How do the Physiochemical Characteristics of Wine affect the Wine's Sensory Quality Rating?

Breonna Moore

Project Design

I used a dataset from UC Irvine Machine Learning Repository to predict sensory wine quality. Insights gained from this model can be used by winemakers to be able to identify and improve poor quality wines in order to protect their brand.

The classes used to describe the wine's quality are poor, normal, and excellent. Modeling for this project was performed using Logistic Regression (LR) and Random Forest (RF). Both models were cross validated across three folds. I found that RF was the best model for predicting poor wine's quality for my dataset. The hyperparameters for the RF were tuned using GridSearchCV. RF increased the recall for the poor quality wines of 0.76 from the baseline model's 0.09, meaning that out of all of the wines truly labeled as poor quality, I was able to predict 76% of them correctly.

The RF model identified features that are the most important for predicting wine quality. The most important feature was percent alcohol content in the wine. Poor quality wines had an average of 10.39 percent alcohol content while normal and excellent quality wines had an average or 10.17 and 12.24 percent alcohol content, respectively. The next two most important features were free sulfur dioxide and volatile acidity. Both alcohol content and free sulfur dioxide are easily manipulated by adding sugar and salt, but in order to manipulate the volatile acidity, the wine has to undergo reverse osmosis and then be mixed with another uncontaminated, lower acidic wine. I would recommend that this model be used in advance to help determine the correct physiochemical properties of the wine before producing it, but if the wine has already been produced, then adjusting the percent alcohol content is the easiest alteration to make that will have the greatest impact on the wine's quality.

Tools

- Python
    - Data Collection: Pandas
    - Data Analysis: Scikit-learn, NumPy
    - Data Visualization: Matplotlib, Seaborn
- Microsoft PowerPoint, Microsoft Word

Data

The dataset from UC Irvine Machine Learning Repository was collected from the Viticulture Commission of the Vinho Verde Region. Originally the data's quality ranged

from 1 to 10, with 10 being the best, but I grouped the data into three categories of poor, normal, and excellent due to the lack of data points in both extreme ends of the quality range.

I used 12 features and 6497 rows of imbalanced data on the red and white wines from the northern Portugal region. The features are physiochemical properties of the wine in the dataset and are listed in Appendix 1.

To remedy the class imbalance in this dataset, I adjusted the class weights so that incorrectly predicting poor and excellent quality wines, the minority classes, came at a higher cost than incorrectly predicting normal quality wines, the majority class. I also adjusted the thresholds of the classes to underpredict. This bettered the prediction of poor quality wines since the majority of poor wines were initially being labeled as normal quality wines, the majority class.

What I Would Improve Upon Next Time

I would have liked to try stacking models to improve the class prediction for both the normal and excellent classes that I did not optimize for in this analysis. I wasted an entire day testing models in a way that was not very useful for my business case. To remedy this, I should seek more guidance sooner. I think that getting the guidance sooner, would have freed up more of my time in the end to be able to produce visuals in tableau or create a webapp with flask. Lastly, I would like to take advantage of the peer presentation practice group to get more practice presenting in front of a live audience and get actionable feedback from my peers.

Appendix 1: Features

Variable	Data Type	Description
Quality (Dependent)	Numerical/Categorical	Based on sensory data and scored 0 – 10, which will be converted to Poor - 0, Normal - 1, or Excellent- 2
Fixed Acidity (Independent)	Numerical	Most acids involved with wine or fixed or nonvolatile, do not evaporate readily
Volatile Acidity (Independent)	Numerical	The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
Citric Acid (Independent)	Numerical	Found in small quantities, citric acid can add 'freshness' and flavor to wines
Residual sugar (Independent)	Numerical	The amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
Chlorides (Independent)	Numerical	The amount of salt in the wine
Free Sulfur Dioxide (Independent)	Numerical	The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
Total Sulfur Dioxide (Independent)	Numerical	The amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes

		evident in the nose and taste of wine
Density (Independent)	Numerical	The density of water is close to that of water depending on the percent alcohol and sugar content
pH (Independent)	Numerical	Measures how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
Sulfates (Independent)	Numerical	A wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
Alcohol (Independent)	Numerical	The percent alcohol content of the wine

Data Sources

1. Data Set:
   P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
   Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.