Predicting the Quality of Wines

Jae-Ho Lee

Abstract

Statistical learning techniques were utilized in predicting the quality of wines given their various properties and attributes. A data set that represents more of the lower-end and higher-end of wine quality should be included to yield a more effective model.

Introduction

Wine has a reputation for being a beverage of wide range and depth of taste, and such public perception of the beverage allows it to distinguish itself from other alcoholic beverages. While this image of the beverage may arouse pride in wine-buffs, it may also serve as deterrent for those interested in entering the world of wine. The rise of interest in wine led to a demand for the numeric rating of the quality of wine, which was popularized in ithe U.S. by Robert Parker in the 1970s [1]. Wine ratings lower the barrier to wine tasting help amateur wine-tasters more easily discover "good wine."

The rating of wine is usually undergone by aggregating the rating of one or more wine critiques. The following analysis performs statistical learning techniques on various physiochemical attributes of wines in order to predict the quality of wines. While the random forest model best performed in predicting wine quality, additional measures in the analysis could be made for improvement.

Methods

Data

The `wine`[2] data set was accessed through the `ucidata` package. It contains 6497 observations with 12 variables. The `quality` of the wine will be used as the response. Predictors include `color` (1599 red, 4898 white) and objective measurements of physicochemical properties of the wines such as `fixed acidity`, `citric acid`, `pH`, and `alcohol`.

[1] Wine Rating
[2] Wine Quality Data Set

Table 1: 5-Fold Cross-Validation

Model	Accuracy
Random Forest	0.6676
Naive Bayes	0.4867
Support Vector Machine	0.5571
Multinomial Regression	0.5419

Models

5-fold cross-validation was used as the validation measure for the trained models.

1. Random Forest

```
rf_mod = train(quality_cat ~ . , wine_trn,
                  method = "ranger",
                  metric = "Accuracy",
                  trControl = cv)
```

2. Naive Bayes

```
nb_mod = train(quality_cat ~ . , wine_trn,
                   method = "nb",
                   metric = "Accuracy",
                   trControl = cv)
```

3. Support Vector Machine

```
svm_mod = train(quality_cat ~ . , wine_trn,
                  method = "lssvmRadial",
                  metric = "Accuracy",
                  trControl = cv,
                  verbose = FALSE)
```

4. Multinomial Regression

```
multinom_mod = train(quality_cat ~ . , wine_trn,
                  method = "multinom",
                  metric = "Accuracy",
                  trControl = cv,
                  trace = FALSE)
```

Results

Accuracy was measured in order to assess the effectiveness of the models.

Table 2: **Random Forest** Accuracy

Test Accuracy
0.7098

Table 3: Test Confusion Matrix

	3	4	5	6	7	8	9	Sum
3	0	0	0	0	0	0	0	0
4	0	6	2	2	0	0	0	10
5	6	22	318	68	8	0	0	422
6	0	9	114	471	101	16	0	711
7	0	0	2	17	108	11	1	139
8	0	0	0	0	0	17	0	17
9	0	0	0	0	0	0	0	0
Sum	6	37	436	558	217	44	1	1299

Discussion

During testing, the random forest model was able to correctly classify the quality of the wines roughly 70.44% of the time. While this result is not bad, we can observe from teh confusion matrix below that the number of lower values (3 and 4) and higher values (8 and 9) are widely under-represented in the data set. Sampling for imbalance in data would yield a better performing model. Additionally, consumers with little knowledge of wine may be content with simply knowing of a wine is "good" or "bad." Manipulating the `quality` value to represent two (good/bad) or three (low/medium/high) values may lead to more accurate models that would be sufficient for the needs of beginner wine-tasters.

Appendix

Data Dictionary

- `fixed acidity`
- `volatile acidity`
- `citric acid`
- `residual sugar`
- `chlorides`
- `free sulfur dioxide`
- `total sulfur dioxide`
- `density`
- `pH`
- `sulphates`

- `alcohol`
- `quality`
  - Score between `0` and `10` based on sensor reading
- `color`
  - `White` or `Red`

Exploratory Data Anlysis

Figure 1: Wine by Color

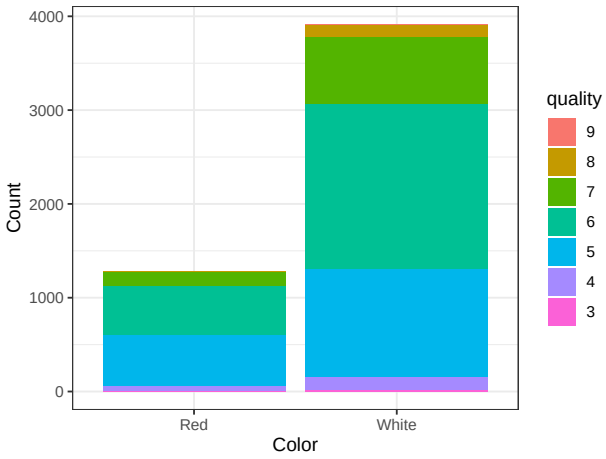

Figure 2: Wine by Quality

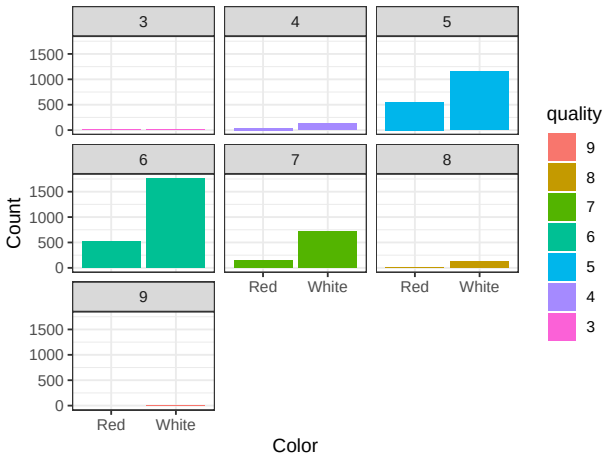

Figure 3: Alcohol Content by Color

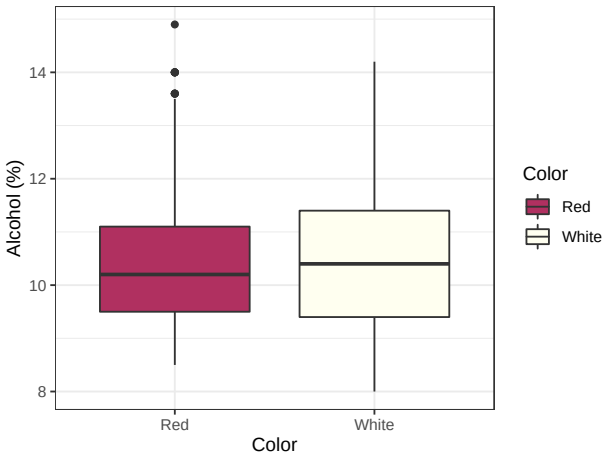

Figure 4: Quality by Alcohol Content