Analysis of Wine Quality

Executive Summary

Bethany Austhof and Adam Pratt

June 3, 2023

Introduction and data collection

Our goal was to **predict the quality of wines** using machine learning. Our data set considered 12 variables, 11 of which were objective variables and one a quality score. The data set contained 6497 observations of various wines, of which 4898 were white and 1599 were red. Data was obtained from the **UCI: Machine Learning Repository: Wine Quality Data Set**.

Variables

The 11 objective variables we considered were:

- **Fixed acidity** naturally occurring in the grapes
- **Volatile acidity** created in the fermentation process
- **Citric acid**, found naturally in grapes, but can also be uused as a preservative
- **Residual sugar** naturally occurring in grapes that remains post-fermentation
- **Chlorides**, which give wine its salinity, enhancing sweet flavors and dulling acidity
- **Free (unbound) sulfur dioxide**, another preservative
- **Total sulfur dioxide content**, which lowers the aroma
- **Density**, which is related to the concentrations of alcohol, sugar, and glycerol
- **pH**, a measure of the acidity of wine
- **Sulphates**, another preservative
- **ABV**, the percentage of alcohol by volume

The remaining variable, our dependent variable, is the **quality score**, a subjective rating of the wine on a scale from 1-10, with 10 being the best.

Our model

For our model, both red and white wine data sets were randomly sampled so that 80% of the data would be used for training and 20% would be for testing. As mentioned above, We had the **11 predictor variables**; our **dependent variable was the quality score**. We used a **Gaussian regression model** to fit the data, and then we used Predict to analyze the accuracy of our model's predictions.

Data visualization

To compare white wine and red wine, we initially created **two histograms** of the frequency of each quality score, separating between the white wines and the red wines. We saw that the upper and lower ends of the score range occurred very infrequently, leading us to eventually **remove these data points from our model** to improve our accuracy. To visualize our results, we created **two heat maps** displaying the fraction of accurately predicted quality scores, again separating between red and white. The two categories of wines ended up being **fairly similar in accuracy**; our two heat maps do a good job of displaying this similarity.

Problems with our analysis

We ran into a few issues in the course of this project. When initially running the model the most appropriate distribution to use seemed to be a **multinomial distribution**; however, with very few data observations for high and low quality scores, we didn't get a reliable model from this method. Our final prediction accuracy was **63.1% for red wine** and **58.4% for white wine**. We intend to look further into adjusting our model to see if we can increase the accuracy of our predictions.