May 2020

Classification Tree

A classification tree is a structural mapping of binary decisions that lead to a conclusion about the class of an object. It is composed of branches that represent attributes, while the leaves represent decisions. It's a popular method in data mining because it's easy to understand.

As we mentioned before, the classification tree is used for categorical response, so we use the dataset without the age and charges column and use Amount as our response, which includes above average and below average. Then we divide 80% of our dataset as training and the rest as testing. Our testing data contains 268 observations and 7 variables. We use k-fold cross-validation to determine the optimal subtree and apply cv.tree function to obtain the size. Based on **Figure 1.1**, size 4 and size 3 has the same lowest cross-validation error rate, which is 100 compared to 437. We decided to use 3 as our optimal size for our classification tree. **Figure 1.2** shows us the tree with three terminal nodes. From the tree, we can see smoke is a primary factor that determines whether the amount is below or above the average. Assuming other elements are the same, people who smoke will incur a higher amount of insurance costs. Another significant factor will be the age group. People who don't smoke but are in the senior group are also more likely to have the amounts that are above the average. We define the senior group if the age is over 47. After we form the tree model, we use a confusion matrix to evaluate it by comparing it to the testing data. From the result in **Figure 1.3**, this model gives us an accuracy of 92%, which indicates a good result for this method.

Random Forest

Random forest is another method used to determine which attributes were deemed important in predicting the accuracy of class prediction. Random forest is a powerful method that reduces variance by randomly selecting the decision trees and then averaging or picking the class that gets the most votes. With this dataset, we are using classification trees in which the "leaves" represent class labels and the "branches" represent the features that lead to the class labels.

Random forest utilizes bagging which repeatedly selects a random sample with replacement of a training set to receive a prediction of the class. Random forest uses a random subset of the variables to increase prediction accuracy and prevent bias for strong predictors. For this dataset, we used an mtry parameter of 2 as it is the square root rounded down of the total number of variables considered. Each node selects from one of the two variables it has randomly chosen from the original 6 attributes based on which optimizes the split. The original 6 attributes used were sex, bmi, children, smoker, region, and (age) group. Once the model was executed, the model received an accuracy of 0.9216418 and an error of 0.07835821. This model also produced the same results for the confusion matrix as the classification tree **Figure 2.1**. After calculating accuracy, we wanted to look at the dataset's important variables according to our random forest model. Using the importance() function in R, we found the highest mean decrease accuracies and mean decrease Gini's were ones from the

attributes “smoker” and “Group” - the third-highest was from the variable “bmi”. The importance() function calculates the mean decrease in out-of-bag prediction error (MeanDecreaseAccuracy) and the mean decrease in the node impurities derived from splitting on the variable, in this case, from the Gini index (MeanDecreaseGini). For visualization purposes, we used the varImpPlot() function to plot these mean decreases **Figure 2.2**. The results of the importance() function tell us the top two significant predictors of the dataset are found in the columns “smoker” and “Group”. It is safe to imply given the results of the classification tree and other data analysis that Yes in “smoker” and Senior Adult in “Group" are the subsets of the dataset that are significant, according to the random forest model.

```
> cv.model
$size
[1] 4 3 2 1

$dev
[1] 100 100 437 437

$k
[1] -Inf    0  209  218

$method
[1] "misclass"

attr(,"class")
[1] "prune"         "tree.sequence"
```

Figure 1.1 Cv. model

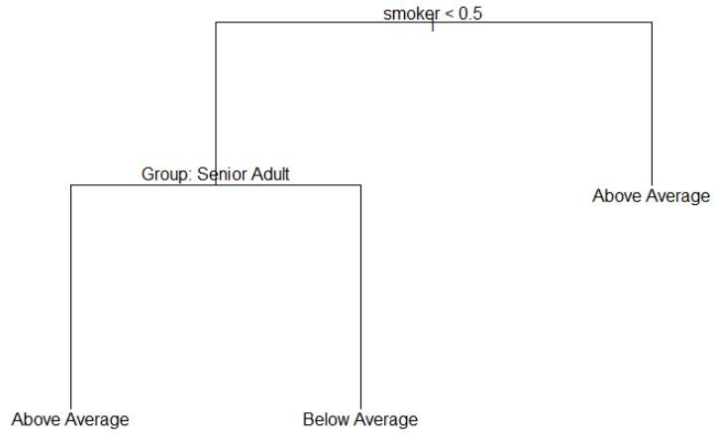

Figure 1.2 Pruned tree

	Amount test	
Prunetree.pred	Above Average	Below Average
Above Average	129	8
Below Average	13	118

Figure 1.3 Confusion Matrix for Classification Tree

	Amount test	
Yhat.rf	Above Average	Below Average
Above Average	129	8
Below Average	13	118

Figure 2.1 Confusion Matrix for Random Forest

importance(rf.in)	MeanDecreaseAccuracy	MeanDecreaseGini
sex	1.874534	5.086366
bmi	6.741533	53.106462
children	1.736063	16.294336
smoker	145.360231	156.065903
region	4.084009	12.821123
Group	121.600263	179.718045

Figure 2.2 Importance() plot for Random Forest

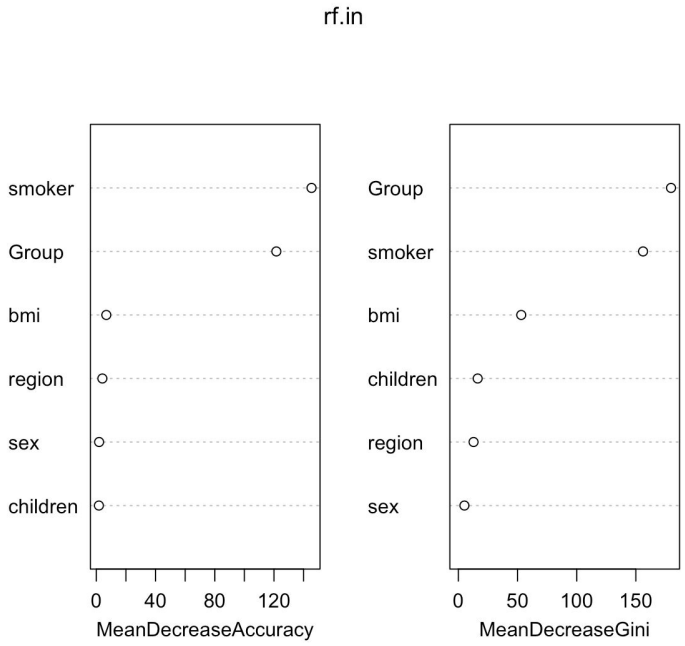

Figure 2.3 VarImpPlot() for Random Forest

References

"RandomForest." *Function | R Documentation*, www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/importance.

"Random Forests Leo Breiman and Adele Cutler." *Random Forests - Classification Description*, www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.