Report: Challenge 4

Ali Akbar Shafi, Samyuktha Sankaran and Sambandh Dhal

Abstract

This challenge consisted of three parts: 1)estimating the quality of white wine using linear regression; 2) Classify the white wine from red wine using Decision tree and 3) Predict the quality of red wine using linear regression model trained on white wine data set. Furthermore the results are interpreted in lieu of classification and regression performances.

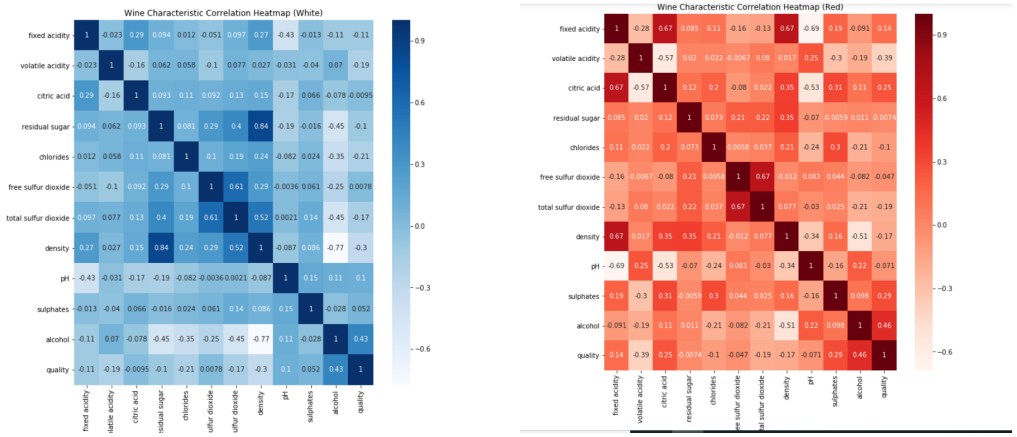
1 Data Analysis and Results

Heat maps [Fig.1], plotted between predictors for white wine and red wine, show the co-relation between different predictors and quality of wines. From the co-relation matrix it is evident that set of parameters that affect the quality for red and white wines are different.

The boxplots for different predictors indicated the presence of outliers in training dataset of both the wines and had to be removed by calculating the studentized residuals after fitting the ordinary least squares model. The linear regression model trained on the white wine training dataset was used to predict the quality of testing datasets for white as well as red wine. The RMSE obtained with predicted values for white wine and red wine were 0.71030 and 0.94286 respectively. Also, a linear regression model was developed using the training dataset for red wine and was used to predict the quality of wine for records in red wine test data. The RMSE obtained in this case was 0.67461.

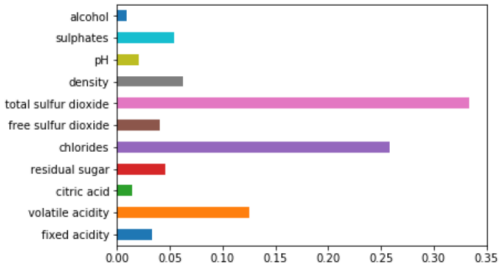
In problem 2 we developed a decision tree classifier on a combined dataset of red wine and white wine. The classifying accuracy of the model turned out to be 100% implying there are certain parameters that clearly bifurcate the type of wine [Red or White].

Figure 2(a) represents the influence of different physiochemical properties as per the Random Forest Classifier. From fig 2(a) we can infer that the total sulfur dioxide, chlorides and volatile

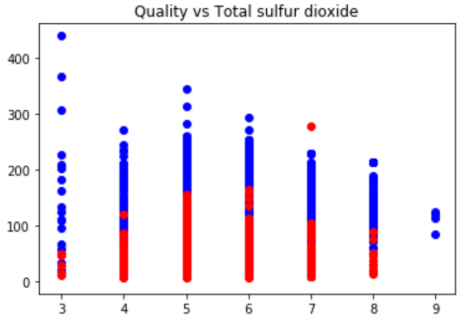


(a) White wine (b) Red wine

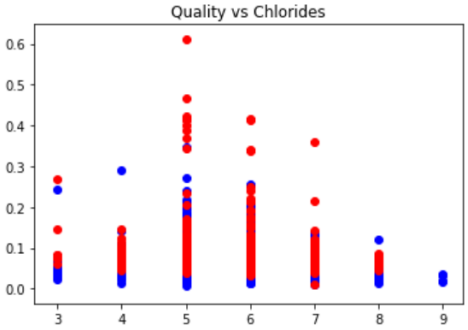
Figure 1



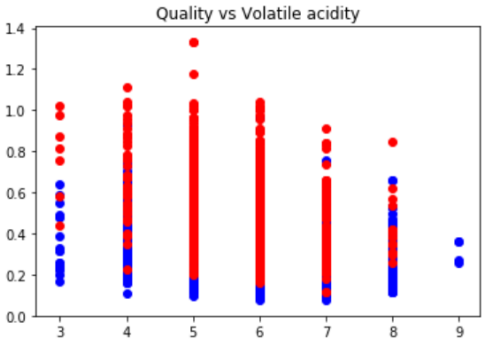
(a)



(b)



(c)



(d)

acidity are top three influencers which allow the classifier to identify between Red wines and white wines. To explore this hypothesis further we plotted the values spread of the above mentioned three predictors with respect to quality for both the wines [Red: Red wine; Blue: White wine]

2 CONCLUSION

From the scatter plots [Fig 2(b,c,d)], the RMSE values of regression models and accuracy of decision tree classifier we conclude that since the decision tree had a better measure of variance of predictors for combined white and red wine data it differentiated between the two wines with great accuracy. When we used trained regression model from white wine onto red wine we tried to predict quality values using not so significant predictors for red wine. Thus, although the transfer learning may save computation resources for training purposes, we need to ensure that the transfer model sufficiently captures the significant predictors for individual classes(in this case red wine and white wine)