Modeling to Predict Wine Quality

Final Project

YoungRi Lee & Jihyun Lee

Last compiled on 2020-11-07

1 Introduction

1.1 Data

This wine quality dataset is from [1]. The dataset includes *vinho verde*, a unique product from the Minho (northwest) region of Portugal. The data were collected from May/2004 to February/2007 using only protected designation of origin samples that were tested at the official certification entity (CVRVV). The CVRVV is an inter-professional organization with the goal of improving the quality and marketing of vinho verde. The data were recorded by a computerized system (iLab), which automatically manages the process of wine sample testing from producer requests to laboratory and sensory analysis.

The outcome variable is `wine quality`, which was measured by a minimum score of three sensory assessors using blind tastes in a scale that ranges from 0 (very bad) to 10 (excellent). There are 11 attributes of the wine based on physicochemical tests: fixed acidity (g(tartaric acid)/dm^3), volatile acidity (g(acetic acid)/dm^3), citric acid (g/dm^3), residual sugar (g/dm^3), chlorides (g(sodium chloride)/dm^3), free sulfur dioxide (mg/dm^3), total sulfur dioxide (mg/dm^3), density (g/dm^3), pH, sulphates (g(potassium sulphate)/dm^3), and alcohol (vol.%). Originally, two datasets were created separately, one for red wine ($n = 1599$) and another for white wine ($n = 4988$). In this report, we use a merged dataset and create a dummy variable to indicate the wine type, `red`. Thus, in total, the dataset includes 11 numerical attributes (covariates), one dummy variable, and one numerical outcome. There is no missing value in this dataset.

Table 1 shows the descriptive statistics of 11 attributes by wine type. [ADD SOME DESCRIPTION]

Figure 1 shows the distribution of wine quality by wine type. Generally, it shows a normal shape distribution and centered around the middle point of the scale. Red wine has fewer observations than

Table 1: Descriptive statistics of 11 attributes

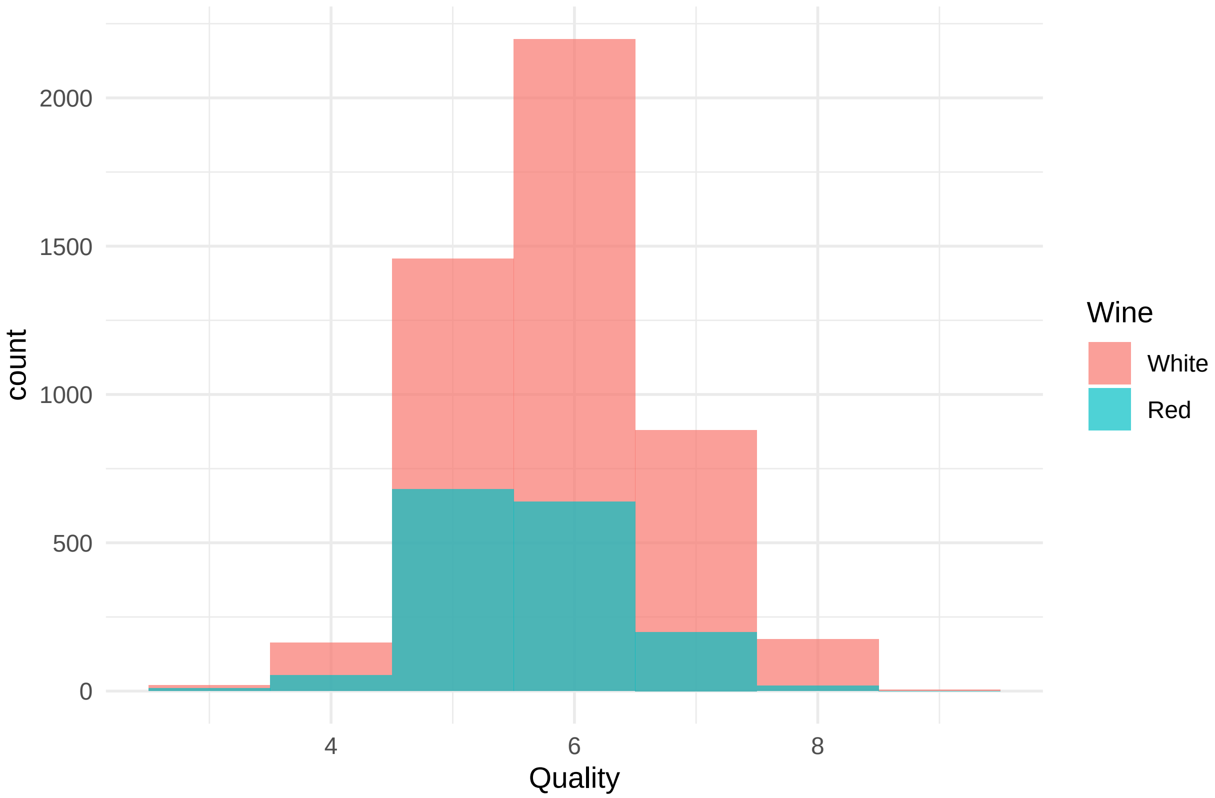
	White			Red		
	Min	Max	Mean	Min	Max	Mean
Fixed acidity	3.8	14.2	6.855	4.6	15.9	8.320
Volatile acidity	0.08	1.10	0.278	0.12	1.58	0.528
Citric acid	0	1.66	0.334	0	1.00	0.271
Residual sugar	0.6	65.8	6.391	0.9	15.5	2.539
Chlorides	0.009	0.346	0.046	0.012	0.611	0.087
Free sulfur dioxide	2	289	35.308	1	72	15.875
Total sulfur dioxide	9	440	138.361	6	289	46.468
Density	0.987	1.039	0.994	0.990	1.004	0.997
pH	2.72	3.82	3.188	2.74	4.01	3.311
Sulphates	0.22	1.08	0.490	0.33	2.00	0.658
Alcohol	8.0	14.2	10.514	8.4	14.9	10.423

Note:

The scale of each attribute is given in the text.

white wine.

Figure 1. Histrogram of wine quality by wine type



1.2 Research Questions

Using this dataset, we would like to (1) build the best model to predict wine quality, (2) examine the most influential set of attributes to predict wine quality.

To achieve this goal, we implemented two classes of techniques: shrinkage approaches (i.e., Ridge and Lasso regressions) and tree-based approaches (i.e., regression tree). This report includes the comparisons of two approaches, model selection procedures (e.g., cross-validation), and proposal of the best model.

[**Possible Implications**]

- Wine producers will use this information to produce the better quality of wine considering the selected attributes.
- Wine consumers will be able to choose a good quality of wine without tasting (e.g., via online shopping) if the physicochemical information of wine is available.

2 Methods

2.1 Shrinkage approaches

2.2 Tree-based approaches

3 Results

4 Discussion

4.1 Limitations

- Only uses the wine data from *vinho verde* region of Portugal. If a test dataset outside of this region is available, we would be able to test to generalize the results to a broader range of wine.

References

[1] P. Cortez et al. “Modeling wine preferences by data mining from physicochemical properties”. In: *Decision Support Systems* 47.4 (2009), pp. 547–553.