Roy Chan
chanroyc@sfu.ca
301202770

**CMPT 459 Fall 2017**

**DataMining**

**Martin Ester**

**TA: Zhilin Zhang**

**Programming Assignment 3**

1. Using the package ‘tree’, a decision tree as learned from the training data. The number of terminal nodes in the tree is 118. The following figure is a plot of the decision tree.

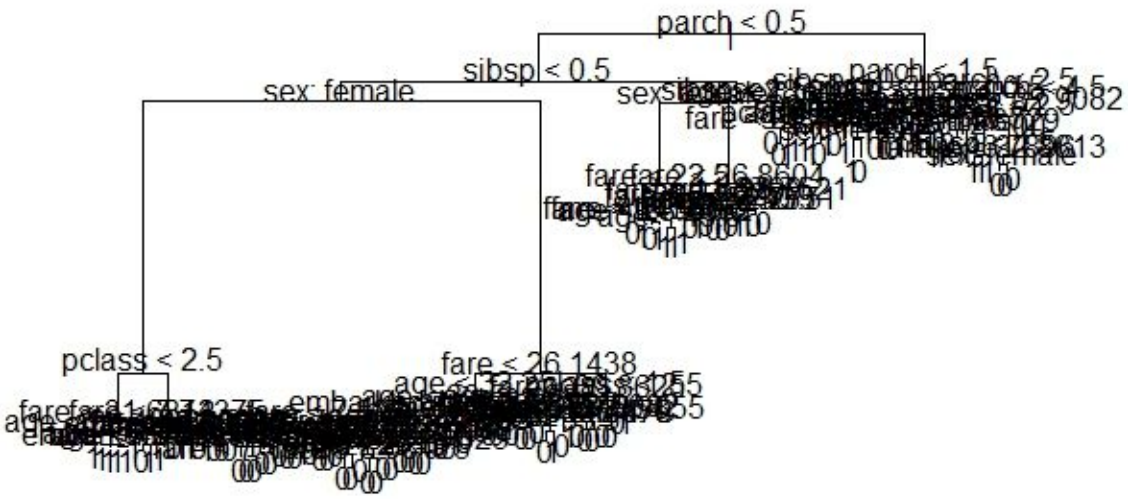


*Figure 1. Decision Tree*

2. Using gini as a split, the top 5 most important attributes in decreasing order are the following: parch, sibsp, sex, pclass and fare. The gini index chooses the attribute with the most information gain as the root, and splits based on the next attribute with the most information gain. By using the decision tree, we can take a look at which combination of values for each attribute to get a general idea of determining the survival of passengers.
3. Using the function cv.tree with prune.misclass as one of the parameters to employ cost complexity pruning, we can determine the optimal level of tree complexity. The size of the best decision tree determined by this method is 13, as shown in the figure below.

Roy Chan
chanroyc@sfu.ca
301202770

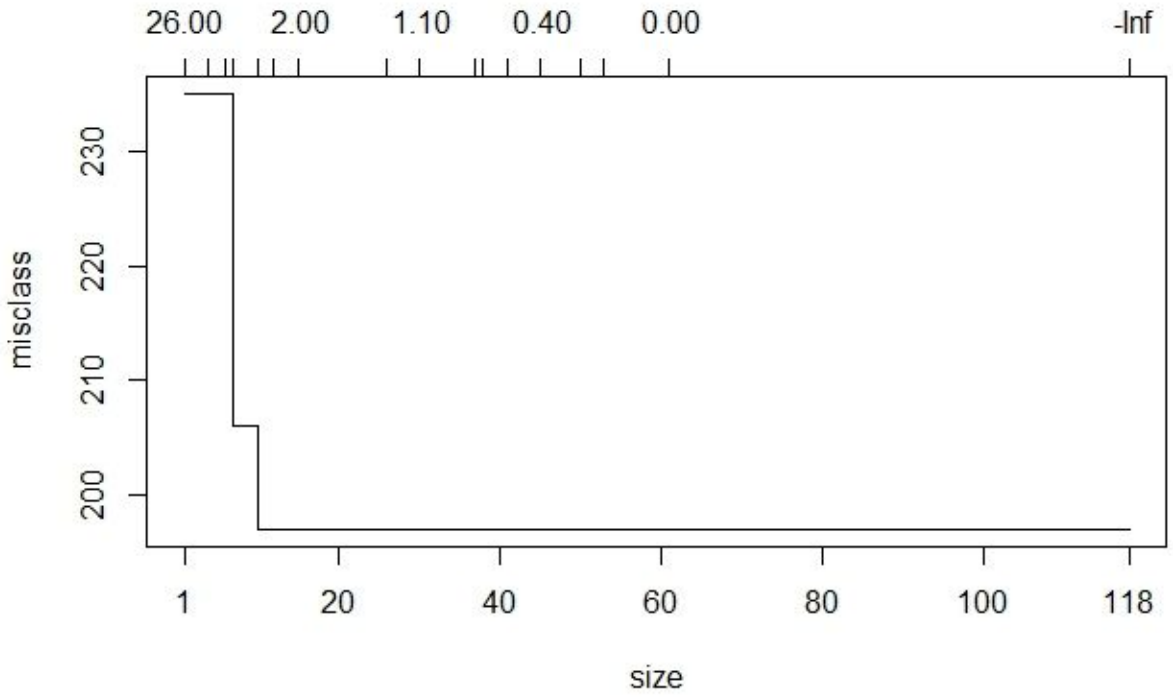


*Figure 2. Number of Misclassifications vs the Size of the Decision Tree*

4. Using the function prune.tree, the following is the best decision tree with the best size of 13 which was computed from the previous task.

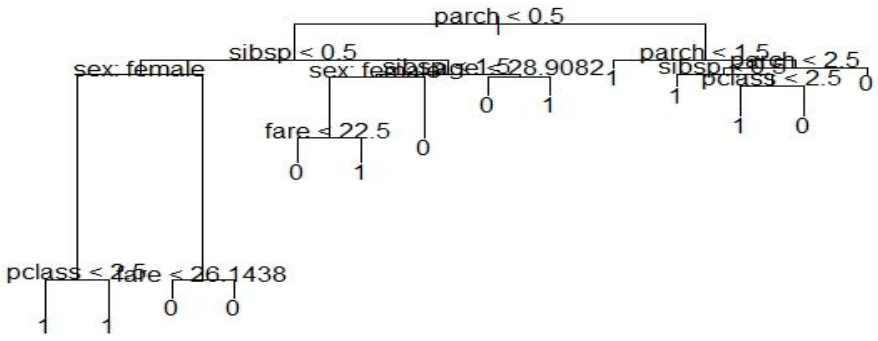


*Figure 3. Pruned Decision Tree with the Best Size determined by previous task*

Roy Chan
chanroyc@sfu.ca
301202770

The decision tree is then applied to predict the class labels of the test data. The following is the confusion matrix using the pruned tree model.

| | 0 | 1 |
|---|---|---|
| 0 | 133 | 30 |
| 1 | 27 | 72 |

*Figure 4. Confusion Matrix of Pruned Decision Tree*

The accuracy of the model is 0.782442748091603. The following is the ROC curve of the best decision tree model.

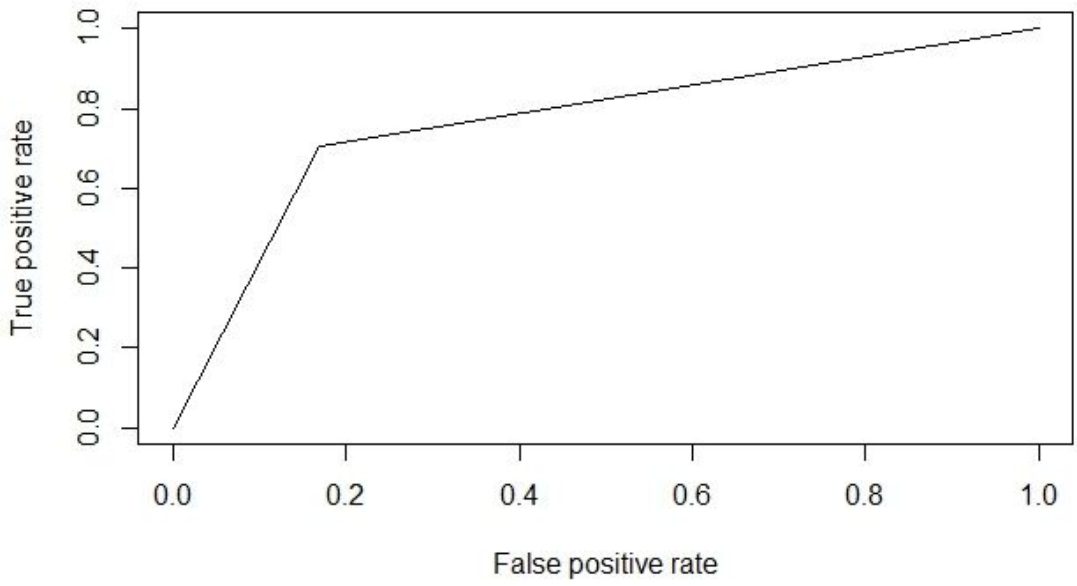


*Figure 5. ROC Curve of the Best Decision Tree Model*

The AUC is 0.7685662.

Roy Chan
chanroyc@sfu.ca
301202770

5. Using the package ‘randomForest, a random forest was learned from the training data with the number of trees set to 100. The random forest model was applied to the train set to predict the class labels of the test data. The following is the confusion matrix using the random forest model.

| | 0 | 1 |
|---|---|---|
| 0 | 149 | 37 |
| 1 | 11 | 65 |

*Figure 6. Confusion Matrix of Random Forest Model*

The accuracy of the model is 0.81679389312977. The following is the ROC curve of the random forest model.

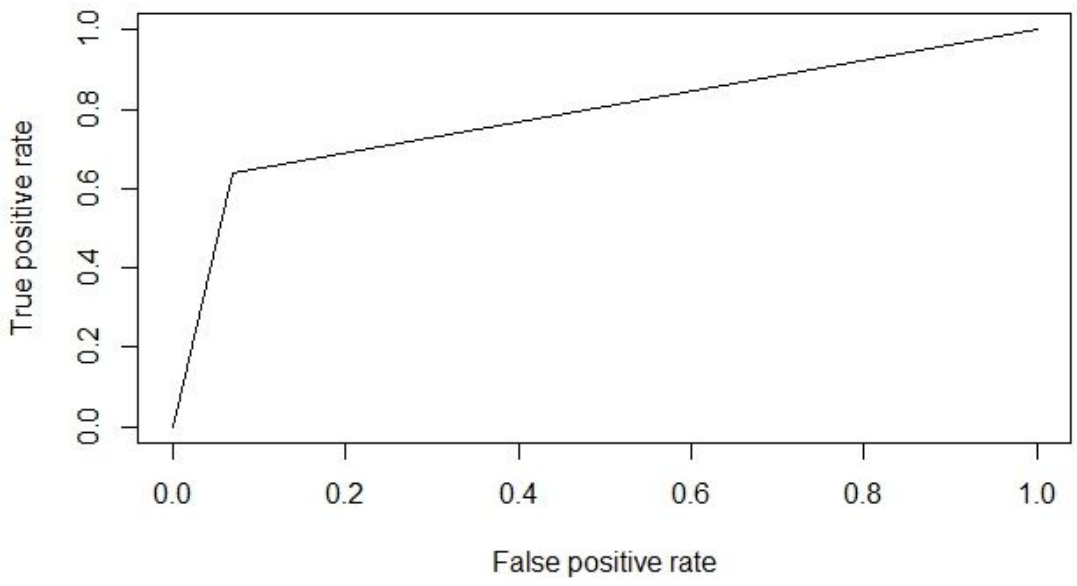


*Figure 7. ROC Curve of Random Forest Model.*

The AUC is 0.7842525.

6. After experimenting with a different number of trees, the higher amount of trees resulted in a higher accuracy and AUC. With that being said, the model does not predict the class labels the same for every run with a static number of trees. For example, with ntrees=1000, one run resulted in an accuracy of 0.8168 and AUC of 0.7824, however,

Roy Chan
chanroyc@sfu.ca
301202770

another run produced an accuracy of 0.8244 and AUC of 0.7922794. Although there isn't a 'best' number of trees, we can assume that the larger number of trees will result in a higher accuracy and AUC. The performance of the best random forest model is an improvement over the best decision tree model of approximately 4%.

7. The function importance() produces the following table.

| | 0 | 1 | MeanDecrease Accuracy | MeanDecrease Gini |
|---|---|---|---|---|
| pclass | 42.196740 | 53.370485 | 62.59876 | 43.60657 |
| sex | 90.493839 | 132.744506 | 129.38598 | 120.28830 |
| age | 23.982154 | 29.039218 | 38.94948 | 55.67666 |
| sibsp | 30.715081 | -3.260604 | 25.47518 | 17.82425 |
| parch | 17.496209 | 14.260271 | 24.39319 | 16.41452 |
| embarked | 6.619156 | 24.486573 | 24.99922 | 16.71012 |

*Figure 8. Importance() Table*

The function varImpPlot() produces the following plot.

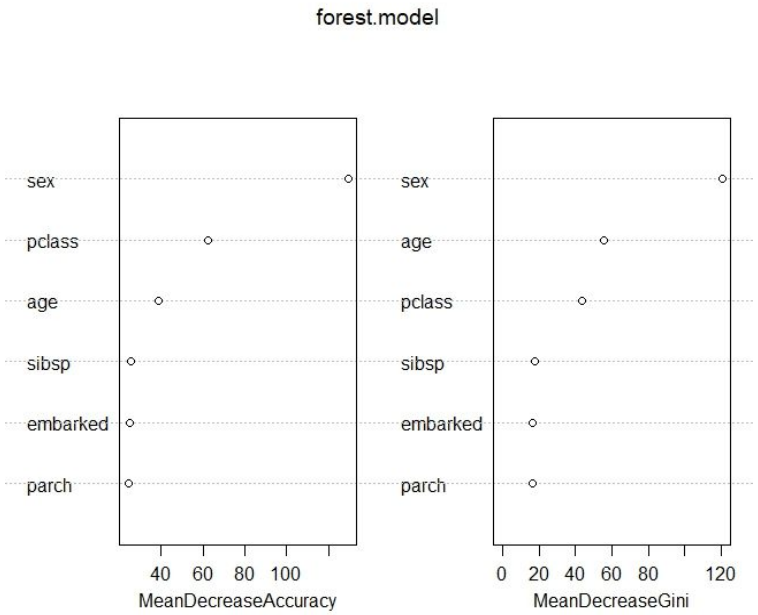


*Figure 9. varImpPlot() plot.*

Roy Chan
chanroyc@sfu.ca
301202770

There are two types of importance measures shown in figure 9, MeanDecreaseAccuracy tests to see how worse the model performs without each variable, so a high decrease in accuracy would be expected for attributes for more predictive variables. The gini one is similar, where it examines the purity of the nodes at the end of the tree. A higher score means the attribute is more important, therefore, the top most important attributes in decreasing order of importance is sex, pclass, age, sibsp, embarked, and parch.