Co-worked with Jieun Jang, Kartika Anjarsari, Minsu Kang

Wine Quality Prediction Model

1. Problem definition

Wine is a fermented grape-based alcoholic beverage. The sugar in the grapes is consumed by yeast, which transforms it to ethanol, carbon dioxide, and heat. It will be intriguing to explore wine's physicochemical characteristics and comprehend their correlations and significance in terms of wine quality and classification. A company aims to evaluate the quality of its wine production. Their goal is to improve their target marketing by modeling the tastes of the niche market they intend to target. Their brand offers both red and white wines. Machine learning can be used to determine which physicochemical characteristics contribute to the quality of a wine.

- Determine if each wine sample is red or white.
- Predict the wine sample's quality, which can be low, medium, or high.

2. Model Description

This study aims to search for the elements which affect wine quality by using multiclass decision classification methods such as K-NN, Logistic Regression, Random Forest, Confusion matrix, and GridSearch.

The dataset is associated with red and white "Vinho Verde" wine variants. Vinho verde is a one-of-a-kind product from Portugal's Minho (northwest) region. Only physicochemical (input) and sensory (output) variables are available due to privacy and logistical concerns (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

The data was obtained from the UCL machine learning repository, and the dataset contains 12 continuous variables.

- Fixed acidity: The total acidity is divided into two groups: the volatile acids and the nonvolatile or fixed acids.
- Volatile acidity: The volatile acidity is a process of wine turning into vinegar.
- Citric acid: Citric acid is one of the fixed acids in wines.
- Residual Sugar: Residual Sugar is the sugar remaining after fermentation stops, or is stopped.
- Chlorides: It can be an important contributor to saltiness in wine.
- Free sulfur dioxide: It is the part of the sulfur dioxide that is added to a wine.
- Total Sulfur Dioxide: It is the sum of the bound and the free sulfur dioxide.

  These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

3. Model results and interpretation

Highlights from Explorative Data Analysis(EDA) by wine type(red, white)

- Medium quality(5~7 from 0 to 10) wines are more frequent in white wines than red wines.
- White wine appears to have substantially higher mean residual sugar and total sulfur dioxide levels than red wine. The mean residual sugar level of white wine(6.39) is more than 2 times higher than that of red wine(2.54). The mean of total sulfur dioxide of white wine(138.36) was three times higher than that of red wine(46.47).
- Red wine appears to have a greater mean value of sulfates, fixed acidity, and volatile acidity than white wine.
- Based on the data, we can conclude that citric acid is more prevalent in white wines than in red wines.
- In general, white wines contain half the chloride concentration of red wines.
- Although the difference in Ph appears to be minor, it is worth noting that it is slightly larger in red wines.
- The level of total sulfur dioxide has the highest correlation(0.70) with the color(red/white) of the wine; free sulfur dioxide(0.47) and residual sugar(0.35) also show relatively high correlations.

Highlights from EDA by quality perspective of wine(low, medium, high):

- It's remarkable how alcohol doesn't inform us much about whether the wine is white or red, yet it makes a big difference in quality. It is worth noting that the average alcohol concentration increases by around 1% at each level of quality. Despite the fact that lesser quality wines have the lowest standard deviation.
- In higher-quality wines, chlorides and volatile acidity are less prevalent and have a lower standard deviation.
- The free sulfur dioxide concentration increases with quality, but the standard deviation decreases.
- Higher quality has less fixed acidity, but the standard deviation in mean quality is slightly higher.
- Alcohol shows the strongest correlation(0.44) with the quality of wine while other factors represent minor levels of correlations.

Highlights from applying machine learning algorithms

- The level of alcohol shows the highest(0.126) level of importance in the regression tree with the highest standard deviation.
- Among three models(Random forest, KNN, Logistic regression) we applied, KNN shows the highest accuracy (0.7065).
- The random forest model accurately predicted most of low quality or high quality wines, but presented some mispredictions in medium quality wines.

4. Business insight and recommendations

We conclude that it is possible to predict the quality of a wine and its type from the physicochemical

attributes. This prediction of quality presents some practical applicabilities, like:

- The result of the model, its interpretation and all the evaluations of EDA provide methods and rules of decision that can help winemakers to determine which physicochemical characteristics contribute to the quality of a wine, therefore increasing the quality of its wine production.
- The result of the model can help winemakers to allocate ingredients more efficiently to control wine production depending on its type and quality.