Estimating the Quality of Wine with Chemical Composition

DATASCI 203 Lab 2, Fall 2022

Theresa Azinge, Wesley Chang, Spandan Garg, and Dakota Potere-Ramos

Introduction

For thousands of years wine has evolved to become an integral part of human culture. While premium wines may still be expensive, specialty wines are no longer rare luxuries due to increased global access through globalization. With more people consuming wine today than ever before, it is important to regulate and control the quality of products in the market. Thus, the production of wine is subject to various global, national, and local regulations which govern aspects of the process such as the use of additives and chemical composition.

The use of additives is necessary for winemakers to produce consistent batches of product despite variations in the quality of grapes. However, due to the subjective nature of human taste, even small changes in chemical components can vastly impact the perceived wine quality.[1] In this study, we try to understand how various chemical components impact the quality of wine as assessed by human evaluators in the Wine Quality Dataset from UC Irvine's ML Repository.[2] We will focus this study on Vinho Verde white wines produced in Portugal. With this experiment we hope to build an understanding of how different chemical components impact wine quality to help wine distributors better assess the quality of wine before purchase.

Data and Methodology

The data in this study was obtained from the University of California Irvine's Machine Learning Repository.[2] The data was compiled and donated for public use by Paulo Cortez et al.[3] The Wine Quality dataset is related to white variants of the Portuguese Vinho Verde wine. Each row in the data represents the results of laboratory physicochemical and sensory analysis on a unique sample of wine. The laboratory analysis provides eleven features related to the chemical composition of the wine. Eight features represent concentrations of various compounds while the other three represent alcohol percentage by volume, density, and pH. The sensory analysis provides a single feature for the quality score of the wine on a 0-10 scale. Each wine sample was evaluated via blind taste testing by a minimum of three sensory assessors, or wine experts. The final quality score value was given by the median score of all evaluations. We performed all exploration and model building on a 30% subsample of the data. The remaining 70%, totaling 3428 rows, was used to generate the statistics in this report.

The operationalization of concepts was straightforward given the dataset was crafted and published for similar analyses using different statistical methods.[1] To operationalize human taste preferences, or wine quality, the quality feature from the sensory analysis was utilized. To operationalize chemical composition, linear combinations of various subsets of the eleven features determined by the laboratory physicochemical analysis were utilized. Exploratory plots of quality versus each of the eleven physicochemical features, with a curve of best fit, were created to visually analyze if any strong relationships existed. Utilizing these plots and intuition from our own taste experiences we selected acetic acid as the primary input variable to explore.

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[2] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[3] 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

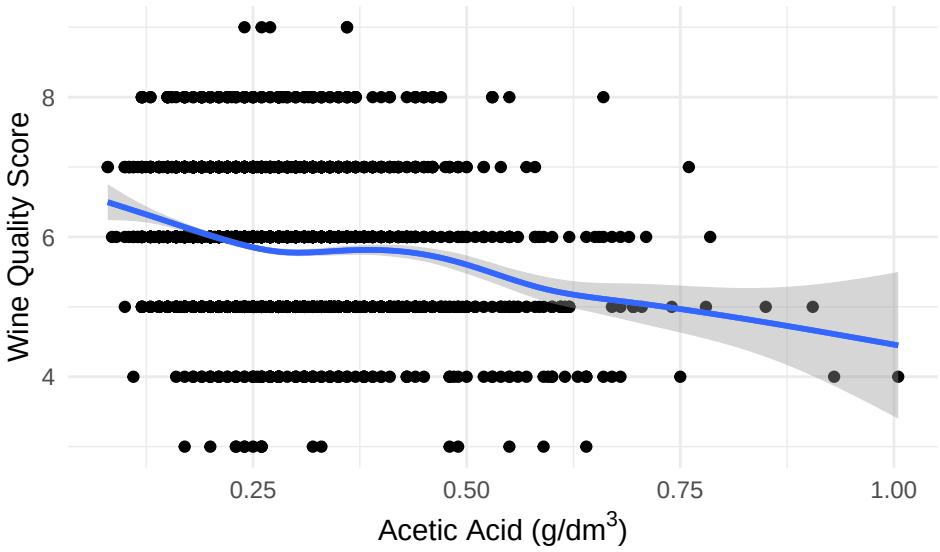

Figure 1: Wine Quality Score as a Function of Acetic Acid

Analysis of figure 1 reveals a roughly negative linear relationship between acetic acid and wine quality; quality score decreases as acetic acid increases. This aligns with our expectations because acetic acid is the main component of vinegar; too much acetic acid can cause degradation in the taste of wine. One should note that the ordinal nature of the quality variable may cause some heteroskedasticity of the residuals. Acknowledging this behavior, we create regression models which assume a positive linear combination of the chemical composition features. In other words, we fit regression of the form:

$$\widehat{quality} = \beta_0 + \beta_1 \cdot \text{acetic acid} + \mathbf{Z}\gamma$$

Where β_1 represents the decrease in wine quality per unit increase of acetic acid (g/dm^3), $\mathbf{Z}$ is a row vector of additional chemical composition covariates, and γ is a column vector of coefficients.

Minimal cleaning, transformation, and omission of entries or features was necessary. There were no NULL or NA values present in the dataset. However, of the eleven features available for regression we opted to omit two from our analysis; pH and free sulfur dioxide. pH is a measure of how acidic or basic a substance is. Given our model includes measures of acidity via acetic acid, citric acid, and tartaric acid we believe pH could be considered an outcome variable or have collinearity with the other acid features. Free sulfur dioxide was omitted because the total sulfur dioxide feature already accounts for free sulfur dioxide in its measure. To improve precision and avoid bias in the estimation of quality we removed these terms from the regression.

Results and Discussion

Table 1 shows the results of four representative regression models. Across these models, the key coefficient on acetic acid (g/dm^3) was statistically significant, with values of the point estimates ranging from -2.04 to -1.71. This means that with a unit increase in the concentration of acetic acid in the wine, there would be an approximately 2 point degradation in quality score assigned to it by human evaluators. This seems to agree with the slope of the plot in Figure 1.

Another statistically significant variable across all models was residual sugar (g/dm^3) with point estimates ranging from 0.03 to 0.06. This is evidence that the quality assigned to white wine increases as the residual sugar in the wine is increased. Since residual sugar is related to the sweetness of the wine, it is logical to expect this relationship.

Table 1: Estimated Regressions

	Output Variable: Wine Quality Score			
	(1)	(2)	(3)	(4)
Acetic acid (g/dm^3)	-1.71^{***}	-2.03^{***}	-2.01^{***}	-2.04^{***}
	(0.16)	(0.14)	(0.14)	(0.14)
Sodium chloride (g/dm^3)		-0.91	-0.73	-0.93
		(0.58)	(0.58)	(0.58)
Potassium sulphate (g/dm^3)		0.41^{***}	0.58^{***}	0.40^{**}
		(0.12)	(0.13)	(0.12)
Alcohol ($vol.\%$)		0.38^{***}	0.28^{***}	0.38^{***}
		(0.01)	(0.03)	(0.01)
Citric acid (g/dm^3)		0.09	0.09	0.09
		(0.10)	(0.10)	(0.11)
Tartaric acid (g/dm^3)		-0.09^{***}	-0.05^{*}	-0.09^{***}
		(0.02)	(0.02)	(0.02)
Residual sugar (g/dm^3)		0.03^{***}	0.06^{***}	0.03^{***}
		(0.003)	(0.01)	(0.003)
Total sulfur dioxide (mg/dm^3)			0.001	0.0002
			(0.0005)	(0.0004)
Density (g/cm^3)			-89.15^{***}	
			(20.81)	
Constant	6.36^{***}	2.76^{***}	91.74^{***}	2.72^{***}
	(0.05)	(0.21)	(20.78)	(0.22)
Observations	3,428	3,428	3,428	3,428
R^2	0.04	0.28	0.29	0.28
Residual Std. Error	0.89 (df = 3426)	0.77 (df = 3420)	0.76 (df = 3418)	0.77 (df = 3419)

Note: HC_1 robust standard errors in parentheses.

Model 3 indicates that density has a very large negative coefficient of -89.15. This value is orders of magnitude greater than other coefficients and caused the base constant (β_0) to be 91.74 compared to a base β_0 coefficient of 2.72 without density. This base β_0 coefficient of 91.74 is quite high in comparison to the quality outcome variable which ranges from 0 to 10. Also, since the density variable refers to the density of the wine, we expect it to be collinear with concentrations of additives, acids and residual components of the wine. Though we did not notice *perfect* collinearity in the model, our reasoning around density combined with the high standard errors led us to drop the variable from the model altogether. Subsequent models showed improvement as no variables exhibited high standard errors.

Taking into consideration our reasoning for omitting density and adding the statistically significant variables, the final iteration of the Vinho Verde model is Model 4 as shown in the regression table. We do need to mention that the R^2 values achieved by all of the models are very low (<0.29), implying that the models are

doing a poor job at capturing the variance in the outcome variable. However, we believe this is a consequence of the ordinal outcome variable and is an inherent limitation of the data itself.

Limitations

Our data corresponds to only the wine produced in the Minho region of Portugal with taste samples taken from May 2004 to Feb 2007. Since all the wine samples examined were taken from this small region and time window, there is a possibility of geographical and temporal clustering due to ingredients like grapes coming from nearby vineyards or being part of the same yearly yield. We expect that if these characteristics do have a large effect within the scope of our specified data, the rest of the estimators will be biased towards zero. However, this is not a concern for us due to the specific conditions needed to produce Vinho Verde white wine. It is also possible that multiple data points correspond to wines made by the same winemakers. However, by the design of our study, we are solely interested in Vinho Verde white wines, it is okay for the data points to not be independent and identically distributed (IID) with respect to all wines produced globally.

While the Cortez et al. study finds that sulfates were a statistically significant predictor of the quality score, we chose to use acetic acid as the primary input variable because of the linear relationship with the outcome variable found during our data exploration. Furthermore, based on our prior knowledge of wines, we know that it is an important component of how wine tastes. Too much acetic acid would cause degradation in wine taste, giving it a vinegar like aftertaste. Of the eleven features available for regression we opted to omit two from our analysis; pH and free sulfur dioxide. We believed that pH would be collinear with acid features. We also omitted free sulfur dioxide because the total sulfur dioxide feature already accounts for free sulfur dioxide.

The most significant limitation of our model comes from the subjectivity of the wine rankings themselves. Human taste perception is extremely subjective to the taster. This poses a significant risk towards replicability in future studies. The original data was compiled by the Cortez et al. study, which makes no mention of the scoring criteria, the ranking process, or any external factors that may influence the judging itself. We are unable to verify the integrity of the scoring process and if there are systematic inconsistencies in scoring between wines; this raises the issue of the accuracy of our model outputs.

Conclusion

This study estimated the wine quality score based on concentrations of key chemical features. For every unit increase of acetic acid (g/dm^3), our models predict a decrease in quality score between 1.71 and 2.04 points. In this model, we are only able to capture the linear relationships between chemical composition and the wine quality. For future experiments, to overcome the limitations introduced by our ordinal dependent variable and to capture non-linear relationships, we could use a more complicated model such as a neural network or a support vector machine (SVM). We believe this should help us achieve higher R^2 values. The ultimate goal of this line of research is to provide a data-driven model that can be generalized to other wines and provide an accurate tool to systematically assess quality of wine for industry professionals.