Analyzing Performance of Model Reuse

Siddharth Ajit, Kanchan Satpute, Priya Balasubramanian
Texas A&M University

Abstract: By analyzing the parameters obtained using a decision tree classifier and linear regression analysis method on the Wine Quality dataset, we have gauged the performance of model reuse. We have two datasets related to white wine and red wine for our analysis. In this report we build a decision tree classifier on the combined dataset to categorize the two wines and then tried to analyze the performance of a linear regression model fitted on the white wine data to do prediction on the red wine data.

I. INTRODUCTION

There are two main datasets pertaining to white wine and red wine. There are 11 attributes which are based on physiochemical tests and an Output variable ‘quality’ which is based on sensory data. Each expert has graded wine quality between 0 (bad) and 10 (excellent). In addition to these two datasets we have also taken into consideration the combined dataset which has all the attribute information from the above data and the output variable is ‘type’ which consists of information regarding the type of wine (white or red).

II. MODEL FORMULATION & ANALYSIS

Our model formulation has two aspects. Firstly, we have taken the combined dataset and built a binary decision tree and a random forest model to classify the white wine and red wine. We have then built a linear regressor on the white wine data and used this model to predict the wine quality of red wine. The challenge now is to analyze the goodness of fit of this model for prediction on the red wine data.

From the decision tree model, we get a very good accuracy on the test data. This shows that the prediction is perfect and the model is very accurately classifying each type of wine. The feature importance graph is Fig.1 shows that the features ‘total sulfur dioxide’, ‘chlorides’, ‘volatile acidity’, 'free sulfur dioxide', 'density' are the important features in depicting the type of wine, listed according to importance. These features are important to depict the difference between both the classes.

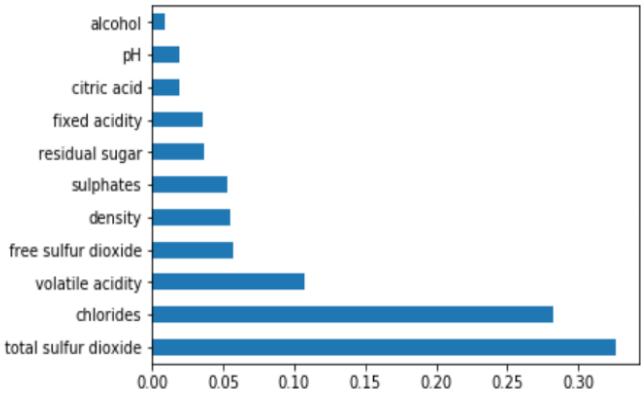

Fig.1 Feature importance

If these features also hold a greater importance in determining the value of quality of wine in linear regression, then it won't be a good idea to reuse the model. That is because the coefficients will largely vary according to the type of wine. We can study the statistics from linear regression analysis to decide which features play an important role in determining the quality of a white wine. From Fig.2, we can see that the

coefficients of 'density', 'volatile acidity', 'sulphates', 'alcohol', 'pH' are higher and their p-values are least, so these features are very significant in determining the quality of wine.

OLS Regression Results

Dep. Variable:	quality	R-squared:	0.984
Model:	OLS	Adj. R-squared:	0.984
Method:	Least Squares	F-statistic:	1.707e+04
Date:	Tue, 16 Oct 2018	Prob (F-statistic):	0.00
Time:	14:02:12	Log-Likelihood:	-3562.2
No. Observations:	3118	AIC:	7146.
Df Residuals:	3107	BIC:	7213.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
fixed acidity	-0.0530	0.019	-2.787	0.005	-0.090	-0.016
volatile acidity	-1.8332	0.143	-12.811	0.000	-2.114	-1.553
citric acid	0.0176	0.121	0.145	0.885	-0.220	0.255
residual sugar	0.0210	0.003	6.481	0.000	0.015	0.027
chlorides	-1.3131	0.694	-1.892	0.059	-2.674	0.048
free sulfur dioxide	0.0041	0.001	3.895	0.000	0.002	0.006
total sulfur dioxide	-0.0007	0.000	-1.468	0.142	-0.002	0.000
density	2.2214	0.444	5.003	0.000	1.351	3.092
pH	0.1756	0.105	1.679	0.093	-0.029	0.381
sulphates	0.3658	0.122	2.990	0.003	0.126	0.606
alcohol	0.3491	0.014	24.716	0.000	0.321	0.377

Fig.2 Linear Regression Results

By comparing the two results, we can observe that the common important features are 'volatile acidity' and 'density'.

When we apply the linear regression model on red wine data to predict the wine quality, the accuracy will be little less compared to the prediction done on white wine data. This is because of the above-mentioned features which have high impact on deciding the difference between wines. But the high impact features like 'total sulfur dioxide' and 'chlorides' (from Fig.1-refer t statistic) do not have much importance in the linear regression model which is a good sign for model reuse. And that is why we get pretty good accuracy even on the red wine data in spite of using a different model.

III. CONCLUSION

In conclusion, the model from white wine can be used to predict on the red wine as long as the important features from classification (Decision tree) and the regression (Linear regression) are different up to certain extent. The p values of the most important feature from decision tree classifier (total sulfur dioxide and chlorides) are not significant at 95% confidence Interval. So, it is reasonable to use the regression model fit on white wine to predict the quality of red wine