DECISION TREE AND ADDING REGRESSION TO TREES

Introduction:

This Project was aimed at learning and using the Machine learning technique of decision trees and regression to classify credit as defaulted or not, to predict the quality of wine using regression technique and finally with the knowledge learnt, use Machine learning technique to classify a news post as popular or unpopular and also to predict the shares.

Data Preparation:

The data was clean for Credit, Wine and did not require any Pre-processing. The Online news data however, needed a binary response variable to classify the popularity. For this, the data was split using the variable share with a median of 1400. The rows with shares less than 1400 were split as unpopular and shares greater than 1400 was split as popular.

Testing and Training Data:

Before modelling, it always a good idea to create a testing and training dataset. However, it is very important that the testing data is not from the training data so as to avoid overfitting. In this project, when working with all the data sets, 90% of the data was used for training and remaining 10% of the data was used for testing.

Model Implementation:

For, the credit data classification, decision tree was used by using the c5.0 package from R. with the help of this, data was classified as defaulted or not defaulted.

For the Wine data, regression to trees was used to predict the quality of wines by using the rpart package from R.

For the online news prediction both the decision tree as well as regression to trees was used. First, the decision tree was used to classify the post as popular or unpopular. Second, regression to trees was used to predict the shares.

Model Evaluation:

Credit

In the credit data analysis, Confusion matrix was used for error calculation.13% were misclassified, giving the model an accuracy of 87%

Wine

In the Wine data, Correlation of the model with test data was taken and it gave a 53 % correlation.

Online news

For the online news popularity prediction, the classification error was 34% giving the decision tree model an accuracy of 66% whereas the regression model to predict the shares had a correlation of 61%.

Conclusion:

Thus, for the credit default Classification, decision tree seemed to be a be performing pretty well with high accuracy compared to the online news popularity prediction. Hence, further with algorithm tuning or learning using different models, we can analyze which model outruns the performance of decision trees.