Group Members:

- **Aditi Mankar (UIN: 667879658)**
- **Prathamesh Bodake (UIN: 663611919)**
- **Kush Thakkar (UIN: 664454403)**

Case Study: Retention Modeling at Scholastic Travel Company (A)

We begin with first understanding the data, our final goal is to determine if the group is retained or not, this can be decided by using the variable 'Retained.in.2012' which is our target variable.

```
my_data <- read_excel("CS2.xlsx")
```

```
## Warning in read_fun(path = enc2native(normalizePath(path)), sheet_i =
sheet, :
## Expecting numeric in A2393 / R2393C1: got 'Data source: Company data
adjusted by
## author using unspecified constants.'
```

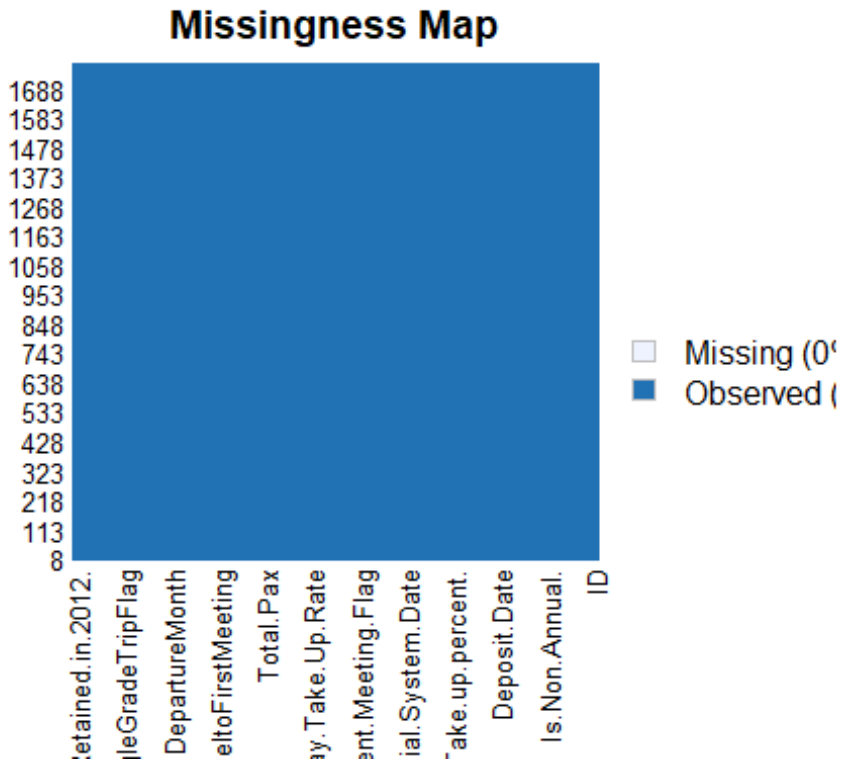
```
my_data
```

```
## # A tibble: 2,392 x 56
##       ID Program.Code From.Grade To.Grade Group.State Is.Non.Annual.  Days
##    <dbl> <chr>        <chr>      <chr>    <chr>                 <dbl> <dbl>
##  1     1 HS           4          4        CA                        0     1
##  2     2 HC           8          8        AZ                        0     7
##  3     3 HD           8          8        FL                        0     3
##  4     4 HN           9          12       VA                        1     3
##  5     5 HD           6          8        FL                        0     6
##  6     6 HC           10         12       LA                        0     4
##  7     7 SG           11         12       MA                        1     6
##  8     8 FN           9          9        MX                        0     8
##  9     9 CC           8          8        AZ                        0     8
## 10    10 HD           8          8        TX                        0     4
## # ... with 2,382 more rows, and 49 more variables: Travel.Type <chr>,
## #   Departure.Date <dttm>, Return.Date <dttm>, Deposit.Date <dttm>,
## #   Special.Pay <chr>, Tuition <dbl>, FRP.Active <dbl>, FRP.Cancelled
<dbl>,
## #   FRP.Take.up.percent. <dbl>, Early.RPL <chr>, Latest.RPL <chr>,
## #   Cancelled.Pax <dbl>, Total.Discount.Pax <dbl>, Initial.System.Date
<chr>,
## #   Poverty.Code <chr>, Region <chr>, CRM.Segment <chr>, School.Type
<chr>,
## #   Parent.Meeting.Flag <dbl>, MDR.Low.Grade <chr>, MDR.High.Grade <chr>,
...
```

```
attach(my_data)
my_data$Retained.in.2012.<- as.factor(Retained.in.2012.)
```

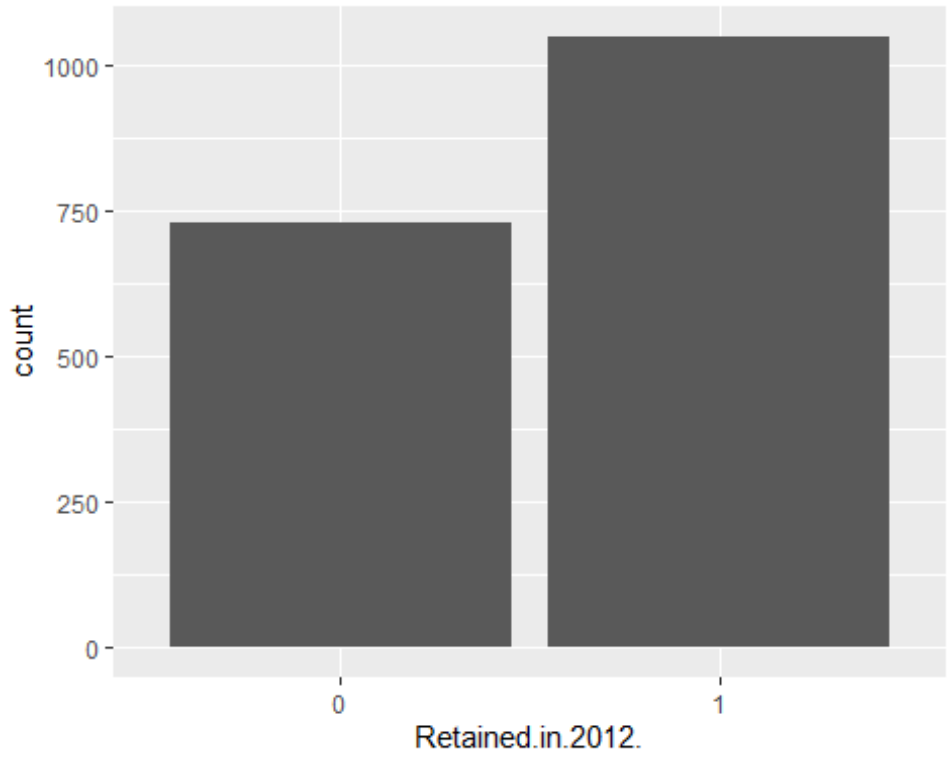
When we look at the summary it is apparent that there are many 'NA' values in our data-set, so we will count how many null values we have in total and then remove all the null values from our dataset

```
sum(is.na(my_data))

## [1] 1081

my_data <- na.omit(my_data)
sum(is.na(my_data))

## [1] 0

missmap(my_data)
```



Now we will look at our target variable to asses our final data without any null values

```
summary(my_data$Retained.in.2012.)

##    0    1
##  729 1049

ggplot(data=my_data) + geom_bar(mapping=aes(x=Retained.in.2012.))
```



From the summary and plot it can be concluded that we have more number of retained groups at 1049 vs not retained groups at 729

We will split our data into two groups 'Categorical' and 'Numerical' as we will perform chi-square test on our categorical data and Target variable. We will discard all the columns which have low p-value (>0.05) as the relationship is not significant.

```
CategoricalVar <- c("Program.Code", "From.Grade", "To.Grade",
                    "Group.State", "Travel.Type","Special.Pay",
                    "Early.RPL", "Latest.RPL", "Initial.System.Date",
                    "Poverty.Code", "Region",
                    "CRM.Segment", "School.Type", "MDR.Low.Grade",
                    "MDR.High.Grade", "Income.Level",
                    "SPR.Product.Type", "SPR.New.Existing", "FirstMeeting",
                    "LastMeeting",
                    "DifferenceTraveltoFirstMeeting",
                    "DifferenceTraveltoLastMeeting", "SchoolGradeTypeLow",
                    "SchoolGradeTypeHigh", "SchoolGradeType",
                    "DepartureMonth", "GroupGradeTypeLow",
                    "GroupGradeTypeHigh",
                    "GroupGradeType",
                    "MajorProgramCode", "FPP.to.School.enrollment",
                    "SchoolSizeIndicator")

NumericVar <- c("ID", "Is.Non.Annual.", "Days","Tuition","FRP.Active",
                "FRP.Cancelled",
                "FRP.Take.up.percent.",
```

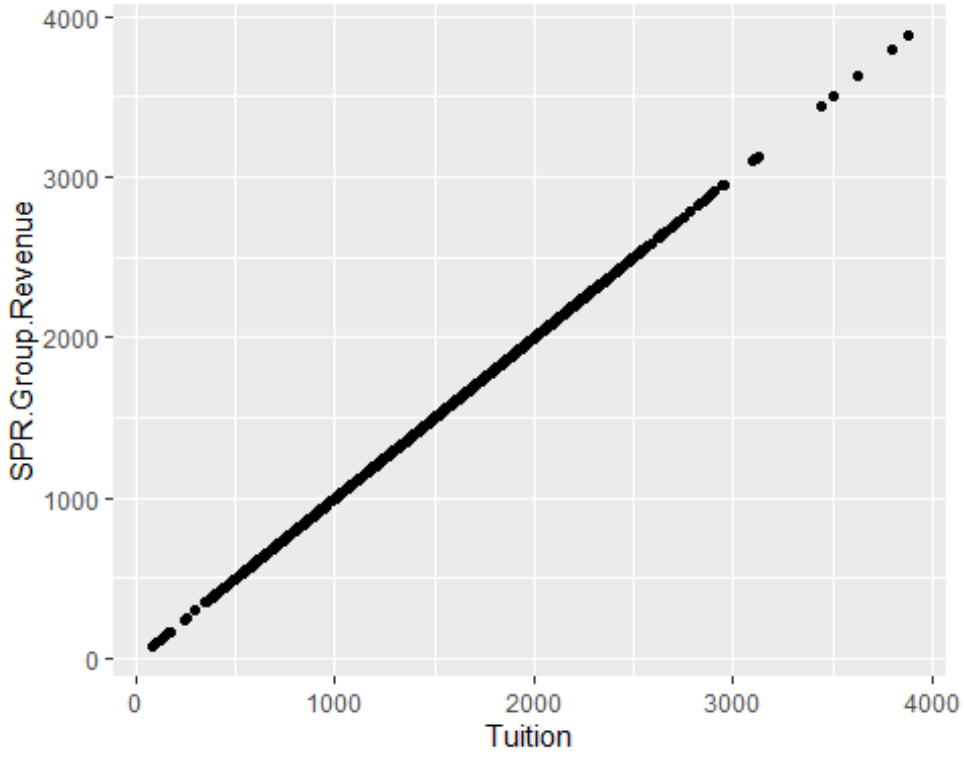
```
                "Cancelled.Pax","Total.Discount.Pax",
                "Parent.Meeting.Flag",
                "Total.School.Enrollment","EZ.Pay.Take.Up.Rate",
                "School.Sponsor","FPP","Total.Pax","SPR.Group.Revenue",
                "NumberOfMeetingswithParents","SingleGradeTripFlag",
                "FPP.to.PAX","Num.of.Non_FPP.PAX","Retained.in.2012.")


myTests <- lapply(my_data[CategoricalVar], function(x)
chisq.test(my_data$Retained.in.2012., x))
do.call(rbind, myTests)[,c(1,3)]

##                                 statistic p.value
## Program.Code                    94.67721  1.151105e-10
## From.Grade                      322.164   3.175513e-63
## To.Grade                        118.3849  2.859717e-21
## Group.State                     113.9997  1.712967e-07
## Travel.Type                     13.5258   0.001155874
## Special.Pay                     8.675492  0.03393151
## Early.RPL                       162.3093  0.01619769
## Latest.RPL                      226.4135  0.02593968
## Initial.System.Date             347.434   0.0007650151
## Poverty.Code                    35.27709  1.324709e-06
## Region                          22.18739  0.0004823436
## CRM.Segment                     118.9961  8.087488e-21
## School.Type                     1.512175  0.4694997
## MDR.Low.Grade                   111.5013  9.181401e-19
## MDR.High.Grade                  128.2195  1.077657e-22
## Income.Level                    57.60673  1.659912e-05
## SPR.Product.Type                46.26352  8.02653e-09
## SPR.New.Existing                201.9295  7.921092e-46
## FirstMeeting                    243.5681  0.002507123
## LastMeeting                     198.805   0.002862807
## DifferenceTraveltoFirstMeeting 348.5603  0.1765277
## DifferenceTraveltoLastMeeting  251.5902  0.1460238
## SchoolGradeTypeLow              71.54132  1.996027e-15
## SchoolGradeTypeHigh             100.2567  1.368648e-21
## SchoolGradeType                 118.2363  7.665548e-22
## DepartureMonth                  49.1222   2.095216e-09
## GroupGradeTypeLow               100.8872  6.36715e-21
## GroupGradeTypeHigh              92.23266  9.374173e-21
## GroupGradeType                  148.3973  3.166953e-26
## MajorProgramCode                41.85616  4.304241e-09
## FPP.to.School.enrollment        1513.543  0.7473017
## SchoolSizeIndicator             75.58358  2.71659e-16
```

Removing insignificant variables such as all dates, School.Type, DifferenceTraveltoFirstMeeting, DifferenceTraveltoLastMeeting and FPP.to.School.enrollment etc from our datset cause they have no strong realtionship with our Target variable

```
my_data <- subset(my_data, select = -c(ID,
School.Type,FPP.to.School.enrollment,DifferenceTraveltoFirstMeeting,Differenc
eTraveltoLastMeeting,
                                       Departure.Date, Return.Date,
Deposit.Date, Initial.System.Date, Parent.Meeting.Flag,
NumberOfMeetingswithParents,
                                       SchoolSizeIndicator, FirstMeeting,
LastMeeting, DifferenceTraveltoFirstMeeting, DifferenceTraveltoLastMeeting,
Days, SchoolGradeType, GroupGradeType, Early.RPL,
                                       Latest.RPL, Group.State) )
```

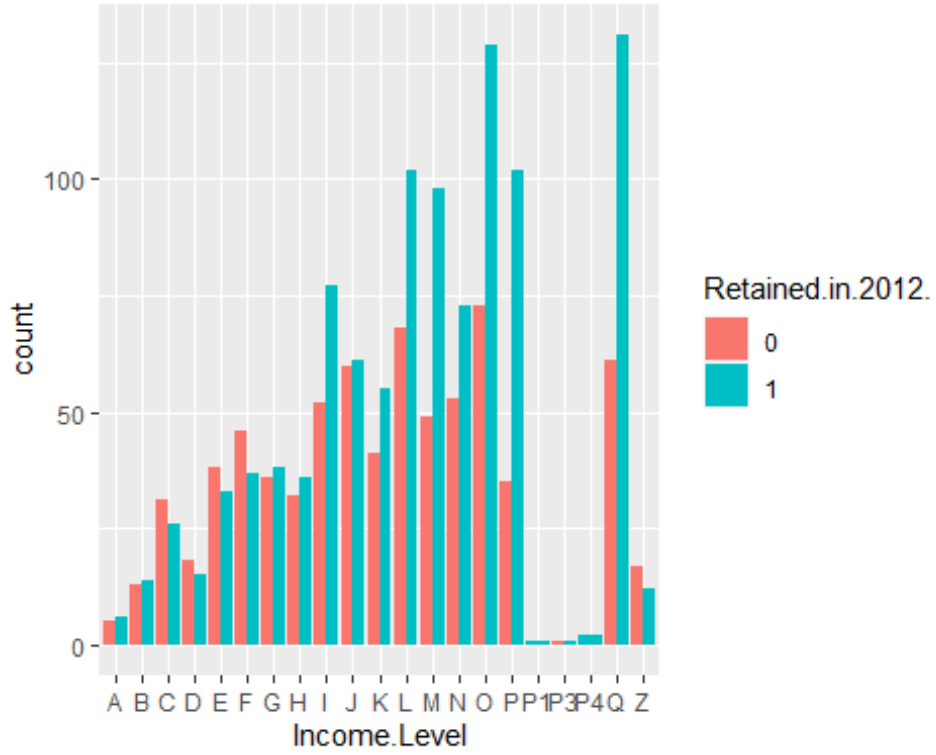
If we consider variables 'Tuition' and 'SPR.Group.Revenue' and plot a scatter plot we can see that the values are exactly same and so we remove 'SPR.Group.Revenue'

```
ggplot(my_data, aes(x=Tuition, y=SPR.Group.Revenue)) + geom_point()
```



If we consider variable Income.Level and compare it with our target variable it can be observed that people from low income levels like 'A','B' and so on have higher number of not retained group. When we look at groups like L, M, O, P we can observe that we have more retained groups as compared to not retained ones

```
ggplot(data=my_data) +
geom_bar(mapping=aes(fill=Retained.in.2012.,x=Income.Level),position="dodge")
```



Hence we can conclude that if the group belongs to groups A, B , C, D, E, F, G, H the group is less likely to return and more likely to return if part of I , K, L M upto Q

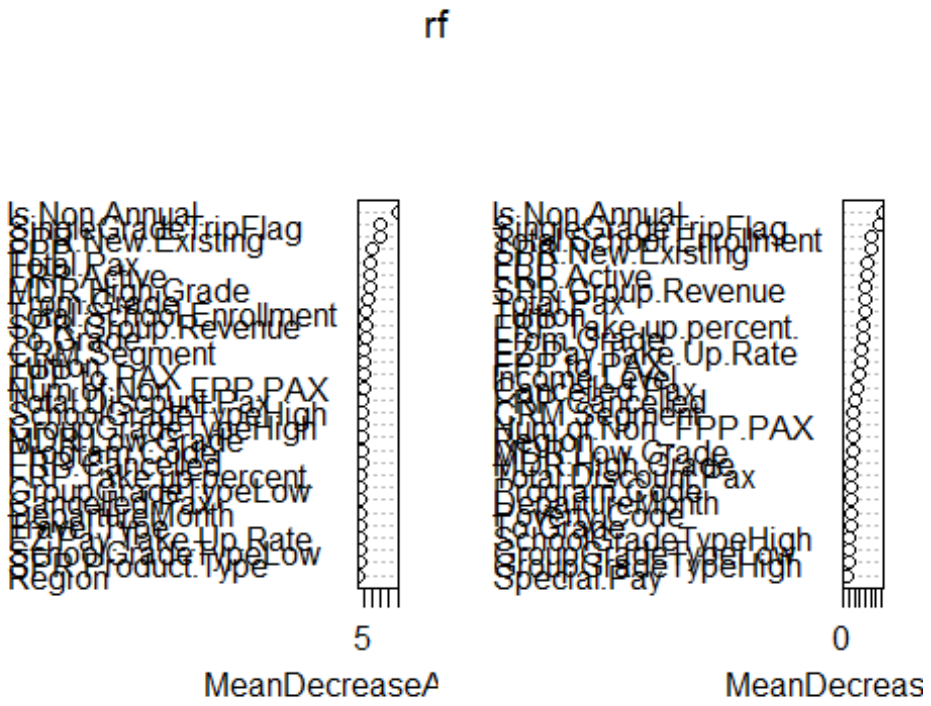
Now we perform Random Forest Model on our dataset

```
set.seed(100)
ntree=100

rf <- randomForest(Retained.in.2012. ~ ., data = my_data, ntree = ntree, mtry
= sqrt(ncol(my_data)-1), proximity = T, importance = T )
print(rf)

##
## Call:
##  randomForest(formula = Retained.in.2012. ~ ., data = my_data,      ntree
= ntree, mtry = sqrt(ncol(my_data) - 1), proximity = T,      importance = T)
##                Type of random forest: classification
##                      Number of trees: 100
## No. of variables tried at each split: 6
##
##         OOB estimate of  error rate: 22.16%
## Confusion matrix:
##     0   1 class.error
## 0 497 232   0.3182442
## 1 162 887   0.1544328
```

```
importance(rf, type =1 ) # type=2
```

```
##                          MeanDecreaseAccuracy
## Program.Code                        3.7583828
## From.Grade                          7.7732625
## To.Grade                            5.9000470
## Is.Non.Annual.                     24.8494314
## Travel.Type                         2.7211371
## Special.Pay                         0.3837641
## Tuition                             5.2771015
## FRP.Active                          8.6893642
## FRP.Cancelled                       3.4294366
## FRP.Take.up.percent.                3.1948465
## Cancelled.Pax                       3.0129270
## Total.Discount.Pax                  4.3102386
## Poverty.Code                        1.6762307
## Region                              2.0444084
## CRM.Segment                         5.3955388
## MDR.Low.Grade                       3.9464244
## MDR.High.Grade                      8.6629070
## Total.School.Enrollment             6.6420087
## Income.Level                        1.9246517
## EZ.Pay.Take.Up.Rate                 2.6522948
## School.Sponsor                      0.3913490
## SPR.Product.Type                    2.3510836
## SPR.New.Existing                   14.3303903
## FPP                                 9.4841895
## Total.Pax                           8.8236672
## SPR.Group.Revenue                   6.5057478
## SchoolGradeTypeLow                  2.4694581
## SchoolGradeTypeHigh                 4.0521729
## DepartureMonth                      2.9398480
## GroupGradeTypeLow                   3.0648446
## GroupGradeTypeHigh                  3.9722519
## MajorProgramCode                    0.5063102
## SingleGradeTripFlag                14.4369510
## FPP.to.PAX                          5.2115114
## Num.of.Non_FPP.PAX                  5.0067464
```

```
varImpPlot(rf)
```




```
rf$err.rate[ntree,1]
```

```
##       OOB
## 0.2215973
```

```
#rf$predicted

CM <- table(rf$predicted, my_data$Retained.in.2012., dnn = c("Predicted",
"Actual"))
CM
```

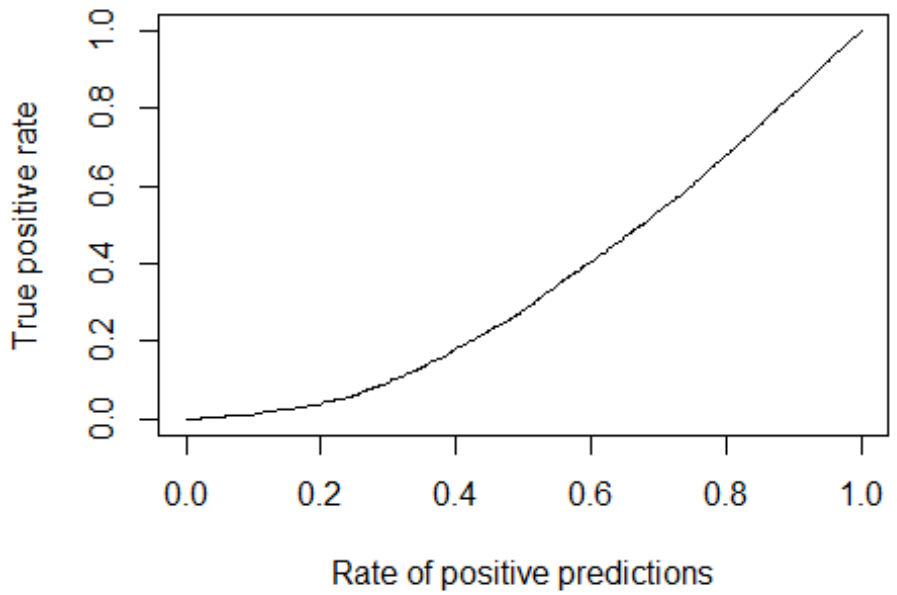
```
##          Actual
## Predicted   0   1
##         0 497 162
##         1 232 887
```

```
confusionMatrix(rf$predicted, my_data$Retained.in.2012.)
```

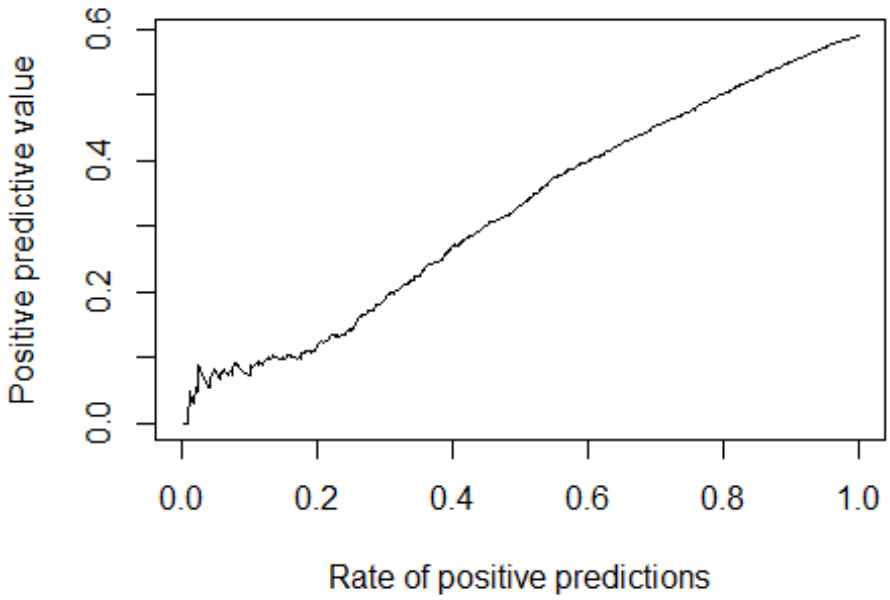
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 497 162
##          1 232 887
##
##                Accuracy : 0.7784
##                  95% CI : (0.7584, 0.7975)
##     No Information Rate : 0.59
```

```
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5352
##
##  Mcnemar's Test P-Value : 0.0005086
##
##             Sensitivity : 0.6818
##             Specificity : 0.8456
##          Pos Pred Value : 0.7542
##          Neg Pred Value : 0.7927
##              Prevalence : 0.4100
##          Detection Rate : 0.2795
##    Detection Prevalence : 0.3706
##       Balanced Accuracy : 0.7637
##
##        'Positive' Class : 0
##
```

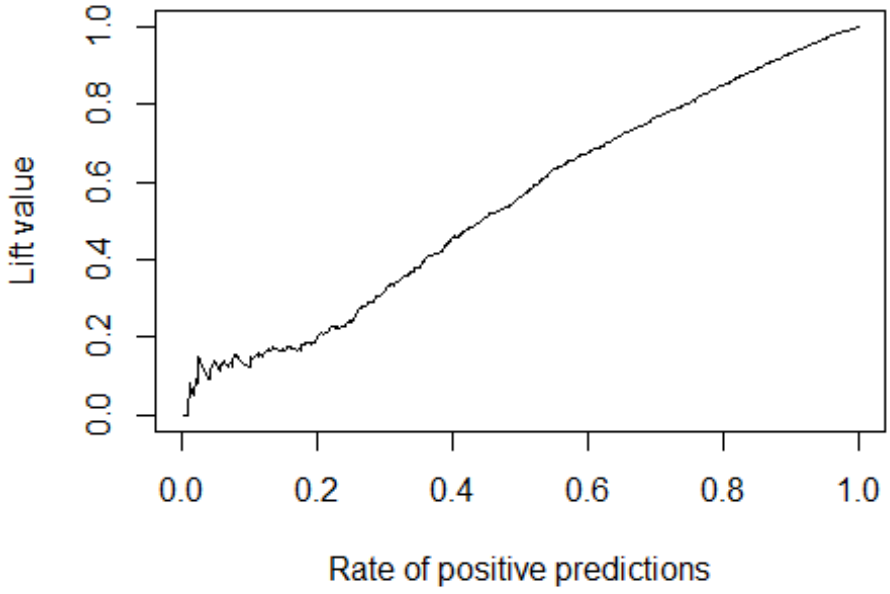
```
pred <- prediction(rf$votes[, 1], my_data$Retained.in.2012.)

#Gain Chart
perf <- performance(pred, "tpr", "rpp")
plot(perf)
```



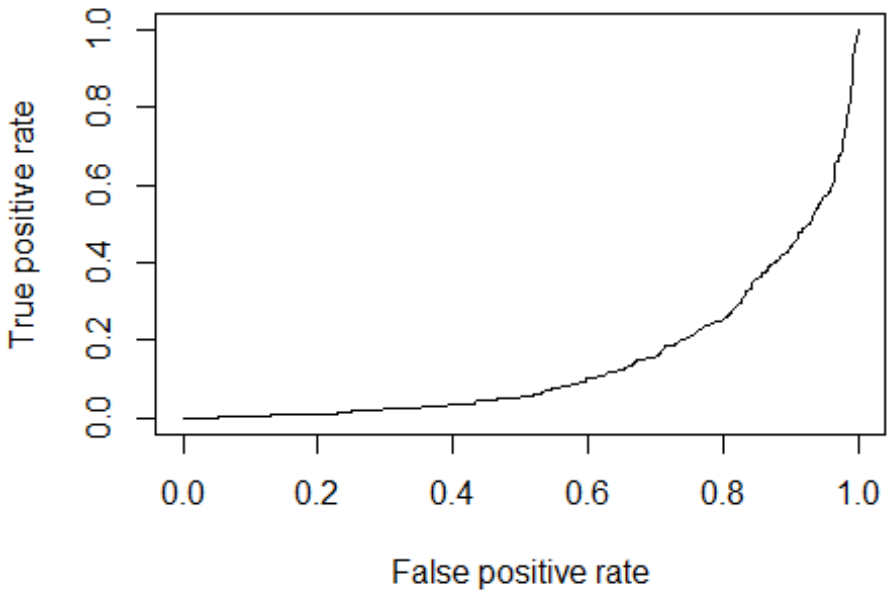

```
#Response Chart
perf <- performance(pred, "ppv", "rpp")
plot(perf)
```




```
#Lift Chart
perf <- performance(pred, "lift", "rpp")
plot(perf)
```



```
#ROC Curve
perf <- performance(pred, "tpr", "fpr")
plot(perf)
```




```
# auc (Area Under Curve)
auc <- performance(pred, "auc")
auc

## A performance instance
##   'Area under the ROC curve'

unlist(slot(auc, "y.values"))

## [1] 0.1482534
```

We can conclude that Random Forest Model gives an accuracy of around 77% and OOB Error rate is around 22%

HW6

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.
3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflict
s() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.1.2
```

```
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 4.1.2
```

```
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.1.2
```

```
library(dplyr)
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.2
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(readxl)
```

```
mydata <- read_excel("C:/Users/Sachin Chavan/Documents/IDS 572 - Data Mining/
Assignments/Homework3/UV7581-XLS-ENG.xlsX")
mydata
```

```
## # A tibble: 2,389 x 56
##       ID Program.Code From.Grade To.Grade Group.State Is.Non.Annual.  Days
##    <dbl> <chr>        <chr>      <chr>    <chr>                <dbl> <dbl>
##  1     1 HS           4          4        CA                       0     1
##  2     2 HC           8          8        AZ                       0     7
##  3     3 HD           8          8        FL                       0     3
##  4     4 HN           9          12       VA                       1     3
##  5     5 HD           6          8        FL                       0     6
##  6     6 HC           10         12       LA                       0     4
##  7     7 SG           11         12       MA                       1     6
##  8     8 FN           9          9        MX                       0     8
##  9     9 CC           8          8        AZ                       0     8
## 10    10 HD           8          8        TX                       0     4
## # ... with 2,379 more rows, and 49 more variables: Travel.Type <chr>,
## #   Departure.Date <dttm>, Return.Date <dttm>, Deposit.Date <dttm>,
## #   Special.Pay <chr>, Tuition <dbl>, FRP.Active <dbl>, FRP.Cancelled <dbl
>,
## #   FRP.Take.up.percent. <dbl>, Early.RPL <chr>, Latest.RPL <chr>,
## #   Cancelled.Pax <dbl>, Total.Discount.Pax <dbl>, Initial.System.Date <ch
r>,
## #   Poverty.Code <chr>, Region <chr>, CRM.Segment <chr>, School.Type <chr>
,
## #   Parent.Meeting.Flag <dbl>, MDR.Low.Grade <chr>, MDR.High.Grade <chr>,
...
```

```
summary(mydata)
```

```
##        ID       Program.Code        From.Grade          To.Grade
##  Min.   :   1   Length:2389        Length:2389        Length:2389
##  1st Qu.: 598   Class :character   Class :character   Class :character
##  Median :1195   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1195
##  3rd Qu.:1792
##  Max.   :2389
##
##  Group.State        Is.Non.Annual.       Days        Travel.Type
##  Length:2389        Min.   :0.000   Min.   : 1.000   Length:2389
##  Class :character   1st Qu.:0.000   1st Qu.: 4.000   Class :character
##  Mode  :character   Median :0.000   Median : 5.000   Mode  :character
##                     Mean   :0.154   Mean   : 4.575
##                     3rd Qu.:0.000   3rd Qu.: 5.000
##                     Max.   :1.000   Max.   :12.000
##
##  Departure.Date                Return.Date
##  Min.   :2011-01-14 00:00:00   Min.   :2011-01-14 00:00:00
##  1st Qu.:2011-04-09 00:00:00   1st Qu.:2011-04-12 00:00:00
##  Median :2011-05-17 00:00:00   Median :2011-05-20 00:00:00
##  Mean   :2011-05-07 18:20:38   Mean   :2011-05-11 11:57:53
##  3rd Qu.:2011-06-07 00:00:00   3rd Qu.:2011-06-10 00:00:00
##  Max.   :2011-06-30 00:00:00   Max.   :2011-07-05 00:00:00
```

```
##
##   Deposit.Date                  Special.Pay           Tuition
##  Min.   :2009-09-25 00:00:00   Length:2389        Min.   :  79
##  1st Qu.:2010-10-15 00:00:00   Class :character   1st Qu.:1174
##  Median :2010-10-28 00:00:00   Mode  :character   Median :1700
##  Mean   :2010-10-24 19:42:37                      Mean   :1615
##  3rd Qu.:2010-11-05 00:00:00                      3rd Qu.:2048
##  Max.   :2011-10-30 00:00:00                      Max.   :4200
##
##    FRP.Active     FRP.Cancelled    FRP.Take.up.percent.  Early.RPL
##  Min.   :  0.00   Min.   : 0.000   Min.   :0.0000       Length:2389
##  1st Qu.:  6.00   1st Qu.: 1.000   1st Qu.:0.4550       Class :character
##  Median : 12.00   Median : 2.000   Median :0.6000       Mode  :character
##  Mean   : 16.87   Mean   : 3.306   Mean   :0.5707
##  3rd Qu.: 23.00   3rd Qu.: 4.000   3rd Qu.:0.7270
##  Max.   :257.00   Max.   :45.000   Max.   :1.0000
##
##   Latest.RPL        Cancelled.Pax    Total.Discount.Pax Initial.System.Dat
e
##  Length:2389        Min.   : 0.000   Min.   : 0.000     Length:2389
##  Class :character   1st Qu.: 2.000   1st Qu.: 1.000     Class :character
##  Mode  :character   Median : 4.000   Median : 2.000     Mode  :character
##                     Mean   : 4.807   Mean   : 2.954
##                     3rd Qu.: 6.000   3rd Qu.: 4.000
##                     Max.   :39.000   Max.   :47.000
##
##  Poverty.Code          Region          CRM.Segment        School.Type
##  Length:2389        Length:2389        Length:2389        Length:2389
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Parent.Meeting.Flag MDR.Low.Grade      MDR.High.Grade
##  Min.   :0.0000      Length:2389        Length:2389
##  1st Qu.:1.0000      Class :character   Class :character
##  Median :1.0000      Mode  :character   Mode  :character
##  Mean   :0.8589
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
##  Total.School.Enrollment Income.Level       EZ.Pay.Take.Up.Rate
##  Min.   :  19.0          Length:2389        Min.   :0.0000
##  1st Qu.: 360.0          Class :character   1st Qu.:0.1000
##  Median : 597.0          Mode  :character   Median :0.2000
##  Mean   : 648.4                             Mean   :0.2079
##  3rd Qu.: 825.8                             3rd Qu.:0.2920
##  Max.   :3990.0                             Max.   :1.7500
##  NA's   :91
```

```
##  School.Sponsor   SPR.Product.Type   SPR.New.Existing        FPP
##  Min.   :0.0000   Length:2389        Length:2389        Min.   :  2.0
##  1st Qu.:0.0000   Class :character   Class :character   1st Qu.: 12.0
##  Median :0.0000   Mode  :character   Mode  :character   Median : 23.0
##  Mean   :0.1059                                        Mean   : 31.3
##  3rd Qu.:0.0000                                        3rd Qu.: 41.0
##  Max.   :1.0000                                        Max.   :286.0
##
##    Total.Pax      SPR.Group.Revenue NumberOfMeetingswithParents
##  Min.   :  2.00   Min.   :  79      Min.   :0.000
##  1st Qu.: 14.00   1st Qu.:1174      1st Qu.:1.000
##  Median : 26.00   Median :1700      Median :1.000
##  Mean   : 34.25   Mean   :1615      Mean   :1.102
##  3rd Qu.: 44.00   3rd Qu.:2048      3rd Qu.:1.000
##  Max.   :313.00   Max.   :4200      Max.   :2.000
##
##  FirstMeeting       LastMeeting        DifferenceTraveltoFirstMeeting
##  Length:2389        Length:2389        Length:2389
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  DifferenceTraveltoLastMeeting SchoolGradeTypeLow SchoolGradeTypeHigh
##  Length:2389                   Length:2389        Length:2389
##  Class :character              Class :character   Class :character
##  Mode  :character              Mode  :character   Mode  :character
##
##
##
##
##  SchoolGradeType    DepartureMonth     GroupGradeTypeLow  GroupGradeTypeHi
gh
##  Length:2389        Length:2389        Length:2389        Length:2389
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  GroupGradeType     MajorProgramCode   SingleGradeTripFlag
##  Length:2389        Length:2389        Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :1.0000
##                                        Mean   :0.5567
##                                        3rd Qu.:1.0000
##                                        Max.   :1.0000
##
##  FPP.to.School.enrollment   FPP.to.PAX     Num.of.Non_FPP.PAX
```

```
##  Length:2389              Min.   :0.6000   Min.   : 0.000
##  Class :character         1st Qu.:0.8824   1st Qu.: 1.000
##  Mode  :character         Median :0.9091   Median : 2.000
##                           Mean   :0.9007   Mean   : 2.954
##                           3rd Qu.:0.9333   3rd Qu.: 4.000
##                           Max.   :1.0000   Max.   :47.000
##
##  SchoolSizeIndicator Retained.in.2012.
##  Length:2389         Min.   :0.0000
##  Class :character    1st Qu.:0.0000
##  Mode  :character    Median :1.0000
##                      Mean   :0.6074
##                      3rd Qu.:1.0000
##                      Max.   :1.0000
##
```

```
mydata$Retained.in.2012. <- as.factor(mydata$Retained.in.2012.)

sum(is.na(mydata))
```

```
## [1] 913
```

```
my_data <- na.omit(mydata)
sum(is.na(mydata))
```

```
## [1] 913
```

```
summary(mydata)
```

```
##        ID       Program.Code        From.Grade          To.Grade
##  Min.   :   1   Length:2389        Length:2389        Length:2389
##  1st Qu.: 598   Class :character   Class :character   Class :character
##  Median :1195   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1195
##  3rd Qu.:1792
##  Max.   :2389
##
##  Group.State        Is.Non.Annual.       Days        Travel.Type
##  Length:2389        Min.   :0.000   Min.   : 1.000   Length:2389
##  Class :character   1st Qu.:0.000   1st Qu.: 4.000   Class :character
##  Mode  :character   Median :0.000   Median : 5.000   Mode  :character
##                     Mean   :0.154   Mean   : 4.575
##                     3rd Qu.:0.000   3rd Qu.: 5.000
##                     Max.   :1.000   Max.   :12.000
##
##  Departure.Date                 Return.Date
##  Min.   :2011-01-14 00:00:00   Min.   :2011-01-14 00:00:00
##  1st Qu.:2011-04-09 00:00:00   1st Qu.:2011-04-12 00:00:00
##  Median :2011-05-17 00:00:00   Median :2011-05-20 00:00:00
##  Mean   :2011-05-07 18:20:38   Mean   :2011-05-11 11:57:53
##  3rd Qu.:2011-06-07 00:00:00   3rd Qu.:2011-06-10 00:00:00
```

```
##  Max.   :2011-06-30 00:00:00   Max.   :2011-07-05 00:00:00
##
##   Deposit.Date                  Special.Pay           Tuition
##  Min.   :2009-09-25 00:00:00   Length:2389        Min.   :  79
##  1st Qu.:2010-10-15 00:00:00   Class :character   1st Qu.:1174
##  Median :2010-10-28 00:00:00   Mode  :character   Median :1700
##  Mean   :2010-10-24 19:42:37                      Mean   :1615
##  3rd Qu.:2010-11-05 00:00:00                      3rd Qu.:2048
##  Max.   :2011-10-30 00:00:00                      Max.   :4200
##
##    FRP.Active     FRP.Cancelled    FRP.Take.up.percent.  Early.RPL
##  Min.   :  0.00   Min.   : 0.000   Min.   :0.0000       Length:2389
##  1st Qu.:  6.00   1st Qu.: 1.000   1st Qu.:0.4550       Class :character
##  Median : 12.00   Median : 2.000   Median :0.6000       Mode  :character
##  Mean   : 16.87   Mean   : 3.306   Mean   :0.5707
##  3rd Qu.: 23.00   3rd Qu.: 4.000   3rd Qu.:0.7270
##  Max.   :257.00   Max.   :45.000   Max.   :1.0000
##
##   Latest.RPL        Cancelled.Pax    Total.Discount.Pax Initial.System.Dat
e
##  Length:2389        Min.   : 0.000   Min.   : 0.000     Length:2389
##  Class :character   1st Qu.: 2.000   1st Qu.: 1.000     Class :character
##  Mode  :character   Median : 4.000   Median : 2.000     Mode  :character
##                     Mean   : 4.807   Mean   : 2.954
##                     3rd Qu.: 6.000   3rd Qu.: 4.000
##                     Max.   :39.000   Max.   :47.000
##
##  Poverty.Code          Region          CRM.Segment        School.Type
##  Length:2389        Length:2389        Length:2389        Length:2389
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Parent.Meeting.Flag MDR.Low.Grade      MDR.High.Grade
##  Min.   :0.0000      Length:2389        Length:2389
##  1st Qu.:1.0000      Class :character   Class :character
##  Median :1.0000      Mode  :character   Mode  :character
##  Mean   :0.8589
##  3rd Qu.:1.0000
##  Max.   :1.0000
##
##  Total.School.Enrollment Income.Level       EZ.Pay.Take.Up.Rate
##  Min.   :  19.0          Length:2389        Min.   :0.0000
##  1st Qu.: 360.0          Class :character   1st Qu.:0.1000
##  Median : 597.0          Mode  :character   Median :0.2000
##  Mean   : 648.4                             Mean   :0.2079
##  3rd Qu.: 825.8                             3rd Qu.:0.2920
##  Max.   :3990.0                             Max.   :1.7500
```

```
##  NA's   :91
##  School.Sponsor   SPR.Product.Type   SPR.New.Existing        FPP
##  Min.   :0.0000   Length:2389        Length:2389        Min.   :  2.0
##  1st Qu.:0.0000   Class :character   Class :character   1st Qu.: 12.0
##  Median :0.0000   Mode  :character   Mode  :character   Median : 23.0
##  Mean   :0.1059                                        Mean   : 31.3
##  3rd Qu.:0.0000                                        3rd Qu.: 41.0
##  Max.   :1.0000                                        Max.   :286.0
##
##    Total.Pax      SPR.Group.Revenue NumberOfMeetingswithParents
##  Min.   :  2.00   Min.   :  79      Min.   :0.000
##  1st Qu.: 14.00   1st Qu.:1174      1st Qu.:1.000
##  Median : 26.00   Median :1700      Median :1.000
##  Mean   : 34.25   Mean   :1615      Mean   :1.102
##  3rd Qu.: 44.00   3rd Qu.:2048      3rd Qu.:1.000
##  Max.   :313.00   Max.   :4200      Max.   :2.000
##
##  FirstMeeting       LastMeeting        DifferenceTraveltoFirstMeeting
##  Length:2389        Length:2389        Length:2389
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  DifferenceTraveltoLastMeeting SchoolGradeTypeLow SchoolGradeTypeHigh
##  Length:2389                   Length:2389        Length:2389
##  Class :character              Class :character   Class :character
##  Mode  :character              Mode  :character   Mode  :character
##
##
##
##
##  SchoolGradeType    DepartureMonth     GroupGradeTypeLow  GroupGradeTypeHi
gh
##  Length:2389        Length:2389        Length:2389        Length:2389
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  GroupGradeType     MajorProgramCode   SingleGradeTripFlag
##  Length:2389        Length:2389        Min.   :0.0000
##  Class :character   Class :character   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median :1.0000
##                                        Mean   :0.5567
##                                        3rd Qu.:1.0000
##                                        Max.   :1.0000
##
```

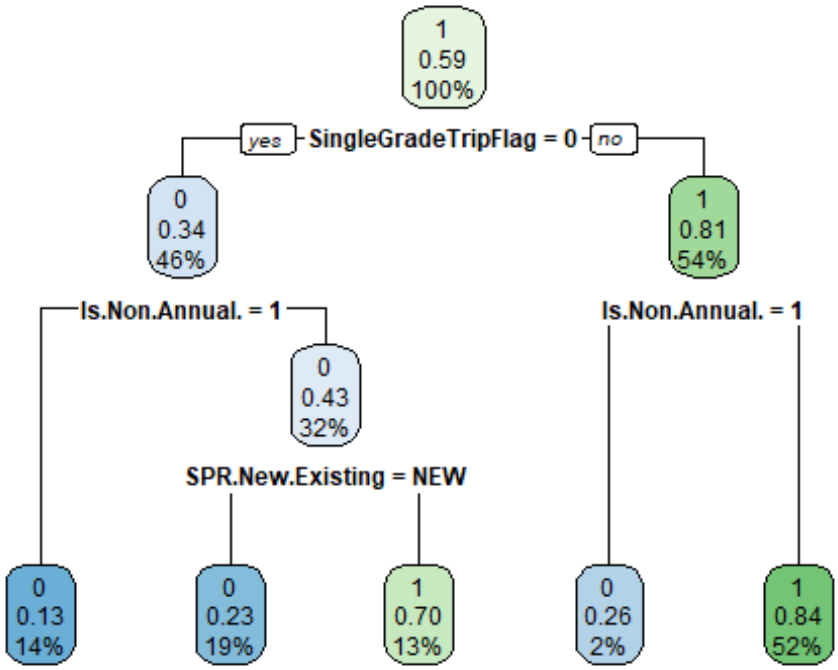
```
##  FPP.to.School.enrollment   FPP.to.PAX     Num.of.Non_FPP.PAX
##  Length:2389              Min.   :0.6000   Min.   : 0.000
##  Class :character         1st Qu.:0.8824   1st Qu.: 1.000
##  Mode  :character         Median :0.9091   Median : 2.000
##                           Mean   :0.9007   Mean   : 2.954
##                           3rd Qu.:0.9333   3rd Qu.: 4.000
##                           Max.   :1.0000   Max.   :47.000
##
##  SchoolSizeIndicator Retained.in.2012.
##  Length:2389         0: 938
##  Class :character    1:1451
##  Mode  :character
##
##
##
##
```

```
mydata
```

```
## # A tibble: 2,389 x 56
##       ID Program.Code From.Grade To.Grade Group.State Is.Non.Annual.  Days
##    <dbl> <chr>        <chr>      <chr>    <chr>                <dbl> <dbl>
##  1     1 HS           4          4        CA                       0     1
##  2     2 HC           8          8        AZ                       0     7
##  3     3 HD           8          8        FL                       0     3
##  4     4 HN           9          12       VA                       1     3
##  5     5 HD           6          8        FL                       0     6
##  6     6 HC           10         12       LA                       0     4
##  7     7 SG           11         12       MA                       1     6
##  8     8 FN           9          9        MX                       0     8
##  9     9 CC           8          8        AZ                       0     8
## 10    10 HD           8          8        TX                       0     4
## # ... with 2,379 more rows, and 49 more variables: Travel.Type <chr>,
## #   Departure.Date <dttm>, Return.Date <dttm>, Deposit.Date <dttm>,
## #   Special.Pay <chr>, Tuition <dbl>, FRP.Active <dbl>, FRP.Cancelled <dbl
>,
## #   FRP.Take.up.percent. <dbl>, Early.RPL <chr>, Latest.RPL <chr>,
## #   Cancelled.Pax <dbl>, Total.Discount.Pax <dbl>, Initial.System.Date <ch
r>,
## #   Poverty.Code <chr>, Region <chr>, CRM.Segment <chr>, School.Type <chr>
,
## #   Parent.Meeting.Flag <dbl>, MDR.Low.Grade <chr>, MDR.High.Grade <chr>,
...
```

```
mydata1 <-  subset(mydata, select=-c(School.Type,FPP.to.School.enrollment,Dif
ferenceTraveltoFirstMeeting,DifferenceTraveltoLastMeeting,
                                     Departure.Date,DifferenceTraveltoFirstMeetin
g, DifferenceTraveltoLastMeeting, Days, SchoolGradeType, GroupGradeType, Earl
y.RPL,
                                     Latest.RPL, Return.Date, Deposit.Date, Incom
```

```
e.Level,Initial.System.Date, Parent.Meeting.Flag, NumberOfMeetingswithParents
,
                                    SchoolSizeIndicator, FirstMeeting, LastMeeti
ng, Poverty.Code,Group.State, CRM.Segment, To.Grade, Program.Code))
set.seed(110)
indx <- sample(2, nrow (mydata1), replace=TRUE, prob = c(0.7,0.3))
train <- mydata1[indx == 1, ]
nrow(train)

## [1] 1675

test <- mydata1[indx == 2, ]
nrow(test)

## [1] 714

myFormula <- Retained.in.2012. ~.
mytree1 <- rpart(myFormula, data = train, method="class")
rpart.plot(mytree1)
```




```
pred_train1<-predict(mytree1,data=train,type="class")
mean_1_train<-mean(train$Retained.in.2012.!=pred_train1)
mean_1_train

## [1] 0.1922388
```

```
pred_test1<-predict(mytree1,newdata=test,type="class")
mean_1_test<-mean(test$Retained.in.2012.!=pred_test1)
mean_1_test
```

```
## [1] 0.2128852
```