Predicting Wine Quality

(Authors) Yang Liu

August 1, 2018

Domain Background

While the origin of wine predates all known human records, it is hypothesized that pre-agricultural humans picked berries and discovered fermentation through the process of attempted storage. While wine typically refers to the alcoholic beverage derived from the fermentation of grapes, it can also refer to similar beverages derived from other types of fruit.
Most wines fall into two major categorizations: red or white wine. Red wine is typically derived from dark colored grapes, in which the dark colored anthocyanins of the grape skins are allowed to color the wort, or liquid mash, in the creation process. In contrast, white wines are typically derived from light colored grapes, or from dark grapes with their skins removed before the mashing process.
Typically, during the fermentation process, no additional sugar or water is added. Elements such as sulfuric compounds, sugar content, acidity all arise from the chemical profile of the grape species and the specificities in fermentation, contributing to the flavor, texture, and fragrance of the wine.

Problem Statement

This project seeks to take a series of wine variables and, through either regression and/or classification, predict the quality of wine as determined by human tasters. It will also be interesting to see which of these families of algorithms performs better of such data.

Datasets and Inputs

The data set, sourced from the UCI machine learning repository, consists of 6497 items, separated into 1599 red wine entries and 4898 white wine entries. Each entry contains 11 input attributes:

- fixed acidity
- volatile acidity

- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol

These attributes may not contribute to the output attribute, quality, which is a discrete value between 0 to 10. This output value was generated by taking the median value of three values given by professional wine tasters. All wines are variants of the Portuguese wine "vinho verde", or "green wine".

Solution Statement

Since wine quality is a highly subjective metric, exacerbated by the metrics pulled from the median values of only 3 judges, there is likely a high degree of noise in the data. As such, highly generalized models are likely to be required to model the data at reasonably high performance.
Both regression and classification techniques will be tested for efficacy in finding relationships between the input variables and wine quality. Since both red and white wines are present, it may be important to separate these as separate problems, due to potential expectations on the part of judges affecting the quality metrics.

Benchmark Model

A non-competitive Kaggle competition with 15 current entrants reports a lowest error metric of 0.28923 (mean absolute error) on the private leaderboard. The competition suggests regression as the primary algorithmic family for prediction.

Project Design

Several algorithms in the supervised learning families of algorithms will be tested, focusing on regression and classification. Since the output value "quality" is an ordinal, discrete value, both regression and classification will likely work for this dataset. Random forests and other ensemble methods may prove effective, as may SVM's due to the possibility of non linear models between attributes. Non linear regression models may also prove effective in classification.