Exploring Wine Data using Linear Regression and Decision Trees

Anirudh Shaktawat
Neehar Yalamarti
Rishi Laddha

Abstract

The Wine Quality Data are data related to white and red wines of the Portuguese "Vinho Verde" wine. The features include only Physico-chemical variables and sensory variables because of the privacy issues of the owner. The problem statement is twofold: Performing linear regression on White wine to predict its quality, and to classify the data into White Wine and Red Wine based on the input features using decision trees and random forests. In addition, the linear regression model for the White Wine is fit on Red Wine to see how the model compares in accuracy when it is fit on a similar (since both are for wines) yet completely different data.

I. Introduction

Wine which was once viewed as an expensive good, is now being extremely enjoyed by a variety of consumers. Some facts for wine being Portugal is in the top ten wine exporting countries, with 3.17 % of the market share. To support its growth, the wine industry is investing in many new technologies for both wine production and selling. Hence Wine quality assessment is a major element for the reason of growth. Certification of the quality prevents the illegal diluting and adulterating of wines to safeguard human health and assures quality for the wine market.

II. Methodology

The first model using is Linear regression which is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables. The case of many explanatory variable is called multiple linear regression. The decision tree algorithm is an advanced tree-based algorithm to perform the classification task to predict whether it's a white or red wine. It works recursively splitting on the features until a criterion is reached to perform prediction. Random forest is another tree-based algorithm which works by combining several decision trees to perform the prediction. The advantage over regular decision trees is that the features here would be de-correlated and hence over fitting would be reduced.

III. Result and conclusion

Part 1: Linear regression model applied on White Wine Data set which predicted the quality of white wine giving an RMSE on training data set of 0.75. Linear regression by Gradient descent was also applied but it predicted on the training data set with an RMSE of 48.28. This shows that simple ordinary least squares linear regression performs perfectly well on the given data set.

Part 2: First a Decision tree algorithm was applied to predict whether wine was white or red (0 or 1) i.e on the combined data set for both the wines. The confusion matrix was plotted for the model. It is observed that there were unacceptable number of false positives and false negatives. In addition, as and when the binary decision tree was applied, it gave very different results showing that there was a very high variance. To reduce variance, there are two things that can be done, one being selecting the subset of features based on their normalized feature importance scores and the other being to move on to other methods like Random Forests. Since there were only 11 features present, feature selection was not done, and ensemble methods like the Random Forest was applied.

For Random Forests, the data set was split into a training set and validation set with 10-fold cross validation technique. The mean train and mean validation accuracy was calculated among all the validation sets. To increase the validation accuracy and to reduce the over-fitting as much as possible, Grid Search was applied for hyper parameter tuning. There was no significant improvement by applying the hyper parameter tuning in validation accuracy and reduction

of overfitting. Hence the conclusion obtained is we need more samples of data i.e enlarging the training data set. But the insight is that the Random Forest had an overfitting of 0.65% while Decision Tree has 1.73%. Hence Random Forest outperformed the Decision tree.

Part 3: The Linear regression model from Part 1 was fit on the Red Wine data set to predict quality and to observe consequences of model reuse, which is further explored in part 4. The RMSE obtained from this reused model was 0.94.

Part 4:

i) The linear regression model from the white wine data set was applied on the red wine data set, as shown in part 3, as well as a decision tree which was trained on the white wine data set was applied on the red wine data set to observe implications of model reuse. It was observed that the linear regression model had white-on-red RMSE of 0.99 while the red-on-red linear regression model had an accuracy of 0.67. For the white-on-red random forest model, the RMSE obtained was 1.19.
Comparing these results, it is concluded that the linear regression model could be a better option for model reuse. But as shown in the figure below, the coefficient values from the red-on-red linear regression model are completely different from the white-on-red linear regression model. This shows that although there is not much difference in RMSE between the two models, and the data set and its features are similar, this model should not be reused.

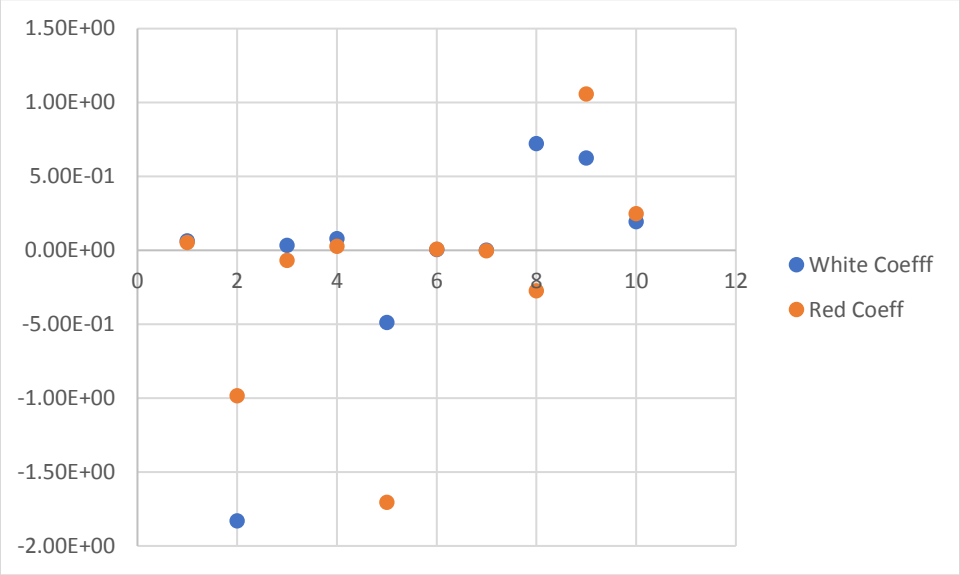

Fig: Variation of the coefficients between the linear regression models.

ii) The random forest classifier which was trained on the combined data set to classify white or red (0 or 1) was then fit on the red wine data set to observe implications of model reuse. It is observed that the training accuracy on the original data set was 99.92 % which had only 4 misclassifications i.e the model was well fit on the training data. When this model was fit on the red wine data set, it predicted with an accuracy of 99.7 %, having 3 misclassifications. When this same model was fit on the white wine data set, it predicted with an accuracy 99.97 % having only 1 misclassification. The insights that can be drawn from this is that the model which was trained on the combined data set will obviously work well on the individual data sets since they are subsets of the combined data set. Therefore, in this context, this model can be reused.

REFERENCES

[1] www.scikit-learn.org, *Random Forest Classifier Documentation*
[2] https://github.com/CourseReps, *ECEN 689, Challenge 4*
[3] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. *Modeling wine preferences by data mining from physicochemical properties*.