Insights from Wine Quality: Linear Regression & Decision Tree Classifier

Kishan Shah, Prabhneet Arora and Shirish Pandagare

***Abstract*:** Wine Quality Data has been used to explore a linear regression fit to predict quality for white wine. Also, Decision Tree classifier was used to classify the two types of wines (red and white). The accuracy score of the data distribution was computed to infer the true class probabilities (wine type) given the predictors of a specific wine. Furthermore, the application of linear regression model trained on white and tested on red was done as an attempt to draw insights into model reuse. Histograms depicting the difference in quality across the same models and model reused for a different dataset is used to draw concrete conclusions on model reuse.

Index terms: Linear Regression, Binary Decision Tree Classifier, Model reuse, Histograms

I. INTRODUCTION

Wine Quality data is a very popular dataset publicly used for research purposes with details being described by Cortez et al [1]. Vinho Verde is a unique product from northwest region of Portugal. Medium in alcohol, its specifically appreciated due to its freshness (especially in summer) [2]. Two datasets pertaining to red and white "vinho verde" wine samples, from the north of Portugal have been used to test regression, using Linear Regression and Classification, using Binary Tree Classification methods.

A very interesting concept of model reuse has been introduced to create a generic and stable model for deployment in various circumstances in comparison to generating a wide variety of more specific and volatile models upon request. Based on this intuition, quality of the red wine was predicted on the linear regression model trained on white wine dataset to obtain an RMSE value of 0.94286 (according to Kaggle score).

II. METHODOLOGY

A. Multivariate Linear regression method is used to obtain a model to predict the quality of the white wine. The model is trained on White Wine training dataset to obtain the coefficient of the model and the accuracy of the model is evaluated based on the RMSE. RMSE is a frequently used criterion for measuring differences between estimated values (from our prediction model) and observed values [3]. Further, using this trained model we have predicted the quality of the white wine based on the white wine testing dataset and we got RMSE of 0.7130 (according to Kaggle score).

B. Decision Tree Classifier is used to classify the red and white wine from the given dataset and the performance of the model was accessed based on the accuracy score which returns the fraction of correctly classified samples. Firstly, we have used Binary Decision Tree classifier to predict the type of wine. The model was trained on the combined dataset of red and white wine and accuracy obtained was 99.29% (according to Kaggle score). Further, a more robust Decision Tree Classifier, Random Forest was used to improve the accuracy of the model and to prevent it from overfitting. After training the Random Forest model, we were able to predict the type of wine with an accuracy of 100% (according to Kaggle score). It can be inferred from these high classification accuracies that these two-wine datasets are significantly apart.

C. The linear regression model trained on white wine dataset was used to predict the quality of the red wine. The RMSE for the predicted quality of red wine obtained is 0.94286, which is higher in comparison to RMSE for red wine quality prediction obtained from the model trained on red wine (0.67461).

III. RESULTS

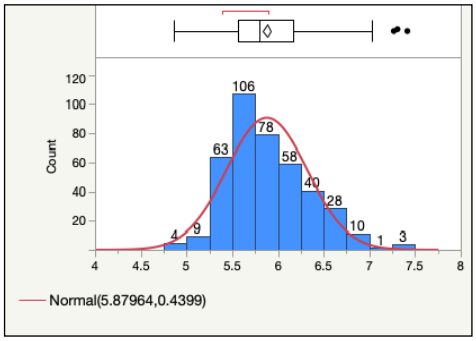

Fig 1(a). Red wine quality using white wine model

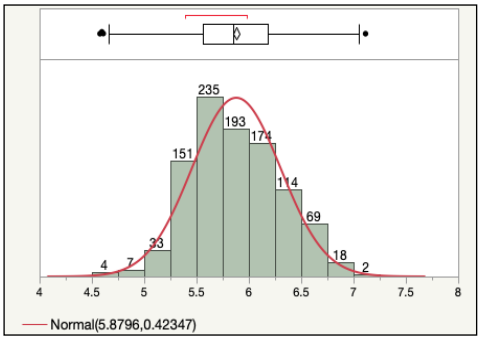

Fig 1(b). White wine quality using white wine model

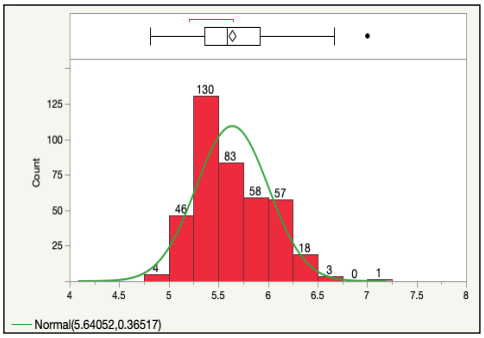

Fig 1(c). Red wine quality using red wine model

From the results of part 1 and part 3 of the challenge, we can say that the white wine model has not performed well on red wine test dataset as RMSE recorded was 0.94286. This contrasted with the red wine testing data predicted on the linear regression model of red wine which resulted in RMSE of 0.67461. This was done instead of comparing with the white wine testing accuracy due to the inconsistency in the test data samples of red and white wine i.e. 400 and 1000 respectively. Empirically, white wine would result in higher RMSE (0.70130) values due to more data samples in white wine testing data as compared to red. The figures above precisely represent the distribution of data for the three cases as discussed above. All three figures shown above represent normal distribution of the data centered around quality of 5.5.

Also, in the classification model, we achieved 100% classification accuracy through Random Forest classifier for classifying red wine and white wine using combined data set. Having such a higher accuracy, we can infer that our model can correctly classify the two datasets for white wine and red wine.

IV. CONCLUSION

From our analysis, it is quite clear that firstly, if the datasets are drawn from similar populations and have little variation in mean and standard deviation then one model can be deployed to predict the response of other population. Here, our histograms clearly signified that the distributions of the two wines was quite similar, so the model of white wine could be deployed for red wine. Understandably, this resulted in increase in RMSE value to 0.9428. In our case, Classification Accuracy from the Random Forests does not clearly indicate model reusability as the Decision Trees tend to perform well on flexible models and it is not a sole measurement of the distinctly different datasets. Therefore, all factors listed above need to be considered before summarizing model reusability.

V. REFERENCES

[1] 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.
[2] C. (n.d.). Retrieved from http://www.vinhoverde.pt/en/
[3] Root-mean-square deviation. (2018, August 28). Retrieved from https://en.wikipedia.org/wiki/Root-mean-square_deviation