Final project

Mengyu Li March 3, 2019

Pre-process data

##

ср

```
dfModelData <- dfRawData[-1] # Exclude the first column
dfModelData <- na.omit(dfModelData) # remove any NAs
colnames(dfModelData) <- c('MonthsSinceLastDonation', 'NumberOfDonations', 'TotalVolumeDonatedCC', 'Mon
dfModelData$MadeDonationInMarch2007 <- as.factor(dfModelData$MadeDonationInMarch2007)</pre>
```

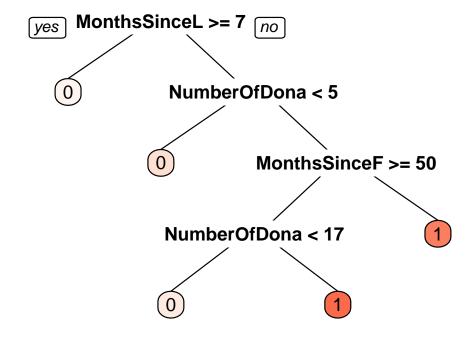
Split data into train and test data

Accuracy

Kappa

```
# Split the data into training and test set
set.seed(123)
# 80% of data are selected as train data
trnSamples <- dfModelData$MadeDonationInMarch2007 %>%
  createDataPartition(p = 0.8, list = FALSE)
trnData <- dfModelData[trnSamples, ]</pre>
testData <- dfModelData[-trnSamples, ]</pre>
control <- trainControl(method = 'repeatedcv', number = 10, repeats = 5)</pre>
# Train the model
model_DT <- train(MadeDonationInMarch2007 ~., data = trnData, method = 'rpart', parms = list(split = "</pre>
# Estimate variable importantce
model_DT
## CART
##
## 462 samples
     4 predictor
##
     2 classes: '0', '1'
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 462, 462, 462, 462, 462, 462, ...
## Resampling results across tuning parameters:
##
```

```
0.00000000 0.7655883 0.3270307
##
##
    0.008675342 0.7858749 0.3631221
##
    0.017350684 0.7921580 0.3757762
    0.026026026 0.7936887
                          0.3615403
##
##
    0.034701368
                0.7943344
                          0.3586897
    0.043376710 0.7966595 0.3702324
##
##
    0.052052052 0.7913900 0.3285536
                0.7829175 0.2672505
##
    0.060727394
##
    0.069402736
                0.7721643
                          0.2019766
##
    ## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.04337671.
prp(model_DT$finalModel, box.palette = "Reds", tweak = 1.2)
```



```
test_pred <- predict(model_DT, newdata = testData)
confusionMatrix(test_pred, testData$MadeDonationInMarch2007 ) #check accuracy

## Confusion Matrix and Statistics
##
## Reference
## Prediction 0 1
## 0 73 18
## 1 14 9
##
## Accuracy : 0.7193</pre>
```

```
95% CI : (0.6274, 0.7994)
##
##
      No Information Rate: 0.7632
      P-Value [Acc > NIR] : 0.8857
##
##
                     Kappa : 0.1817
##
   Mcnemar's Test P-Value : 0.5959
##
##
##
               Sensitivity: 0.8391
               Specificity : 0.3333
##
            Pos Pred Value : 0.8022
##
            Neg Pred Value: 0.3913
##
                Prevalence: 0.7632
##
            Detection Rate: 0.6404
##
##
     Detection Prevalence: 0.7982
##
         Balanced Accuracy : 0.5862
##
##
          'Positive' Class : 0
```

##