Receiver Operating Characteristic Curve for Decision Tree and Random Forest Classification Model

Overview: ***Student Id:*** *17230755*

This task involves plotting the Receiver Operating Curve (ROC curve) and finding Area Under ROC Curve for the held-out data set (Test data set) for the previously designed Machine Learning Models. In this case the **Decision Tree** and **Random Forest** models are chosen for the prediction of **test result** column in illness data. Illness data set is partitioned into 70% training data and 30% test data.

Tools and Packages Used: R**,** RStudio**,** ROCR Package

Several R packages can be used for plotting ROC curve. Out of that I have used ROCR package. ROCR is a flexible tool for creating cut off-parameterized 2D performance curves by freely combining two from over 25 performance measures. The parameterization can be visualized by printing cut off values at the corresponding curve positions, or by **colouring the curve according to cut-offs**. Despite its flexibility, ROCR is easy to use, with only three commands and reasonable default values for all optional parameters. Below are the most steps followed to plot ROC curve. ***[1]***

- Prediction-Class and Prediction: For class prediction and function to create prediction objects. prediction (predictions, labels): This function gives: TP, FP, FN, TN, for different cut off (alpha) values.
- Performance: Function to create performance objects (measures: “tpr”, “fpr” (for ROC) and “auc” for Area Under the curve.).
- plot(performance(prediction, ‘fpr’,’tpr’): Plots the ROC curve.
- To use prediction() function, predicted values should be in probability, hence we convert predicted class to probability using predict(CV Model, type = ‘prob’). ***[1]***

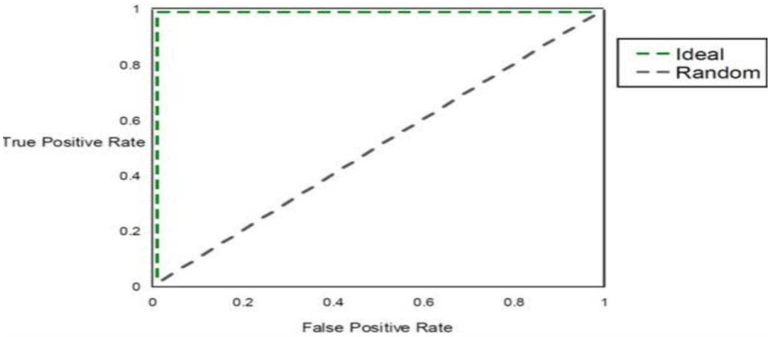
ROC Curve and AUROC Overview:

ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier (containing two classes) system **as its discrimination threshold is varied (Cut-offs)**. ROC curve is plotted as **True Positive Rate (tpr)** in y axis and **False Positive Rate (fpr)** also called as 1-Specificity in X axis. ***[2]***

Sensitivity = TPR = TP/P = TP/(TP+FN)
Specificity = TNR = TN/N = TN/(TN+FP)
1-Specificity = FPR

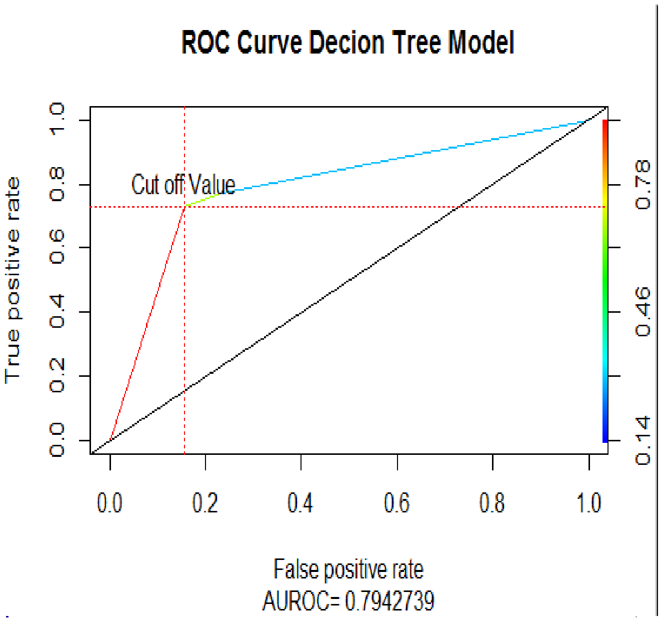
TPR = True Positive Rate; TP = No of True Positives.
TNR = True Negative Rate; TN = No of True Negatives.
FP = No of False Positives; FN = No of False Negatives. ***[2]***

ROC Curve: For True Positive if the predicted value is positive the curve moves one step upward along y axis. For True Negative if the predicted value is positive (False Positive) the curve moves one step horizontally along x axis. The accuracy of the classifier depends on how well the model able to distinguish the classes. An area of 1 represents a perfect classification. The Ideal ROC curve should be as below (Ideal and Random Model). ***[3]***



Model Evaluation based on AUROC as below:
0.90-1 = Excellent Model
0.80-0.90 = Good Model
0.70-0.80 = Fair Model
0.60-0.70 = Poor Model
0.50-0.60 = Fail ***[3][4]***

ROC Curve for Decision Tree and Random Forest Model:

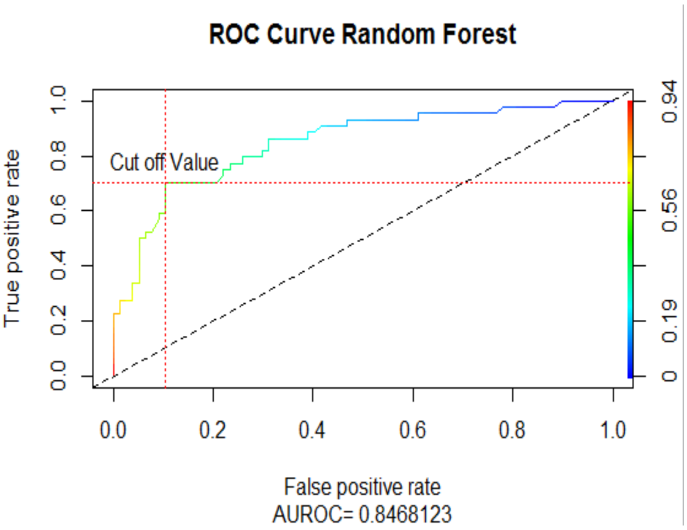


Three Cases of cut off or threshold colourised value is demonstrated in colourised graph.

Observation 1: Blue Region cut off 0.14. This tells in normally distributed data if we move our cut off value near to 0 in x axis, FPR increases i.e. for True Negative classification, our model predicts positive.

Observation 2: Green region cut off 0.46. FPR is less and sensitivity is better. i.e. our model can distinguish positive class more accurately. However, if we want our model to distinguish both the Sensitivity and Specificity. However, small FP and FN in other words small misclassification will be present.

Observation 3: Red Region cut off 0.78, model may predict positive class accurately. However, because of cut off moved towards 0.78 in normal distribution, our model may predict negative for True Positive. Depending on the area of application optimum threshold need to be selected.

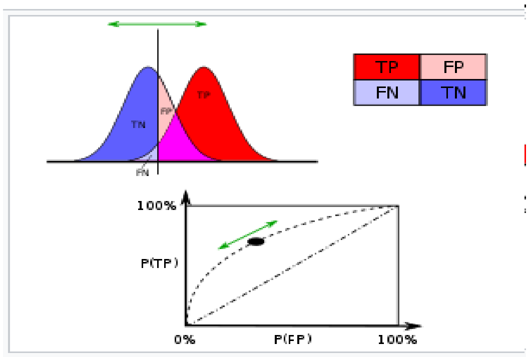


Similar logic applies to ROC of random forest as well. We can move the threshold depending upon the area of application and type of prediction that matters the most. Greater the area under the curve, better the prediction of our model.

Note: More number of steps we can observe in ROC of random forest. This means that we have more cut off value to select from.

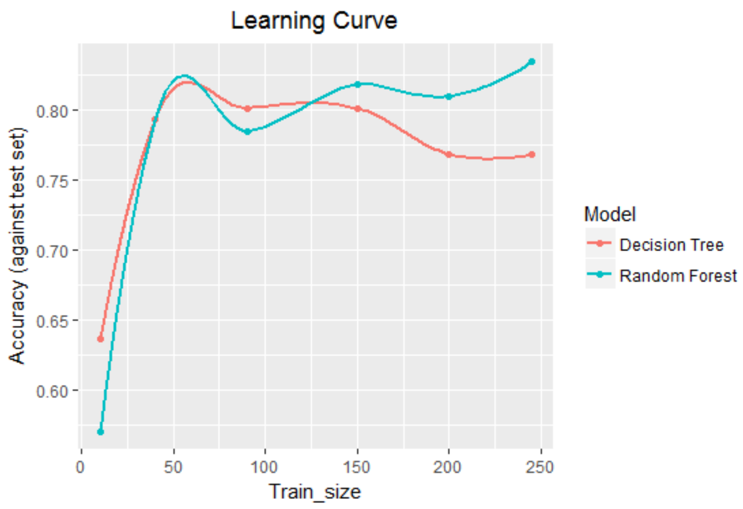
Conclusion: Based on the ROC curve from both the models, AUC obtained is better in random forest than decision tree. This shows random forest able to distinguish the classes better than Decision Tree. Moreover, cut off values are selected depending on area of application. If sensitivity and Specificity is of equal importance, we need to find the best cut off i.e. max of (Sensitivity + Specificity). In addition, from ROC of both the models we can say sensitivity and specificity in case of Random forest is better than Decision Tree. Values of optimum cut off and AUROC is mentioned in below table. Cut off value is also mentioned in ROC Curve.

Threshold or Cut off movement *[5]:*



Models	AUROC	Optimum Cut off	Sensitivity (tpr)	Specificity (1-fpr)
Decision Tree	0.7942	0.671	0.727	0.844
Random Forest	0.8468	0.504	0.704	0.896

Extras: Learning Curve



Learning Curve is plotted Accuracy on test set on y axis and training data size on x axis. By observing graph, we can say that Accuracy increases as we increase the training size i.e. Models Learning ability increases. However, this learning curve will be saturated after some point where increase in training data doesn't contribute in learning. In this case learning curve might increase after 250. This indicates if we have more training data we can still increase the performance of the model.

Learning curve is mainly used to decide how much training data is sufficient to train the model to get better performance of the model. Moreover, small variation i.e. dip in learning curve may due to **noise** in the training set. ***[6]***

Note: Complete R Code Available on Request.

References/Citations:

[1] *R Documentation:* https://cran.r-project.org/web/packages/ROCR/ROCR.pdf

[2] *Wikipedia: https://en.wikipedia.org/wiki/Receiver_operating_characteristic*

[3] *Area Under ROC Curve: http://gim.unmc.edu/dxtests/roc3.htm*

[4] *Ideal ROC Curve:*https://www.researchgate.net/figure/263014244_fig4_AUC-The-figure-shows-the-ROC-curve-and-corresponding-area-under-the-ROC-curve-AUC-The

[5] *Wikipedia: https://en.wikipedia.org/wiki/Receiver_operating_characteristic*

[6] *Learning Curve:* https://en.wikipedia.org/wiki/Learning_curve

[7] *Other reference: https://www.medcalc.org/manual/roc-curves.php*

[8] *ROC and AUC Data School: http://www.dataschool.io/roc-curves-and-auc-explained/*