29*

[3] *Antonio, Capurso. The Six Attributes of Quality in Wine. Wine And Other Stories. 24 Jan. 2020*

[4] *Kwak, Young-Sik, Yoon-Jung Nam, Jae-Won Hong. Effect of Online Collective Intelligence in Wine Industry: Focus on Correlation between Wine Quality Ratings and on-Premise Prices*

[5] *Badole, Mayur. Wine Quality Prediction Using Machine Learning: Predicting Wine Quality. Analytics Vidhya, 25 July 2022*

[6] *Yogesh Gupta. Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science. 2018; 125: 305-312*

[7] *Niggl, Dennis. Predict Red Wine Quality. Kaggle*

[8] *Tingwei, Zhou. Red Wine Quality Prediction through Active Learning. Journal of Physics: Conference Series 1966, no. 1 (2021): 012021.*

[9] *Shaw, B., Suman, A.K., Chakraborty, B. (2020). Wine Quality Analysis Using Machine Learning. In: Mandal, J., Bhattacharya, D. (eds) Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, vol 937*

[10] *Kothawade, Rohan Dilip. Wine Quality Prediction Model Using Machine Learning Techniques, n.d.*