Assignment 11: Predictive Modeling - Part 2

1. Install the package mlbench and use the follows to import the data

library(mlbench)  
data(PimaIndiansDiabetes)  
df <- PimaIndiansDiabetes  
  
head(df)

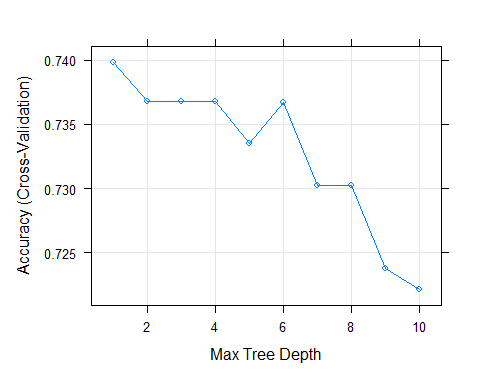
## pregnant glucose pressure triceps insulin mass pedigree age diabetes  
## 1 6 148 72 35 0 33.6 0.627 50 pos  
## 2 1 85 66 29 0 26.6 0.351 31 neg  
## 3 8 183 64 0 0 23.3 0.672 32 pos  
## 4 1 89 66 23 94 28.1 0.167 21 neg  
## 5 0 137 40 35 168 43.1 2.288 33 pos  
## 6 5 116 74 0 0 25.6 0.201 30 neg

* Set seed to be 2020.
* The target variable is diabetes
* Partition the data into 80% training and 20% testing.

names(df)[9] <- 'target'  
  
library(caret)  
set.seed(2020)  
splitIndex <- createDataPartition(df$target, p = .80,   
 list = FALSE)  
df\_train <- df[ splitIndex,]  
df\_test <- df[-splitIndex,]

1. Tuning Decision Tree: Use cross-validation with 10 k-folds to find the maxdepth with the greatest accuracy. Plot the accuracy associated with different maxdepths against the maxdepths. The range to search for maxdepth is from 1 to 10.

# Decide the range of the maxdepth to search for the best  
tuneGrid = expand.grid(maxdepth = 1:10)  
# Tell caret to do Approach 2, i.e. Cross-Validation  
trControl = trainControl(method = "cv",  
 number = 10)  
# Do Approach 2   
tree\_approach2 <- train(target~., data=df\_train,   
 method = "rpart2",   
 trControl = trControl,  
 tuneGrid = tuneGrid)  
  
# Plot the results  
plot(tree\_approach2)



1. Make the final decision to select the maxdepth for your decision tree. Is your selected maxdepth the same as the maxdepth found in 2?
2. Calculate the accuracy of your decision tree (the decision tree with your selected maxdepth in 3) on the test data.

pred <- predict(tree\_approach2, df\_test)  
cm <- confusionMatrix(data = pred, reference = df\_test$target, positive = "pos")  
cm$overall[1]

## Accuracy   
## 0.7254902