STA 210: Final Project

Uncorking the Relationship Between Physicochemical Properties and Perceived Wine Quality

Matthew Rui • Rohit Suresh

Analysis

Introduction and Data

Wine quality evaluations play an important role in numerous parts of the wine supply chain. They are important for quality assurance, market differentiation, and consumer guidance. Generally, a wine quality evaluation analyzes appearance, aroma, flavor, structure, and overall balance to generate a score from 3 to 9. Physicochemical properties of wine samples are readily available to most manufacturers, and quantifying the relationship between physicochemical properties and wine quality evaluations could be of great interest to manufacturers. It may help them with quality control, product development, and cost efficiency [1].

However, it's unclear what makes a "high quality" wine, whether quality is a meaningful designation, and if quality ratings are based on the physicochemical properties of wine or rooted in some other arbitrary qualitative judgment. As such, we're interested in studying the relationship between color, alcohol content, and other physicochemical properties and the quality rating of wines. **We hypothesize that color and wine have a linear relationship with the log-odds of being rated as "high quality."** Furthermore, **we're interested in predicting the quality rating of a wine as accurately as possible**, as a highly accurate classifier model implies that "quality" ratings are labeled in a manner consistent enough for us to predict. If we're unable to fit an accurate prediction model, it implies that perhaps our data is too noisy to predict, and perhaps that quality ratings are assigned inconsistently.

For this study, we utilize data from the UC Irvine Machine Learning Repository [2]. Datasets for white and red wine were provided separately from this source, which we combined into one dataset, in the process creating a new variable for the color of wine (color). The other variables in the dataset refer to specific physicochemical properties of wine commonly analyzed by manufacturers. Of greatest interest to our analysis, quality refers to the median of three quality ratings from 3 (poor) to 9 (excellent) by humans. Alcohol (% vol) refers to the alcohol content of the wine. A complete list of variable names and associated summaries are listed in the appendix [Appendix A].

We find that red wines, on average, have lower quality scores than white wines.

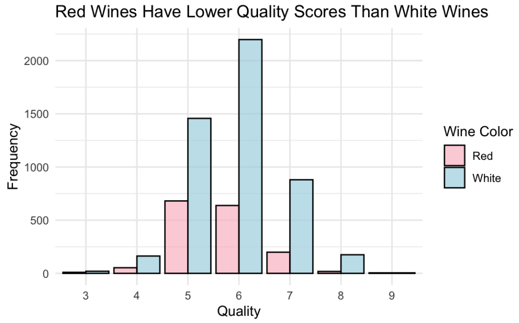


Figure 1: Distribution of Quality Ratings by Color

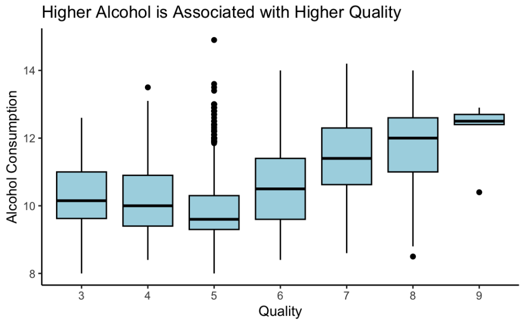


Figure 2: Box Plot of Alcohol and Quality

Figure 2 seems to indicate that different quality ratings are associated with differing alcohol levels, with a generally increasing trend from quality 5 and onward.

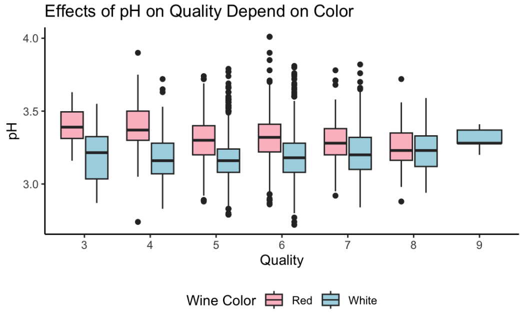


Figure 3: Box Plot of pH, Color, and Quality

pH also declines with increasing quality for red wine, but pH increases with increasing quality for white wine (not enough data was present to plot red wine with quality 9). This suggests that **we should also study the interaction between our various parameters**. While this will result in a more complex model, it might result in a more accurate picture of the data.

For our study, we establish a threshold for high quality vs low quality by analyzing the distribution of quality scores.

Analysis

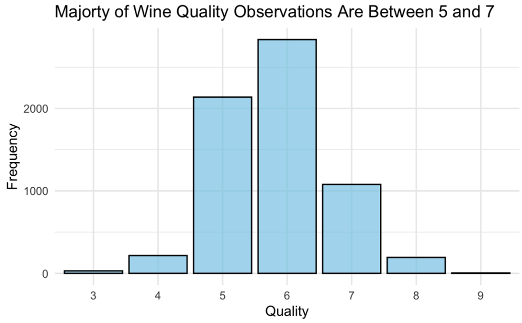


Figure 4: Distribution of Quality Ratings

The majority of samples had a quality rating below or equal to 6. Samples with quality scores greater than 6 were classified as high quality, signifying that this wine was rated better than the median wine. Samples with quality scores less than or equal to 6 were classified as low quality. There were 1277 samples marked as high quality, and there were 5220 samples marked as low quality.

Once again, simple exploratory analysis hints that the relationship between quality and alcohol might explain some of the distinction.

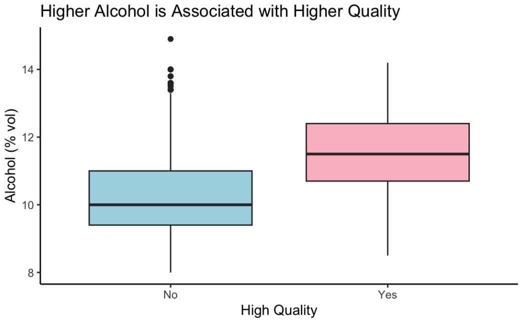


Figure 5: Box Plot of Alcohol and High Quality

Methodology

With color of wine (categorical), alcohol, and the physicochemical properties of wine (fixed and volatile acidity, citric acid, residual sugar, ln(chlorides), ln(free sulfur dioxide), total sulfur dioxide, density, pH, and sulfates) as predictors, we fit a logistic regression to estimate the log odds of any given wine being either high or low quality. We used empirical logit plots to assess the assumptions of this logistic model [Appendix B].

To create our classifier, we utilize the AOC curve of this logistic model to identify the threshold that maximizes the sensitivity and specificity. We also fit another logistic model with an interaction term between color and every other predictor and similarly find the optimal threshold for a classifier with the AOC curve.

Finally, we train and tune a random forest model to classify wines into high/low quality with the same predictors as the logistic models. We tune this model with k-fold cross validation with 10 folds and 500 decision trees.

Utilizing the same tuning methodology and predictors, we train another random forest to classify wines into its original quality category, rather than high/low quality.

Key Findings and Discussion

Alcohol and Color have a linear relationship with the log-odds of being "High Quality"

We first fit a logistic regression to predict whether a given observation is classified as "high quality", defined as receiving a quality rating of above 6.

Predictor	Odds Ratio	P-value
Fixed Acidity	1.624	**<0.001**
Volatile Acidity	0.025	**<0.001**
Citric Acid	0.775	0.460
Residual Sugar	1.232	**<0.001**
ln(Chlorides)	0.598	**<0.001**
ln(Free Sulfur Dioxide)	1.516	**<0.001**
Total Sulfur Dioxide	0.995	**<0.001**
Density	<0.001	**<0.001**
pH	12.541	**<0.001**
Sulfates	11.001	**<0.001**
Alcohol	1.606	**<0.001**
Color (White)	0.440	**<0.001**

Figure 6: Logistic Regression Estimates

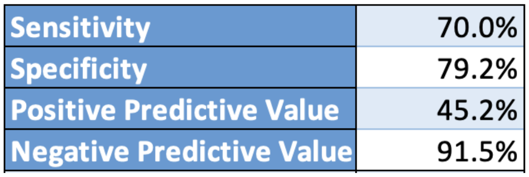
We find that every predictor except for citric acid has a linear relationship with the log-odds of a wine being classified as "high quality" at a 0.05 significance level. Interestingly, white wine has an odds-ratio lower than 1, indicating that **red wine, holding the physicochemical content and alcohol of wine constant, is generally classified as higher quality than white wine**. This directly contradicts our exploratory analysis (Figure 1), which seemed to indicate that white wines were generally rated higher quality, confirming our suspicion to further examine and study interactions between predictors.

pH and sulfates were associated with the largest odds-ratios. This is unsurprising, as pH measures the acidity of wine, which impacts the perceived taste and thus "quality" of wine. In fact, wines only range from 2.9 to 4.2 in pH, with winemakers preferring wines with pH values between 3.0 and 3.5. [3] Thus, while the odds-ratio is large,

Analysis

pH is unlikely to provide meaningful insight into how "quality" is determined.

Sulfate having a large odds-ratio is also unsurprising. Yeast naturally produces sulfates during fermentation and they are also used as a preservative, specifically to protect against oxidation [4]. Excessive oxidation can distort the color and taste of wine, clearly hampering the "quality" of wine [5]. While this large odds-ratio is expected, it provides more meaningful insights than pH, as sulfate values range from 0 (sulfate-free wines) to 350 mg, the legal maximum in the United States [6].



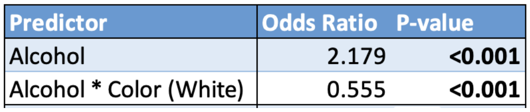
Sensitivity	70.0%
Specificity	79.2%
Positive Predictive Value	45.2%
Negative Predictive Value	91.5%

Figure 7: Logistic Classifier Accuracy

We use this logistic model to create a binary classifier. With a positive predictive value of 45.2%, a predicted "high quality" wine is actually more likely to be "low quality", implying a large number of false positives in this model. The decently high specificity and negative predictive value suggests that, unlike predicting high quality wines, this binary classifier excels at predicting low quality wines.

We suspect that this might be due to the **imbalanced outcome classes of the data**, as there are over 4 times more "low quality" wines than "high quality" counterparts. We suggest future research directions to address this issue later in this paper.

The mixed results of this classifier suggest that further refinements are necessary to identify high quality wines. As such, we examine how the color of wine might affect the relationship between other predictors and the log odds of being high quality.

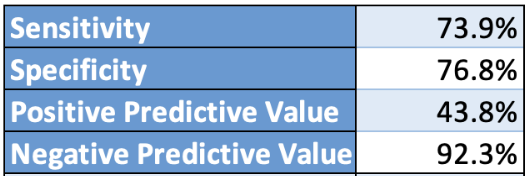


Predictor	Odds Ratio	P-value
Alcohol	2.179	**<0.001**
Alcohol * Color (White)	0.555	**<0.001**

Figure 8: Selected Logistic-Interaction Estimates

We find that the **relationship between alcohol and the log odds of being predicted high quality depends on the color of the wine**. This indicates that the standards of being "high quality" depends on the color of the wine. This goes beyond our suspicions from Figure 1 and findings from the non-interaction logistic model: **"quality" does not only simply depend on the color of the wine, but rather also that the desired attributes of "high quality" wine also depend on the color of the wine**. This interaction effect can also be seen in other predictor variables, such as sulfates, pH, and density. However, for many other predictors, such as acidity and residual sugar, there was insufficient evidence to conclude a significant interaction effect [Appendix C].

Just as before, we create a binary classifier using this logistic-interaction model. It performs marginally better than the previous logistic binary classifier in sensitivity and negative predictive value, but worse in specificity and positive predictive value.



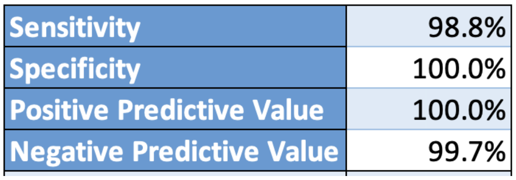
Sensitivity	73.9%
Specificity	76.8%
Positive Predictive Value	43.8%
Negative Predictive Value	92.3%

Figure 9: Logistic-Interaction Classifier Accuracy

Once again, the interaction model presents a poor positive predictive value. This marginal change from the logistic classifier questions whether this dataset might be perhaps too noisy to classify wines significantly better than these logistic classifiers. While these logistic models have provided insights into the specific predictors and interactions that are linearly related to the log-odds of being rated "high quality, " we conclude that another approach is necessary to most accurately predict the quality of wines in this dataset.

Random Forest suggests consistent "quality" ratings are consistently assigned

For the remainder of this study, we utilize random forest models to construct a classifier to serve as comparison for our logistic-based classifiers.



Sensitivity	98.8%
Specificity	100.0%
Positive Predictive Value	100.0%
Negative Predictive Value	99.7%

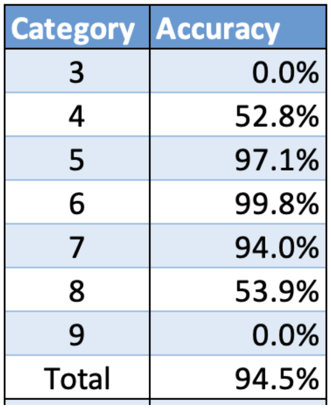
Figure 10: Random Forest Binary Classifier Accuracy

Unfortunately for our logistic classifier, **the random forest achieves near-perfect accuracy**. From the sensitivity and specificity, the random forest classifier correctly predicts nearly all high quality wines and all low quality wines. The positive and negative predictive values suggest that among the wines predicted to be a certain quality, they actually were that quality. This near-perfect accuracy indicates that wine quality is assigned in a consistent manner, at least consistent enough for prediction models to be accurately trained.

We provide two possible explanations for this much improved performance: First, from our previous interaction model, we concluded that there were many predictor variables for which we had insufficient evidence to conclude interactions effects to be significant. **A random forest classifier is able to naturally capture interactions between terms**, which could explain some of the improved accuracy demonstrated by this model. Moreover, we only consider interactions with wine color in the logistic model, whereas the **random forest considers all possible interaction effects**.

Second, and more interestingly, this might indicate that **decision trees provide a much more accurate way of classifying wine quality**. This makes intuitive sense, as human judgment is much more similar to a decision tree, rather than a logistic model's regression approach, where each predictor variable incrementally increases the odds of being "high quality."

The random forest performs so well, we wonder if the same random forest classifier can correctly identify the original quality category of each wine, not just whether they're rated "high" or "low" quality.



Category	Accuracy
3	0.0%
4	52.8%
5	97.1%
6	99.8%
7	94.0%
8	53.9%
9	0.0%
Total	94.5%

Figure 11: Random Forest Classifier Accuracy

While the total accuracy of 94.5% may seem impressive, the random forest classifier predicts no observations to be rated 3 or 9, and is only marginally better than a coin flip for classes 4 and 8. This is primarily attributed to the skewed quality classes of this imbalanced dataset. Future studies could focus on applying up/downweighting techniques to improve the accuracy of any classifier. We're also optimistic that stratified (between the majority and minority classes) k-fold cross validation could lead to better parameter tuning for the random forests.

Conclusion

The analysis presented within this work serves to highlight the complex physicochemical relationships with quality ratings of wines. Our exploratory analysis indicated that there might exist a relationship between the color of wine and quality rating, as we noticed higher quality scores for red wines compared to white wines. We also hypothesized that color could be an interaction factor, as we noticed that the relationships between pH and quality seemed dependent on color.

In our analysis, we utilize logistic regression models with and without interaction terms for wine color. From the non-interaction model, we were able to conclude a **relationship between sulfates and pH with wine quality** consistent with scientific explanations for the taste of wine. We also found a **relationship between color and quality of wine**. Despite this, this model had a poor sensitivity and positive predictive value when attempting to classify wines. The interaction model did not perform much better at classification; however, it enabled us to conclude that the **relationship between certain predictors, such as sulfates and pH, and wine quality depend on wine color**.

Understanding that the relationships between the predictors in the dataset might be complex, we moved to use a random forest classifier, which achieved near perfect accuracy. This could be attributed to complex relationships between interaction terms uncaptured by our logistic models and hints that **decision trees might inherently better capture and emulate human decision making**. Taking it a step further, we looked at the accuracy of the random forest when predicting the original quality labels. While it was able to very accurately predict between ratings 5 and 7, the random forest struggled outside of that range. We believe that this analysis could benefit from a more balanced dataset that could improve model performance.

Despite this, our analysis offers an accurate model to predict whether a wine is high quality and has begun to reveal some insights about the interactions between color and physicochemical predictors. Future studies could focus on further elucidating interactions between predictors and developing better models to predict the original quality rating of a wine.

References & Appendix

1. Bhardwaj, Piyush, et al. “A machine learning application in wine quality prediction.” *Machine Learning with Applications*, vol. 8, June 2022, p. 100261, https://doi.org/10.1016/j.mlwa.2022.100261.
2. “Wine Quality.” *UCI Machine Learning Repository*, archive.ics.uci.edu/dataset/186/wine+quality.
3. Cole-Palmer. “Testing the Ph Value of Wine.” *Testing the pH Value of Wine*, www.coleparmer.com/tech-article/measuring-ph-in-wine-making. Accessed 30 Apr. 2024.
4. Ajmera, Rachael. “Sulfites in Wine: Uses and Side Effects.” *Healthline*, Healthline Media, 9 Sept. 2019, www.healthline.com/nutrition/sulfites-in-wine.
5. JJ Buckley Fine Wines. “What Is Oxidization and What Is It Doing to Your Wine?” *What Is Oxidization and What Is It Doing to Your Wine?*, www.jjbuckley.com/wine-knowledge/blog/what-is-oxidization-and-what-is-it-doing-to-your-wine-/1015#. Accessed 29 Apr. 2024.
6. Wine Folly. “The Bottom Line on Sulfites in Wine.” *The Deal with Sulfites in Wine*, winefolly.com/deep-dive/sulfites-in-wine/. Accessed 29 Apr. 2024.

References & Appendix

A: Summary of Dataset Variables.

- Fixed acidity: Estimation of fixed acids with the concentration of tartaric acid in $g_{\text{tartaric acid}}/dm^3$
- Volatile acidity: Estimation of volatile acids with the concentration of acetic acid in $g_{\text{acetic acid}}/dm^3$
- Citric acid: Concentration of citric acid in g/dm^3
- Residual sugar: Concentration of sugars in g/dm^3
- Chlorides: Estimation of chlorides with the concentration of sodium chlorides in $g_{\text{sodium chloride}}/dm^3$
- Free sulfur dioxide: Concentration of free sulfur dioxide in mg/dm^3
- Total sulfur dioxide: Concentration of total sulfur dioxide in mg/dm^3
- Density: Density of liquid in g/cm^3
- pH: Standard measure of pH
- Sulfates: Estimation of sulfates with the concentration of potassium sulfate in $g_{\text{potassium sulfate}}/dm^3$
- Alcohol: Concentration of alcohol in *% vol.*
- Quality: Median of three quality ratings from 3 (poor) to 9 (excellent) by humans
- Color: Red or white wine

B: Examining the assumptions of the logistic model.

We fit empirical logit plots for each predictor of our logistic regression. We chose to log transform Chlorides and Free Sulfur Dioxide to better adhere to our assumption of linearity.

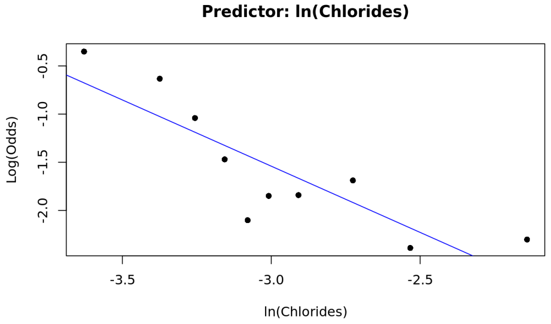


Figure 12: Empirical Logit Plot of ln(Chlorides) for Logistic Model

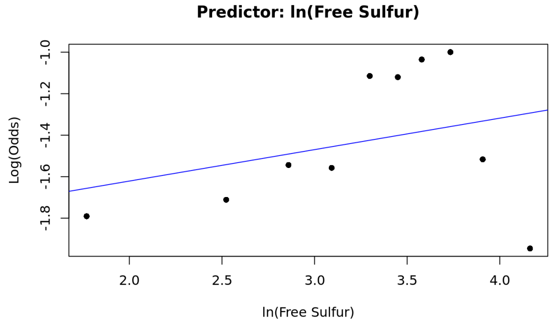


Figure 13: Empirical Logit Plot of ln(Free Sulfur) for Logistic Model

Citric Acid, Residual Sugar, Total Sulfur Dioxide, and Sulfates had questionable empirical logit plots. However, transformations were not applied as the transformed empirical logit plots satisfied the linearity assumptions of logistic regression worse than the untransformed empirical logit plots. The omitted empirical logit plots can be reproduced from the source code of this study.

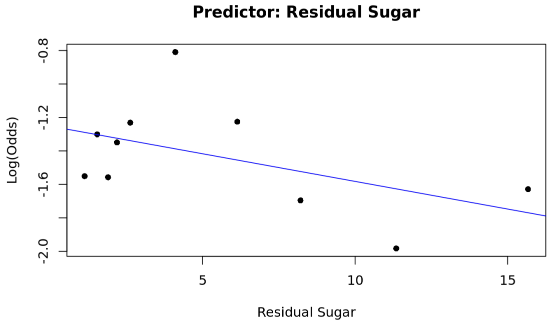


Figure 14: Empirical Logit Plot of Residual Sugar for Logistic Model

C: Complete Estimates for Logistic-Interaction Model

Predictor	Odds Ratio	P-value
Fixed Acidity	1.307	**0.031**
Volatile Acidity	0.071	**< 0.001**
Citric Acid	1.579	0.579
Residual Sugar	1.264	**0.002**
ln(Chlorides)	0.353	**0.002**
ln(Free Sulfur Dioxide)	1.204	0.293
Total Sulfur Dioxide	0.982	**< 0.001**
Density	< 0.001	**0.055**
pH	1.168	0.876
Sulfates	37.338	**< 0.001**
Alcohol	2.179	**< 0.001**
Color (White)	> 1e16	**0.009**
Fixed Acidity * Color (White)	1.307	0.082
Volatile Acidity * Color (White)	0.429	0.356
Citric Acid * Color (White)	0.298	0.188
Residual Sugar * Color (White)	0.963	0.647
ln(Chlorides) * Color (White)	1.600	0.200
ln(Free Sulfur Dioxide) * Color (White)	1.547	**0.033**
Total Sulfur Dioxide * Color (White)	1.014	**0.005**
Density * Color (White)	< 0.001	**0.009**
pH * Color (White)	21.115	**0.005**
Sulfates * Color (White)	0.228	**0.019**
Alcohol * Color (White)	0.555	**< 0.001**

Figure 15: Complete Logistic-Interaction Estimates