Move over Sommelier, Make way for a Statistician

Statistical Analysis of the Physicochemical Characteristics of Award-Winning Wines

Abstract

This study examines the relationship between physicochemical properties and wine quality ratings using data from the Vinho Verde Region. Through logistic regression analysis of both red and white wines, we identify key physicochemical characteristics that predict high-quality ratings from expert wine tasters. The models achieve 89.37% and 80.53% accuracy for red and white wines respectively, suggesting that objective physicochemical measurements can effectively predict subjective quality assessments. These findings have implications for wine production, quality control, and consumer education in the $330 billion global wine industry. Our results indicate that different chemical properties influence quality perception in red versus white wines, with factors like sulphates and chlorides playing crucial roles in both varieties.

1 Introduction

The 2024 Decanter World Wine Awards saw over 18,000 wines from 57 different countries competing to take home the title of best in show in their respective fields.[1] Wine is an extremely popular alcoholic beverage that is deeply entrenched in culinary and social traditions. Early archaeological discoveries place the origins of wine-making anywhere between six to eight thousand years ago, and since then the fight to find the "best tasting" wines has been fierce. Valued at over $330 billion dollars, the global wine making industry is massive and therefore competitions such as the illustrious Decanter World Wine Awards are extremely important for establishing a wine brand's supremacy.[2] The judges are industry experts including winemakers, retailers, and wine writers. However, due to the personal nature of rating high quality wines, it can appear to be a nebulous black box with which judges use to determine the best wines.

There are so many facets to a wine that give it its identity: the grapes used, the fermentation processes, and even the final chemical characteristics which can include the alcohol content, pH, sugar concentration, acidity, and a bevy of other characteristics. This research project specifically asks, "which physicochemical characteristics of a wine can best predict whether a wine will be considered high quality?" The study relies on the Wine Quality dataset available at the University of California Irvine.[3] This study takes a 10% random sample of the dataset and uses a logistic regression model with a binary response variable to see how well quality can be predicted.

2 Materials and Methods

This data was found in the University of California Irvine's Machine Learning Repository. Since wine characteristics vary so heavily across region, production method, and grapes used, this study will examine *vinho verde*, a specific type of wine from the northwestern Minho region of Portugal. While *vinho verde* comes in both red and white varieties, the majority of *vinho verde* that is sold consists of white wines. *Vinho Verde* is known for its crisp acidity, more subtle carbonation, and lower alcohol content which makes it a popular choice during the summer months. This data was collected from May 2004 to February 2007 only using protected designation of origin samples that were tested at official certification levels by the CVRVV, a regional Portuguese commission that promotes and ensures the quality of *vinho verde*.

The data was collected by the CVRVV and made available for research use through a team of scientists led by Dr. Paolo Cortez at University of Minho, Portugal. Since, white wines and red wines have distinct characteristics, Cortez et al. separated the data into white and red datasets in order to determine the relative importance of each chemical characteristic for each type of wine. This step is important as to not muddle the data and accurately determine how perception of quality changes across red and white wines.

The data consists of both qualitative and quantitative components. Quantitative data includes physicochemical characteristics such as pH levels, citric acid content, residual sugar levels, and alcohol content. Qualitative data was the assessments of "sensory assessors" that included blind taste tests from at least three different assessors, with the median score being taken. These quality ratings were placed on a continuous scale between 0 (very bad) to 10 (excellent).

Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar
Chlorides	Free Sulfur Dioxide	Total Sulfur Dioxide	Density
pH	Sulphates	Alcohol	

Figure 1: Physicochemical Properties

[1] Ofgang, Erik. "The Best U.S. Wine-According to the Decanter World Wine Awards 2024." Forbes, June 21, 2024.

[2] "Global Wine Strategic Business Report 2024: Market to Surpass $525 Billion by 2030" Business Wire, May 17, 2024.

[3] Cortez, Paulo, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. 2009. Wine Quality. UCI Machine Learning Repository.

3 Model

A logistic regression model was used by first initializing a new binary variable called *highquality* — which tests whether the original wine had a rating of 7 or higher. The choice of 7 as the cutoff was because the ratings for the wines clustered around 6 (which was both the median and the upper quartile). A small but significant number of both red and white wines were rated 7 or above – making it a revealing cut-off for analysis. The model predicts the probability, given a specific set of physicochemical characteristics, of the wine being high quality.

The logistic regression model started with 11 predictor variables mentioned above. For red wine, it was found that there were seven predictors of significance: fixed acidity, citric acid, chlorides, free sulfur dioxide, pH, sulphates, and alcohol content. Interestingly, the white wines had different predictors. Two new predictors, density and residual sugars were introduced whereas alcohol and citric acid were removed.

After the models were created and modified to balance predictive power, interpretability, and prevent overfitting, they were tested against the remaining split sets of data. The models were about 89.37% accurate for red wines and 80.53% accurate for white wines in predicting whether a wine would be considered high quality or not.

Red Wines	**Actually High Quality**	**Actually Not**	**White Wines**	**Actually High Quality**	**Actually Not**
Predicted High Quality	24	7	Predicted High Quality	104	65
Predicted Not	44	405	Predicted Not	221	1079

Figure 2: Confusion Matrices

The final equations for the red wine model was:

$$p(\text{highquality}=1)=\frac{1}{1+e^{-(\beta_0+\beta_1\text{fixed.acidity}+\beta_2\text{citric.acid}+\beta_3\text{chlorides}+\beta_4\text{free.sulfur.dioxide}+\beta_5\text{pH}+\beta_6\text{sulphates}+\beta_7\text{alcohol})}}$$

With only minor modifications, the final equations for white wine model was:

$$p(\text{highquality}=1)=\frac{1}{1+e^{-(\beta_0+\beta_1\text{fixed.acidity}+\beta_2\text{residual.sugar}+\beta_3\text{chlorides}+\beta_4\text{free.sulfur.dioxide}+\beta_5\text{density}+\beta_6\text{pH}+\beta_7\text{sulphates})}}$$

The correlation coefficients for each of the predictors were as follows:

Model	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Red	8.00	-0.77	6.14	-29.97	-0.13	-5.29	3.67	0.92
White	-15.40	0.07	0.03	-19.84	0.02	1.12	-0.09	0.75

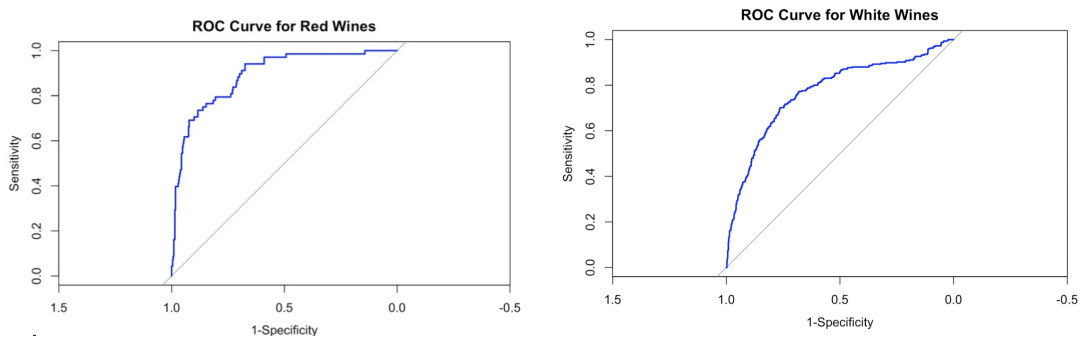

Figure 3: ROC Curves for Red and White Wine Analyses

Finally, a Receiver Operator Characteristic (ROC) curve was constructed for both models to show the diagnostic ability of the binary classifier. Since the area above the curve is clearly positive (Red Wine AUC: 0.89, White Wine AUC: 0.771), it can be inferred that the model can successfully differentiate between positive and negative classes.

4 Discussion

Intuitively, many of these characteristics make sense in the context of wine making. For example, sulphates in wine are important due to their antioxidant properties. Oxidation in wine causes a loss in flavor, aroma, and color. Furthermore, sulphates have antimicrobial properties which prevents unwanted bacteria from growing in the fermentation process. This allows winemakers to get the taste they want from their batches. Other characteristics such as chlorides add to the minerality of the wine, often found in more coastal areas. This added "saltiness" can actually draw out the sweetness of the wine. When examining the difference in predictors across red and white wines, residual sugars and density are likely more significant predictors of white wine due to the unique flavor profile of white wines. These wines tend to be sweeter and have more sugar. As a result, their densities are likely higher. On the other hand for red wines, sugar plays a smaller role in the perception of quality as they are predominantly dryer and characteristics such as acidity take precedence.

Wineries can draw several insights from this study. First, the current system of wine testing in Portugal and other countries relies heavily on human judges whose evaluations are prone to subjective biases. A physicochemical approach can increase objectivity of wine ratings and can be used to replace and supplement human judgement. Second, a physicochemical approach and the model can be used to improve the training on oenology students. Third, as Cortez et al. posit, the physicochemical preferences of a particular niche (and profitable) market can be documented through free wine sampling and rating, helping wineries improve product market fit. Finally, from a marketing perspective, communicating the effect on taste of these physicochemical properties can help consumers become more discerning about their own palates.

While the study offers valuable insights, its geographic scope—limited to a single wine region—may not capture the full spectrum of wine variations. Since growing conditions, traditions, and practices differ substantially across wine regions, conclusions drawn from this localized sample may not hold true elsewhere. Expanding the analysis to encompass multiple wine regions would strengthen the broader applicability of the findings. Additionally, the binary classification was based on a rating threshold of 7 or higher to define "high quality." This threshold is, of course, subjective and could affect the generalizability of the model. An in-depth sensitivity analysis on different thresholds might provide a more robust understanding.

Future researchers can extend this study in a few different directions. First, a future study may involve incorporating additional wines from other regions and exploring if these results stay consistent. The null hypothesis should be that the chemical characteristics should not change too drastically. Second, since this data was collected in 2004-2007, it would be interesting to learn how physicochemical preferences have evolved since then. Third, a study featuring subjective ratings by casual customers rather than experts in wine might reveal a curious divide in taste preferences. Finally, while this study examined the effect of physicochemical properties on subjective taste, it would be beneficial to know more about the environmental factors (growing conditions, soil composition, and climate) that influence these properties. Despite these limitations and possibilities of improvement, this study effectively demonstrates that physicochemical properties can serve as reliable predictors of a wine's perceived quality, enabling wineries and consumers to make well-informed decisions.

References

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553. doi:10.1016/j.dss.2009.05.016

[2] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Wine Quality Dataset. UCI Machine Learning Repository. doi:10.24432/C56S3T

[3] Ofgang, E. (2024, June 21). The Best U.S. Wine—According to the Decanter World Wine Awards 2024. *Forbes*. Retrieved from `https://www.forbes.com/sites/erikofgang/2024/06/19/the-best-us-wine-according-to-the-decanter-world-wine-awards-2024/`

[4] Business Wire. (2024, May 17). Global Wine Strategic Business Report 2024: Market to Surpass $525 Billion by 2030. Retrieved from `https://www.businesswire.com/news/home/20240517170997/en`