Predictive Analytics Term 4, 2020

Associate Professor Ole Maneesoonthorn

Predicting Wine quality

Wine making is a complex task, with many attributes contributing to the quality of the final product. These include the variety quality of the grapes, the climate in which the wine is made and stored, the production process, the properties of wine containers and the bottling process. Some of these attributes are measurable but some are not.

Even though some of these factors are not measurable, there are certain physicochemical properties of wine that can be measured. Examples of these include various acid levels, pH level, sugar content, alcohol content and other chemical compounds that can be measured with laboratory equipment.

The measurement of wine quality is highly dependent on the subjective judgement of independent wine experts. A score ranging from 0 to 10 is given following a thorough quality assessment and tasting procedure. These quality scores are widely available to consumers and retailers, and can have significant impact on retail pricing, advertisements, shelfing and ultimately sales of the wine itself.

Since the wine quality score is unknown to the wine maker until the wines are released, a predictive model would be desirable. These predictive outcomes can be used in multiple folds:

1. To try to gauge at what factors may contribute to the final quality score.
    - This would wine makers understand how to make interventions during the production process to improve the quality score of their wines.
    - It would also contribute to the planning of the wine making process in the upcoming season to optimize the balance between costs and quality score.

2. To obtain the potential quality score for a new wine, which would be useful in:
    - Bottling/package planning – wine with higher quality score can be bottled with premium packaging, while wine with poor quality score may be bottled and packaged differently (larger volume or bulk).
    - Marketing planning – wine with higher potential quality score may be allocated larger marketing budget, with more sophisticated marketing campaign.
    - Pricing – if the quality score is potentially high, a higher margin level can be factored into the pricing of the wine.

The data files “red.csv” and “white.csv”, contains the physicochemical properties and quality score of red and white wines, respectively, produced in Portugal. (Data source: UCI Machine Learning Repository.) The descriptions of the data are given below:

- FA = fixed acidity (g of tartaric acid per dm^3)
- VA = volatile acidity (g of acetic acid per dm^3)
- CA = citric acid (g per dm^3)
- RS =
- residual sugar (g per dm^3)
- Ch = chlorides (g of sodium chloride per dm^3)
- FSD = free sulfur dioxide (mg per dm^3)
- TSD = total sulfur dioxide (mg per dm^3)
- Density = Density of the wine (g per cm^3)
- pH = pH level
- Sulphates = Sulphates level (g of potassium sulphate per dm^3)
- Alc = alcohol (vol. %)
- QS = quality score (range from 0 to 10)

Predictive Analytics Syndicate Task #1

Choose whether you want to explore the red or white wine range. You will be exploring the wine of your choice for the entire course, so choose carefully!

1. Explore the properties of the data, including
   a. Average of each variable
   b. Histogram of the quality score variable
   c. Relationship between the physicochemical properties and the wine quality score using the correlation function.
2. Split the data into training and test sets. Use random sampling, with 80% allocated to the training set and 20% allocated to the test set. Use the linear regression to build a predictive model for the wine quality score using the training set data. Explore the following:
   a. Multiple linear regression model with all physicochemical properties.
   b. Stepwise regression.
   c. Any potential nonlinear relationships.
3. Obtain the predictions for wine quality score for the test set data. Compute the predictive accuracy metrics for the test set. Comment on which model in your consideration set performs best.
4. Investigate the best performing predictive model you have chosen. Discuss what your chosen model reveals about the relationships between the wine quality score and the physicochemical properties.
5. What are the limitations of your analysis? Discuss.

R-square: how well fit
in training data
→ doesnt mean it can predict

Produce a 3-page report that summarizes your analysis.

This task is due at 6pm on Saturday 3rd October 2020.

lack of data

hidden var;
tester (1 vs a panel) } → data →

linear is not the right tool ⇒ the technique
maybe there is more method
linear is structure and not flexible; grouping
and cluster