Analysis and Prediction of Wine Quality

Zhen-Yan Chen

December 2023

1 Abstract

In this project, we explore the relationship between wine quality and its chemical properties, with a specific focus on both the red and white variants of Portuguese 'Vinho Verde' wine. Our analysis addresses **skewed data**, **collinearity**, and **variable selection**, incorporating **interaction effects** to capture the collaborative influence of variables on wine quality based on ordinal regression. The result is the ordinal regression predictive model we built, achieving accuracies of 60.7% for red wine and 52.7% for white wine slightly lower than those of SVM, which is a minor interpretability-accuracy tradeoff.

Furthermore, the introduction of a LASSO penalty highlights the critical importance of fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, and alcohol level in determining wine quality. Consequently, we are able to recommend wine producers focus on optimizing these key factors for wine production. Through the fundamental analysis, our project contributes to the pursuit of excellence in wine-making by emphasizing the significance of specific chemical components in achieving superior wine quality.

Overall, we develop an explainable and predictive model with high prediction accuracy close to the classical 'black box' SVM. Based on our results, we give actionable recommendations for wine producers to improve quality related to its chemical properties.

2 Problem Description

2.1 Theme, Motivation, and Goals

- What is the relationship between quality and chemical variables? How can we affect the quality by its chemical content?
- If we can find the relationships between chemical variables and quality, can we give some suggestions to wine producer for improving the wine quality?
- How can we build a predictive model for wine producer to predict the evaluation from wine experts?

2.2 Statistical Challenge

- Skewed data
- Multi-collinearity
- Imbalance data: quality are most in 5∼7 (see histograms in Appendix 8.1).

3 Data Description

The two datasets from UCI Machine Learning Repository, are related to the red and white variants of Portuguese 'Vinho Verde' wine. `https://archive.ics.uci.edu/dataset/186/wine+quality`. The dataset of red wine contains 1599 observations and the dataset of white wine contains 4898 observations. Both have the same 12 variables. Each observation represents a particular wine. The 11 independent variables represent the chemical properties of the wines. The response variable, quality, is determined from the median scores given by at least three wine experts, ranging from 0 to 10.

In our data, the independent variables are all continuous and the response variable is integer from 3 to 9. All variables have no missing value.

- **Fixed acidity**: acid from grape into wine, not easy to evaporate. (g/dm^3)
- **Volatile acidity**: acid easy to evaporate, mainly CH_3COOH having an unpleasant smell. (g/dm^3)
- **Citric acid**: a type of fixed acid supplement, boosts the acidity of the wine, typically found in small quantities. (g/dm^3)

- **Residual sugar**: sugar remaining after fermentation stops, > 45 are considered sweet. (g/dm^3)
- **Chlorides**: sodium chloride, depends on soil and grape type. (g/dm^3)
- **Free sulfur dioxide**: prevents microbial growth and the oxidation of wine. > 50 ppm, SO_2 becomes evident in the nose and taste of wine. (mg/dm^3)
- **Total sulfur dioxide**: total amount of free and bound forms of SO_2. (mg/dm^3)
- **Density**: depending on the percent alcohol and sugar content, wine is higher than water. (g/cm^3)
- **pH**: most wines are between 3-4.
- **Sulphates**: contribute to sulfur dioxide (SO_2) levels. (g/dm^3)
- **Alcohol**: generally between 5–15%. High alcohol allow more flavors and aromas easily captured, but too high will mask the other aromas and flavors of the wine.

You can find some exploratory data analysis in Appendix 8.1. As for the issue of **outliers**, even though the majority of quality ratings fall between 5 and 7, we believe the remaining data still provides valuable information of real-world wine assessment. Therefore, we choose to conduct the following analysis using the entire dataset.

4 Methods

Since the quality ranges from 0 to 10, we use the **cumulative model in ordinal regression** to estimate the coefficients. Considering the similarity in AIC values for the logit, probit, and cloglog link functions (3090, 3114, and 3111, respectively), we simply choose the **logit link** function for the subsequent analysis.

The cumulative model for ordinal regression:

$$\log \frac{P(Y_i \leq r)}{P(Y_i > r)} = \theta_r + X_i\beta, \quad r = 1, ..., c+1$$

for the response variable $Y_i \in \{1, ..., c+1\}$, the covariates $X = \begin{bmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{nk} \end{bmatrix} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$, and $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$.

4.1 Model 1: Directly fit ordinal regression and detect multi-collinearity by VIF

First, we use the original data to fit ordinal regression. Through iterative removal of the covariate with the highest VIF until all VIFs are less than 5, we exclude density in both red and white wine (VIF = 16.8 and 38.1, respectively). The resulting AIC values are 3106 for red wine and 10997 for white wine, correspondingly. The results show in the following tables:

Red wine	Value	Std. Error	p value
fixed acidity	0.0557	0.0507	0.2715
volatile acidity	-3.4512	0.4002	0.0000
citric acid	-0.8037	0.4628	0.0824
residual sugar	0.0554	0.0382	0.1473
chlorides	-5.2800	1.3547	0.0001
free sulfur dioxide	0.0145	0.0068	0.0320
total sulfur dioxide	-0.0114	0.0024	0.0000
pH	-1.2322	0.4926	0.0124
sulphates	2.8101	0.3581	0.0000
alcohol	0.8994	0.0595	0.0000

White wine	Value	Std. Error	p value
fixed acidity	-0.1208	0.0379	0.0015
volatile acidity	-5.1417	0.3063	0.0000
citric acid	-0.0180	0.2421	0.9408
residual sugar	0.0625	0.0065	0.0000
chlorides	-2.6461	1.3642	0.0524
free sulfur dioxide	0.0147	0.0022	0.0000
total sulfur dioxide	-0.0028	0.0010	0.0038
pH	0.5098	0.2106	0.0155
sulphates	1.1825	0.2465	0.0000
alcohol	0.9491	0.0314	0.0000

4.2 Model 2: Transform data and detect multi-collinearity by VIF

Observing skewness in variables in Figure 1, we apply the $f(x) = \log x + 1$ transformation to variables with skewness greater than 0.5 and fit ordinal regression. Through iterative examination of the highest VIF, we delete density (VIF = 88.5) in red wine data and density (VIF = 28.7) and chlorides (VIF = 25.3) in white wine data. The resulting AIC values are 3100 for red wine and 10860 for white wine, respectively. The results show in the following tables:

Red wine	Value	Std. Error	p value
fixed acidity	0.8451	0.4938	0.0870
volatile acidity	-5.4025	0.6215	0.0000
citric acid	-1.0929	0.4579	0.0170
residual sugar	0.2247	0.2013	0.2644
chlorides	-5.8507	1.6104	0.0003
free sulfur dioxide	0.3890	0.1368	0.0045
total sulfur dioxide	-0.5655	0.1328	0.0000
pH	-1.1012	0.4976	0.0269
sulphates	5.5021	0.6440	0.0000
alcohol	10.4710	0.7122	0.0000

White wine	Value	Std. Error	p value
fixed acidity	-0.6237	0.3039	0.0402
volatile acidity	-6.5603	0.4129	0.0000
citric acid	0.1474	0.3364	0.6612
residual sugar	0.4531	0.0457	0.0000
free sulfur dioxide	0.9033	0.0721	0.0000
total sulfur dioxide	-0.0065	0.0010	0.0000
pH	0.5851	0.2117	0.0057
sulphates	1.9614	0.3804	0.0000
alcohol	0.9547	0.0298	0.0000

4.3 Model 3: Transform data and detect multi-collinearity by LASSO

Finally, we apply a LASSO penalty to the model using the transformed data. Through backward selection, we choose lambda values of $\lambda_{red} = 0.00676$ and $\lambda_{white} = 0.00634$. This leads to AIC values of 3192 for red wine and 10934 for white wine. The results are presented in the following tables:

Red wine	Value	Std. Error	p value
fixed acidity	1.1340	0.3119	0.0003
volatile acidity	-5.7038	0.4999	0.0000
residual sugar	-0.0668	0.1963	0.7335
free sulfur dioxide	0.0057	0.0832	0.9456
alcohol	11.3940	0.6549	0.0000

White wine	Value	Std. Error	p value
fixed acidity	-1.1644	0.2646	0
volatile acidity	-7.2592	0.3959	0
residual sugar	0.3826	0.0446	0
free sulfur dioxide	0.6461	0.0593	0
alcohol	1.0050	0.0287	0

4.4 Model Assumptions Checking

The model assumptions of cumulative model for ordinal regression are:

1. The dependent variable are ordered: our data is already satisfied.
2. At least one of the independent variables are either continuous, categorical or ordinal: our data is already satisfied.
3. No multi-collinearity: in model 1 and 2, we have detected and removed the collinearity term via VIF; in model 3, we use LASSO to deal with collinearity by its shrinkage effect.
4. Proportional odds: test the proportional odds assumption by Brant's test. Proportional odds means that the relationship between each pair of outcome groups has to be the same, so it is also called parallel regression assumption. Our all models passed Brant's test with the supporting evidence in Appendix 8.3.

4.5 Model 4: Interaction Effect

To examine the interaction effects, we separately add possible interactions to Model 2. (You can find the results in Appendix 8.2) After assessing the significance of each interaction term, we add them simultaneously. By selecting the model with the minimum AIC, the final model for red wine includes interactions between free sulfur dioxide and total sulfur dioxide, as well as the interaction between chlorides and sulphates, resulting in an AIC of 3079 and an accuracy rate of 60.7%. For white wine, we add the interaction between free sulfur dioxide and total sulfur dioxide, resulting in an AIC of 10757 and an accuracy rate of 52.8%. The equations below shows the final models:

For red wine:

$$\begin{aligned}\log\frac{P(Y_i \le r)}{P(Y_i > r)} = \theta_r &+ 0.3580\ \text{fixed.acidity} - 5.3445\ \text{volatile.acidity} - 0.7833\ \text{citric.acid} \\ &+ 0.2982\ \text{residual.sugar} + 10.0947\ \text{chlorides} + 2.3170\ \text{free.sulfur.dioxide} \\ &+ 0.7137\ \text{total.sulfur.dioxide} - 1.5126\ \text{pH} + 7.5918\ \text{sulphates} + 10.5444\ \text{alcohol} \\ &- 0.5206\ \text{free.sulfur.dioxide} \times\ \text{total.sulfur.dioxide} - 23.7085\ \text{chlorides} \times\ \text{sulphates}\end{aligned}$$

For white wine:

$$\begin{aligned}\log\frac{P(Y_i \le r)}{P(Y_i > r)} = \theta_r &- 0.7166\ \text{fixed.acidity} - 6.6226\ \text{volatile.acidity}0.2727\ \text{citric.acid} \\ &+ 0.4781\ \text{residual.sugar} + 2.3329\ \text{free.sulfur.dioxide} + 0.0338\ \text{total.sulfur.dioxide} \\ &+ 0.5162\ \text{pH} + 1.8302\ \text{sulphates} + 0.9481\ \text{alcohol} \\ &- 0.0114\ \text{free.sulfur.dioxide} \times\ \text{total.sulfur.dioxide}\end{aligned}$$

5 Results

To solve our problems for relationships between quality and chemical variables and to predict wine quality, we select over the models above. After finding the relationships, we interpret the results and draw conclusions that can help wine producer improve wine quality.

5.1 Model Selection

	Red wine	White wine
Model 1	3106	10997
Model 2	3100	10860
Model 3	3192	10934
Model 4	3079	10757

Table 1: AIC of all three models

- For prediction: the results above suggest that we should opt for the model 4 with the setting of AIC as the criterion. The below contingency table indicates that our model tends to be more conservative (mostly concentrated between 5 to 7), which is reasonable because of the imbalanced distribution of response variable. However, the accuracy rate within one unit deviation is mostly above 90%.

real \ pred	5	6	7
3	9	1	0
4	39	14	0
5	505	173	3
6	202	410	26
7	7	136	56
8	0	11	7
±0	74%	64%	28%
±1	99%	100%	96%

real \ pred	3	4	5	6	7
3	1	0	11	8	0
4	0	3	112	47	1
5	0	2	770	676	9
6	0	2	405	1608	183
7	0	0	58	619	203
8	0	0	8	107	60
9	0	0	0	3	2
±0	5%	2%	53%	73%	30%
±1	5%	71%	99%	100%	93%

Table 2: Contingency table of real classes and predicted classes; left is for red wine and right is for white wine

The overall resulting accuracies ($\frac{\#\text{right prediction}}{\text{all}}$) are 60.7% for red wine and 52.7% for white wine. In a study by Ángela Nebot, Francisco Mugica, and Antoni Escobet (2015), using SVM, the overall accuracies were reported as 62.4% for red wine and 64.6% for white wine when considering only the correctly classified classes (T = 0.5). Compared to our result, our overall accuracies are slightly smaller.

- For the relationships between chemical variables and quality: we can choose the model with LASSO penalty (model 3) since the third model has much fewer variables, and the difference in resulting AIC between the second and third models is less than 100.

5.2 Interpretation and Conclusion

We conclude from LASSO model that fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, and alcohol are more critical than the other variables in determining wine quality. Moreover, from their p-value, only the coefficients of fixed acidity, volatile acidity, and alcohol are significant under 0.05 confidence level in red wine.

Our interpretation for this result is that volatile acidity provides an unpleasant smell, so it is reasonable that it has negative effect to quality in both kinds of wine; high alcohol allow more flavors and aromas easily captured and our data, "Vinho Verde" wine, is famous for tropical fruity aroma, so it is reasonable that it has positive effect on quality in both kinds of wine. For white wine, fixed acidity has negative effect and residual sugar has positive effect on quality, since people prefer sweeter white wine. While red wine is not, fixed acidity will balance with other elements. And the positive effect of free sulfur dioxide may depend on grape type, since sulfur dioxide prevents microbial growth and the oxidation of wine.

Therefore, we recommend the wine producer to increase alcohol through distillation and decrease volatile acidity via precise control of fermentation generally. For red wine, they can increase fixed acidity through some acidity supplement such as citric acid. For white wine, they can increase the sweetness, more residual sugar and less fixed acidity, by sugar supplement or better grape and with better grape, we may add more antioxidants such as sulphates that will produce higher free sulfur dioxide.

6 Discussion

Remark on the interaction effect model, it would not pass collinearity assumption, but it has lower AIC, so we only utilize it for prediction.

Due to the imbalance in the response, our analysis encounter challenges in predicting the boundary. To address this, we are considering the application of oversampling the minority class and undersampling the majority, such as Synthetic Minority Oversampling Technique (SMOTE). SMOTE (2002) examines minority class instances, uses k-nearest neighbors to select a random nearest neighbor, and creates a synthetic instance randomly in feature space. Further research will concentrate on handling the unbalanced response data to improve prediction results and a better understanding of wine quality.

7 Reference

- **Data source**: `https://archive.ics.uci.edu/dataset/186/wine+quality`
- Àngela Nebot, Francisco Mugica & Antoni Escobet (2015). Modeling Wine Preferences from Physico-chemical Properties using Fuzzy Techniques. doi: 10.5220/0005551905010507
- Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall & W. Philip Kegelmeyer (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16 P.321–357

8 Appendix

8.1 Exploratory Data Analysis

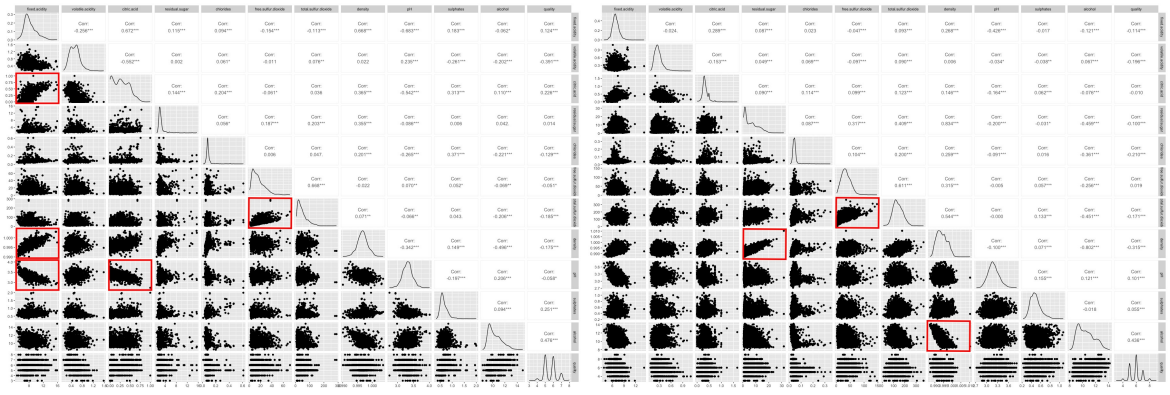

Figure 1: Comparison of variables

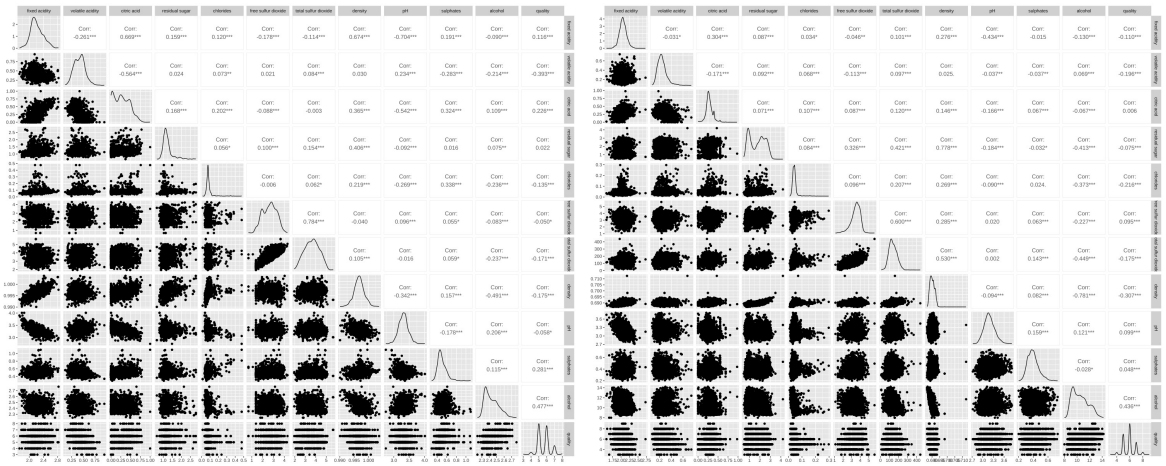

Figure 2: Comparison of transformed variables

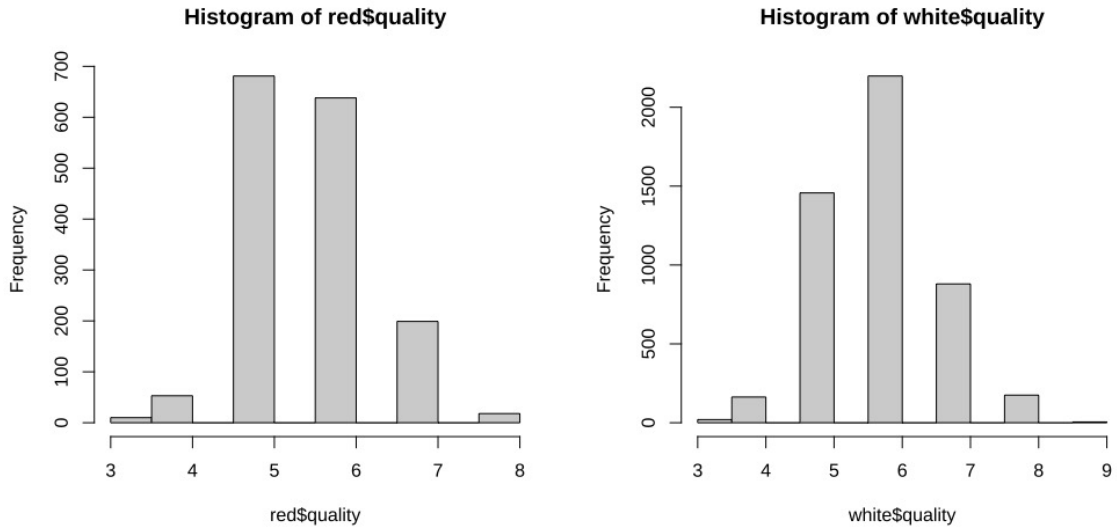

Figure 3: Distribution of Quality

8.2 Interaction Analysis

The results of adding interactions to Model 2 separately are as follows:

Red wine	Value	Std. Error	p value
fixed.acidity:volatile.acidity	-1.1742	2.8631	0.6817
citric.acid:sulphates	-7.0653	2.9334	0.0160
chlorides:sulphates	-23.5602	7.6292	0.0020
free.sulfur.dioxide:total.sulfur.dioxide	-0.5168	0.1272	0.0000

White wine	Value	Std. Error	p value
fixed.acidity:volatile.acidity	-7.2406	0.2064	0.0000
citric.acid:sulphates	-12.3155	0.1738	0.0000
residual.sugar:alcohol	0.0061	0.0361	0.8649
free.sulfur.dioxide:total.sulfur.dioxide	-0.0114	0.0011	0.0000

8.3 Test for Parallel Regression Assumption

Test for	X2	df	probability
Omnibus	119.16	40	0
fixed.acidity	8.21	4	0.08
volatile.acidity	10.93	4	0.03
citric.acid	5.46	4	0.24
residual.sugar	9.95	4	0.04
chlorides	6.38	4	0.17
free.sulfur.dioxide	3.92	4	0.42
total.sulfur.dioxide	18.64	4	0
pH	11.88	4	0.02
sulphates	2.24	4	0.69
alcohol	16.2	4	0

H0: Parallel Regression Assumption holds

Test for	X2	df	probability
Omnibus	140.96	40	0
fixed.acidity	10.86	4	0.03
volatile.acidity	12.48	4	0.01
citric.acid	6.87	4	0.14
residual.sugar	9.7	4	0.05
chlorides	7.34	4	0.12
free.sulfur.dioxide	5.6	4	0.23
total.sulfur.dioxide	23.7	4	0
pH	13.93	4	0.01
sulphates	3.99	4	0.41
alcohol	14.58	4	0.01

H0: Parallel Regression Assumption holds

Test for	X2	df	probability
Omnibus	89.99	20	0
fixed.acidity	4.3	4	0.37
volatile.acidity	14	4	0.01
residual.sugar	6.03	4	0.2
free.sulfur.dioxide	16.98	4	0
alcohol	35.2	4	0

H0: Parallel Regression Assumption holds

Test for	X2	df	probability
Omnibus	121.5	50	0
fixed.acidity	93.6	5	0
volatile.acidity	36.75	5	0
citric.acid	6.73	5	0.24
residual.sugar	5.75	5	0.33
chlorides	29.9	5	0
free.sulfur.dioxide	-24.26	5	1
total.sulfur.dioxide	6.61	5	0.25
pH	25.41	5	0
sulphates	8.52	5	0.13
alcohol	55.03	5	0

H0: Parallel Regression Assumption holds

Test for	X2	df	probability
Omnibus	264.03	45	0
fixed.acidity	40.35	5	0
volatile.acidity	34.91	5	0
citric.acid	6.72	5	0.24
residual.sugar	2.58	5	0.76
free.sulfur.dioxide	20.71	5	0
total.sulfur.dioxide	16.66	5	0.01
pH	28.64	5	0
sulphates	9.64	5	0.09
alcohol	60.28	5	0

H0: Parallel Regression Assumption holds

Test for	X2	df	probability
Omnibus	208.51	25	0
fixed.acidity	17.8	5	0
volatile.acidity	38.02	5	0
residual.sugar	7.09	5	0.21
free.sulfur.dioxide	42.35	5	0
alcohol	68.12	5	0

H0: Parallel Regression Assumption holds

Figure 4: Test for Parallel Regression Assumption