Advances in Food Science and Human Nutrition (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/afshn.2022.040101
ISSN 2616-258X Vol. 4 Num. 1

Predicting Wine Score based on Physicochemical Properties

Congrui Li

School of Statistics, Beijing Normal University, Beijing, 100875, China

Keywords: Random Forest, Regression Tree, Wine Score, physicochemical properties

Abstract: Existing wine scoring systems are based on subjective feelings, which are difficult to quantify. Thus, the score of a particular wine has little guidance on the brewing process. In this paper, regression tree and random forest are used to predict wine score by using the physicochemical properties of wine. According to the results of the model, higher quality wine can be produced by controlling the physical and chemical properties, thus reducing the waste of raw materials.

1. Introduction

As brands and types of wine grow, many consumers refer to the score of wine when choosing one. Wines with high scores are favoured, while wines without or with relatively low scores are at risk of unselling.

At present, the common wine scoring system is mainly divided into two types. One is the scoring system established by famous wine commentators, and the other is the scoring system established by famous wine magazines. These two systems are based on the sensory of the wine evaluator, that is, for the brewer, there is no objective control standard for whether a wine is a high score wine or a defective wine in the brewing process. If the inferior products are brewed, they will face loss problems and bring great waste.

Therefore, this paper attempts to establish a link between the physical and chemical properties of wine and its score. Using the physical and chemical properties of wine to predict its score by regression tree and random forest. In this way, in the brewing process by controlling the physical and chemical indicators to achieve the purpose of brewing high grade wine, reduce the rate of inferior products, and reduce the waste of raw materials.

2. Notation

The data comes from UCI [1][2]. It has 11 input variables and 1 output variable, as shown in table 1.

Table 1: Notation

Symbols	Meanings
Fix	Fixed Acidity
Vol	Volatile Acidity

Cit	Citric Acid
Res	Residual Sugar
Chl	Chlorides
Fre	Free Sulfur Dioxide
Tot	Total Sulfur Dioxide
Den	Density
Ph	Ph
Sul	Sulphates
Alc	Alcohol
Qua	Quality

3. The regression tree

The process of constructing regression tree: Split the predictive variable space into K nonoverlapping regions $R_1, R_2, \dots \dots, R_k$. Make the same prediction for each observation falling into the same region R_i, and the predicted value equal to the simple arithmetic average of response value of the training set on R_k. [3]

Firstly, the data are pre-processed and it is found that the overall distribution of the data is very uneven. Most wines scored ordinary, just a little wine have low or high scores. This conclusion can be confirmed by box-plotting wine scores. Figure 1 shows the result.

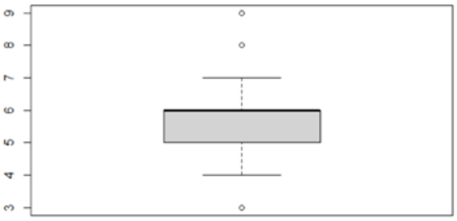

Figure 1: Boxplot of wine score

From the figure above, we can see that the high score wine with 8 points and 9 points and the low score wine with 3 points are outliers. But these data are of great importance to the objectives of this study because we want to know what wines are more likely to get high scores.

Data screening showed that 180 samples scored 8 or 9 and 20 samples scored 3, and there are 4698 samples scored in the range [4, 7]. This shows that if the regression tree algorithm is used directly to the original data set, high-quality wine will be buried in ordinary wine. The inner node of the regression tree will not truly reflect the characteristics of high-score wine, and the terminal node will also tend to the average score of ordinary wine.

To verify the above conjecture, I made the regression tree images using all the data as follow, the result is shown in Figure 2.

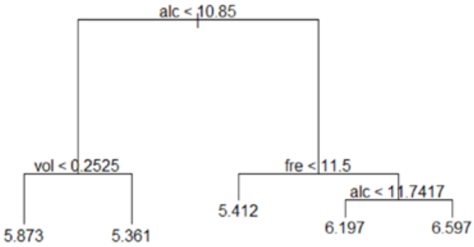

Figure 2: Regression tree using all data

The results were the same as expected. The data of the terminal node (leaf node) were all centralized around 5 and 6, the low- and high-score wines were not reflected at all.

In order to avoid this phenomenon, we cannot use all samples of ordinary wine. Instead, I sample 100 wine from the ordinary wine samples, and form a new data set with all high- and low-score wine samples to construct a regression tree, and obtained the following result (Figure 3).

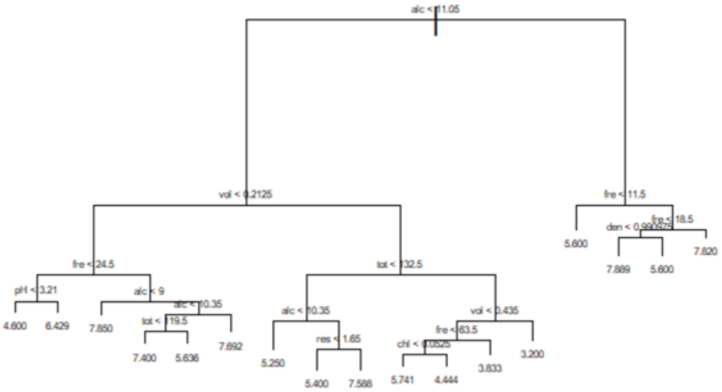

Figure 3: The new regression tree

At this time, the regression tree's terminal node has higher scores and lower scores. There are too many internal nodes, thus the terminal nodes overlapped slightly. To clearly show the terminal node value, the data is organized into the table 2.

Table 2: Terminal Node Value of Regression Tree

Number	Value	Number	Value
1	4.600	10	5.741
2	6.429	11	4.444
3	7.850	12	3.833
4	7.400	13	3.200
5	5.636	14	5.600
6	7.692	15	7.889
7	5.250	16	5.600
8	5.400	17	7.820
9	7.588		

At this time, there are values like 7.889 and 3.200 appeared at the terminal nodes of the regression tree, which means the samples of low-alcohol and high-alcohol are fully considered in the model.

However, the regression tree has obvious problems. The amount of leaf nodes is too large, the tree is too complex, that means there might be overfitting problem. Therefore, the regression tree is pruned as follow (Figure 4).

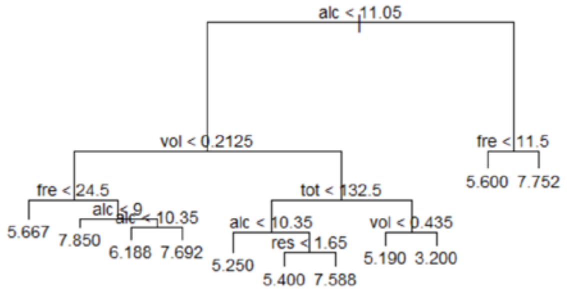

Figure 4: The pruned regression tree

It can be seen that the major factors are alcohol, volatile acid, free sulfur dioxide and total sulfur dioxide.

4. Random forest

Random forest is an integrated algorithm, which uses bagging method to form a training set of each tree and establish a series of decision trees.

Similarly, considering that most of the wine in the original data set have medium scores, low- and high-score wine will be submerged. I extracted 100 samples from the medium score wine, and formed a new data set with low score wine and high score wine. The new data set was divided into training set and test set according to 7:3.

To use the information in the dataset as much as possible, the process was repeated 1000 times and save each test error to a vector.

For each time, the test error is calculated by the formula below:

$$\frac{1}{n}\sum_{i=1}^{n}(predict_i - test_i)^2 \qquad （1）$$

Where n is the number of sample in test set

Taking the mean value of each element of the vector as the final test error. The result is 1.729. To avoid extreme values, output the maximum element of the vector, the result is 2.208, which is acceptable.

Output the importance of variables, we can get the table 3.

Table 3: Importance of variable

Variable	IncNodePurity
Fixed acidity	37.526
Volatile acidity	61.730
Citric acid	14.470
Residual sugar	24.406
Chlorides	19.078
Free sulfur dioxide	80.195
Total sulfur dioxide	71.048
Density	48.444
pH	13.356
Sulphates	18.281
Alcohol	66.713

Put the above results in a bar chart, as shown in the figure 5.

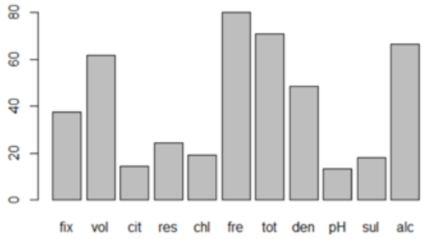

Figure 5: The importance of variables

We can ger that the major factors are free sulfur dioxide, total sulfur dioxide, alcohol and volatile acid.

5. Results

The results from regression trees and random forests need to be analysed in conjunction with the wine scoring process. Wine score is an assessment of four aspects of wine: appearance (colour), aroma, taste and overall impression score. [3]

The important variables analysed by regression tree and random forest will affect some of the aspects above. Volatile acid affects colour, taste and aroma of wine; sulfur dioxide affects aroma of wine; alcohol affects taste and aroma of wine.

6. Discussion

According to the results above, random forest and regression tree can be used to predict wine score by physical and chemical properties, and the prediction error is acceptable. Controlling these physical and chemical properties during brewing is more likely to produce higher quality wines. However, the model still has some limitations: 1. the model is built based on the data of white wine. Haven't applied the model on the data of red wine. 2. There may be some potential variables not included in the model, such as tannin content.

In further research, the model for red wine can be built, and more physical and chemical properties can be considered to improve the present model.

References

[1] Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2]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.

[3] https://www.wine-world.com/culture/pj/20150605173918670