Predicting Wine Quality Based off their physicochemical features

By: Tanuj Chawla

Introduction

- Wine is typically reviewed by wine critics or experts who have developed a refined palate and extensive knowledge of wine. They will taste the wine, taking note of its appearance, aroma, flavor, and mouthfeel. They will then assign a score or rating to the wine, often using a 100-point scale, to indicate its quality and overall enjoyment.

- The problem with this process is that it is subjective and based on the individual critic's personal preferences and biases. Different critics may have different opinions on the same wine, and their ratings may not always align with the preferences of the general public. Additionally, the use of a 100-point scale can be misleading, as it may give the impression that a wine with a higher score is objectively better than one with a lower score, when in reality it is simply the critic's personal preference. This can lead to confusion and inconsistency in the wine review process.

- Data Source - Wine Quality data from the University of California, Irvine Machine Learning Repository.Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The data contain 4897 observations in white wine and 1599 observations in red wine data with 11 features in each. The physicochemical features are lab based. Quality is an ordinal variable with a possible ranking from 1 (worst) to 10 (best). Each variety of wine is blind tasted by three independent tasters and the final rank assigned is the median rank given by the tasters. There are no missing or null values. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).Data is available at: https://archive.ics.uci.edu/ml/datasets/Wine+Quality.

Key Questions

- What physicochemical features or combination of features impact the quality of wine?
- What are the features of a good wine or a bad wine? Why does this matter?
- Can we predict the quality based off what makes up the wine to increase business efficiency and productivity?

Who Might Care?

Wine Manufacturers

E. & J. Gallo Winery

Constellation Brands

THE WINE GROUP

Wine Distributors

SOUTHERN GLAZER'S WINE & SPIRITS

RNDC Republic National Distributing Company

BREAKTHRU BEVERAGE GROUP

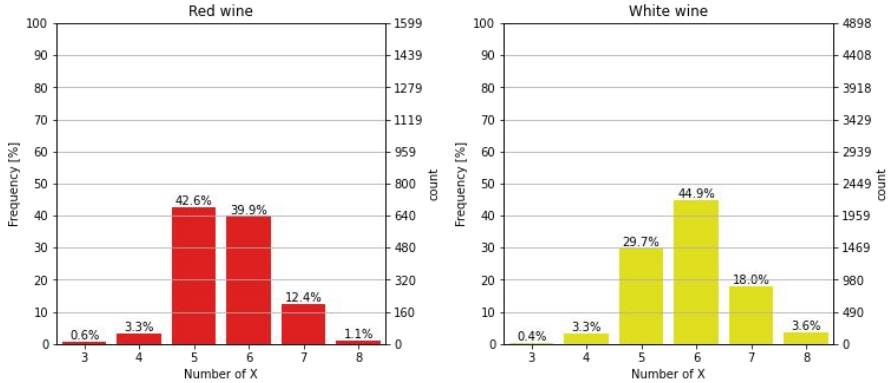
Wine Tasters

Backcountry Wine Tours

What Factors affect Wine Quality?

- **wine type** - 1599 Red and 4897 White wine are being analyzed
- **fixed acidity** - Most acids involved with wine are fixed or nonvolatile (Tartaric Acid - g/dm^3)
- **volatile acidity** - The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste (Acetic Acid - g/dm^3)
- **citric acid** - acts as a preservative to increase acidity (small quantities add freshness and flavor to wines) (g/dm^3)
- **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 (g/dm^3)
- **chlorides** - The amount of salt in the wine (Sodium Chloride - mg/dm^3)
- **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 (mg/dm^3)
- **total sulfur dioxide** - Amount of free and bound forms of S02 becomes evident in the nose and taste of wine (mg/dm^3)
- **density** - the density of water is close to that of water depending on the percent alcohol and sugar content (g/cm^3)
- **pH** - Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
- **sulphates** - a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant (Potassium Sulphate g/dm^3)
- **alcohol** - the percent alcohol content of the wine
- **quality** - score between 0 and 10 given by wine experts

Exploratory Data Analysis

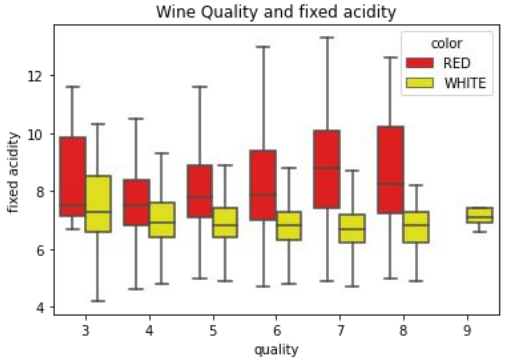


- We can clearly see the distribution of our target variable is imbalanced. We can see the most value is 5 in red wine and 6 in white wine, and all class values are in between 3 to 8 in both wines. Our focus will be on high quality wines (rating 7 & rating 8).

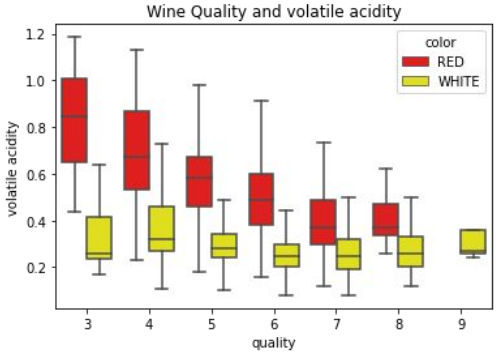
Exploratory Data Analysis cont...

Understanding Wine Attributes and Properties: **Acidity**

- Acids are one of 4 fundamental traits in wine and give wine tart and sour taste. Acidity adds a sense of vitality and refreshment to the wine and also make the flavor of the wine more prominent. Acid has an appetizing effect, so you want to have another cup. In Low levels, it adds to the complexity of flavor but in High levels it causes the wine to spoil.

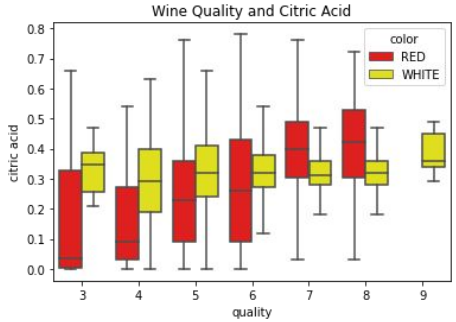


- Fixed acidity and volatile acidity seem to be higher in red wine as compared to white wine.

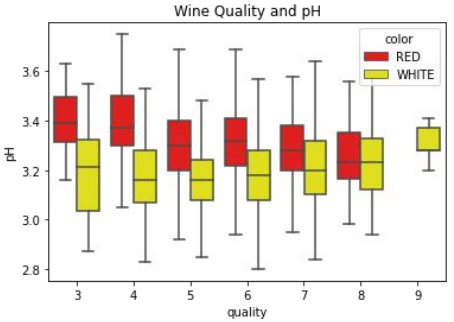


Exploratory Data Analysis cont...

Understanding Wine Attributes and Properties: **Acidity cont..**



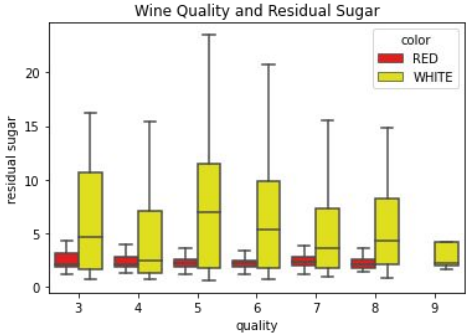
From all numbers, we can observe that citric acid is more present in white than red wines on average. On the other hand, Red wines, typically have higher pH so they are less acidic than white wines.



Exploratory Data Analysis cont...

Understanding Wine Attributes and Properties: **Sweetness**

- Residual Sugar (or RS) is from natural grape sugars leftover in a wine after the alcoholic fermentation finishes. It can improve one of 4 fundamental traits in wine, the sweetness. Sweetness in wines is not a sugary artificial flavor, but more of a natural fruit-based flavor. This sensation is typically felt on the tip of the tongue. When a wine is dry, the sweetness perceived is related to the fruit flavors found in wine.It's common to taste a subtle sweet flavor and not know exactly how to describe it. Sometimes it helps to think of fruits associated with wine. White wines have citrus flavors, and red wines have flavors like raspberries, blueberries, plums, cherries, blackberries, or jam.

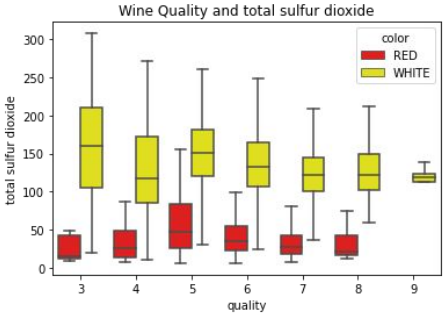


White wines are much sweeter than red wines. Red wines tend to concentrate in low levels of sugar.

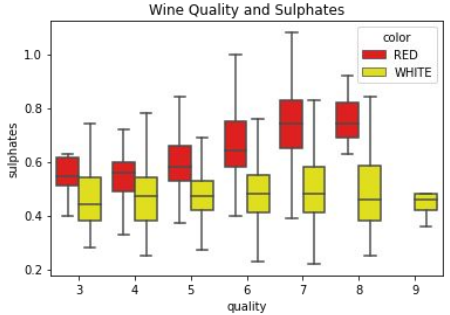
Exploratory Data Analysis cont...

Understanding Wine Attributes and Properties: **TOTOL SULFUR DIOXIDE AND SULPHATES**

- Sulfites preserve wine and slow chemical reactions, which cause the wine to go bad. Stop bacteria and yeasts from growing, keeping the fresh taste of the wine. However, the high sulfur dioxide content will produce an unpleasant smell like a rotten egg, which may cause health problems after drinking.



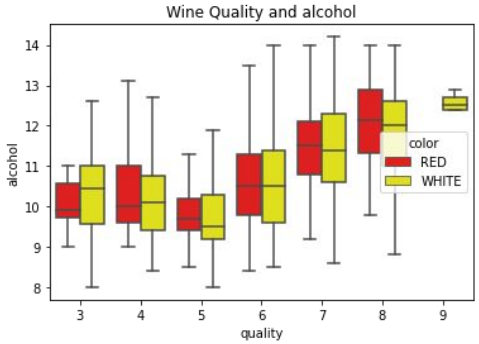
- The amount of sulphates is less in white wine than red wine but the total amount of sulfer dioxide is higher in white wine than in red wine.



Exploratory Data Analysis cont...

Understanding Wine Attributes and Properties: **Alcohol**

- Alcohol is also one of 4 fundamental traits. Alcohol is formed as a result of yeast converting sugar during the fermentation process. The percentage of alcohol can vary from wine to wine. We interpret alcohol using many different taste receptors which is why it can taste bitter, sweet, spicy, and oily all at once. Your genetics actually plays a role in how bitter or sweet alcohol tastes. Regardless, we can all sense alcohol towards the backs of our mouths in our throats as a warming sensation.



We can see a very positive correlation with alcohol and quality. Both wines behave very similar with alcohol

Machine Learning - Red Wine

Compared 3 models: Logistic Regression, Random Forest, and K-Nearest Neighbors, grid searched for respective hyperparameter space for values that maximized algorithm performance and ordered relative performance of the algorithms by calculating the Receiver Operating Characteristic Area Under the Curve (ROC AUC) scores. Random Forest performed the best with a ROC AUC Score of 0.91% and overall accuracy of

TABLE 1: ALGORITHM PERFORMANCE - RED WINE

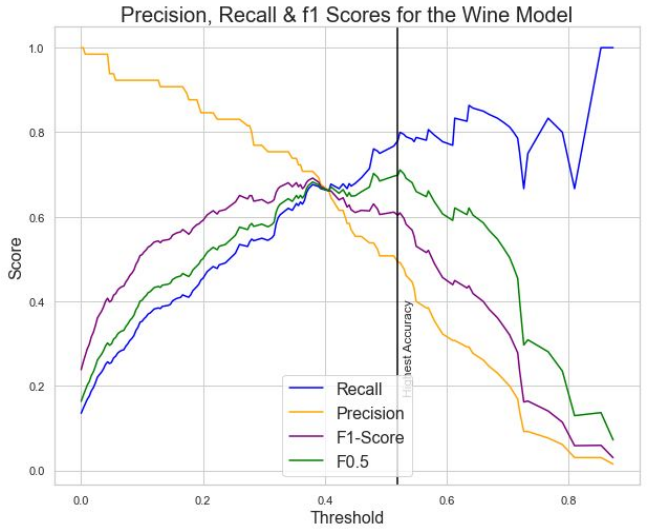
Classifier	ROC-AUC Score	Best Hyperparameter Value
Logistic Regression	0.864	C = 0.1
K-Nearest Neighbors	0.908	N_neighbors = 44, Weights = distance
Random Forest	0.912	Max_depth = 25 Max_features = 3, Min_samples_leaf: 1 Min_samples_split: 2 N_estimators: 300

Improving Classifications

The default cutoff probability of 0.5 for threshold of individual cases lends equal weight to both precision and recall. However, this may not be the optimal model and different scenarios potentially would favor one over the other. To demonstrate this, I adjusted the threshold for two different scenarios:

Scenario 1: Accuracy

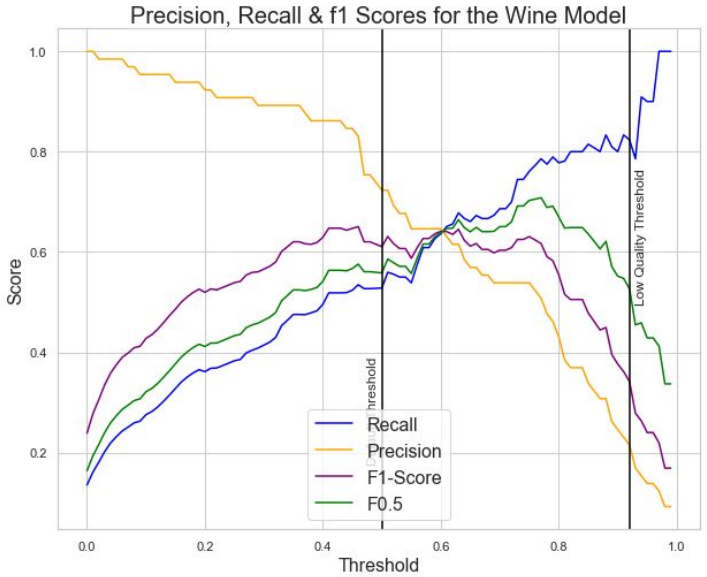
Essentially we would want to look at the ability of our model to accurately identify good quality wines and bad quality wines for wine manufacturers and distributors to efficiently and effectively be able to output only those good quality wines. In order to avoid overlooking good wine, we would want to focus on minimizing false positives. In order to simulate such a scenario, I adjusted the beta parameter of the F-beta metric to emphasize precision (Beta < 1) using beta = 0.5 and then quantified F-beta for probability threshold values ranging between 0-1. The threshold value producing the largest F-beta score of 0.9146 is .52. This resulted in higher false positives by 12, increase in True high quality wines by 13 and slightly lower false negatives 13.



Improving Classifications cont..

Scenario 2: Focus on Low Quality Wines

Let's say we have a list of wines that we would like to start manufacturing and we want to make sure we are only producing the best wine possible. We would want to make sure we are predicting high quality wines as high and making sure no low quality wines are being misabled as high quality. In this instance, we also would not care for False negatives as we want to make sure we are only serving the best wine. To do so I examined the trade off between precision and recall and found the optimal threshold for this case would be 0.90.



Machine Learning - White Wine

Compared 3 models: Logistic Regression, Random Forest, and K-Nearest Neighbors, grid searched for respective hyperparameter space for values that maximized algorithm performance and ordered relative performance of the algorithms by calculating the Receiver Operating Characteristic Area Under the Curve (ROC AUC) scores. Random Forest performed the best with a ROC AUC Score of 0.91% and overall accuracy of 0.879

TABLE 1: ALGORITHM PERFORMANCE - RED WINE

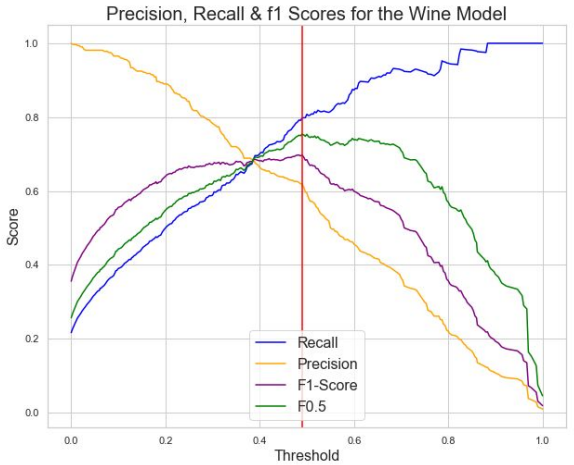
Classifier	ROC-AUC Score	Best Hyperparameter Value
Logistic Regression	0.7431	C = 0.1
K-Nearest Neighbors	0.8468	N_neighbors = 48, Weights = distance
Random Forest	0.908	Max_depth = 20 Max_features = 3, Min_samples_leaf: 1 Min_samples_split: 2 N_estimators: 300

Improving Classifications

The default cutoff probability of 0.5 for threshold of individual cases lends equal weight to both precision and recall. However, this may not be the optimal model and different scenarios potentially would favor one over the other. To demonstrate this, I adjusted the threshold for two different scenarios:

Scenario 1: Accuracy

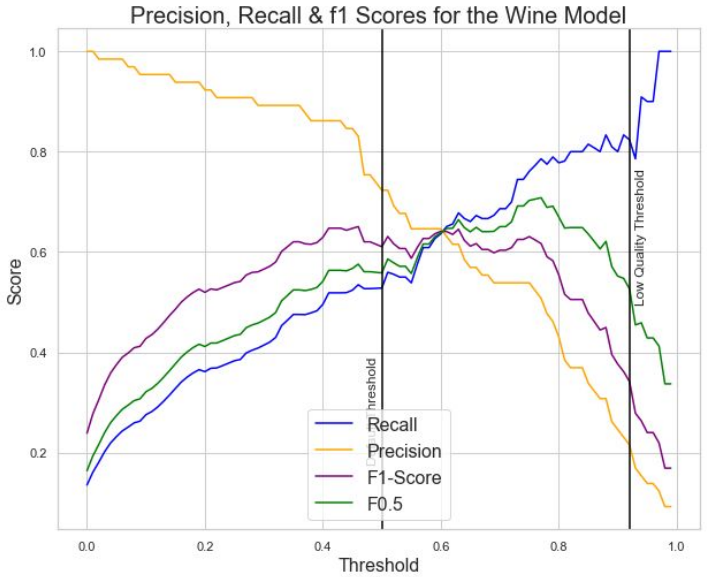
Essentially we would want to look at the ability of our model to accurately identify good quality wines and bad quality wines for wine manufacturers and distributors to efficiently and effectively be able to output only those good quality wines. In order to avoid overlooking good wine, we would want to focus on minimizing false positives. In order to simulate such a scenario, I adjusted the beta parameter of the F-beta metric to emphasize precision (Beta < 1) using beta = 0.5 and then quantified F-beta for probability threshold values ranging between 0-1. The threshold value producing the largest F-beta score of 0.9146 is .505. This resulted in lower false positives by 5, increase in True high quality wines by 5 and slightly lower false negatives 7.



Improving Classifications cont..

Scenario 2: Focus on Low Quality Wines

Let's say we have a list of wines that we would like to start manufacturing and we want to make sure we are only producing the best wine possible. We would want to make sure we are predicting high quality wines as high and making sure no low quality wines are being misabled as high quality. In this instance, we also would not care for False negatives as we want to make sure we are only serving the best wine. To do so I examined the trade off between precision and recall and found the optimal threshold for this case would be 0.90.



In summary

- Predicting the quality of wine based on its chemical composition can help winemakers make more informed decisions about which wines to produce and how to blend them
- It can also help wine merchants and distributors choose which wines to stock and sell, potentially leading to better sales and customer satisfaction
- Understanding the chemical factors that contribute to wine quality can also help winemakers optimize their production processes and improve the overall quality of their wines
- Some important chemical factors to consider when predicting wine quality include pH, alcohol content, acidity, and the presence of certain compounds like tannins and antioxidants
- It may also be useful to consider sensory factors such as flavor and aroma, as these can have a significant impact on wine quality and customer perception
- Overall, creating a model that can accurately predict wine quality based on chemical composition can have numerous benefits for the wine industry and for consumers.