Optimizing Wine Preferences with Machine Learning (working title…)

Benjamin Berkman (bjb433), Zixiao Chen (zc2194), Stanley Sukanto (ss14358), David Trakhtenberg (dt2229)

*Center for Data Science | New York University*

***Abstract-***Wine sales have dramatically increased during the coronavirus pandemic. According to the Washington Post, Americans purchased nearly 30 percent more wine in June 2020 than in March 2020.[[1]](#footnote-1) Given the increased demand, quick and accurate classification of a wine’s quality is important in order ensure a positive consumer experience and fair price at market. This paper will study if the physicochemical properties of wine can sufficiently predict its quality through supervised learning approaches. As we will discuss, this has the potential to aide wine makers (oenologists), evaluators, and consumers. [brief overview of the findings of the model].

**1 Business Understanding**

Wine is a unique good that has objective, easily measured physicochemical properties yet a somewhat subjective quality rating that is displayed to consumers. These physicochemical properties are readily measured via laboratory tests and include measures of alcohol percentage, pH, and residual sugar, among other features that will be explained in this paper. The median value of three assessors’ blind tests determines a wine’s quality score (Cortez, Cerdeira and Almeida). These scores are marketed to the public, but there is often little transparency in what contributes to a good score. Thus, there exists a need to model the effect of these properties on a wine’s quality score, and to be able to predict a wine’s score given its properties.

A machine learning solution has the potential to address several business problems. First, a supervised learning model will be able to identify which physicochemical properties of wine most directly influence its quality. This may allow oenologists to directly increase certain physicochemical properties in the wine making process. For example, residual sugar – one of the properties to be studied – can be controlled via sugar fermentation in yeasts (Cortez, Cerdeira and Almeida). Additionally, this can serve as a quality control mechanism through the wine scoring process. If the median value of the three assessors’ scores significantly differs from the projected score of the model, it could serve as reason for re-evaluating the quality. (Ultimately, human evaluators may be rendered unnecessary given enough training feedback into the supervised learning model). Moreover, a machine learning solution may help test if wine assessors in training are scoring wine on par with the model. Finally, this paper will employ a brief unsupervised clustering approach to group similar wines. This may benefit consumers via targeted marketing. A vendor could employ a recommendation system, recommending wines with similar physicochemical properties given a customer enjoyed a specific wine.

As we will discuss in the next section, the data in this paper is limited to Portuguese white and red vinho verde wines introduced in 2009.[[2]](#footnote-2) However, this opportunity has the potential to be expanded to all wines that have both measured physiochemical properties and evaluation scores. With technological increases in the wine industry, these properties are increasingly feasible to measure. Additionally, as wine is no longer seen exclusively as a luxury good, a variety of applications, such as CellarTracker[[3]](#footnote-3) provide millions of user ratings for millions of wines. There does not appear to exist a database that maps these user ratings to the physiochemical properties of their respective wines, but this would serve as a valuable future research topic.

The state of the art in solving this problem has largely been limited to ad-hoc studies of small data sets of varying physiochemical properties such as the Portuguese data set in this discussion. Hopfer et al. measured five quality proxies against sensory attributes, volatile compounds, and elemental composition in Californian wine (Hopfer, Nelson and Ebeler). In other cases, hierarchical clustering was applied to both physicochemical and sensory assessments of Brazilian wine (de Castilhosa, Cattelana and Conti-Silva). However, there does not appear to be a large-scale machine learning model in production that predicts wine quality given its physiochemical properties.

**2 Data Understanding**

….

Clustering, an unsupervised learning technique, was also employed to understand the similarity of the wines in the data set. Red wines and white wines were clustered separately using both hierarchical clustering and k-means clustering. For this exercise, the quality of the wines was not considered, only the physiochemical properties. The hierarchical clustering demonstrated that the red wines could efficiently be clustered into two clusters, while the white wines fit nicely into three clusters. K-means clustering then separated the red and white wines into two and three clusters, respectively.

As shown in the end note, when plotted against density and pH, the clusters do form relatively strong boundaries. [[4]](#endnote-1) As demonstrated by Provost and Fawcett in their discussion of whiskey analytics, this clustering can help inform recommendations (Provost and Fawcett). If a customer enjoyed a specific wine, a vendor could Histogram

Description automatically generatedrecommend a wine with similar physiochemical properties (and therefore within the same dendrogram node or k-means cluster) which in turn may have a similar taste profile. If a vendor had a

limited budget, it could stock only wines within one cluster (to serve as a specialized vendor), or consciously pick wines from each cluster (to ensure a diverse selection).

**3 Data Preparation**

One major consideration involved whether to split up the red and white wine datasets. On one hand, they have different tasting notes so they should be analyzed separately. On the other hand, a model that can successfully aggregate the two could serve as a more robust prediction system. [Insert decision.] The target variable is defined as the quality of each wine, scored by wine assessors on a scale from 0 (very bad) to 10 (excellent).

A picture containing histogram

Description automatically generated

One note to make is the plots in figure 1 are normally distributed, indicating that there are far more normal scores than low or high ones. The target is broken into groups of “good” and “bad” wine for binary classification learners. A quality of 6 and above being good equitably splits up the datasets.

Scikit-learn’s built-in Decision Tree Classifier was utilized to estimate the normalized mutual information (MI) of each feature. Additionally, each feature was plotted to get its individual AUC (one plot per red and white). Alcohol has the highest MI and AUC for each dataset. Sulphates has a strong AUC for red wine while density seems to to be a good indicator for white wine quality. Alcohol is directly proportional to the amount of sugars fermented by yeast. It stands to reason that the volume percent of alcohol in a wine will affect the sensory taste. [Insert citation]

[Feature engineering]…

1. https://www.washingtonpost.com/lifestyle/food/wine-sales-have-surged-during-the-pandemic--but-not-for-small-producers/2020/06/19/c7c49cba-b0b6-11ea-8f56-63f38c990077\_story.html [↑](#footnote-ref-1)
2. https://archive.ics.uci.edu/ml/datasets/wine+quality [↑](#footnote-ref-2)
3. https://www.cellartracker.com/ [↑](#footnote-ref-3)
4. K-means plots, with two (red wines) and three (white wines) clusters: Chart, scatter chart

   Description automatically generated Chart, scatter chart

   Description automatically generated

   # Works Cited

   Cortez, Paulo, et al. "Modeling wine preferences by data mining from physicochemical properties." *Decision Support Systems* (2009): 4800-058.

   de Castilhosa, Maurício, et al. "Influence of two different vinification procedures on the physicochemical and sensory properties of Brazilian non-Vitis vinifera red wines." *LWT - Food Science and Technology* (2013): 360-366.

   Hopfer, Helen, et al. "Correlating Wine Quality Indicators to Chemical and Sensory Measurements ." *Molecules* (2015): 8453-8483.

   Provost, Foster and Tom Fawcett. *Data Science for Business*. O'Reilly, 2013. [↑](#endnote-ref-1)