WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Using ROC-curve to Illustrate the Use of Binomial Logistic Regression Model in Wine Quality Analysis

Samrajya Bikram Thapa [a]*

[a]*School of Engineering, University of California-Merced, Merced, 95343, USA*

DOI: https://doi.org/10.55248/gengpi.2022.3.10.67

A B S T R A C T

The major objective of this paper is to predict wine quality based on chemical attributes using binary logistic regression model. The evaluation of wine quality in this study is analyzed based on statistical methods, correlation analysis, and regression analysis. This study highlights several features are not relevant (only six essential variables were enough) to predict wine quality. The result showed that this method is robust enough to predict wine quality with an overall accuracy of more than 75%. Similar techniques or machine learning methods can help in predicting more accurately.

Keywords:Logistic regression, ROC, Accuracy, Precision, Sensitivity, Specificity.

1. INTRODUCTION

Today, various customers appreciate wine and hence consumption has increased rapidly over the last few years. In the current competitive market, wine quality is crucial for both consumers and producers. Quality of wine is a characteristic involving the combination of different physiochemical components of the wine, such as pH, alcohol content, total sulfur, and anthocyanin levels, which are all important in wine quality certification [6]. This type of quality certification helps ensure the quality of wines on the market. Today, all wine industries are applying new techniques and implementing new technologies and applying these in all areas. There have been other studies conducted to analyze different factors that contribute to the quality of wine. Sun et al. [13] predicted 6 geographic wine origins based on a neural network with 15 input variables and reported a 100% prediction rate. Dahal et al. [7] compared performance of various machine learning methods and reported Gradient Boosting Regression (GBR) surpassed all other models' performance. Chen et al. [5] proposed a method for predicting wine quality using human flavor reviews and obtained about 76% accuracy using a hierarchical clustering approach and an association rule algorithm. Building off of prior research, this study focuses on white wine dataset accessed from UC Irvine Machine Learning Repository [2], where the goal is to predict the quality of the wine based on physiochemical (predictor) variables. In this study, binomial logistic regression model is implemented to predict the wine quality using the physiochemicaltests features, and to understand the relationships between predictor variables and their effectiveness on the quality of wine. Moreover, the predictions are also made for wine quality based on important variables selected according to their dependencies. Model selection is compared through training and test data set and where applicable, the Akaike Information Criterion (AIC). The remaining sections are arranged as follows: study data and methodology are explained in section 2. In section 3, outcomes of the statistical and regression analysis are explained and reported. Section 4 highlights the conclusion and future directions.

2. MATERIAL AND METHOD

For this study, publicly available wine quality dataset was extracted from the UCI Machine Learning Repository [2]. There are 11 variables in the dataset, all of which are based on physiochemical tests (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol) and 'Quality' as an output variable. The 'Quality' rating is based on sensory test and scaled in 11 quality classes from 0- very bad to 10- excellent. 'Quality' variable was recoded into dichotomous variable with 0 (quality score 5 or below) and 1 (quality score above 5). In analyzing the data, the data was partitioned randomly into training and testing data in ratio of 4:1 respectively. To identify the relationship between the response and predictor variables, trained data were used. Data partition is primarily done to avoid over performance of model in training dataset and under performance in testing dataset.

For the calculations, R software packages [10] were used. The analysis process includes the following steps: (i) calculation of correlation and removal of correlated predictors; (ii) modelling using binary logistic regression model; (iii) use of ROC-curves to evaluate the results. As analyzing a summary statistic, some variables are widely spread. For example, the values of 'residual.sugar', 'fixed.acidity', 'free.sulfur.dioxide' and 'total.sulfur.dioxide' are extremely large compared to other variables. Such a large value can have a dominance over other quantities during the training process in regression models. One way of dealing with such problem is transforming data by standardization and normalization. Variables with exceptional high values were transformed using formula:

$$z = (x - mean)/sd \quad (1)$$

where z, x, mean, and sd are standardized input, input, mean, and standard deviation respectively.

Before passing the data into regression model, correlation coefficient was calculated and variables with high correlation coefficient were dropped to avoid multicollinearity using 'vif' function. Variable with more than 8 vif score were not passed into the regression model (Density variable was dropped from further analysis at this point at it had vif score of 19). Having highly correlated variables in a model cause problem when fitting the model and interpreting the results. In this paper, binomial logistic regression model is used to determine dependency of wine quality on predictor variables. Binomial logistic regression is used to model the relationship between a dependent and one or more independent variable. It is employed to predict a variable's value based on the values of two or more other variables. The response variable (Quality) is handled as a binomial with values of 0 and 1, depending on whether the wine is of poor or good quality respectively.

$$\ln(p/(1-p)) = \beta_0 + \sum_{i=1}^{n} \beta_1 * x_i \quad (2)$$

where p is probability of occurrence of good quality wine, β_0 is the intercept, β_1 is the regression coefficient, and x_i is an independent variable. Same framework has been used to predict response variable in numerous studies in wide range of research. Jafari et al. [9] compared binary and multinomial logistic regression model to produce soil class maps and explained how binary logistic regression model helps in selection of better covariates. Sam et al. [11] used stepwise conditional binomial logistic regression model to estimate whether pixels for a given fire have burned in any low, moderate, or high severity classes. Another study applied binary logistic regression model to explain effect of place attachment and interest in longleaf pine to apply prescribed fire [14].

The AIC score was applied to choose between the fitted models (Akaike 1974). This favors sparse models by adjusting the residual deviation for the number of predictors. The model that had the lowest AIC and residual deviation was therefore chosen. For model performance and evaluation, the accuracy, precision, sensitivity, specificity numbers as well as Receiver Operating Characteristics – Area Under the Curve (ROC-AUC) score, were all calculated and used to assess performance of the model. (Sidey-Gibbons &Sidey-Gibbons, 2019).

Accuracy is defined as the value that is predicted when the sum of the True Positive Rate (TPR) and True Negative Rate (TNR) values of a confusion matrix is divided by the sum of the True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR) values. [3].

$$\text{Accuracy}=(TPR+TNR)/(TPR+TNR+FPR+FNR) \quad (3)$$

Precision refers to the value obtained when True Positive is divided by the sum of True Positive and False Positive values of a confusion matrix [4].

$$\text{Precision} = TPR / (TPR + FPR) \quad (4)$$

The true positive rate, also known as sensitivity, is the model's ability to identify all positive events from the total of true positives and false negative values of a confusion matrix [8].

$$\text{Sensitivity} = TPR / (TPR + FNR) \quad (5)$$

The true negative rate, also known as specificity, is the number of accurate negative class predictions divided by sum of True Negative and False Positive of a confusion matrix [4].

$$\text{Specificity} = TNR / (TNR + FPR) \quad (6)$$

The ROC curve shows the sensitivity on y-axis and the (1-specificity) on x-axis for varying cut-off points. The Area Under the Curve (AUC) which is calculated by integrating the areas under the line segments, is a popular measure of a variable's ability to predict outcomes in ROC analysis. The AUC is a useful combined measure of sensitivity and specificity for evaluating the model's intrinsic validity.

3. RESULTS

A binomial logistic regression was employed to find the best predictors of the response variable. Six of the eleven variables were significant in the final model are considered important predictors of quality as p-value for these predictors is less than 0.05 (95% confidence interval). Table 1 represents the regression summary of the final model for response variable. Final model explained only 25% of the variance in determining the wine quality (as explained by adjusted $R^2 = 0.2515$), which shows there were more than one predictor variable which were not predicting the quality of wine. The standard regression coefficient for "Volatile acidity" is -3.82833, which indicates that if all other predictors are constant (controlled) then

increment of one unit of volatile acidity decrease the quality score by 3.82833. The same statement can be made for other predictors. Here, negative sign in standard regression coefficient indicates decrease and positive sign indicate increase in wine quality.

Table 1: Final binomial logistic regression model with only statistically significant variables.

Significant variables	Standard regression coefficient	Standard error	p-value
Volatile acidity	-3.82833	0.44750	0.000000*
Sulphates	0.68251	0.26724	0.001282*
Alcohol	1.07352	0.04571	0.000000*
Fixed.Acidity	-0.12218	0.03913	0.001797*
Residual sugar	0.65279	0.07374	0.000000*
Free sulfur dioxide	0.13904	0.04195	0.000919*

* Significant at p <0.05

Table 2 represents prediction, accuracy, sensitivity, and specificity values of 'quality' using binary logistic regression analysis for white wine dataset (train and test data) which is calculated based on confusion matrix using equation 3,4,5,6 as explained in method section. Confusion matrix takes two parameters one is predicted values by the model and other is actual values. Table 2 proves that binary logistic regression model gives an accurate prediction for selected features rather than all predictor variables.

Table 2: Model performance metrics obtained using training and testing datasets

	Train Data (%)	Test Data (%)
Accuracy	75.60	75.41
Precision	88.80	88.40
Sensitivity	77.8	77.20
Specificity	68.4	70.16

Fig. 1 shows the ROC curve with AUC score of 0.797 for the selected binomial logistic regression model of wine quality. High AUC value in this study is a useful parameter for evaluating the overall value of an indicator variable. Best model was selected based on highest AUC score (0.797) and lowest AIC score (3975.24).

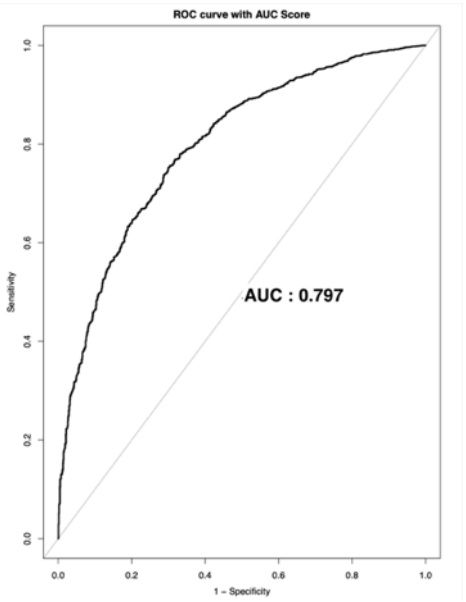

Figure 1: ROC curve with AUC score of wine quality binomial logistic regression model.

4. CONCLUSION

The wine industry has seen an increase in interest in recent years, necessitating growth. This study demonstrated that binary logistic regression model can be used to analyze the parameters in the existing dataset to determine the wine quality. This study demonstrates that, rather than considering all predictors when making a prediction, most significant or important predictors should be taken in account. Model presented in this paper have accurately labeled 75.6% of the test data and 75.4% of train data, and accuracy could be increased possibly by using different machine learning

algorithm. Wine producers can use this finding and manufacture wine with properties that more consumers would appreciate and that are of higher qualities. Large wine datasets can be used for studies in the future and different machine learning algorithms may be explored for wine quality prediction.

REFERENCES

[1] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, pp. 716–23

[2] Asuncion, A. & Newman, D. (2007), UCI Machine Learning Repository, University of California, Irvine, [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[3] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research. 13. pp. 281–305.

http://dx.doi.org/10.5555/2188385.2188395.

[4] Bhardwaj, P., Tiwari, P., Olejar Jr., K., Parr, W., & Kulasiri, D. (2022). A machine learning application in wine quality prediction. *Machine Learning with Application, 8*, 100261.

[5] Chen, B., Rhodes, C., Crawford, A., & Hambuchen, L. "Wine informatics: applying data mining on wine sensory reviews processed by the computational wine wheel," IEEE International conference on Data Mining workshop, 2014.

[6] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4),pp.547–553. http://dx.doi.org/10.1016/j.dss.2009.05.016.

[7] Dahal, K. R., Dahal, J. N., Banjade, H., & Gaire, S. (2021). Prediction of wine quality using machine learning algorithms. Open Journal of Statistics, 11, pp. 278–289.

http://dx.doi.org/10.4236/ojs.2021.112015.

[8] Haury, A. C., Gestraud, P., & Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS ONE, 6(12), e28210. http://dx.doi.org/10.1371/journal.pone.0028210.

[9] Jafari, A., Finke, P. A., Van De Wauw, J., Ayoubi, S., & Khademi, H. (2012). Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: comparing logistic regression approaches to predict diagnostic horizons and soil types. *European Journal of Soil Science. 63, pp.* 284-298.

https://doi.org/10.1111/j.13652389.2012.01425.x

[10] R Core Team 2021 R: A language and environment for Statistical Computing (Vienna: R Foundation for Statistical Computing) (available at: www.r-project.org/)

[11] Sam, J. A., Baldwin, W. J., Westerling, A. L., Preisler, H. K., Xu, Q., Hurteau, B. M., Sleeter, B. M., & Thapa, S. B. (2022). Simulating burn severity maps at 30 meters in two forested regions in California. Environmental Research Letters, 17(10), 105004.

[12] Sidey-Gibbons, J., & Sidey-Gibbons, C. (2019). Machine learning in medicine: a practical introduction. BMC Medical Research Methodology, 19.

http://dx.doi.org/10.1186/s12874-019-0681-4.

[13] Sun, L., Danzer, K., & Thiel, G. (1997) "Classification of wine samples by means of artificial neural networks and discrimination analytical methods". Fresenius Journal of Analytical Chemistry 359 (2), pp. 143–149.

[14] Thapa, S. B. (2018). Perceptions regarding longleaf pine ecosystem restoration using prescribed fire. Master's thesis, Mississippi State University, [Online]. Available: https://scholarsjunction.msstate.edu/td/3493