Capstone Project: Wine Quality Project

Andrew Wessel
May 21st 2022

Problem outline

Determining features that make a wine high quality can increase profits and market share

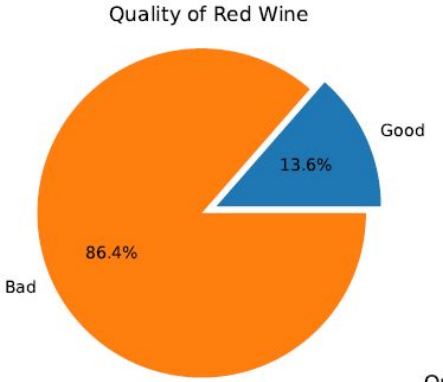
Chemical and physical properties of the wines can be measured and optimized for increased quality

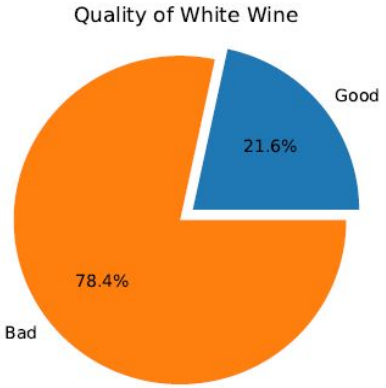
Is there a way to predict which features of red and white wines most influence quality?

Is it possible to determine if a wine has characteristics of a red or white wine based on its chemical and physical properties?

Project & Data - Wine Quality Prediction

- First Desired Outcome
  - Use physicochemical wine data to predict quality of red and white wines
- Dataset
  - 1599 red wine entries, 4898 white wine entries of the Vihno Verde variety from Portugal
  - Information about physicochemical data on amounts of compounds in wine or physical characteristics of the wine
  - Output variable: Quality of wine (integer: 0-10)
    - Wines of quality ≥ 7 labeled “good”
    - Wines of quality < 7 labeled “bad”



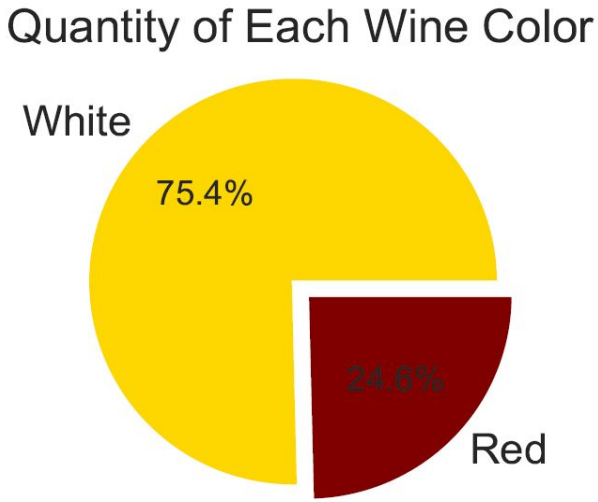


https://archive.ics.uci.edu/ml/datasets/wine+quality

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Project & Data - Wine Color Prediction

- Second Desired Outcome
  - Use physicochemical wine data to predict the color of wine (or characteristics of a given color of wine)
- Dataset
  - 6497 total Vinho Verde wine entries
    - 1599 red wine entries
    - 4898 white wine entries
  - Analysis uses combination of red and white data sets
  - Output variable: Wine Color (“red”, “white”)



Exploratory Data Analysis - Data variables

Input variables (based on physicochemical tests):

1. *fixed acidity* - amount of fixed acidity in the wine (numeric)
2. *volatile acidity* - amount of volatile acidity in the wine (numeric)
3. *citric acid* - the amount of citric acid in the wine (numeric), a preservative and acidifying reagent
4. *residual sugar* - the amount of residual sugars in the wine after fermentation (numeric)
5. *chlorides* - amount of chlorides in the wine (numeric), a proxy measure of sodium in the wine
6. *free sulfur dioxide* - the amount of free sulfur dioxide dissolved in the wine (numeric), preservative
7. *total sulfur dioxide* - the total sulfur dioxide dissolved in the wine (numeric), preservative
8. *density* - the density of the wine (numeric)
9. *pH* - the pH of the wine (numeric)
10. *sulphates* - (a.k.a. sulfates) the amount of sulfates in the wine (numeric), a measure of hardness of the water
11. *alcohol* - the amount of alcohol in the wine (numeric)

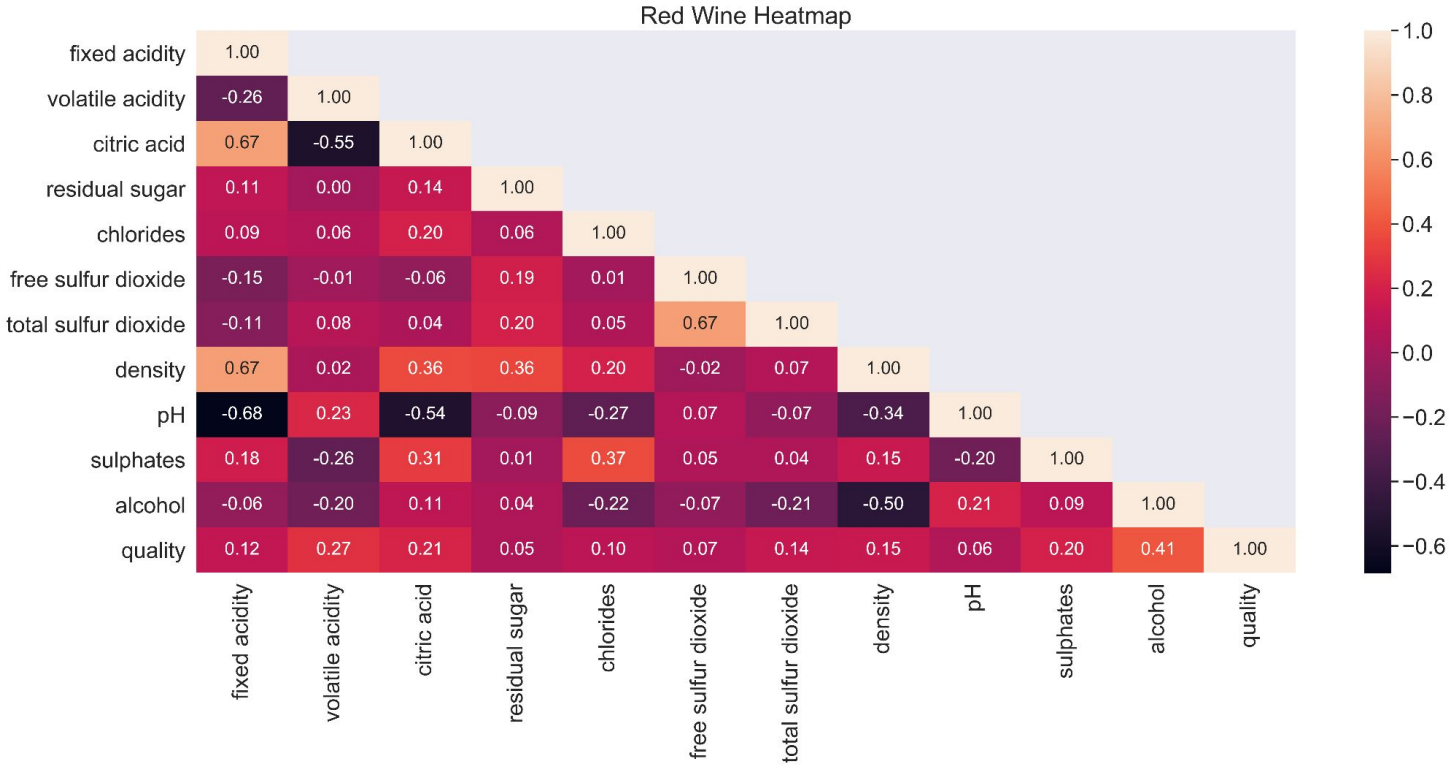
Output variables (based on sensory data):

12. *quality* - a quality score between 0 and 10 (numeric)
13. *color* - color of the wine (string: “red”, “white”)

Note:
The “color” feature was not included in the red and white wine quality analyses.

The “quality” feature was included in the color of wine analysis.

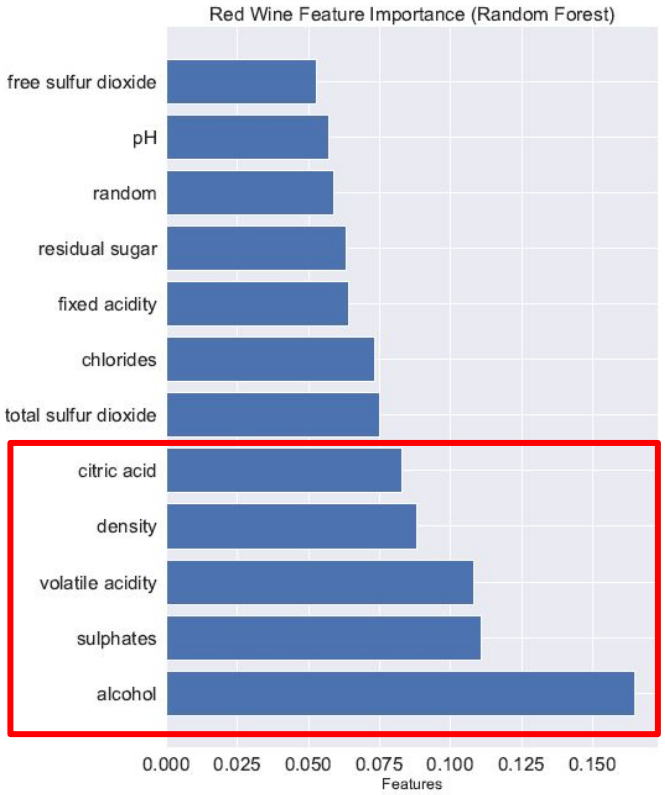
Exploratory Data Analysis - Red Wine Heat Map



Heat Map of Features Generated from Pearson's R

Exploratory Data Analysis - Red Wine Feature Importance

- Top five most important features for red wine quality
  - Alcohol content is most important
  - Sulfates and volatile acidity
  - Density and citric acid
- These features describe a dryer, full-bodied, strong, red wine
- The pH and free sulfur dioxide were less influential than the random background thus aren't as important in building the model

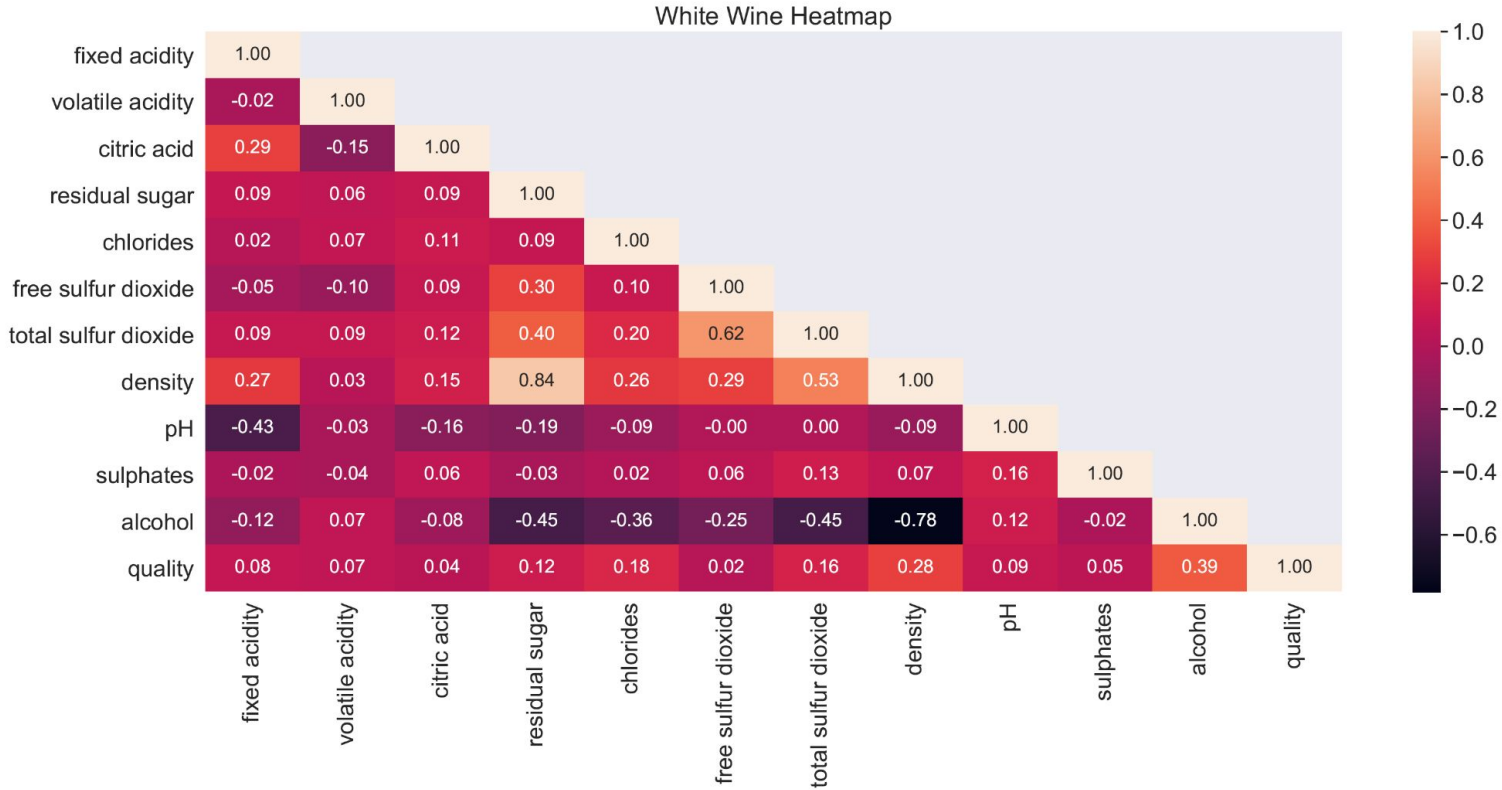


Modeling - Quality of Red Wine

- Three classification models were chosen
  - Random Forest
  - Stochastic Gradient Descent
  - Support Vector Classifier
- Hyperparameters were chosen using GridSearch
- ROC-AUC scores ranged from 0.82 to 0.92 with Random Forest leading

Classifier	ROC-AUC Score	Best Hyperparameter Values
Random Forest	0.924	{'bootstrap': True, 'max_depth': 55, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 275}
Stochastic Gradient Descent Classifier	0.892	{'alpha': 0.1, 'epsilon': 0.001, 'l1_ratio': 0.1, 'loss': 'log', 'max_iter': 1000.0}
Support Vector Classifier	0.829	{'C': 0.3162277660168379, 'gamma': 'scale', 'kernel': 'poly', 'probability': True}

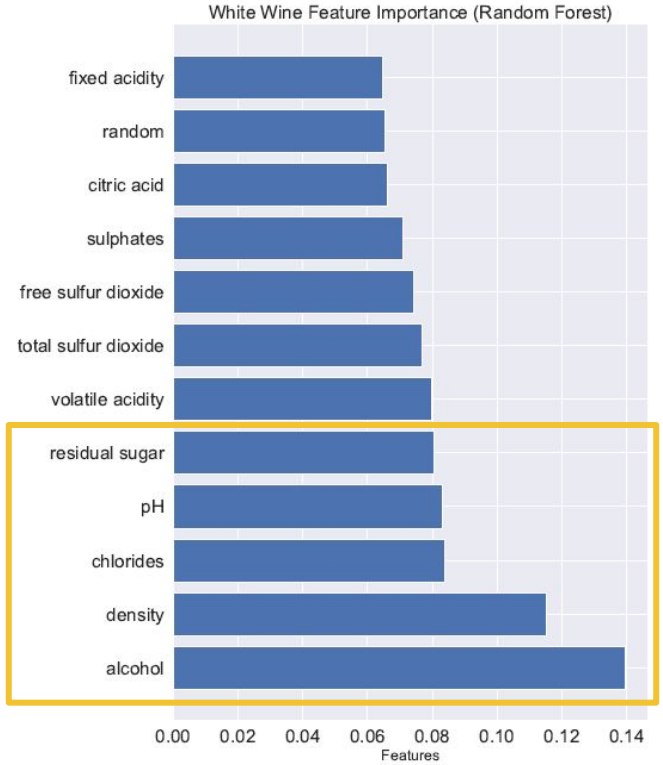
Exploratory Data Analysis - White Wine Heat Map



Heat Map of Features Generated from Pearson's R

Exploratory Data Analysis - White Wine Feature Importance

- Top five most important features for white wine quality
  - Alcohol content is again most important
  - Density is next important
  - Chlorides, residual sugar, and pH are next
- These features describe a sweet, acidic, strong, white wine
- The fixed acidity was less influential than the random background thus aren't as important in building the model



Modeling - Quality of White Wine

- Analysis performed was identical to Red Wine Quality
- Hyperparameters were chosen using GridSearch
- ROC-AUC scores ranged from 0.78 to 0.91 with Random Forest leading
- More samples of white wine could have led to more variability

Classifier	ROC-AUC Score	Best Hyperparameter Values
Random Forest	0.912	{'bootstrap': True, 'max_depth': 55, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500}
Stochastic Gradient Descent Classifier	0.784	{'alpha': 0.001, 'epsilon': 0.001, 'l1_ratio': 0.1, 'loss': 'log', 'max_iter': 1000.0}
Support Vector Classifier	0.837	{'C': 1.3894954943731375, 'gamma': 'scale', 'kernel': 'rbf', 'probability': True}

Exploratory Data Analysis - Determination of the Color of Wine

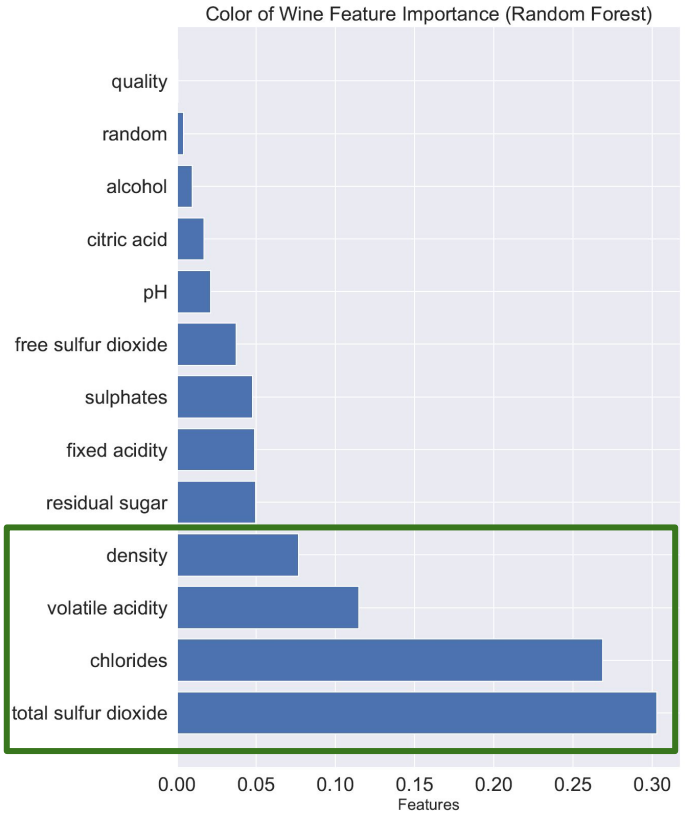
Average and Median Values for Wine Features

- Some features have obvious differences between red and white wines
  - Chlorides
  - Sulfur Dioxide (free and total)
  - Volatile Acidity
  - Residual Sugar
- Some features are more similar statistics
  - Density
  - Alcohol

	Red Wine		White Wine	
	Average	Median, 50%	Average	Median, 50%
Fixed Acidity	8.32	7.90	6.85	6.80
Volatile Acidity	0.53	0.52	0.28	0.26
Citric Acid	0.27	0.26	0.33	0.32
Residual Sugar	2.54	2.20	6.37	5.20
Chlorides	0.087	0.079	0.046	0.043
Free Sulfur Dioxide	15.9	14.0	35.3	34.0
Total Sulfur Dioxide	46.5	38.0	138	134
Density	0.997	0.997	0.994	0.994
pH	3.31	3.31	3.19	3.18
Sulphates	0.658	0.620	0.490	0.470
Alcohol	10.42	10.20	10.51	10.40

Exploratory Data Analysis - Determination of the Color of Wine Feature Importance

- Top four most important features for determining the color of wine
  - Total sulfur dioxide and chlorides have the most impact
  - Volatile acidity and density have some moderate impact
- These features highlight the major differences between the characteristics of red and white wines
- The wine quality was less influential than the random background thus aren't as important in building the model



Modeling - Determination of the Color of Wine

Classifier	ROC-AUC Score	Best Hyperparameter Values
Random Forest	0.999	{'bootstrap': True, 'max_depth': 55, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500}

- Analysis performed was similar to Red Wine and White Wine Quality
  - Hyperparameters were chosen using GridSearch
  - Only Random Forest was chosen because it performed best with previous models
- ROC-AUC scores with Random Forest
  - Training data = 1.0
  - Test data = 0.999

Recommendations for Wine Quality and Color Success

- Improving Red Wine Quality
  - Focusing on minimizing chloride and boosting alcohol levels to raise quality
  - Using citric acid as an additive while reducing volatile acidity may raise quality
- Improving White Wine Quality
  - Focusing on minimizing chloride and boosting alcohol levels to raise quality
  - Having smaller consistent ranges for total sulfur dioxide may boost quality
- Determining Color Characteristics of a Wine
  - Very prominent differences between red and white wine styles
  - Notably amounts of sulfur dioxide and chlorides present can determine color similarity of a given wine

Considerations, Improvement, and Future Work

- Strong correlations in heat maps
  - Try to reduce risk of collinearities within the data set
- Include more feature data in wine data set
  - Sales, grape type, and other proprietary feature information was not included in data set
  - Different types of wines from different areas may perform differently than the wines from the Vihno Verde variety; try wines from different locations
  - Adding subtypes (e.g. cabernet, pinot gris, etc.) to add more flexibility of the data set
- Reduce features needed for determination of color of wine
  - Attempting to use the top four features while eliminating extra features may improve model training time and eliminate possible collinearity
  - Looking at the wine can also determine color of the wine but integrating wines of intermediate types (e.g. rosé) may show more characteristics of red over white