Milestone Report
Wine Quality Data Set

Introduction

I once had a sommelier tell me that the simplest way to select a good wine is by looking at the animal that was on the label. His theory was that the wine that had the deadliest animal on the label tasted the best. Scientific? No. Funny? Maybe.. Depends who you ask. But! Is there any better way to determine what makes a great wine? The quality of our favorite wines is determined by the chemical composition of its makeup. Factors such as pH, acidities, sulphur and alcohol contents greatly affect the flavor, aroma, and mouth feel of these wines. Comparing known components of a selection of wines and the quality scores that were given to these wines, we will find the common characteristics of a great wine and a poor wine.

Data learned from this project would be useful to wineries and producers looking to produce higher quality wines. The ability to fine tune which chemical compounds are used in producing the grapes allows the winemakers to tailor their wines for more enjoyability. Consumers could also use this data to select wines for purchase that will be in line with their tastes.

The Dataset

The wine quality dataset comes from the UCI Machine Learning Repository. It consists of 2 datasets for the Portuguese Vinho Verde wine. First is the red wine data set that has 1599 wines with quality rankings from 3 to 8. The second set is the white wine dataset of 4898 wines with quality ranking of 3 to 9. Each of these datasets include the quantities of 12 variables : fixed acidity, volatile acidity,citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality. There is no data in regards to grape type, brand name, or price.

The data were clean and did not need any wrangling. There were no missing values for any of the variables. The distribution of the ratings is a normalized bell curve with many more 'normal' wines and few 'poor' or 'excellent' wines.

Quality

The wines had quality ranking values of 3 through 8. I separated the dataset of each kind of wine into 'High Quality' and 'Low Quality' categories. The High quality wines had quality ratings of 6 or higher and the Low quality wines had ratings of 5 or lower. These groupings allowed me to gain a better perspective of what factors make a wine high quality or low quality.

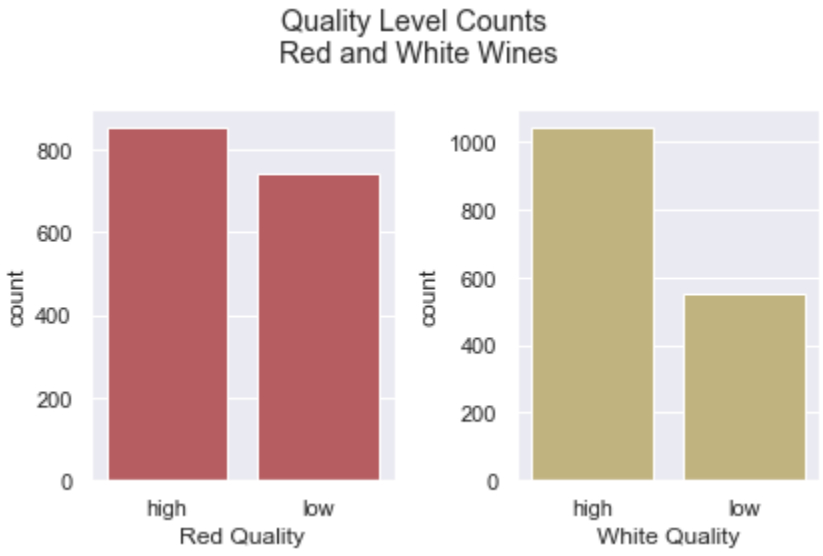

Figure 1. Quality counts of Red and White Wines

EDA

Alcohol

In Figure 1 we see that for both white and red wines, higher alcohol content is associated with a higher wine quality in general. Alcohol content is what gives wine its 'body'. Alcohol is more viscous than water and it greatly affects the mouth feel of the wine. This is especially true with red wines that have tannins. The alcohol content helps to balance against the earthiness of the tannins. After performing a t-test, it was found that the difference in mean volatile acidity between high and low quality wine is statistically significant for both white and red wines ($p << .001$).

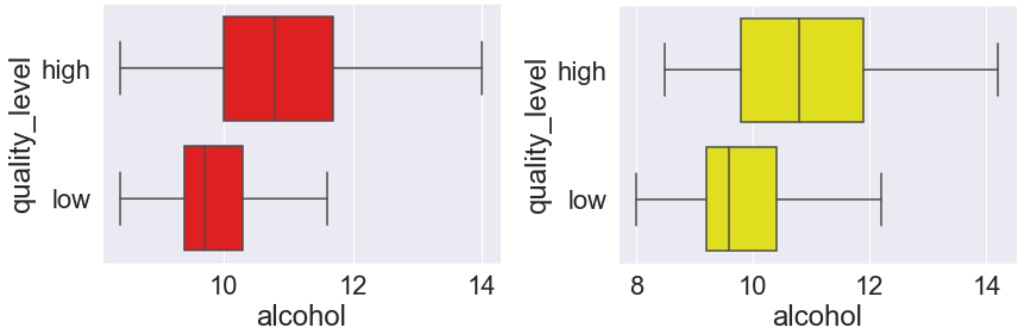

Figure 2. Alcohol Content vs Quality of Red and White Wines

Volatile acidity

In Figure 2 we see that wines with lower levels of volatile acidity are associated with higher quality flavors. Volatile acidity is generally the level of acetic acid present. Acetic acid, at high levels, can produce a vinegar-like flavor, so it would make sense that lower levels favor better tasting wines. After performing a t-test, it was found that the difference in mean volatile acidity between high and low quality wine is statistically significant for both white and red wines ($p << .001$).

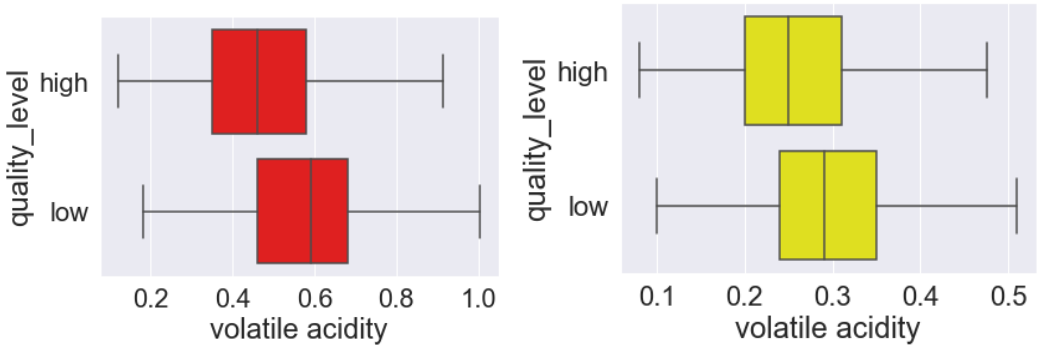

Figure 3. Volatile acidity vs Quality of Red and White Wines

Citric Acid

In figure 3 we can see that high levels of citric acid are associated with lower quality wines. In small quantities, citric acid can create a 'crispness' or 'fresh' flavor. In red wines, too much of this flavor will clash with the flavor of the tannins present and give an off flavor. In white wines, this crispness often helps balance against the alcohol flavor. Typically, it's understood thatwhite wines do better with a higher citric acid level. However, after performing a t-test, citric acid was not found to be statistically significantly different between high and low quality wines for white wine (p-value = .96). However, it was found that the difference in mean citric acid between high and low quality wine is statistically significant for red wines ($p << .001$).

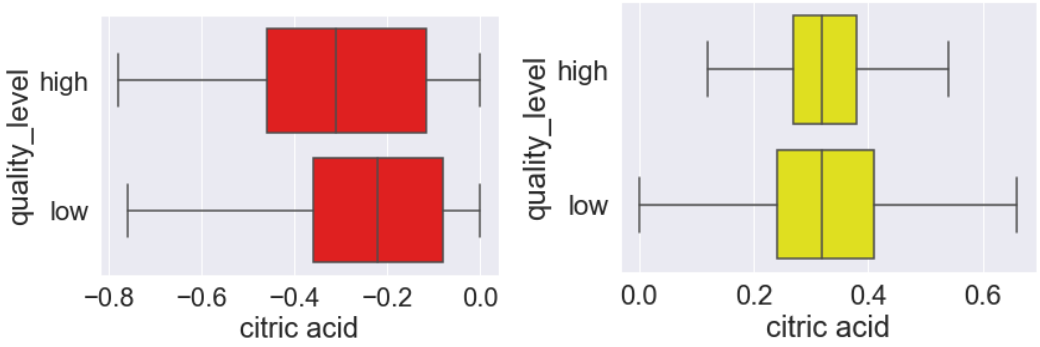

Figure 4. Citric acid vs Quality of Red and White Wines

Chloride

Chloride levels are a measurement of the amount of salt in the wine. The salt is usually found in the skins of the grapes. We can see in figure 4 that too much salt will have a negative effect on the wine quality. After performing a t-test, it was found that the difference in mean volatile acidity between high and low quality wine is statistically significant for both white and red wines ($p << .001$).

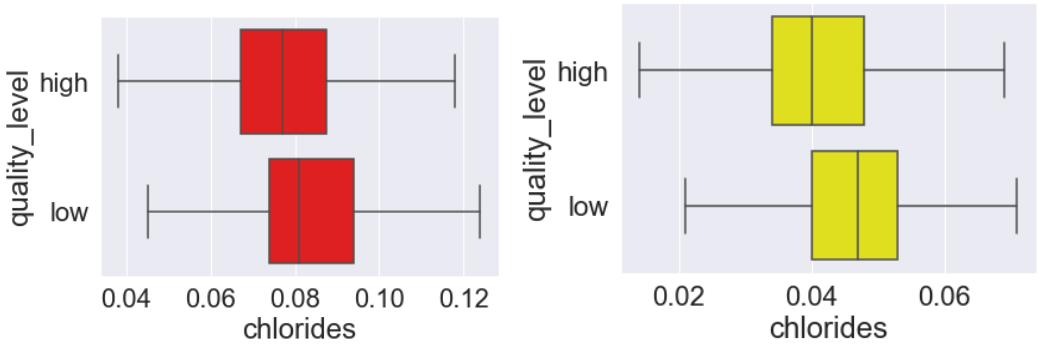

Figure 5. Chlorides vs Quality of Red and White Wines

Density

The density of wine is the mass per unit volume of wine at 20°C. This affects the way the wine feels in the mouth. Generally, the more dense a wine is, the less pleasant it can feel. In figure 5 we can see that higher density is associated with lower quality wines. After performing a t-test, it was found that the difference in mean volatile acidity between high and low quality wine is statistically significant for both white and red wines ($p << .001$).

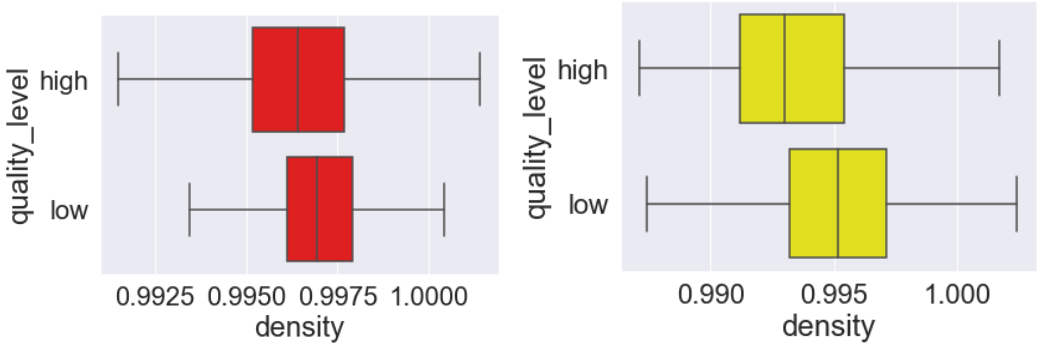

Figure 6. Density vs Quality of Red and White Wines

Sulphates

Sulphates act as an antimicrobial and an antioxidant. In figure 6 it shows that wines with higher sulphate levels rate higher. It would make sense that wines that are clean would taste better and rate as better quality. After performing a t-test, it was found that the difference in mean volatile acidity between high and low quality wine is statistically significant for both white and red wines ($p << .001$).

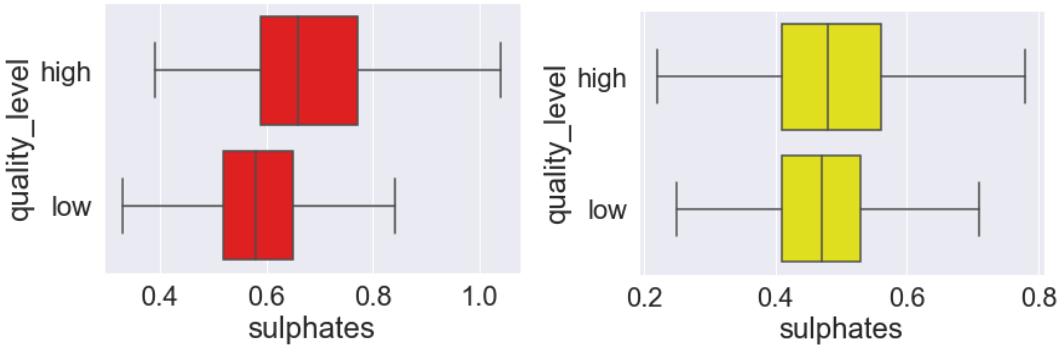

Figure 7. Sulphates vs Quality of Red and White Wines

Feature Correlations

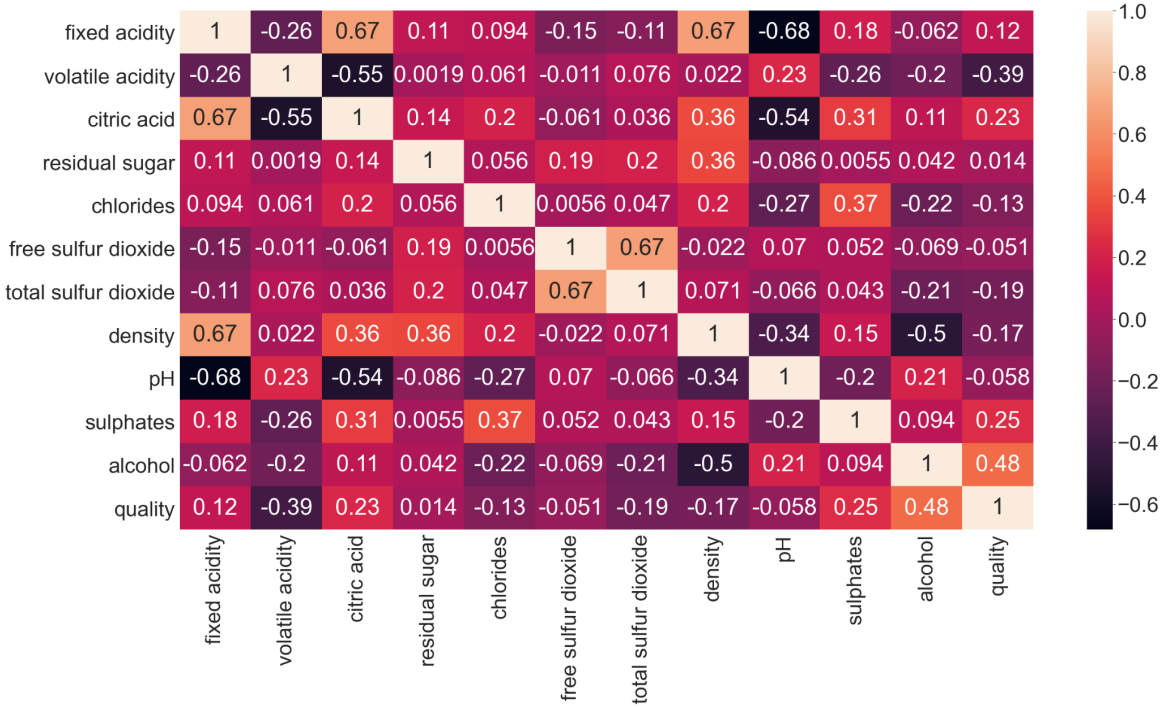

Figure 8. Heatmap correlation of red wine components

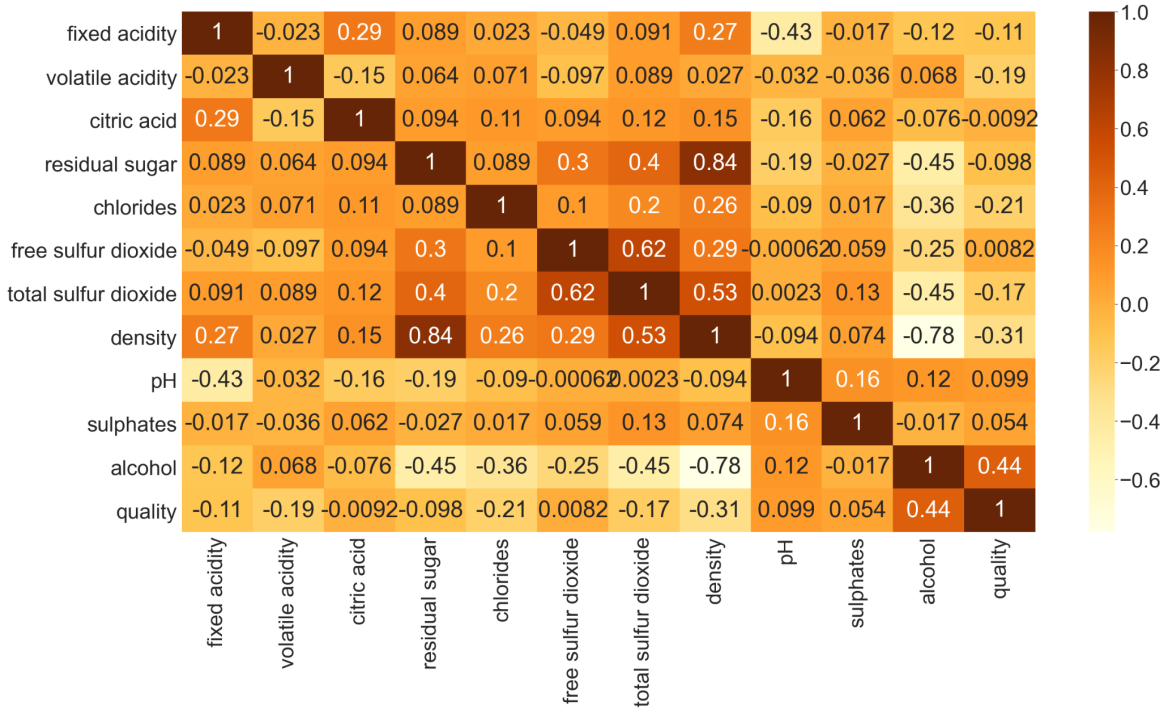

Figure 9. Heatmap correlations of white wine components

The heat map charts in figures 8 and 9 show the correlations between the 12 different features in the wines. Here we can see that combinations of different features will produce different results. We can see that in regards to the quality metric in both red and white wines, alcohol has, by far, the largest correlation. Red wines also have a strong positive correlation with citric acid and sulphates, and a strong negative correlation with volatile acidity. White wines have a strong negative correlation with density, chlorides, and volatile acidity.

Feature Importance

Using the feature importances method of the Random Forest Classifier, I am able to show which features of the wine data set are the most important factors for determining the quality of the wine. With red wines we can see that, as expected, alcohol is the most important feature for producing a quality red wine. volatile acidity, sulphates, and total sulfur dioxide are also very important in creating a quality red wine.

Looking at white wines, there is not as clear a divide between the most important and less important features, but we find that alcohol is once again the most important feature for producing a quality wine. Density and volatile acidity are also major factors in the outcome of white wine, though interestingly sulphates don't appear as particularly important in contrast to red wine.

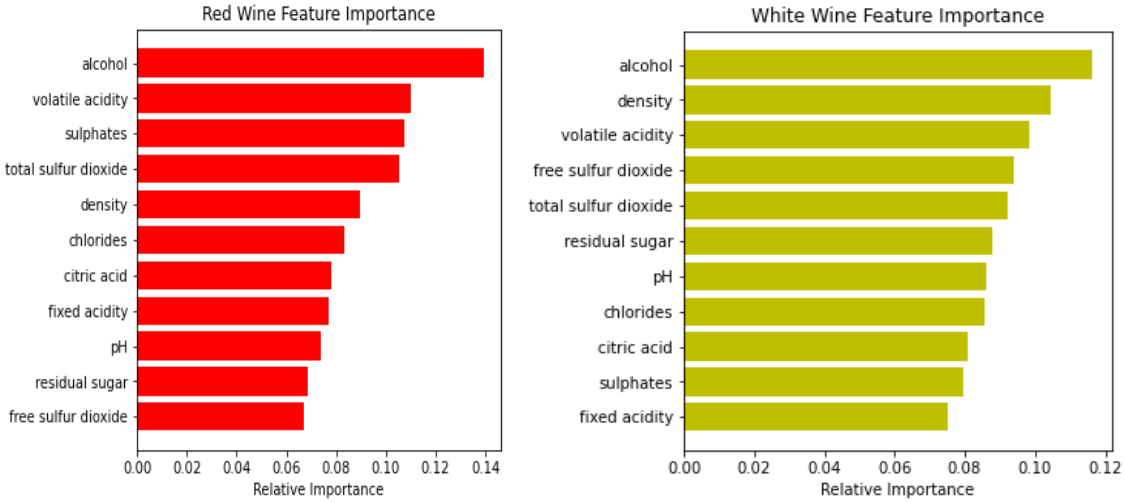

Figure 10. Red Wine and White WIne Feature Importance

Model Selection

Model Selection Red WInes		
Model	ROC-AUC Score	Optimal Parameters
Random Forest	0.8239	N_estimators: 500 Max depth: 6 Min_samples_leaf: 1 Min_samples_split: 10 Bootstrap: True
K Neighbors Classifier	0.7799	N_neighbors: 324
Logistic Regression	0.8086	C: 1 Max_iter: 100

Figure 11. Model Selection

Model Selection White Wines		
Model	ROC-AUC Score	Optimal Parameters
Random Forest	0.8512	N_estimators: 500 Max depth: 10 Min_samples_leaf: 4 Min_samples_split: 2 Bootstrap: True
K Neighbors Classifier	0.7961	N_neighbors: 24
Logistic Regression	0.7999	C: 100 Max_iter: 100

Figure 12. Model Selection

In choosing a classifier, I compared the performance of three different models: K-Nearest Neighbors, Logistic Regression and Random Forest. I used Sci-kit learn's GridSearchCV to choose the hyperparameters that would yield the best results. GridSearchCV computed a receiver operating characteristic area under the curve (ROC AUC) score, which was used to properly compare the models. With a ROC AUC score of 0.8757 for red wines and 0.8512 for white wines, Random Forest was the better performing classification model.

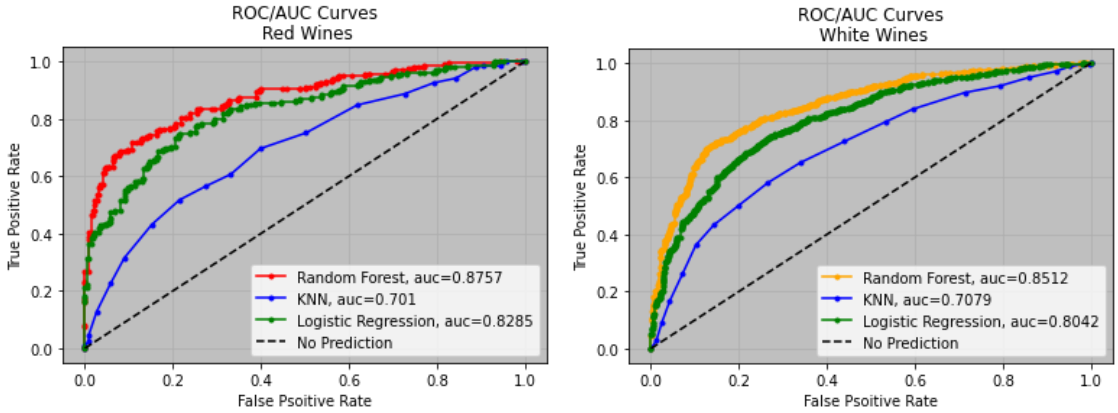

Figure 13. ROC/AUC Curves - All Model

Thresholding for the Business Case

This model is intended to be used by winemakers to determine which wines they want to produce. For this reason, precision is more important than recall. Our clients here would be more concerned with making sure a wine they produce really will be high quality than they might be concerned about missing out on potential opportunities.

My next step was to plot precision and recall by threshold to see what threshold would make the most sense for this business case. I did this for the red wines below in Fig 13.

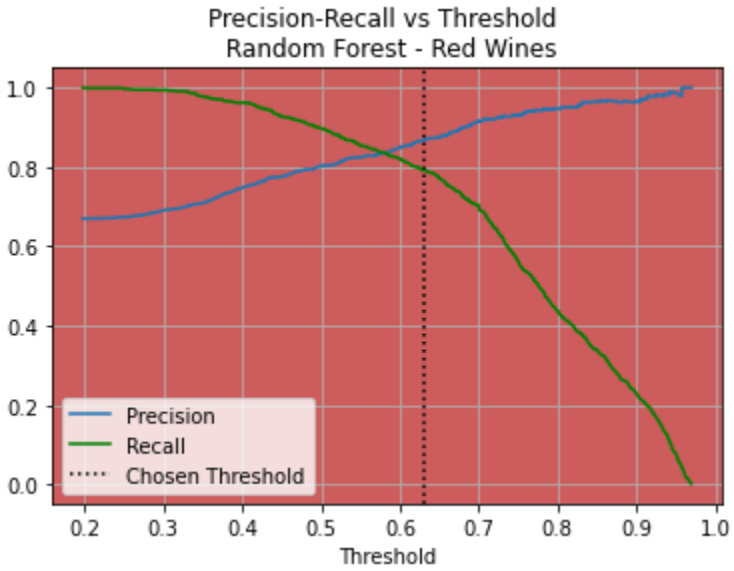

Figure 14 .Precision-Recall vs Threshold Random Forest - Red Wines

For red wines, choosing a threshold of 0.63 returns a precision value of 0.87 and a recall value of 0.80. With prioritizing precision over recall, this threshold gives the best precision score while still maintaining a good recall score.

Classification Report - Red Wines			
Random Forest			
	Precision	Recall	Threshold
True	**0.87**	**0.80**	**0.63**

Figure 15. Classification Report- Red Wines

Confusion Matrix -Red Wines		
Random Forest		
	Predicted 0 (Low Quality)	Predicted 1 (High Quality)
Actual 0 (Low Quality)	**173**	**50**
Actual 1 (High Quality)	**52**	**205**

Figure 16. Confusion Matrix - Red Wines

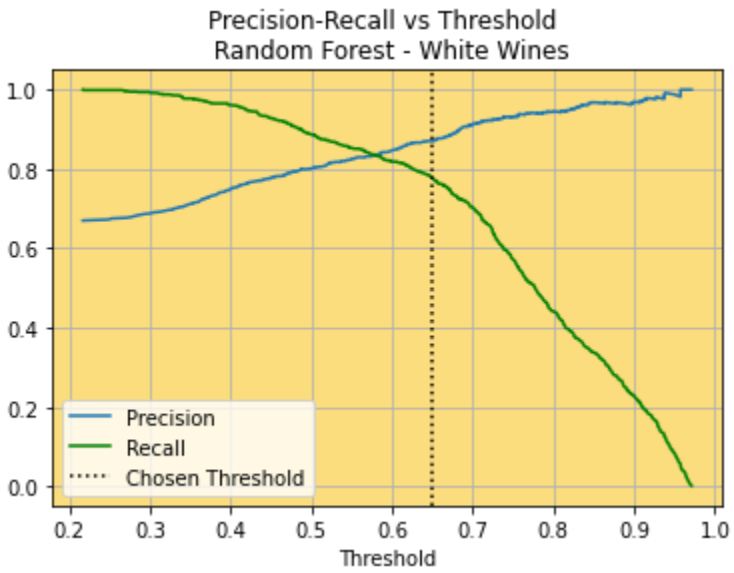

Figure 17 .Precision-Recall vs Threshold Random Forest - White Wines

Figure 16 shows the precision recall curve for white wines. I found that a similar threshold value of 0.64 produces a precision value of 0.87 and recall value of 0.79. Once again providing the best precision score while maintaining an acceptable recall score.

Classification Report White Wines			
Random Forest			
	Precision	Recall	Threshold
True	**0.87**	**0.79**	**0.64**

Figure 18. Classification Report - White Wines

Confusion Matrix -White Wines		
Random Forest		
	Predicted 0 (Low Quality)	Predicted 1 (High Quality)
Actual 0 (Low Quality)	**173**	**50**
Actual 1 (High Quality)	**52**	**205**

Figure 19. Confusion Matrix. - White Wines

Conclusion

We can conclude that the strongest indicator of a quality wine is its alcohol content. The way that a wine “feels” in our mouth is due to the amount of alcohol in the wine. This feeling is often characterized as the body of the wine. Having a strong body in a wine allows it to balance out other components. The balance of the other components with the body of the wine is what determines the wine’s flavor. In red wines, volatile acidity, sulphates and total sulfur dioxide are other major components that determine the quality of the wine. Proper levels of these four components raise the level of quality for a red wine.

In white wines, there are fewer determining components. Besides alcohol, a white wine's density and volatile acidity are the next biggest determining components of its quality. Density is another factor in the mouth feel of the wine. Having a low density helps produce the light crispness that is famous in fine white wines. Volatile acidity, a product of acetic acid affects the flavor of the wine. High acetic acid produces a vinegar like flavor, so a reduced level of volatile acidity helps to produce a balanced flavor