Red Wine Quality Analysis

By: Jayani Tripathi, Shelby Graf, Jackson Brigham

December 10, 2023

Table of Contents

Project Description

The overall objective of our project is to identify which red wine characteristics have the greatest impact on wine quality. To start, we sourced a dataset on Kaggle called “Red Wine Quality” that contains information about red variants of the Portuguese "Vinho Verde" wine.

In a business context, our analysis could be helpful to anyone interested in the wine industry and learning more about the relationship between wine characteristics and overall quality. Wine collecting is a lucrative business, so the results of our predictions could be useful for someone looking to create a high-quality wine portfolio in the evolving wine industry. Ultimately, our analysis is targeting winemakers, distributors, vendors, and sommeliers. Our results will give them a better understanding of what red wine attributes are most important to consumers.

We plan to perform both a logistic regression and a multiple linear regression. Logistic regression provides easily interpretable results that can be applied in a business context. If a winemaker can see which variables are strongly correlated with low or high wine quality, they would know what characteristics of their wine to prioritize. They could ask questions like “Does alcohol content have more of an impact on quality than sulfates?” to create more high-quality wines.

We will assess the accuracy of the logistic regression model using error metrics. Even if the model makes some incorrect classifications, it is important to note that it would be better for the model to mistake a high-quality wine as low quality. High-quality wine typically costs more and consumers would be more upset about paying for a high-quality wine and receiving low quality, as opposed to paying a low price for high-quality wine.

The purpose of multiple linear regression is to fit a relationship between a dependent variable and a set of independent variables within the dataset. The goal of this predictive modeling technique is to accurately predict new cases. Because we transformed our output variable, quality, into a dummy variable, a multiple linear regression will make a prediction of what will be a good or bad quality red wine. The information produced by this model would be most helpful to someone looking to build a robust wine portfolio. Wine collectors likely will be collecting red wines to age them for years to come, and therefore would use the information from our analysis to predict what wines would be best to invest in, as they will not be tasting them.

Data Preprocessing

Cleaning, Transforming, and Partitioning

Our original dataset was clean and organized and we did not encounter any missing rows or values. The dataset has about 1,600 rows of data and includes variables related to red wine such as pH, sulfates, and alcohol content. We did perform some transformations to the data to simplify it before performing our classification and prediction.

Our output variable, quality, was originally formatted in the dataset as a column containing discrete values from 1 to 10 which represented the score given to the wine based on sensory data. First, we decided to use the “reduce categories” feature in Analytic Solver to transform quality into a categorical dummy variable. Having a binary output variable will ultimately simplify the classification and prediction models we plan to run. We categorized any quality scores from zero to six as a 0, to indicate a poor quality red wine. We assigned quality scores from six to ten as a 1, indicating a high quality red wine. The new output variable column is called Reduced_quality.

Since the average person likely does not know what is considered a low, medium, or high alcohol content in wine, we believed it would be more intuitive to categorize this variable. We researched alcohol content in red wines and found that the general consensus is: wines that consist of a low alcohol content level have an ABV at or below 12.5%; medium alcohol wines have an ABV between 12.5 to 13.5%; and high alcohol wines have an ABV above 13.5. We used the same category reduction process discussed above to turn alcohol content into a categorical variable. We assigned low alcohol values as 0, medium as 1, and high as 2. The new column name is Alcohol Level.

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5

Figure 1.1: Original Columns Before Preprocessing

Record ID2	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	Alcohol Level	Reduced_Quality
Record 1	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	0	0
Record 2	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	0	0
Record 3	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	0	0

Figure 1.2: Columns After Data Preprocessing

Finally, we partitioned the dataset into 60% training and 40% validation. There are 959 records in our training set and 640 in the validation set. Data partitioning is an essential step in supervised machine learning. Partitioning the data helps to prevent overfitting and helps the model perform better because it divides the dataset into smaller, more manageable subsets often used for training, validation, and testing. Partitioning allows us to analyze and cross-validate our model's performance on the different sets to investigate if the model shows metrics/evidence indicative of overfitting. This is pertinent because an unfit model would not be suitable for generalized use on unseen data, which would not be optimal given our business context.

Visualizations and Correlations

Based on the classification coefficient output that we will further analyze in the results section of our report, we chose two important predictor variables to visualize: density and sulfates. First, we can see that density and quality are inversely related. As density decreases, the quality score increases. On the other hand, sulfates and quality are positively correlated. As sulfates increase, the quality score also increases.

Lastly, we wanted to see a count of how many total good and bad quality wines we have in our dataset. Good-quality wines classified as 1s only make up 14% of our data, while bad-quality wines classified as 0s make up the other 86%. Because our goal is to understand what traits make up a good-quality wine, this statistic is important to keep in mind as we analyze the results of our models. The uneven distribution of good and bad-quality wines in the dataset may skew our results, indicating that this may not be the best dataset to use when assessing good-quality red wines.

Figure 1.3: Average Density per Quality Score

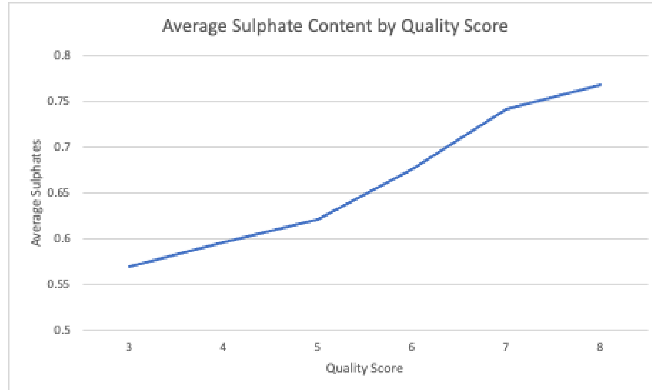

Figure 1.4: Average Sulfate Content per Quality Score

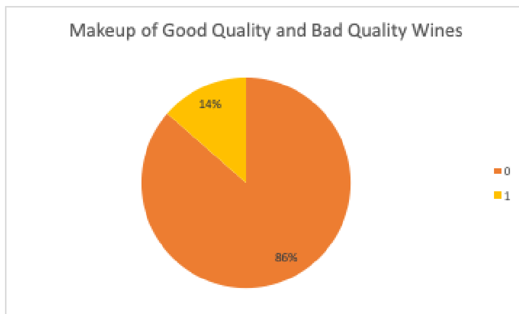

Figure 1.5: Total Makeup of Good & Bad Quality Wines

Models and Analyses

In our analysis of wine quality classification, we opted to utilize both logistic regression and multiple linear regression using Excel's Analytic Solver. These regression techniques cater to distinct scenarios and business problems.

Model #1: Logistic Regression

Logistic regression is useful for our analysis because the dependent variable is binary, representing two classes (0 or 1). It models the probability of wine being classified as good or bad quality. The output of logistic regression is a probability transformed using the logistic function (sigmoid function) to ensure it falls between 0 and 1. Logistic regression assumes a linear relationship between the independent variables and the "log-odds" of the event. It also assumes that the observations are independent and that there is no perfect multicollinearity (we addressed this by using category reduction instead of creating multiple dummy variables from one column). Example applications include predicting probability based on influencing factors, which aligns well with our project objectives. Logistic regression provides results and metrics in terms of odds ratios, making it easy for us to interpret the impact of each predictor variable (ex: sulfates) on the likelihood of a wine being of good quality.

Logistic regression calculates the probability of a specific event occurring, which, in this case, is the probability of a red wine being classified as good quality. This probability can be useful for decision-making, especially in the context of winemaking where understanding the factors contributing to quality is crucial for the selection of wines for vendors or other contexts. Logistic regression can help us identify the most influential predictor variables in determining wine quality, which is highly valuable for winemakers or vendors who may want to prioritize specific characteristics in the production/procurement process.

Creating a clear decision boundary between the two classes, logistic regression makes it straightforward to classify wines as either good or poor quality based on a given set of predictor variables, we chose to utilize all variables. Logistic regression provides metrics such as accuracy, precision, recall, and the area under the ROC curve, which are essential for evaluating the model's performance as well as determining the model's ability to perform on unseen data.

In summary, logistic regression is an optimal technique for our analysis as it works well with the binary output variable. Logistic regression provides interpretable results that can guide winemakers in improving the quality of their red wines. If the model cannot perform well on unseen data, we cannot assure our stakeholders that the model should be used to make business decisions.

Model #2: Linear Regression:
We chose to utilize linear regression, as it is useful when the dependent variable is continuous (quality) and can take any value within a given range (0 -1 for our binary categorical variable of quality). We utilized linear regression because it is also suitable for predicting numeric outcomes/probabilities. Our model predicts a linear relationship between the independent variables (all variables) and the continuous outcome (categorical quality). Linear regression assumes that the relationship between the independent and dependent variables is linear and that the residuals are normally distributed. Example applications include predicting outcomes based on historical data, which is highly applicable to our context and to our goals.

In the context of analyzing red wine quality, linear regression may not be the most suitable technique, however, if we consider our business context, linear regression could have the following useful purposes:

Linear regression can be used to model the relationship between the independent variables (wine characteristics) and the continuous outcome (reduced_quality), providing insights into how changes in the predictor variables are associated with changes in the outcome. By examining the coefficients in the linear regression equation, we can also identify which wine characteristics have a significant impact on the predicted outcome. This is valuable for understanding the importance of different red wine attributes associated with quality.

Performance metrics such as Mean Squared Error (MSE) or R-squared can be used to evaluate the goodness of fit of the linear regression model. These metrics provide insights into how well the model explains the variability in the continuous outcome (quality).

<u>Predictors included in both models:</u>

1. Fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

2. Volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3. Citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines
4. Residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter, and wines with greater than 45 grams/liter are considered sweet
5. Chlorides: the amount of salt in the wine
6. Free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
7. Total sulfur dioxide: the amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
8. Density: the density of water is close to that of water depending on the percent alcohol and sugar content
9. pH: describes how acidic or basic a given wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
10. Sulfates: a wine additive that can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
11. Alcohol level: the alcohol content of the wine

Feature Selection

We did not perform a feature selection on either model. This might have been helpful if we thought we had irrelevant or redundant variables that would not have much of an impact on our output variable, wine quality. However, there are no input variables we initially assumed would have no impact on the outcome. Our dataset is not very big, with only 11 input variables. We concluded that it would be valuable to see each variable's impact on the outcome as we believe they are all relevant.

Results and Discussion

Model #1: Logistic Regression

For utilization of our categorical classification of red wine quality, logistic regression is the most useful technique to predict the classification of a given wine's quality based on the selected variables. The figures below show the performance of the model on the partitioned training and validation datasets:

Training: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	799	29
1	93	38

Error Report

Class	# Cases	# Errors	% Error
0	828	29	3.502415459
1	131	93	70.99236641
Overall	959	122	12.72158498

Metrics

Metric	Value
Accuracy (#correct)	837
Accuracy (%correct)	87.27841502
Specificity	0.964975845
Sensitivity (Recall)	0.290076336
Precision	0.567164179
F1 score	0.383838384
Success Class	1
Success Probability	0.5

Figure 3.1:Training Classification Summary

Validation: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	535	19
1	57	29

Error Report

Class	# Cases	# Errors	% Error
0	554	19	3.429602888
1	86	57	66.27906977
Overall	640	76	11.875

Metrics

Metric	Value
Accuracy (#correct)	564
Accuracy (%correct)	88.125
Specificity	0.965703971
Sensitivity (Recall)	0.337209302
Precision	0.604166667
F1 score	0.432835821
Success Class	1
Success Probability	0.5

Figure 3.2:Validation Classification Summary

To begin, we performed a pre-attentive quantitative assessment of the metrics to analyze how we should approach the analysis.

True Positive Rate vs. False Positive Rate
First, we analyzed the classification summary for the validation data. Figure 3.2 shows the confusion matrix, which includes the true positive rate (sensitivity) and false positive rate (1 – specificity), using the validation to ensure accurate metrics. From this matrix, we can see that the model is classifying 535 of 554 low-quality wines correctly as low quality, which represents a percent error of only 3.43%. The model is classifying only 29 out of 86 cases as high quality, with a percent error of 66.38%, indicating that this model is not great at accurately classifying high-quality red wines.

These metrics are indicative that our model has a good FPR for our business context, but the TPR may be subjective. However, it is important to acknowledge that there is often a trade-off between TPR and FPR for classification models.

We decided that the most important metric for our business context is TPR, or sensitivity/recall. TPR represents the ratio of true positive predictions to the total number of actual positive instances. It measures the model's ability to correctly classify high-quality wines (1s), as high-quality wines (1s). FPR is also important and represents the ratio of false positive predictions to the total number of actual low-quality classifications. FPR allows us to interpret the model's tendency to incorrectly predict high-quality wines as low-quality wines.

The matrix shows that our model has an overall strong performance, but we still need to analyze the meanings behind the metrics besides TPR and FPR and the context/influencers behind the results to see if the model is optimal for use on unseen data to predict quality.

Understanding Performance Metrics
The poor performance of the TPR might seem to be a bad thing at first, but given our business context, it is intuitive to optimize the model to focus on correctly predicting bad quality, as there is a marginally larger consequence for mispredicting low-quality wines as high-quality. It is much less consequential for our stakeholders/target audience to predict a high-quality wine as low compared to predicting a low-quality wine as high. This is because high-quality wine typically costs more and consumers would likely be more upset about paying for high-quality wine and receiving low quality, as opposed to paying a low price for high-quality wine.

It is safe to assume that there are evident sunk costs associated with misclassifying ~75% of high-quality wines as low quality, however, this will keep the end consumer happy and potentially assist in mitigating risks with customer/end-user satisfaction (offering good value). These metrics are also representative of the fact that in the wine industry, some wines with characteristics of low-quality wine will still be classified as high-quality wine, and this is evident in the scores of the dataset we chose.

There are some cheaper wines that consumers would still classify as being a high-quality score due to their taste/satisfaction. There are plenty of discount wines with lower-quality chemical makeup that still receive high-quality ratings from consumers. Nearly all wine consumers have a favorite cheap wine that they would consider high quality although its variables associated with it would be indicative of a low quality classification. There are also certainly high-quality wines that a majority of consumers may consider to taste like lower quality, or maybe they consider it low quality for value. Both of these scenarios can explain why an optimal model should have very few false negatives.

The intuitive assumption that consumers could potentially give quality scores that would categorize wines as lower quality when perceived as expensive or lacking in value adds complexity to the classification task. This extra layer of analysis emphasizes the significance of comprehending consumer perceptions from market research/tacit knowledge of the industry when interpreting our metrics related to wine quality.

Further Analyzing Metrics

After examining the performance of the model on the training dataset, we can deduce that the model performs similarly on both the validation and training sets. This indicates that our model is not severely overfitting, but we still need to perform a more in-depth analysis of the validation metrics to determine if there are other pertinent metrics to compare between both sets that may be indicative of the model overfitting results.

The number of correctly classified instances in the validation set is 564 of 640 records. These are instances where the model's predictions align with the actual outcomes. Most of these belong to the 0 class.

Our model achieved an accuracy of 88.125%, indicating relatively high overall correctness in predictions. However, it is still crucial to consider context when analyzing this metric. The accuracy performance of 1s and 0s is much different. The specificity or True Negative rate of 96.57% suggests a high accuracy in identifying instances of low-quality wine, which is ideal for our business context. The performance for the sensitivity of 33.72% indicates that the model captures only about 1/3 of the high-quality wines in the validation set, however, the tradeoff must be considered. A precision of 60.42% suggests that when the model predicts a wine as high quality, it is correct approximately 60.42% of the time. The F1 score of 43.28% indicates a trade-off between precision and recall, considering both false positives and false negatives.

In summary, while the accuracy is relatively high, the trade-off between precision and recall, as reflected in the F1 score, suggests we need to further examine the model's ability to capture high-quality wines. Adjusting the threshold probability or exploring additional features may help enhance the model's performance.

Further TPR Analysis

A high sensitivity (close to 1) is desirable in scenarios where the cost of missing a positive case (false negative) is high, and the goal is to minimize the chances of overlooking positive instances. This is true of our context, but having a low FPR is more pertinent to our application's success.

Sensitivity is often in trade-off with specificity (True Negative Rate). Improving sensitivity might lead to an increase in false positives, and vice versa. The choice between sensitivity and

specificity depends on the goals and constraints of specific applications, and our application has higher consequences for having a bad FPR as opposed to a lower TPR.
Although a balanced set of metrics provides a more nuanced evaluation, it is acceptable to have a low level of sensitivity; it is entirely dependent on the specific application and the consequences of false negatives, which is high in our case of application.

Based on these findings, we would highly recommend that any potential user of this model does an in-depth analysis of which metric(s) is/are most pertinent for their business context. If you are a wine sommelier and the most important thing for your applications is to find the best quality wines, we would advise that the model should be reoptimized for a high performance of TPR. However, as stated above, there is almost always a trade-off between TPR and FPR.

Predictor Variable Analysis

Predictor	Estimate	Magnitude	P-Value
density	-809.5439242	809.5439242	2.339E-12
Intercept	786.1077017	786.1077017	2.2736E-12
chlorides	-7.161506511	7.161506511	0.03499094
sulphates	4.861521603	4.861521603	1.0698E-10
pH	2.767380908	2.767380908	0.01985282
citric acid	1.669925839	1.669925839	0.11262594
volatile acidity	-1.430469951	1.430469951	0.13108272
fixed acidity	0.738645769	0.738645769	7.4838E-07
residual sugar	0.44808735	0.44808735	3.9939E-08
Alcohol Level	0.212536509	0.212536509	0.52279361
total sulfur dioxide	-0.014481157	0.014481157	0.00877052
free sulfur dioxide	-0.001139557	0.001139557	0.93792461

Next, we used the logistic regression output to assess which predictor variables were statistically significant and their strength of correlation with the output variable.

Statistically Significant Predictor Variables ($p<0.05$): Density, intercept, chlorides, sulfates, pH, residual sugar, and total sulfur dioxide.

Figure 3.3: Classification Coefficient Output

Predictor variables in order of strength of correlation with output variable (descending):
1) density, 2) intercept, 3) chlorides, 4) sulfates, 5) pH, 6. citric acid, 7) volatile acidity, 8) fixed acidity, 9) residual sugar, 10) alcohol level, 11) total sulfur dioxide, and 12) free sulfur dioxide

Density appears to be the most strongly correlated predictor variable with wine quality, as the magnitude of its correlation is very high and its p-value is very close to zero. The coefficient value of -809 indicates the magnitude of the effect. In this case, a one unit decrease in density is associated with a large decrease in the log-odds. Its negative value means it is inversely correlated with wine quality, meaning that as the density increases, quality decreases.

The two other predictor variables with the biggest impact on wine quality are chlorides and sulfates. The magnitude of their effect on quality is much smaller than that of density, but they are very statistically significant variables. They are positively correlated with wine quality, meaning that when chlorides and sulfates increase, so does the quality of the wine.

Lift/Gain Chart Analysis - ROC Curve Data
A Gains chart, also known as a lift chart, is a visual representation of the effectiveness of a classification model in predicting the positive class. It is particularly useful in binary classification problems. The Gains chart helps to understand how well a model is performing in

terms of identifying positive instances as compared to a random guess. It provides insights into the model's ability to prioritize instances with higher predicted probabilities.

For our analysis, it is important to consider contextual information to determine whether it is useful to our analysis to delve into the metrics of the lift/gain charts and interpret the graphs, associated metrics, and the AUC of the ROC curve.

In our case, lift charts are not very useful. Analyzing lift chart metrics is only useful when trying to predict values that can be added together with a goal of maximization. This would be applicable for contexts such as revenue per sale or profit, however, our context falls under the context type that are bad candidates for analyzing lift charts, applicable contexts include predicting the likelihood of customer churn, loan defaults, credit scores, final grades, etc.

ROC Curve Data, AUC=0.84678

# Records	FPR	TPR (Model)	TPR (Random)	TPR (Optimum)
0	0	0	0	1
1	0.001805	0	0.001805054	1
2	0.001805	0.011627907	0.001805054	1

We briefly looked at our ROC metric to determine that our model has very good performance overall. In general, an AUC (Area Under the ROC Curve) value closer to 1.0 indicates better model performance, while an AUC closer to 0.5 suggests a weaker performance, comparable to random chance. Because our AUC of the validation lift chart is between 0.8 and 0.9, we can deduce that the model has an excellent ability to effectively distinguish between the positive and negative classes.

Model #2: Linear Regression

Metric	Value
SSE	60.19921368
MSE	0.094061271
RMSE	0.306694101
MAD	0.207021244
R2	0.191346303

Figure 3.4: Validation Prediction Summary

Metric	Value
SSE	85.58533818
MSE	0.089244357
RMSE	0.29873794
MAD	0.203603256
R2	0.243312873

Figure 3.5: Training Prediction Summary

Metric	Value
Residual DF	947
R2	0.243312873
Adjusted R2	0.234523476
Std. Error Estimate	0.300624725
RSS	85.58533818

Figure 3.6: Regression Summary

Figures 3.4 and 3.5 display the error metrics for both the training and validation data. The mean absolute deviation (MAD) is the most meaningful error metric when assessing the quality of the model. The MAD for the training and validation data is very similar, which does not indicate overfitting in this model.

To assess the quality of the model, it is most important to look at the validation error metrics. Because we want the SSE to be as low as possible, a value of 60.2 is not ideal. The R^2 value should also be as close to 1 as possible, so a value of 0.19 tells us that the proportion of variability in Y that is explained by the model is not very high. Overall, these metrics tell us that the model is not sufficient enough for us to recommend our target audience use it.

Figure 3.7: Prediction Coefficient Output

Predictor	Estimate	Abs. Value of Estimate	P-Value
density	-78.1137948	78.11379476	9.05563E-16
Intercept	76.2767117	76.27671172	6.90881E-16
chlorides	-0.76235592	0.762355918	0.003100787
sulphates	0.48462721	0.484627213	1.9157E-11
pH	0.23671449	0.236714489	0.021027179
citric acid	0.13276343	0.132763431	0.133072067
volatile acidity	-0.12658136	0.126581363	0.08686175
Alcohol Level	0.09028554	0.090285544	0.025729791
fixed acidity	0.07786913	0.077869131	1.24194E-08
residual sugar	0.04423397	0.044233974	3.16547E-08
free sulfur dioxide	-0.00106309	0.001063093	0.396703087
total sulfur dioxide	-0.00073076	0.000730765	0.082971661

Similar to our logistic regression model, our linear regression model tells us that the most statistically significant and highly correlated variables with quality are density, chlorides, and sulfates. However, the linear regression model displays a negative correlation between chlorides and quality. This conflicts with the results of our logistic regression, and ultimately we believe that the results of this prediction model are not of the best quality and should not be used to make any major real-life decisions.

Summary

As stated previously, neither our logistic regression model nor our linear regression model produced the highest quality results. While they did provide some useful insights about red wine characteristics, we would not actively recommend that our key audience use either of these models as the basis for any high-stakes decisions or large investments. Realistically, these models would be best suited for sommeliers or anyone interested in learning more about red wine taste and quality. Had the dataset contained more high-quality wines, our models may have produced more refined results that we could confidently recommend this information to winemakers, distributors, and inventory procurement analysts.

After analyzing our logistic and linear regression models, we believe the key characteristics that winemakers should focus on are density, chlorides, and sulfates. Upon doing further industry research to verify our findings (ex: the significance of sulfates & their effects), we can conclude that the variables with p-values < 0.05 are truly statistically significant in determining a consumer perception of wine quality. Sulfates have a heavy contribution to taste and freshness, which is the principal reason they are extremely statistically significant contributors to wine quality. The negative correlation of chlorides with wine quality indicates that consumers prefer a less salty wine.

From this project, we learned how machine learning techniques like classification and prediction can be useful to a variety of industries and business contexts; analyzing the wine industry is just a slice of the potential these models have to predict and forecast unknown outcomes on unseen data for use on real-world data. We also learned how to better assess the quality of performance metrics and measures from classification and prediction models.

The skills we have acquired will enable us to further advance our skills in analyzing large-scale datasets, specifically utilizing machine learning techniques in SAP, Excel, and other business programs.

Works Cited

"Wine's Alcohol Levels Explained." *Cult Wines*, www.wineinvestment.com/us/learn/magazine/2019/06/wines-alcohol-levels-explained/. Accessed 10 Dec. 2023.