Homework 3

BUSN 41204 - 2023

- Aman Krishna
- Christian Pavilanis
- Jingwen Li
- Yazmin Ramirez Delgado

In [ ]:
```
knitr::opts_chunk$set(eval = FALSE)
```

Due: end of day Saturday, February 4

Submission instructions: Submit one write-up per group on gradescope.com.

IMPORTANT:

- Write names of everyone that worked on the assignment on the submission.
- Specify every member of the group when submitting on Gradescope (https://help.gradescope.com/article/m5qz2xsnjy-student-add-group-members)

For this homework, we will be using the case *Retention Modeling at Scholastic Travel Company*. Read:

- Case: Retention Modeling at Scholastic Travel Company (A);
- Supplement: Retention Modeling at Scholastic Travel Company (B);

which are available on Canvas.

Your goal is to help David build a model for retention.

The following code will get you started.

Load relevant libraries

In [ ]:
```
library(dplyr)
library(caret)
library(glmnet)
```

Load the data

Here we will load the data from the CSV data file, examine its structure, and fix the data types incorrectly identified by R when importing from CSV.

In [ ]:
```
STCdata_A<-read.csv('travelData.csv')
STCdata_A<-STCdata_A[,-1]
```

You can use the function `str` to quickly check the internal structure of an R object. Here we are using it to investigate type of data in each column of the loaded data.

In [ ]:
```
str(STCdata_A)
```

```
'data.frame':	2389 obs. of  55 variables:
 $ Program.Code                  : chr  "HS" "HC" "HD" "HN" ...
 $ From.Grade                    : int  4 8 8 9 6 10 11 9 8 8 ...
 $ To.Grade                      : int  4 8 8 12 8 12 12 9 8 8 ...
 $ Group.State                   : chr  "CA" "AZ" "FL" "VA" ...
 $ Is.Non.Annual.                : int  0 0 0 1 0 0 1 0 0 0 ...
 $ Days                          : int  1 7 3 3 6 4 6 8 8 4 ...
 $ Travel.Type                   : chr  "A" "A" "A" "B" ...
 $ Departure.Date                : chr  "1/14/2011" "1/14/2011" "1/15/2011" "1/15/2011" ...
 $ Return.Date                   : chr  "1/14/2011" "1/21/2011" "1/17/2011" "1/17/2011" ...
 $ Deposit.Date                  : chr  "8/30/2010" "11/15/2009" "10/15/2010" "1/7/2011" ...
 $ Special.Pay                   : chr  NA "CP" NA NA ...
 $ Tuition                       : int  424 2350 1181 376 865 2025 1977 3379 2200 1428 ...
 $ FRP.Active                    : int  25 9 17 0 40 9 16 10 30 51 ...
 $ FRP.Cancelled                 : int  3 9 6 0 8 4 4 0 0 1 ...
 $ FRP.Take.up.percent.          : num  0.424 0.409 0.708 0 0.494 0.9 0.64 0.769 0.577 0.773 ...
 $ Early.RPL                     : chr  "3/29/2010" "10/20/2009" "4/29/2010" NA ...
 $ Latest.RPL                    : chr  "8/12/2010" "8/10/2010" "8/16/2010" NA ...
 $ Cancelled.Pax                 : int  3 11 6 1 9 3 5 1 0 1 ...
 $ Total.Discount.Pax            : int  4 3 3 0 8 1 2 1 4 6 ...
 $ Initial.System.Date           : chr  "3/26/2010" "10/2/2009" "1/28/2010" "10/19/2010" ...
 $ Poverty.Code                  : chr  "B" "C" "C" "" ...
 $ Region                        : chr  "Southern California" "Other" "Other" "Other" ...
 $ CRM.Segment                   : int  4 10 10 7 10 8 8 7 5 5 ...
 $ School.Type                   : chr  "PUBLIC" "PUBLIC" "PUBLIC" "CHD" ...
 $ Parent.Meeting.Flag           : int  1 1 1 0 1 1 1 1 1 1 ...
 $ MDR.Low.Grade                 : chr  "K" "7" "6" "" ...
 $ MDR.High.Grade                : int  5 8 8 NA 8 12 12 NA 12 8 ...
 $ Total.School.Enrollment       : int  927 850 955 NA 720 939 225 NA 500 635 ...
 $ Income.Level                  : chr  "Q" "A" "O" "" ...
 $ EZ.Pay.Take.Up.Rate           : num  0.17 0.091 0.042 0 0.383 0.1 0.08 0 0.231 0.136 ...
 $ School.Sponsor                : int  1 0 0 0 0 0 0 0 0 0 ...
 $ SPR.Product.Type              : chr  "CA History" "East Coast" "East Coast" "East Coast" ...
 $ SPR.New.Existing              : chr  "EXISTING" "EXISTING" "EXISTING" "EXISTING" ...
 $ FPP                           : int  59 22 24 18 81 10 25 13 52 66 ...
 $ Total.Pax                     : int  63 25 27 18 89 11 27 14 56 72 ...
 $ SPR.Group.Revenue             : int  424 2350 1181 376 865 2025 1977 3379 2200 1428 ...
 $ NumberOfMeetingswithParents   : int  1 2 1 0 1 1 1 1 1 1 ...
 $ FirstMeeting                  : chr  "8/12/2010" "11/17/2009" "9/13/2010" NA ...
 $ LastMeeting                   : chr  "8/12/2010" "8/27/2010" "9/13/2010" NA ...
 $ DifferenceTraveltoFirstMeeting: int  155 423 124 NA 145 91 63 138 143 146 ...
 $ DifferenceTraveltoLastMeeting : int  155 140 124 NA 145 91 63 138 143 146 ...
 $ SchoolGradeTypeLow            : chr  "Elementary" "Middle" "Middle" "High" ...
 $ SchoolGradeTypeHigh           : chr  "Elementary" "Middle" "Middle" "High" ...
 $ SchoolGradeType               : chr  "Elementary->Elementary" "Middle->Middle" "Middle->Middle" "High->High" ...
 $ DepartureMonth                : chr  "January" "January" "January" "January" ...
 $ GroupGradeTypeLow             : chr  "K" "Middle" "Middle" "Undefined" ...
 $ GroupGradeTypeHigh            : chr  "Elementary" "Middle" "Middle" "Undefined" ...
 $ GroupGradeType                : chr  "K->Elementary" "Middle->Middle" "Middle->Middle" "Undefined->Undefined" ...
 $ MajorProgramCode              : chr  "H" "H" "H" "H" ...
 $ SingleGradeTripFlag           : int  1 1 1 0 0 0 0 1 1 1 ...
 $ FPP.to.School.enrollment      : num  0.0636 0.0259 0.0251 NA 0.1125 ...
 $ FPP.to.PAX                    : num  0.937 0.88 0.889 1 0.91 ...
 $ Num.of.Non_FPP.PAX            : int  4 3 3 0 8 1 2 1 4 6 ...
 $ SchoolSizeIndicator           : chr  "L" "L" "L" "" ...
 $ Retained.in.2012.             : int  1 1 1 0 0 1 0 0 1 1 ...
```

Notice that some columns are identified as numerical or integer, but really the should be factors.

For instance, we have that column `From.Grade`

In [ ]:
```
n_distinct(STCdata_A$From.Grade, na.rm = FALSE)    ## n_distinct is a function from dplyr package
```

11

only has 11 levels. It might be a better idea to treat it as a factor instead.

You can fix incorrectly classified data types as follows:

In [ ]:
```
STCdata_A <- mutate_at(STCdata_A, vars(From.Grade), as.factor)
```

We can check that indeed the column represents a factor:

In [ ]:
```
str( STCdata_A$From.Grade )
```

```
 Factor w/ 10 levels "3","4","5","6",..: 2 6 6 7 4 8 9 7 6 6 ...
```

Fix other columns that are numeric at the moment, but could be converted to factors. The following line first finds numeric columns and then identifies the number of unique elements in each one.

In [ ]:
```
( unique.per.column <- sapply( dplyr::select_if(STCdata_A, is.numeric), n_distinct ) )
```

To.Grade: 11 **Is.Non.Annual.:** 2 **Days:** 12 **Tuition:** 1230 **FRP.Active:** 93 **FRP.Cancelled:** 29 **FRP.Take.up.percent.:** 476 **Cancelled.Pax:** 34 **Total.Discount.Pax:** 26 **CRM.Segment:** 12 **Parent.Meeting.Flag:** 2 **MDR.High.Grade:** 13 **Total.School.Enrollment:** 894 **EZ.Pay.Take.Up.Rate:** 371 **School.Sponsor:** 2 **FPP:** 146 **Total.Pax:** 159 **SPR.Group.Revenue:** 1230 **NumberOfMeetingswithParents:** 3 **DifferenceTraveltoFirstMeeting:** 343 **DifferenceTraveltoLastMeeting:** 252 **SingleGradeTripFlag:** 2 **FPP.to.School.enrollment:** 1910 **FPP.to.PAX:** 306 **Num.of.Non_FPP.PAX:** 26 **Retained.in.2012.:** 2

Let us convert every column that has less than 15 unique values into a factor. The following line identify names of such columns.

In [ ]:

```
( column.names.to.factor <- names(unique.per.column)[unique.per.column < 15] )
```

'To.Grade' · 'Is.Non.Annual.' · 'Days' · 'CRM.Segment' · 'Parent.Meeting.Flag' · 'MDR.High.Grade' · 'School.Sponsor' · 'NumberOfMeetingswithParents' · 'SingleGradeTripFlag' · 'Retained.in.2012.'

From this, we can see that the columns `To.Grade`, `Is.Non.Annual.`, `Days`, `CRM.Segment`, `Parent.Meeting.Flag`, `MDR.High.Grade`, `School.Sponsor`, `NumberOfMeetingswithParents`, `SingleGradeTripFlag` can be converted to factors. We can also convert the output `Retained.in.2012.`

Convert these columns into factors.

In [ ]:

```
STCdata_A <- mutate_at(STCdata_A, column.names.to.factor, as.factor)
```

Now let's take care of date columns.

In [ ]:

```
date.columns = c('Departure.Date', 'Return.Date', 'Deposit.Date', 'Early.RPL', 'Latest.RPL',
                 'Initial.System.Date', 'FirstMeeting', 'LastMeeting')
STCdata_A <- mutate_at(STCdata_A, date.columns, function(x) as.Date(x, format = "%m/%d/%Y"))
```

And finally we change all the character columns to factors as well.

In [ ]:

```
STCdata_A <- mutate_if(STCdata_A, is.character, as.factor)
```

Let's see what we have:

In [ ]:

```
str(STCdata_A)
```

```
'data.frame':	2389 obs. of  55 variables:
 $ Program.Code                  : Factor w/ 28 levels "CC","CD","CN",..: 15 6 7 12 7 6 25 5 1 7 ...
 $ From.Grade                    : Factor w/ 10 levels "3","4","5","6",..: 2 6 6 7 4 8 9 7 6 6 ...
 $ To.Grade                      : Factor w/ 10 levels "3","4","5","6",..: 2 6 6 10 6 10 10 7 6 6 ...
 $ Group.State                   : Factor w/ 54 levels "AB","AK","AL",..: 7 5 11 49 11 20 21 29 5 47 ...
 $ Is.Non.Annual.                : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 2 1 1 1 ...
 $ Days                          : Factor w/ 12 levels "1","2","3","4",..: 1 7 3 3 6 4 6 8 8 4 ...
 $ Travel.Type                   : Factor w/ 4 levels "A","B","N","T": 1 1 1 2 4 1 1 1 1 1 ...
 $ Departure.Date                : Date, format: "2011-01-14" "2011-01-14" ...
 $ Return.Date                   : Date, format: "2011-01-14" "2011-01-21" ...
 $ Deposit.Date                  : Date, format: "2010-08-30" "2009-11-15" ...
 $ Special.Pay                   : Factor w/ 4 levels "","CP","FR","SA": NA 2 NA NA NA NA NA NA 2 NA ...
 $ Tuition                       : int  424 2350 1181 376 865 2025 1977 3379 2200 1428 ...
 $ FRP.Active                    : int  25 9 17 0 40 9 16 10 30 51 ...
 $ FRP.Cancelled                 : int  3 9 6 0 8 4 4 0 0 1 ...
 $ FRP.Take.up.percent.          : num  0.424 0.409 0.708 0 0.494 0.9 0.64 0.769 0.577 0.773 ...
 $ Early.RPL                     : Date, format: "2010-03-29" "2009-10-20" ...
 $ Latest.RPL                    : Date, format: "2010-08-12" "2010-08-10" ...
 $ Cancelled.Pax                 : int  3 11 6 1 9 3 5 1 0 1 ...
 $ Total.Discount.Pax            : int  4 3 3 0 8 1 2 1 4 6 ...
 $ Initial.System.Date           : Date, format: "2010-03-26" "2009-10-02" ...
 $ Poverty.Code                  : Factor w/ 7 levels "","0","A","B",..: 4 5 5 1 6 5 1 1 1 1 ...
 $ Region                        : Factor w/ 6 levels "Dallas","Houston",..: 6 4 4 4 4 4 4 4 4 2 ...
 $ CRM.Segment                   : Factor w/ 11 levels "1","2","3","4",..: 4 10 10 7 10 8 8 7 5 5 ...
 $ School.Type                   : Factor w/ 4 levels "CHD","Catholic",..: 3 3 3 1 3 3 2 1 1 4 ...
 $ Parent.Meeting.Flag           : Factor w/ 2 levels "0","1": 2 2 2 1 2 2 2 2 2 2 ...
 $ MDR.Low.Grade                 : Factor w/ 13 levels "","1","10","2",..: 12 9 8 1 8 3 11 1 8 13 ...
 $ MDR.High.Grade                : Factor w/ 12 levels "1","2","3","4",..: 5 8 8 NA 8 12 12 NA 12 8 ...
 $ Total.School.Enrollment       : int  927 850 955 NA 720 939 225 NA 500 635 ...
 $ Income.Level                  : Factor w/ 23 levels "","A","B","C",..: 22 2 16 1 4 10 8 1 12 12 ...
 $ EZ.Pay.Take.Up.Rate           : num  0.17 0.091 0.042 0 0.383 0.1 0.08 0 0.231 0.136 ...
 $ School.Sponsor                : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1 ...
 $ SPR.Product.Type              : Factor w/ 6 levels "CA History","Costa Rica",..: 1 3 3 3 3 3 6 3 3 3 ...
 $ SPR.New.Existing              : Factor w/ 2 levels "EXISTING","NEW": 1 1 1 1 1 2 1 1 1 1 ...
 $ FPP                           : int  59 22 24 18 81 10 25 13 52 66 ...
 $ Total.Pax                     : int  63 25 27 18 89 11 27 14 56 72 ...
 $ SPR.Group.Revenue             : int  424 2350 1181 376 865 2025 1977 3379 2200 1428 ...
 $ NumberOfMeetingswithParents   : Factor w/ 3 levels "0","1","2": 2 3 2 1 2 2 2 2 2 2 ...
 $ FirstMeeting                  : Date, format: "2010-08-12" "2009-11-17" ...
 $ LastMeeting                   : Date, format: "2010-08-12" "2010-08-27" ...
 $ DifferenceTraveltoFirstMeeting: int  155 423 124 NA 145 91 63 138 143 146 ...
 $ DifferenceTraveltoLastMeeting : int  155 140 124 NA 145 91 63 138 143 146 ...
 $ SchoolGradeTypeLow            : Factor w/ 4 levels "Elementary","High",..: 1 3 3 2 3 2 2 2 3 3 ...
 $ SchoolGradeTypeHigh           : Factor w/ 4 levels "Elementary","High",..: 1 3 3 2 3 2 2 2 3 3 ...
 $ SchoolGradeType               : Factor w/ 9 levels "Elementary->Elementary",..: 1 7 7 5 7 5 5 5 7 7 ...
 $ DepartureMonth                : Factor w/ 6 levels "April","February",..: 3 3 3 3 3 3 3 3 3 2 ...
 $ GroupGradeTypeLow             : Factor w/ 6 levels "Elementary","High",..: 3 4 4 6 4 2 2 6 4 5 ...
 $ GroupGradeTypeHigh            : Factor w/ 4 levels "Elementary","High",..: 1 3 3 4 3 2 2 4 2 3 ...
 $ GroupGradeType                : Factor w/ 13 levels "Elementary->Elementary",..: 5 9 9 13 9 4 4 13 8 12 ...
 $ MajorProgramCode              : Factor w/ 4 levels "C","H","I","S": 2 2 2 2 2 2 4 3 1 2 ...
 $ SingleGradeTripFlag           : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 2 2 2 ...
 $ FPP.to.School.enrollment      : num  0.0636 0.0259 0.0251 NA 0.1125 ...
 $ FPP.to.PAX                    : num  0.937 0.88 0.889 1 0.91 ...
 $ Num.of.Non_FPP.PAX            : int  4 3 3 0 8 1 2 1 4 6 ...
 $ SchoolSizeIndicator           : Factor w/ 5 levels "","L","M-L","S",..: 2 2 2 1 3 2 4 1 5 3 ...
 $ Retained.in.2012.             : Factor w/ 2 levels "0","1": 2 2 2 1 1 2 1 1 2 2 ...
```

Pretty good!!!

Data preprocessing

The data contains a number of columns with missing values. Let's investigate. The following tells us the number of missing values in each column.

In [ ]:

```
sapply(STCdata_A, function(x) sum(is.na(x)))
```

Program.Code: 0 **From.Grade:** 127 **To.Grade:** 150 **Group.State:** 0 **Is.Non.Annual.:** 0 **Days:** 0 **Travel.Type:** 0 **Departure.Date:** 0 **Return.Date:** 0 **Deposit.Date:** 0 **Special.Pay:** 1917 **Tuition:** 0 **FRP.Active:** 0 **FRP.Cancelled:** 0 **FRP.Take.up.percent.:** 0 **Early.RPL:** 673 **Latest.RPL:** 19 **Cancelled.Pax:** 0 **Total.Discount.Pax:** 0 **Initial.System.Date:** 8 **Poverty.Code:** 0 **Region:** 0 **CRM.Segment:** 4 **School.Type:** 0 **Parent.Meeting.Flag:** 0 **MDR.Low.Grade:** 0 **MDR.High.Grade:** 68 **Total.School.Enrollment:** 91 **Income.Level:** 0 **EZ.Pay.Take.Up.Rate:** 0 **School.Sponsor:** 0 **SPR.Product.Type:** 0 **SPR.New.Existing:** 0 **FPP:** 0 **Total.Pax:** 0 **SPR.Group.Revenue:** 0 **NumberOfMeetingswithParents:** 0 **FirstMeeting:** 337 **LastMeeting:** 337 **DifferenceTraveltoFirstMeeting:** 337 **DifferenceTraveltoLastMeeting:** 337 **SchoolGradeTypeLow:** 0 **SchoolGradeTypeHigh:** 0 **SchoolGradeType:** 0 **DepartureMonth:** 0 **GroupGradeTypeLow:** 0 **GroupGradeTypeHigh:** 0 **GroupGradeType:** 0 **MajorProgramCode:** 0 **SingleGradeTripFlag:** 0 **FPP.to.School.enrollment:** 91 **FPP.to.PAX:** 0 **Num.of.Non_FPP.PAX:** 0 **SchoolSizeIndicator:** 0 **Retained.in.2012.:** 0

Dealing with missing values is a challenging problem, which could occupy a quarter of its own. The purpose of this homework is not to investigate in-depth approaches to dealing with missing values, but rather to investigate classification. For that reason, we take the following simple approach.

The function `fixNAs` below fixes missing values. The function defines reactions:

- adds a new category "FIXED_NA" for a missing value of a categorical/factor variable;
- fills zero value for a missing value of a numeric variable;
- fills "1900-01-01" for a missing value of a date variable.

Then it loops through all columns in the dataframe, reads their types, and loops through all the values, applying the defined reaction to any missing data point. In addition, the function creates a surrogate dummy variable for each column containing at least one missing value (for example, `Special.Pay_surrogate`), which takes a value of 1 whenever the original variable (`Special.Pay`) has a missing value, and 0 otherwise.

In [ ]:

```
# Create a custom function to fix missing values ("NAs") and
# preserve the NA info as surrogate variables
fixNAs <- function(data_frame){
  # Define reactions to NAs
  integer_reac <- 0
  factor_reac <- "FIXED_NA"
  character_reac <- "FIXED_NA"
  date_reac <- as.Date("1900-01-01")

  # Loop through columns in the data frame
  # and depending on which class the
  # variable is, apply the defined reaction and
  # create a surrogate

  for (i in 1:ncol(data_frame)) {
    if (class(data_frame[,i]) %in% c("numeric","integer")) {
      if (any(is.na(data_frame[,i]))) {
        data_frame[,paste0(colnames(data_frame)[i],"_surrogate")] <-
          as.factor(ifelse(is.na(data_frame[,i]),"1","0"))
        data_frame[is.na(data_frame[,i]), i] <- integer_reac
      }
    } else
      if (class(data_frame[,i]) %in% c("factor")) {
        if (any(is.na(data_frame[,i]))){
          data_frame[,i]<-as.character(data_frame[,i])
          data_frame[,paste0(colnames(data_frame)[i],"_surrogate")] <-
            as.factor(ifelse(is.na(data_frame[,i]),"1","0"))
          data_frame[is.na(data_frame[,i]),i]<-factor_reac
          data_frame[,i]<-as.factor(data_frame[,i])
        }
      } else {
        if (class(data_frame[,i]) %in% c("character")) {
          if (any(is.na(data_frame[,i]))){
            data_frame[,paste0(colnames(data_frame)[i],"_surrogate")]<-
              as.factor(ifelse(is.na(data_frame[,i]),"1","0"))
            data_frame[is.na(data_frame[,i]),i]<-character_reac
          }
        } else {
          if (class(data_frame[,i]) %in% c("Date")) {
            if (any(is.na(data_frame[,i]))){
              data_frame[,paste0(colnames(data_frame)[i],"_surrogate")]<-
                as.factor(ifelse(is.na(data_frame[,i]),"1","0"))
              data_frame[is.na(data_frame[,i]),i]<-date_reac
            }
          }
        }
      }
  }

  return(data_frame)
}
```

We apply the above defined function to our data frame.

In [ ]:

```
STCdata_A<-fixNAs(STCdata_A)
```

We can see that the columns do not have any missing values any more.

In [ ]:

```
any( sapply(STCdata_A, function(x) sum(is.na(x))) > 0)
```

```
FALSE
```

Next, we combine the rare categories. Levels that do not occur often during training tend not to have reliable effect estimates and contribute to over-fit.

Let us check for rare categories in the variable `Group.State`.

In [ ]:

```
table(STCdata_A$Group.State)
```

```
             AB              AK              AL              AR              AZ 
              1               5              21              10              53 
        Bermuda              CA              CO              CT  Cayman Islands 
              1             718              89              15               1 
             FL              GA              HI              IA              ID 
             62              22               9              35              14 
             IL              IN              KS              KY              LA 
            104              43              26              16              31 
             MA              MD              ME              MI              MN 
             36              15               7              71              51 
             MO              MS              MT              MX              NC 
             43               9               6               3              16 
             ND              NE              NH              NJ              NM 
              5              42               7               6              20 
             NV              NY              OH              OK              OR 
             20              19              53              33              51 
             PA              PR              RI              SC              SD 
              5               1               3              10              11 
             TN              TX              UT              VA              VT 
             38             308               9              18               1 
             WA              WI              WV              WY 
            147              46               1               2 
```

Let us create a custom function to combine rare categories. The function again loops through all the columns in the dataframe, reads their types, and creates a table of counts for each level of the factor/categorical variables. All levels with counts less than the `mincount` are combined into "other." The function combines rare categories into "Other."+the name of the original variable (for example, `Other.State`). This function has two arguments:

- the name of the dataframe; and
- the count of observations in a category to define "rare."

In [ ]:

```
combinerarecategories<-function(data_frame,mincount){
  for (i in 1:ncol(data_frame)) {
    a<-data_frame[,i]
    replace <- names(which(table(a) < mincount))
    levels(a)[levels(a) %in% replace] <-
      paste("Other", colnames(data_frame)[i], sep=".")
    data_frame[,i]<-a
  }
  return(data_frame)
}
```

Let us combine categories with < 10 values in `STCdata` into "Other." Ultimately, it is going to depend on the person doing the analysis on what they decide to call ``rare''.

In [ ]:

```
STCdata_A<-combinerarecategories(STCdata_A,10)
```

Let us look at `Group.State` again.

In [ ]:

```
table(STCdata_A$Group.State)
```

```
Other.Group.State                AL                 AR                 AZ
               82                21                 10                 53
               CA                CO                 CT                 FL
              718                89                 15                 62
               GA                IA                 ID                 IL
               22                35                 14                104
               IN                KS                 KY                 LA
               43                26                 16                 31
               MA                MD                 MI                 MN
               36                15                 71                 51
               MO                NC                 NE                 NM
               43                16                 42                 20
               NV                NY                 OH                 OK
               20                19                 53                 33
               OR                SC                 SD                 TN
               51                10                 11                 38
               TX                VA                 WA                 WI
              308                18                147                 46
```

You can investigate other columns to see if everything looks fine.

Split the data into training and testing sets

This is a very important step, both conceptually and technically. Conceptually, because the goal of predictive modeling is not to build a model that fits well the data it trains on, but rather one that would best predict the new data. A test set is in this sense the best representation of what the "new data" may look like. Technically, to facilitate comparison between different models, we need to maintain the same IDs in the corresponding sets at all times. We will accomplishes this through two "tricks":

- a random seed ensures that the random-number generator is initialized identically in each run; and
- the `inTrain` vector is created once and can then be applied anytime the data needs to be split.

By default, the code sets 500 data points in the test set, and the remainder 1,889 into the training set.

In [ ]:

```
# set a random number generation seed to
# ensure that the split is the same every time
set.seed(233)

inTrain <- createDataPartition(
  y = STCdata_A$Retained.in.2012.,
  p = 1888/2389,
  list = FALSE)
df.train <- STCdata_A[ inTrain,]
df.test <- STCdata_A[ -inTrain, ]
```

Let us check that both the training and test sets have a similar proportion of positive and negative cases.

In [ ]:

```
print('Training set proportion:')
table(df.train$Retained.in.2012.) / nrow(df.train)
print('Test set proportion:')
table(df.test$Retained.in.2012.) / nrow(df.test)
```

```
[1] "Training set proportion:"
        0         1
0.3928004 0.6071996
[1] "Test set proportion:"
    0     1
0.392 0.608
```

Fitting a logistic regression model

Let us fit a logistic regression model with all the variables included on the training set.

In [ ]:

```
lgfit.all <- glm(Retained.in.2012.~ .,
                 data=df.train,
                 family="binomial")
summary(lgfit.all)
```

```
Warning message:
"glm.fit: fitted probabilities numerically 0 or 1 occurred"
```

```
Call:
glm(formula = Retained.in.2012. ~ ., family = "binomial", data = df.train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7206  -0.5092   0.2285   0.5545   3.1577  

Coefficients: (44 not defined because of singularities)
                                                                   Estimate
(Intercept)                                                      -1.699e+02
Program.CodeCD                                                    7.495e-01
Program.CodeOther.Program.Code                                    4.205e-01
Program.CodeHC                                                    2.585e-01
Program.CodeHD                                                    2.816e-01
Program.CodeHG                                                   -8.370e-01
Program.CodeHN                                                    5.577e-01
Program.CodeHO                                                    4.023e-01
Program.CodeHS                                                   -9.511e-01
Program.CodeHVP                                                  -7.901e-02
Program.CodeIC                                                   -2.083e+01
Program.CodeSC                                                   -4.916e-01
Program.CodeSG                                                    1.934e+01
Program.CodeSK                                                    1.633e+01
Program.CodeSM                                                    1.672e+01
Program.CodeST                                                    1.757e+01
From.Grade11                                                      1.542e+00
From.Grade12                                                     -2.486e+00
From.GradeOther.From.Grade                                       -1.280e+01
From.Grade4                                                       3.236e+00
From.Grade5                                                       2.445e+00
From.Grade6                                                       2.174e+00
From.Grade7                                                       2.343e+00
From.Grade8                                                       2.767e+00
From.Grade9                                                       2.810e-01
From.GradeFIXED_NA                                                2.145e+00
To.Grade11                                                        3.170e-01
To.Grade12                                                       -2.254e-01
To.GradeOther.To.Grade                                            2.718e+01
To.Grade4                                                        -1.836e+00
To.Grade5                                                        -1.816e+00
To.Grade6                                                        -1.523e+00
To.Grade7                                                        -1.094e+00
To.Grade8                                                        -2.053e+00
To.Grade9                                                         3.513e-01
To.GradeFIXED_NA                                                 -1.078e+00
Group.StateAL                                                    -2.477e-01
Group.StateAR                                                    -1.034e+00
Group.StateAZ                                                    -2.317e-01
Group.StateCA                                                     6.674e-01
Group.StateCO                                                     1.003e-01
Group.StateCT                                                    -1.273e-01
Group.StateFL                                                     1.053e-01
Group.StateGA                                                    -1.002e+00
Group.StateIA                                                     2.276e-01
Group.StateID                                                    -1.946e+00
Group.StateIL                                                     2.166e-01
Group.StateIN                                                    -1.502e+00
Group.StateKS                                                    -4.659e-01
Group.StateKY                                                    -5.784e-01
Group.StateLA                                                    -8.454e-01
Group.StateMA                                                    -6.635e-01
Group.StateMD                                                    -2.134e+00
Group.StateMI                                                    -4.080e-01
Group.StateMN                                                     5.950e-01
Group.StateMO                                                    -9.632e-02
Group.StateNC                                                    -7.515e-01
Group.StateNE                                                     6.577e-01
Group.StateNM                                                     1.799e-01
Group.StateNV                                                     1.357e+00
Group.StateNY                                                    -5.341e-01
Group.StateOH                                                    -9.493e-01
Group.StateOK                                                    -6.534e-02
Group.StateOR                                                     1.818e-01
Group.StateSC                                                    -3.473e+00
Group.StateSD                                                     9.883e-01
Group.StateTN                                                    -6.390e-01
Group.StateTX                                                     7.316e-01
Group.StateVA                                                     1.879e+00
Group.StateWA                                                    -3.834e-01
Group.StateWI                                                     3.284e-01
Is.Non.Annual.1                                                  -2.936e+00
Days2                                                            -1.337e-01
Days3                                                            -2.684e-01
Days4                                                            -5.705e-01
Days5                                                            -1.257e+00
Days6                                                            -1.015e+00
Days7                                                            -1.938e+00
Days8                                                            -2.652e+00
DaysOther.Days                                                   -4.741e+00
Days11                                                            1.339e+01
Travel.TypeB                                                      2.628e-01
Travel.TypeOther.Travel.Type                                      1.651e-01
Departure.Date                                                   -4.044e-01
```

```
Return.Date                                                        4.154e-01
Deposit.Date                                                     -3.985e-03
Special.PayCP                                                      2.099e+01
Special.PayFIXED_NA                                                1.898e+01
Special.PayFR                                                      1.897e+01
Special.PaySA                                                      2.004e+01
Tuition                                                           -3.349e-04
FRP.Active                                                         2.981e-02
FRP.Cancelled                                                     -3.807e-02
FRP.Take.up.percent.                                              -1.823e-02
Early.RPL                                                         -1.035e-03
Latest.RPL                                                         1.388e-03
Cancelled.Pax                                                      3.243e-02
Total.Discount.Pax                                                 8.370e-02
Initial.System.Date                                                1.078e-03
Poverty.CodeOther.Poverty.Code                                    -1.444e+00
Poverty.CodeA                                                     -7.845e-01
Poverty.CodeB                                                     -9.065e-01
Poverty.CodeC                                                     -9.367e-01
Poverty.CodeD                                                     -1.510e+00
Poverty.CodeE                                                      3.821e-01
RegionHouston                                                     -3.970e-01
RegionNorthern California                                         -5.908e-01
RegionOther                                                               NA
RegionPacific Northwest                                                   NA
RegionSouthern California                                                 NA
CRM.Segment10                                                      1.104e+00
CRM.Segment11                                                      1.468e+00
CRM.Segment2                                                       2.461e-01
CRM.Segment3                                                       8.149e-01
CRM.Segment4                                                       3.399e+00
CRM.Segment5                                                       1.078e+00
CRM.Segment6                                                       2.439e+00
CRM.Segment7                                                      -1.231e-01
CRM.Segment8                                                      -4.386e-01
CRM.SegmentOther.CRM.Segment                                      -1.558e+00
School.TypeCatholic                                               -7.301e-02
School.TypePUBLIC                                                  2.718e-01
School.TypePrivate non-Christian                                   8.510e-01
Parent.Meeting.Flag1                                               3.494e+01
MDR.Low.GradeOther.MDR.Low.Grade                                   9.882e+00
MDR.Low.Grade3                                                     1.100e+01
MDR.Low.Grade4                                                     9.286e+00
MDR.Low.Grade5                                                     9.826e+00
MDR.Low.Grade6                                                     1.098e+01
MDR.Low.Grade7                                                     1.112e+01
MDR.Low.Grade8                                                     1.115e+01
MDR.Low.Grade9                                                     5.388e+00
MDR.Low.GradeK                                                     1.078e+01
MDR.Low.GradePK                                                    1.070e+01
MDR.High.Grade12                                                  -8.651e-01
MDR.High.Grade5                                                   -1.514e+01
MDR.High.Grade6                                                   -1.615e+01
MDR.High.Grade7                                                   -1.311e+01
MDR.High.Grade8                                                   -1.295e+01
MDR.High.Grade9                                                   -4.093e-01
MDR.High.GradeFIXED_NA                                                    NA
Total.School.Enrollment                                           -2.455e-05
Income.LevelA                                                      6.971e-01
Income.LevelB                                                      1.139e+00
Income.LevelC                                                      4.301e-01
Income.LevelD                                                      4.618e-01
Income.LevelE                                                      4.157e-01
Income.LevelF                                                      6.435e-01
Income.LevelG                                                      1.681e-01
Income.LevelH                                                      7.329e-01
Income.LevelI                                                      3.560e-01
Income.LevelJ                                                      7.167e-02
Income.LevelK                                                      7.383e-01
Income.LevelL                                                      3.106e-01
Income.LevelM                                                      4.103e-01
Income.LevelN                                                      1.388e-01
Income.LevelO                                                      6.390e-02
Income.LevelP                                                      1.189e+00
Income.LevelOther.Income.Level                                     7.638e-01
Income.LevelQ                                                     -3.348e-01
Income.LevelZ                                                      8.476e-01
EZ.Pay.Take.Up.Rate                                                1.040e-01
School.Sponsor1                                                   -1.133e-01
SPR.Product.TypeCosta Rica                                                NA
SPR.Product.TypeEast Coast                                         1.679e-02
SPR.Product.TypeOther.SPR.Product.Type                            -5.707e-01
SPR.Product.TypeInternational                                      6.952e+00
SPR.Product.TypeScience                                           -1.693e+01
SPR.New.ExistingNEW                                               -1.706e+00
FPP                                                               -6.908e-03
Total.Pax                                                                 NA
SPR.Group.Revenue                                                         NA
NumberOfMeetingswithParents1                                      -9.165e-02
NumberOfMeetingswithParents2                                              NA
FirstMeeting                                                      -1.317e-04
LastMeeting                                                       -7.191e-04
DifferenceTraveltoFirstMeeting                                    -1.460e-03
```

```
DifferenceTraveltoLastMeeting                                            NA
SchoolGradeTypeLowHigh                                                   NA
SchoolGradeTypeLowMiddle                                                 NA
SchoolGradeTypeLowUndefined                                              NA
SchoolGradeTypeHighHigh                                                  NA
SchoolGradeTypeHighMiddle                                                NA
SchoolGradeTypeHighUndefined                                             NA
SchoolGradeTypeOther.SchoolGradeType                             -1.978e-01
SchoolGradeTypeElementary->Middle                                        NA
SchoolGradeTypeHigh->High                                                NA
SchoolGradeTypeMiddle->High                                      -2.318e+00
SchoolGradeTypeMiddle->Middle                                            NA
SchoolGradeTypeMiddle->Undefined                                         NA
SchoolGradeTypeUndefined->Undefined                                      NA
DepartureMonthFebruary                                            2.289e+00
DepartureMonthOther.DepartureMonth                                9.903e-02
DepartureMonthJune                                               -6.427e-01
DepartureMonthMarch                                               6.314e-01
DepartureMonthMay                                                 3.860e-03
GroupGradeTypeLowHigh                                             6.263e+00
GroupGradeTypeLowK                                                       NA
GroupGradeTypeLowMiddle                                                  NA
GroupGradeTypeLowPK                                                      NA
GroupGradeTypeLowUndefined                                               NA
GroupGradeTypeHighHigh                                           -1.316e+01
GroupGradeTypeHighMiddle                                                 NA
GroupGradeTypeHighUndefined                                              NA
GroupGradeTypeElementary->Middle                                  1.498e+00
GroupGradeTypeHigh->High                                                 NA
GroupGradeTypeK->Elementary                                       1.429e-01
GroupGradeTypeK->High                                            -5.283e-01
GroupGradeTypeK->Middle                                                  NA
GroupGradeTypeMiddle->High                                        3.706e-01
GroupGradeTypeMiddle->Middle                                             NA
GroupGradeTypePK->Elementary                                     -5.049e-01
GroupGradeTypePK->High                                                   NA
GroupGradeTypePK->Middle                                                 NA
GroupGradeTypeUndefined->Undefined                                       NA
MajorProgramCodeH                                                -3.836e-01
MajorProgramCodeI                                                -1.696e+01
MajorProgramCodeS                                                        NA
SingleGradeTripFlag1                                              8.017e-01
FPP.to.School.enrollment                                          8.266e-01
FPP.to.PAX                                                        1.868e+00
Num.of.Non_FPP.PAX                                                       NA
SchoolSizeIndicatorL                                              1.552e+00
SchoolSizeIndicatorM-L                                            1.301e+00
SchoolSizeIndicatorS                                              5.591e-01
SchoolSizeIndicatorS-M                                            1.634e+00
From.Grade_surrogate1                                                    NA
To.Grade_surrogate1                                                      NA
Special.Pay_surrogate1                                                   NA
Early.RPL_surrogate1                                             -4.157e+01
Latest.RPL_surrogate1                                             5.639e+01
Initial.System.Date_surrogateOther.Initial.System.Date_surrogate  4.294e+01
CRM.Segment_surrogateOther.CRM.Segment_surrogate                  5.486e-01
MDR.High.Grade_surrogate1                                                NA
Total.School.Enrollment_surrogate1                                       NA
FirstMeeting_surrogate1                                                  NA
LastMeeting_surrogate1                                                   NA
DifferenceTraveltoFirstMeeting_surrogate1                                NA
DifferenceTraveltoLastMeeting_surrogate1                                 NA
FPP.to.School.enrollment_surrogate1                                      NA
                                                                 Std. Error
(Intercept)                                                       1.307e+03
Program.CodeCD                                                    1.070e+00
Program.CodeOther.Program.Code                                    1.456e+00
Program.CodeHC                                                    1.702e+00
Program.CodeHD                                                    1.696e+00
Program.CodeHG                                                    2.096e+00
Program.CodeHN                                                    1.763e+00
Program.CodeHO                                                    2.065e+00
Program.CodeHS                                                    1.957e+00
Program.CodeHVP                                                   1.788e+00
Program.CodeIC                                                    2.693e+03
Program.CodeSC                                                    1.996e+00
Program.CodeSG                                                    1.283e+03
Program.CodeSK                                                    1.283e+03
Program.CodeSM                                                    1.283e+03
Program.CodeST                                                    1.283e+03
From.Grade11                                                      1.154e+00
From.Grade12                                                      1.613e+00
From.GradeOther.From.Grade                                        1.098e+03
From.Grade4                                                       3.172e+00
From.Grade5                                                       2.917e+00
From.Grade6                                                       2.818e+00
From.Grade7                                                       2.814e+00
From.Grade8                                                       2.779e+00
From.Grade9                                                       1.046e+00
From.GradeFIXED_NA                                                2.664e+00
To.Grade11                                                        1.253e+00
To.Grade12                                                        9.853e-01
To.GradeOther.To.Grade                                            2.639e+03
To.Grade4                                                         3.243e+00
To.Grade5                                                         3.035e+00
To.Grade6                                                         2.903e+00
To.Grade7                                                         2.803e+00
To.Grade8                                                         2.742e+00
To.Grade9                                                         9.735e-01
To.GradeFIXED_NA                                                  2.606e+00
Group.StateAL                                                     9.085e-01
Group.StateAR                                                     1.191e+00
Group.StateAZ                                                     6.455e-01
Group.StateCA                                                     4.650e-01
Group.StateCO                                                     5.293e-01
Group.StateCT                                                     1.224e+00
Group.StateFL                                                     6.149e-01
Group.StateGA                                                     8.267e-01
Group.StateIA                                                     7.929e-01
Group.StateID                                                     1.330e+00
Group.StateIL                                                     5.819e-01
Group.StateIN                                                     6.923e-01
Group.StateKS                                                     8.382e-01
Group.StateKY                                                     1.390e+00
Group.StateLA                                                     7.268e-01
Group.StateMA                                                     7.948e-01
Group.StateMD                                                     9.803e-01
Group.StateMI                                                     6.477e-01
Group.StateMN                                                     7.243e-01
Group.StateMO                                                     6.553e-01
Group.StateNC                                                     9.003e-01
Group.StateNE                                                     7.659e-01
Group.StateNM                                                     8.827e-01
Group.StateNV                                                     8.468e-01
Group.StateNY                                                     9.610e-01
Group.StateOH                                                     6.946e-01
Group.StateOK                                                     7.255e-01
Group.StateOR                                                     6.281e-01
Group.StateSC                                                     1.213e+00
Group.StateSD                                                     9.473e-01
Group.StateTN                                                     7.194e-01
Group.StateTX                                                     4.907e-01
Group.StateVA                                                     1.479e+00
Group.StateWA                                                     4.978e-01
Group.StateWI                                                     7.301e-01
Is.Non.Annual.1                                                   2.487e-01
Days2                                                             6.526e-01
Days3                                                             1.114e+00
Days4                                                             1.203e+00
Days5                                                             1.303e+00
Days6                                                             1.453e+00
Days7                                                             1.653e+00
Days8                                                             1.875e+00
DaysOther.Days                                                    2.375e+00
Days11                                                            9.916e+02
Travel.TypeB                                                      5.016e-01
Travel.TypeOther.Travel.Type                                      1.184e+00
Departure.Date                                                    2.188e-01
Return.Date                                                       2.183e-01
Deposit.Date                                                      2.422e-03
Special.PayCP                                                     1.276e+03
Special.PayFIXED_NA                                               1.276e+03
Special.PayFR                                                     1.276e+03
Special.PaySA                                                     1.276e+03
Tuition                                                           4.845e-04
FRP.Active                                                        1.379e-02
FRP.Cancelled                                                     4.269e-02
FRP.Take.up.percent.                                              4.648e-01
Early.RPL                                                         2.632e-03
Latest.RPL                                                        1.640e-03
Cancelled.Pax                                                     3.334e-02
Total.Discount.Pax                                                7.863e-02
Initial.System.Date                                               2.081e-03
Poverty.CodeOther.Poverty.Code                                    1.402e+00
Poverty.CodeA                                                     6.718e-01
Poverty.CodeB                                                     6.246e-01
Poverty.CodeC                                                     6.340e-01
Poverty.CodeD                                                     8.452e-01
Poverty.CodeE                                                     1.116e+00
RegionHouston                                                     3.919e-01
RegionNorthern California                                         2.990e-01
RegionOther                                                              NA
RegionPacific Northwest                                                  NA
RegionSouthern California                                                NA
CRM.Segment10                                                     4.365e-01
CRM.Segment11                                                     1.409e+00
CRM.Segment2                                                      6.929e-01
CRM.Segment3                                                      1.152e+00
CRM.Segment4                                                      8.429e-01
CRM.Segment5                                                      4.699e-01
CRM.Segment6                                                      9.071e-01
CRM.Segment7                                                      9.013e-01
CRM.Segment8                                                      9.851e-01
CRM.SegmentOther.CRM.Segment                                      1.139e+00
School.TypeCatholic                                               4.634e-01
School.TypePUBLIC                                                 5.429e-01
School.TypePrivate non-Christian                                  5.022e-01
Parent.Meeting.Flag1                                              2.563e+02
MDR.Low.GradeOther.MDR.Low.Grade                                  9.235e+02
MDR.Low.Grade3                                                    9.235e+02
MDR.Low.Grade4                                                    9.235e+02
MDR.Low.Grade5                                                    9.235e+02
MDR.Low.Grade6                                                    9.235e+02
MDR.Low.Grade7                                                    9.235e+02
MDR.Low.Grade8                                                    9.235e+02
MDR.Low.Grade9                                                    9.235e+02
MDR.Low.GradeK                                                    9.235e+02
MDR.Low.GradePK                                                   9.235e+02
MDR.High.Grade12                                                  1.694e+00
MDR.High.Grade5                                                   9.235e+02
MDR.High.Grade6                                                   9.235e+02
MDR.High.Grade7                                                   9.235e+02
MDR.High.Grade8                                                   9.235e+02
MDR.High.Grade9                                                   1.778e+00
MDR.High.GradeFIXED_NA                                                   NA
Total.School.Enrollment                                           3.737e-04
Income.LevelA                                                     1.786e+00
Income.LevelB                                                     1.671e+00
Income.LevelC                                                     1.601e+00
Income.LevelD                                                     1.620e+00
Income.LevelE                                                     1.591e+00
Income.LevelF                                                     1.591e+00
Income.LevelG                                                     1.600e+00
Income.LevelH                                                     1.572e+00
Income.LevelI                                                     1.569e+00
Income.LevelJ                                                     1.572e+00
Income.LevelK                                                     1.577e+00
Income.LevelL                                                     1.571e+00
Income.LevelM                                                     1.568e+00
Income.LevelN                                                     1.575e+00
Income.LevelO                                                     1.559e+00
Income.LevelP                                                     1.562e+00
Income.LevelOther.Income.Level                                    2.503e+00
Income.LevelQ                                                     1.570e+00
Income.LevelZ                                                     1.695e+00
EZ.Pay.Take.Up.Rate                                               5.187e-01
School.Sponsor1                                                   3.579e-01
SPR.Product.TypeCosta Rica                                               NA
SPR.Product.TypeEast Coast                                        1.429e+00
SPR.Product.TypeOther.SPR.Product.Type                            1.781e+00
SPR.Product.TypeInternational                                     3.393e+03
SPR.Product.TypeScience                                           1.283e+03
SPR.New.ExistingNEW                                               1.945e-01
FPP                                                               9.622e-03
Total.Pax                                                                NA
SPR.Group.Revenue                                                        NA
NumberOfMeetingswithParents1                                      3.056e-01
NumberOfMeetingswithParents2                                             NA
FirstMeeting                                                      6.526e-03
LastMeeting                                                       2.690e-03
DifferenceTraveltoFirstMeeting                                    6.132e-03
DifferenceTraveltoLastMeeting                                            NA
SchoolGradeTypeLowHigh                                                   NA
SchoolGradeTypeLowMiddle                                                 NA
SchoolGradeTypeLowUndefined                                              NA
SchoolGradeTypeHighHigh                                                  NA
SchoolGradeTypeHighMiddle                                                NA
SchoolGradeTypeHighUndefined                                             NA
SchoolGradeTypeOther.SchoolGradeType                              2.126e+00
SchoolGradeTypeElementary->Middle                                        NA
SchoolGradeTypeHigh->High                                                NA
SchoolGradeTypeMiddle->High                                       2.604e+00
SchoolGradeTypeMiddle->Middle                                            NA
SchoolGradeTypeMiddle->Undefined                                         NA
SchoolGradeTypeUndefined->Undefined                                      NA
DepartureMonthFebruary                                            1.105e+00
DepartureMonthOther.DepartureMonth                                1.434e+00
DepartureMonthJune                                                5.909e-01
DepartureMonthMarch                                               3.535e-01
DepartureMonthMay                                                 4.309e-01
GroupGradeTypeLowHigh                                             3.167e+00
GroupGradeTypeLowK                                                       NA
GroupGradeTypeLowMiddle                                                  NA
GroupGradeTypeLowPK                                                      NA
GroupGradeTypeLowUndefined                                               NA
GroupGradeTypeHighHigh                                            9.235e+02
GroupGradeTypeHighMiddle                                                 NA
GroupGradeTypeHighUndefined                                              NA
GroupGradeTypeElementary->Middle                                  1.806e+00
GroupGradeTypeHigh->High                                                 NA
GroupGradeTypeK->Elementary                                       2.655e+00
GroupGradeTypeK->High                                             7.873e-01
GroupGradeTypeK->Middle                                                  NA
GroupGradeTypeMiddle->High                                        6.303e-01
GroupGradeTypeMiddle->Middle                                             NA
GroupGradeTypePK->Elementary                                      2.663e+00
GroupGradeTypePK->High                                                   NA
GroupGradeTypePK->Middle                                                 NA
GroupGradeTypeUndefined->Undefined                                       NA
```

```
MajorProgramCodeH                                                 1.425e+00
MajorProgramCodeI                                                 2.400e+03
MajorProgramCodeS                                                        NA
SingleGradeTripFlag1                                              4.350e-01
FPP.to.School.enrollment                                          1.380e+00
FPP.to.PAX                                                        1.915e+00
Num.of.Non_FPP.PAX                                                       NA
SchoolSizeIndicatorL                                              7.783e-01
SchoolSizeIndicatorM-L                                            7.121e-01
SchoolSizeIndicatorS                                              6.801e-01
SchoolSizeIndicatorS-M                                            6.856e-01
From.Grade_surrogate1                                                    NA
To.Grade_surrogate1                                                      NA
Special.Pay_surrogate1                                                   NA
Early.RPL_surrogate1                                              1.062e+02
Latest.RPL_surrogate1                                             6.638e+01
Initial.System.Date_surrogateOther.Initial.System.Date_surrogate  8.379e+01
CRM.Segment_surrogateOther.CRM.Segment_surrogate                  1.774e+00
MDR.High.Grade_surrogate1                                                NA
Total.School.Enrollment_surrogate1                                       NA
FirstMeeting_surrogate1                                                  NA
LastMeeting_surrogate1                                                   NA
DifferenceTraveltoFirstMeeting_surrogate1                                NA
DifferenceTraveltoLastMeeting_surrogate1                                 NA
FPP.to.School.enrollment_surrogate1                                      NA
                                                                 z value
(Intercept)                                                       -0.130
Program.CodeCD                                                     0.701
Program.CodeOther.Program.Code                                     0.289
Program.CodeHC                                                     0.152
Program.CodeHD                                                     0.166
Program.CodeHG                                                    -0.399
Program.CodeHN                                                     0.316
Program.CodeHO                                                     0.195
Program.CodeHS                                                    -0.486
Program.CodeHVP                                                   -0.044
Program.CodeIC                                                    -0.008
Program.CodeSC                                                    -0.246
Program.CodeSG                                                     0.015
Program.CodeSK                                                     0.013
Program.CodeSM                                                     0.013
Program.CodeST                                                     0.014
From.Grade11                                                       1.336
From.Grade12                                                      -1.541
From.GradeOther.From.Grade                                        -0.012
From.Grade4                                                        1.020
From.Grade5                                                        0.838
From.Grade6                                                        0.771
From.Grade7                                                        0.833
From.Grade8                                                        0.996
From.Grade9                                                        0.269
From.GradeFIXED_NA                                                 0.805
To.Grade11                                                         0.253
To.Grade12                                                        -0.229
To.GradeOther.To.Grade                                             0.010
To.Grade4                                                         -0.566
To.Grade5                                                         -0.598
To.Grade6                                                         -0.525
To.Grade7                                                         -0.390
To.Grade8                                                         -0.749
To.Grade9                                                          0.361
To.GradeFIXED_NA                                                  -0.413
Group.StateAL                                                     -0.273
Group.StateAR                                                     -0.868
Group.StateAZ                                                     -0.359
Group.StateCA                                                      1.435
Group.StateCO                                                      0.190
Group.StateCT                                                     -0.104
Group.StateFL                                                      0.171
Group.StateGA                                                     -1.212
Group.StateIA                                                      0.287
Group.StateID                                                     -1.463
Group.StateIL                                                      0.372
Group.StateIN                                                     -2.170
Group.StateKS                                                     -0.556
Group.StateKY                                                     -0.416
Group.StateLA                                                     -1.163
Group.StateMA                                                     -0.835
Group.StateMD                                                     -2.177
Group.StateMI                                                     -0.630
Group.StateMN                                                      0.822
Group.StateMO                                                     -0.147
Group.StateNC                                                     -0.835
Group.StateNE                                                      0.859
Group.StateNM                                                      0.204
Group.StateNV                                                      1.603
Group.StateNY                                                     -0.556
Group.StateOH                                                     -1.367
Group.StateOK                                                     -0.090
Group.StateOR                                                      0.290
Group.StateSC                                                     -2.863
Group.StateSD                                                      1.043
Group.StateTN                                                     -0.888
Group.StateTX                                                      1.491
Group.StateVA                                                      1.271
Group.StateWA                                                     -0.770
Group.StateWI                                                      0.450
Is.Non.Annual.1                                                  -11.806
Days2                                                             -0.205
Days3                                                             -0.241
Days4                                                             -0.474
Days5                                                             -0.965
Days6                                                             -0.699
Days7                                                             -1.173
Days8                                                             -1.414
DaysOther.Days                                                    -1.996
Days11                                                             0.014
Travel.TypeB                                                       0.524
Travel.TypeOther.Travel.Type                                       0.139
Departure.Date                                                    -1.848
Return.Date                                                        1.903
Deposit.Date                                                      -1.645
Special.PayCP                                                      0.016
Special.PayFIXED_NA                                                0.015
Special.PayFR                                                      0.015
Special.PaySA                                                      0.016
Tuition                                                           -0.691
FRP.Active                                                         2.162
FRP.Cancelled                                                     -0.892
FRP.Take.up.percent.                                              -0.039
Early.RPL                                                         -0.393
Latest.RPL                                                         0.846
Cancelled.Pax                                                      0.973
Total.Discount.Pax                                                 1.064
Initial.System.Date                                                0.518
Poverty.CodeOther.Poverty.Code                                    -1.030
Poverty.CodeA                                                     -1.168
Poverty.CodeB                                                     -1.451
Poverty.CodeC                                                     -1.477
Poverty.CodeD                                                     -1.787
Poverty.CodeE                                                      0.342
RegionHouston                                                     -1.013
RegionNorthern California                                         -1.976
RegionOther                                                           NA
RegionPacific Northwest                                               NA
RegionSouthern California                                             NA
CRM.Segment10                                                      2.530
CRM.Segment11                                                      1.042
CRM.Segment2                                                       0.355
CRM.Segment3                                                       0.708
CRM.Segment4                                                       4.032
CRM.Segment5                                                       2.295
CRM.Segment6                                                       2.689
CRM.Segment7                                                      -0.137
CRM.Segment8                                                      -0.445
CRM.SegmentOther.CRM.Segment                                      -1.369
School.TypeCatholic                                               -0.158
School.TypePUBLIC                                                  0.501
School.TypePrivate non-Christian                                   1.695
Parent.Meeting.Flag1                                               0.136
MDR.Low.GradeOther.MDR.Low.Grade                                   0.011
MDR.Low.Grade3                                                     0.012
MDR.Low.Grade4                                                     0.010
MDR.Low.Grade5                                                     0.011
MDR.Low.Grade6                                                     0.012
MDR.Low.Grade7                                                     0.012
MDR.Low.Grade8                                                     0.012
MDR.Low.Grade9                                                     0.006
MDR.Low.GradeK                                                     0.012
MDR.Low.GradePK                                                    0.012
MDR.High.Grade12                                                  -0.511
MDR.High.Grade5                                                   -0.016
MDR.High.Grade6                                                   -0.017
MDR.High.Grade7                                                   -0.014
MDR.High.Grade8                                                   -0.014
MDR.High.Grade9                                                   -0.230
MDR.High.GradeFIXED_NA                                                NA
Total.School.Enrollment                                           -0.066
Income.LevelA                                                      0.390
Income.LevelB                                                      0.682
Income.LevelC                                                      0.269
Income.LevelD                                                      0.285
Income.LevelE                                                      0.261
Income.LevelF                                                      0.404
Income.LevelG                                                      0.105
Income.LevelH                                                      0.466
Income.LevelI                                                      0.227
Income.LevelJ                                                      0.046
Income.LevelK                                                      0.468
Income.LevelL                                                      0.198
Income.LevelM                                                      0.262
Income.LevelN                                                      0.088
Income.LevelO                                                      0.041
Income.LevelP                                                      0.761
Income.LevelOther.Income.Level                                     0.305
Income.LevelQ                                                     -0.213
Income.LevelZ                                                      0.500
EZ.Pay.Take.Up.Rate                                                0.200
School.Sponsor1                                                   -0.317
SPR.Product.TypeCosta Rica                                            NA
SPR.Product.TypeEast Coast                                         0.012
SPR.Product.TypeOther.SPR.Product.Type                            -0.320
SPR.Product.TypeInternational                                      0.002
SPR.Product.TypeScience                                           -0.013
SPR.New.ExistingNEW                                               -8.770
FPP                                                               -0.718
Total.Pax                                                             NA
SPR.Group.Revenue                                                     NA
NumberOfMeetingswithParents1                                      -0.300
NumberOfMeetingswithParents2                                          NA
FirstMeeting                                                      -0.020
LastMeeting                                                       -0.267
DifferenceTraveltoFirstMeeting                                    -0.238
DifferenceTraveltoLastMeeting                                         NA
SchoolGradeTypeLowHigh                                                NA
SchoolGradeTypeLowMiddle                                              NA
SchoolGradeTypeLowUndefined                                           NA
SchoolGradeTypeHighHigh                                               NA
SchoolGradeTypeHighMiddle                                             NA
SchoolGradeTypeHighUndefined                                          NA
SchoolGradeTypeOther.SchoolGradeType                              -0.093
SchoolGradeTypeElementary->Middle                                     NA
SchoolGradeTypeHigh->High                                             NA
SchoolGradeTypeMiddle->High                                       -0.890
SchoolGradeTypeMiddle->Middle                                         NA
SchoolGradeTypeMiddle->Undefined                                      NA
SchoolGradeTypeUndefined->Undefined                                   NA
DepartureMonthFebruary                                             2.072
DepartureMonthOther.DepartureMonth                                 0.069
DepartureMonthJune                                                -1.088
DepartureMonthMarch                                                1.786
DepartureMonthMay                                                  0.009
GroupGradeTypeLowHigh                                              1.977
GroupGradeTypeLowK                                                    NA
GroupGradeTypeLowMiddle                                               NA
GroupGradeTypeLowPK                                                   NA
GroupGradeTypeLowUndefined                                            NA
GroupGradeTypeHighHigh                                            -0.014
GroupGradeTypeHighMiddle                                              NA
GroupGradeTypeHighUndefined                                           NA
GroupGradeTypeElementary->Middle                                   0.829
GroupGradeTypeHigh->High                                              NA
GroupGradeTypeK->Elementary                                        0.054
GroupGradeTypeK->High                                             -0.671
GroupGradeTypeK->Middle                                               NA
GroupGradeTypeMiddle->High                                         0.588
GroupGradeTypeMiddle->Middle                                          NA
GroupGradeTypePK->Elementary                                      -0.190
GroupGradeTypePK->High                                                NA
GroupGradeTypePK->Middle                                              NA
GroupGradeTypeUndefined->Undefined                                    NA
MajorProgramCodeH                                                 -0.269
MajorProgramCodeI                                                 -0.007
MajorProgramCodeS                                                     NA
SingleGradeTripFlag1                                               1.843
FPP.to.School.enrollment                                           0.599
FPP.to.PAX                                                         0.976
Num.of.Non_FPP.PAX                                                    NA
SchoolSizeIndicatorL                                               1.994
SchoolSizeIndicatorM-L                                             1.827
SchoolSizeIndicatorS                                               0.822
SchoolSizeIndicatorS-M                                             2.383
From.Grade_surrogate1                                                 NA
To.Grade_surrogate1                                                   NA
Special.Pay_surrogate1                                                NA
Early.RPL_surrogate1                                              -0.392
Latest.RPL_surrogate1                                              0.849
Initial.System.Date_surrogateOther.Initial.System.Date_surrogate   0.512
CRM.Segment_surrogateOther.CRM.Segment_surrogate                   0.309
MDR.High.Grade_surrogate1                                             NA
Total.School.Enrollment_surrogate1                                    NA
FirstMeeting_surrogate1                                               NA
LastMeeting_surrogate1                                                NA
DifferenceTraveltoFirstMeeting_surrogate1                             NA
DifferenceTraveltoLastMeeting_surrogate1                              NA
FPP.to.School.enrollment_surrogate1                                   NA
                                                                 Pr(>|z|)
(Intercept)                                                       0.89651
Program.CodeCD                                                    0.48347
Program.CodeOther.Program.Code                                    0.77274
Program.CodeHC                                                    0.87934
Program.CodeHD                                                    0.86813
Program.CodeHG                                                    0.68960
Program.CodeHN                                                    0.75177
Program.CodeHO                                                    0.84552
Program.CodeHS                                                    0.62703
Program.CodeHVP                                                   0.96475
Program.CodeIC                                                    0.99383
Program.CodeSC                                                    0.80544
Program.CodeSG                                                    0.98798    
Program.CodeSK                                                    0.98984    
Program.CodeSM                                                    0.98960    
Program.CodeST                                                    0.98907    
From.Grade11                                                      0.18167    
From.Grade12                                                      0.12342    
From.GradeOther.From.Grade                                        0.99070    
From.Grade4                                                       0.30755    
From.Grade5                                                       0.40203    
From.Grade6                                                       0.44050    
From.Grade7                                                       0.40503    
From.Grade8                                                       0.31949    
From.Grade9                                                       0.78820    
From.GradeFIXED_NA                                                0.42083    
To.Grade11                                                        0.80035    
To.Grade12                                                        0.81904    
To.GradeOther.To.Grade                                            0.99178    
To.Grade4                                                         0.57140    
To.Grade5                                                         0.54964    
To.Grade6                                                         0.59987    
To.Grade7                                                         0.69627    
To.Grade8                                                         0.45409    
To.Grade9                                                         0.71822    
To.GradeFIXED_NA                                                  0.67924    
Group.StateAL                                                     0.78511    
Group.StateAR                                                     0.38550    
Group.StateAZ                                                     0.71966    
Group.StateCA                                                     0.15121    
Group.StateCO                                                     0.84969    
Group.StateCT                                                     0.91718    
Group.StateFL                                                     0.86401    
Group.StateGA                                                     0.22552    
Group.StateIA                                                     0.77404    
Group.StateID                                                     0.14356    
Group.StateIL                                                     0.70974    
Group.StateIN                                                     0.03004 *  
Group.StateKS                                                     0.57832    
Group.StateKY                                                     0.67740    
Group.StateLA                                                     0.24478    
Group.StateMA                                                     0.40386    
Group.StateMD                                                     0.02951 *  
Group.StateMI                                                     0.52876    
Group.StateMN                                                     0.41134    
Group.StateMO                                                     0.88314    
Group.StateNC                                                     0.40390    
Group.StateNE                                                     0.39048    
Group.StateNM                                                     0.83849    
Group.StateNV                                                     0.10902    
Group.StateNY                                                     0.57842    
Group.StateOH                                                     0.17177    
Group.StateOK                                                     0.92824    
Group.StateOR                                                     0.77218    
Group.StateSC                                                     0.00420 ** 
Group.StateSD                                                     0.29683    
Group.StateTN                                                     0.37440    
Group.StateTX                                                     0.13595    
Group.StateVA                                                     0.20386    
Group.StateWA                                                     0.44109    
Group.StateWI                                                     0.65285    
Is.Non.Annual.1                                                   < 2e-16 ***
Days2                                                             0.83762    
Days3                                                             0.80968    
Days4                                                             0.63546    
Days5                                                             0.33441    
Days6                                                             0.48474    
Days7                                                             0.24091    
Days8                                                             0.15726    
DaysOther.Days                                                    0.04595 *  
Days11                                                            0.98922    
Travel.TypeB                                                      0.60029    
Travel.TypeOther.Travel.Type                                      0.88913    
Departure.Date                                                    0.06457 .  
Return.Date                                                       0.05706 .  
Deposit.Date                                                      0.09996 .  
Special.PayCP                                                     0.98687    
Special.PayFIXED_NA                                               0.98813    
Special.PayFR                                                     0.98814    
Special.PaySA                                                     0.98746    
Tuition                                                           0.48943    
FRP.Active                                                        0.03059 *  
FRP.Cancelled                                                     0.37252    
FRP.Take.up.percent.                                              0.96873    
Early.RPL                                                         0.69411    
Latest.RPL                                                        0.39754    
Cancelled.Pax                                                     0.33075    
Total.Discount.Pax                                                0.28710    
Initial.System.Date                                               0.60440    
Poverty.CodeOther.Poverty.Code                                    0.30321    
Poverty.CodeA                                                     0.24291    
Poverty.CodeB                                                     0.14667    
Poverty.CodeC                                                     0.13956    
Poverty.CodeD                                                     0.07393 .  
Poverty.CodeE                                                     0.73206    
```

```
RegionHouston                                                     0.31104
RegionNorthern California                                         0.04817 *
RegionOther                                                            NA
RegionPacific Northwest                                                NA
RegionSouthern California                                              NA
CRM.Segment10                                                     0.01142 *
CRM.Segment11                                                     0.29740
CRM.Segment2                                                      0.72249
CRM.Segment3                                                      0.47923
CRM.Segment4                                                     5.52e-05 ***
CRM.Segment5                                                      0.02174 *
CRM.Segment6                                                      0.00717 **
CRM.Segment7                                                      0.89139
CRM.Segment8                                                      0.65618
CRM.SegmentOther.CRM.Segment                                      0.17115
School.TypeCatholic                                               0.87481
School.TypePUBLIC                                                 0.61660
School.TypePrivate non-Christian                                  0.09015 .
Parent.Meeting.Flag1                                              0.89158
MDR.Low.GradeOther.MDR.Low.Grade                                  0.99146
MDR.Low.Grade3                                                    0.99049
MDR.Low.Grade4                                                    0.99198
MDR.Low.Grade5                                                    0.99151
MDR.Low.Grade6                                                    0.99052
MDR.Low.Grade7                                                    0.99039
MDR.Low.Grade8                                                    0.99037
MDR.Low.Grade9                                                    0.99535
MDR.Low.GradeK                                                    0.99068
MDR.Low.GradePK                                                   0.99076
MDR.High.Grade12                                                  0.60958
MDR.High.Grade5                                                   0.98692
MDR.High.Grade6                                                   0.98605
MDR.High.Grade7                                                   0.98867
MDR.High.Grade8                                                   0.98881
MDR.High.Grade9                                                   0.81789
MDR.High.GradeFIXED_NA                                                 NA
Total.School.Enrollment                                           0.94762
Income.LevelA                                                     0.69630
Income.LevelB                                                     0.49551
Income.LevelC                                                     0.78819
Income.LevelD                                                     0.77555
Income.LevelE                                                     0.79394
Income.LevelF                                                     0.68592
Income.LevelG                                                     0.91635
Income.LevelH                                                     0.64096
Income.LevelI                                                     0.82049
Income.LevelJ                                                     0.96364
Income.LevelK                                                     0.63974
Income.LevelL                                                     0.84328
Income.LevelM                                                     0.79363
Income.LevelN                                                     0.92978
Income.LevelO                                                     0.96731
Income.LevelP                                                     0.44652
Income.LevelOther.Income.Level                                    0.76025
Income.LevelQ                                                     0.83114
Income.LevelZ                                                     0.61706
EZ.Pay.Take.Up.Rate                                               0.84115
School.Sponsor1                                                   0.75159
SPR.Product.TypeCosta Rica                                             NA
SPR.Product.TypeEast Coast                                        0.99063
SPR.Product.TypeOther.SPR.Product.Type                            0.74861
SPR.Product.TypeInternational                                     0.99837
SPR.Product.TypeScience                                           0.98947
SPR.New.ExistingNEW                                               < 2e-16 ***
FPP                                                               0.47281
Total.Pax                                                              NA
SPR.Group.Revenue                                                      NA
NumberOfMeetingswithParents1                                      0.76427
NumberOfMeetingswithParents2                                           NA
FirstMeeting                                                      0.98390
LastMeeting                                                       0.78924
DifferenceTraveltoFirstMeeting                                    0.81179
DifferenceTraveltoLastMeeting                                          NA
SchoolGradeTypeLowHigh                                                 NA
SchoolGradeTypeLowMiddle                                               NA
SchoolGradeTypeLowUndefined                                            NA
SchoolGradeTypeHighHigh                                                NA
SchoolGradeTypeHighMiddle                                              NA
SchoolGradeTypeHighUndefined                                           NA
SchoolGradeTypeOther.SchoolGradeType                              0.92586
SchoolGradeTypeElementary->Middle                                      NA
SchoolGradeTypeHigh->High                                              NA
SchoolGradeTypeMiddle->High                                       0.37341
SchoolGradeTypeMiddle->Middle                                          NA
SchoolGradeTypeMiddle->Undefined                                       NA
SchoolGradeTypeUndefined->Undefined                                    NA
DepartureMonthFebruary                                            0.03825 *
DepartureMonthOther.DepartureMonth                                0.94495
DepartureMonthJune                                                0.27674
DepartureMonthMarch                                               0.07409 .
DepartureMonthMay                                                 0.99285
GroupGradeTypeLowHigh                                             0.04799 *
GroupGradeTypeLowK                                                     NA
```

```
GroupGradeTypeLowMiddle                                                NA
GroupGradeTypeLowPK                                                    NA
GroupGradeTypeLowUndefined                                             NA
GroupGradeTypeHighHigh                                            0.98863
GroupGradeTypeHighMiddle                                               NA
GroupGradeTypeHighUndefined                                            NA
GroupGradeTypeElementary->Middle                                  0.40690
GroupGradeTypeHigh->High                                               NA
GroupGradeTypeK->Elementary                                       0.95708
GroupGradeTypeK->High                                             0.50220
GroupGradeTypeK->Middle                                                NA
GroupGradeTypeMiddle->High                                        0.55652
GroupGradeTypeMiddle->Middle                                           NA
GroupGradeTypePK->Elementary                                      0.84964
GroupGradeTypePK->High                                                 NA
GroupGradeTypePK->Middle                                               NA
GroupGradeTypeUndefined->Undefined                                     NA
MajorProgramCodeH                                                 0.78779
MajorProgramCodeI                                                 0.99436
MajorProgramCodeS                                                      NA
SingleGradeTripFlag1                                              0.06534 .
FPP.to.School.enrollment                                          0.54929
FPP.to.PAX                                                        0.32917
Num.of.Non_FPP.PAX                                                     NA
SchoolSizeIndicatorL                                              0.04618 *
SchoolSizeIndicatorM-L                                            0.06768 .
SchoolSizeIndicatorS                                              0.41103
SchoolSizeIndicatorS-M                                            0.01717 *
From.Grade_surrogate1                                                  NA
To.Grade_surrogate1                                                    NA
Special.Pay_surrogate1                                                 NA
Early.RPL_surrogate1                                              0.69542
Latest.RPL_surrogate1                                             0.39562
Initial.System.Date_surrogateOther.Initial.System.Date_surrogate  0.60834
CRM.Segment_surrogateOther.CRM.Segment_surrogate                  0.75715
MDR.High.Grade_surrogate1                                              NA
Total.School.Enrollment_surrogate1                                     NA
FirstMeeting_surrogate1                                                NA
LastMeeting_surrogate1                                                 NA
DifferenceTraveltoFirstMeeting_surrogate1                              NA
DifferenceTraveltoLastMeeting_surrogate1                               NA
FPP.to.School.enrollment_surrogate1                                    NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2531.2  on 1888  degrees of freedom
Residual deviance: 1409.6  on 1693  degrees of freedom
AIC: 1801.6

Number of Fisher Scoring iterations: 15
```

The model is overfit. It has too many insignificant variables.

Let us fit a much simpler model. We will use stepwise regressions.

Recall stepwise regression from BUS 41100 Applied regression course. See, for example, Week 9 slides. You can also check Section 6.1.2 of the ISLR book.

There are three approaches to running stepwise regressions: backward, forward and both. We need to specify criterion for inclusion/exclusion of variables. We will use one based on Bayesian information criteria.

Observe the process of variables being added to the model, (labeled by "+" in the output), gradual expansion of the model, and improvement of BIC.

In [ ]:

```
# Start from a null model with intercept only, and add one covarite at a time until maximum BIC.
lgfit.null <- glm(Retained.in.2012.~ 1,
                 data=df.train, family="binomial")

lgfit.selected <- step(lgfit.null,                  # the starting model for our search
                       scope=formula(lgfit.all),   # the largest possible model that we will consider.
                       direction="forward",
                       k=log(nrow(df.train)),       # by default step() uses AIC, but by
                                                    # multiplying log(n) on the penalty, we get BIC.
                                                    # See ?step -> Arguments -> k

                       trace=1)
```

```
Start:  AIC=2538.74
Retained.in.2012. ~ 1

                                           Df Deviance    AIC
+ SingleGradeTripFlag                       1   2129.3 2144.4
+ Is.Non.Annual.                            1   2236.0 2251.1
+ From.Grade                               10   2196.2 2279.2
+ SPR.New.Existing                          1   2265.7 2280.8
+ Total.Pax                                 1   2357.4 2372.5
+ FPP                                       1   2358.5 2373.6
+ FRP.Active                                1   2387.8 2402.8
+ Total.Discount.Pax                        1   2399.9 2415.0
+ Num.of.Non_FPP.PAX                        1   2399.9 2415.0
+ SchoolGradeTypeHigh                       3   2415.0 2445.2
+ SchoolGradeType                           7   2390.5 2450.8
+ To.Grade                                 10   2396.0 2479.0
+ DepartureMonth                            5   2446.3 2491.6
+ SchoolGradeTypeLow                        3   2466.2 2496.3
+ CRM.Segment                              10   2416.2 2499.2
+ Return.Date                               1   2488.2 2503.3
+ Departure.Date                            1   2488.7 2503.8
+ GroupGradeTypeLow                         5   2459.9 2505.1
+ GroupGradeTypeHigh                        3   2480.1 2510.3
+ FPP.to.PAX                                1   2498.6 2513.7
+ MajorProgramCode                          3   2488.3 2518.5
+ Tuition                                   1   2503.4 2518.5
+ SPR.Group.Revenue                         1   2503.4 2518.5
+ SchoolSizeIndicator                       4   2483.2 2521.0
+ DifferenceTraveltoFirstMeeting            1   2506.6 2521.7
+ School.Sponsor                            1   2506.9 2521.9
+ MDR.High.Grade                            7   2463.6 2523.9
+ GroupGradeType                           11   2436.1 2526.6
+ SPR.Product.Type                          5   2482.8 2528.1
+ Special.Pay_surrogate                     1   2515.1 2530.2
+ DifferenceTraveltoLastMeeting             1   2519.0 2534.1
+ FPP.to.School.enrollment                  1   2519.3 2534.4
+ Deposit.Date                              1   2519.9 2535.0
+ Special.Pay                               4   2498.6 2536.3
+ Total.School.Enrollment                   1   2522.3 2537.4
+ Early.RPL                                 1   2522.6 2537.7
+ Early.RPL_surrogate                       1   2522.7 2537.7
<none>                                          2531.2 2538.7
+ FRP.Cancelled                             1   2523.7 2538.8
+ MDR.Low.Grade                            10   2456.2 2539.2
+ NumberOfMeetingswithParents               2   2517.1 2539.7
+ Travel.Type                               2   2517.3 2540.0
+ Poverty.Code                              6   2488.7 2541.4
+ Cancelled.Pax                             1   2528.6 2543.7
+ CRM.Segment_surrogate                     1   2529.1 2544.2
+ Parent.Meeting.Flag                       1   2529.9 2545.0
+ FirstMeeting_surrogate                    1   2529.9 2545.0
+ LastMeeting_surrogate                     1   2529.9 2545.0
+ DifferenceTraveltoFirstMeeting_surrogate  1   2529.9 2545.0
+ DifferenceTraveltoLastMeeting_surrogate   1   2529.9 2545.0
+ LastMeeting                               1   2529.9 2545.0
+ FirstMeeting                              1   2529.9 2545.0
+ School.Type                               3   2514.9 2545.1
+ From.Grade_surrogate                      1   2530.2 2545.3
+ MDR.High.Grade_surrogate                  1   2530.3 2545.4
+ Latest.RPL_surrogate                      1   2530.5 2545.6
+ Latest.RPL                                1   2530.6 2545.7
+ EZ.Pay.Take.Up.Rate                       1   2530.7 2545.8
+ Total.School.Enrollment_surrogate         1   2530.7 2545.8
+ FPP.to.School.enrollment_surrogate        1   2530.7 2545.8
+ FRP.Take.up.percent.                      1   2530.7 2545.8
+ To.Grade_surrogate                        1   2530.8 2545.8
+ Initial.System.Date_surrogate             1   2531.0 2546.1
+ Initial.System.Date                       1   2531.1 2546.2
+ Region                                    5   2503.2 2548.4
+ Days                                      9   2493.5 2568.9
+ Program.Code                             15   2453.9 2574.6
+ Income.Level                             19   2457.9 2608.7
+ Group.State                              35   2438.2 2709.8

Step:  AIC=2144.42
Retained.in.2012. ~ SingleGradeTripFlag

                                           Df Deviance    AIC
+ SPR.New.Existing                          1   1996.6 2019.2
+ Is.Non.Annual.                            1   1996.7 2019.3
+ Total.Pax                                 1   2052.8 2075.4
+ FPP                                       1   2054.8 2077.5
+ Total.Discount.Pax                        1   2056.0 2078.7
+ Num.of.Non_FPP.PAX                        1   2056.0 2078.7
+ FRP.Active                                1   2060.4 2083.0
+ SchoolGradeTypeHigh                       3   2084.7 2122.4
+ SchoolGradeTypeLow                        3   2086.3 2124.0
+ To.Grade_surrogate                        1   2101.7 2124.3
+ From.Grade_surrogate                      1   2109.3 2132.0
+ Departure.Date                            1   2113.3 2135.9
+ Return.Date                               1   2113.5 2136.2
+ SchoolSizeIndicator                       4   2092.3 2137.5
+ DifferenceTraveltoFirstMeeting            1   2118.7 2141.3
+ Total.School.Enrollment                   1   2119.7 2142.3
+ DepartureMonth                            5   2089.8 2142.6
+ GroupGradeTypeHigh                        3   2106.0 2143.7
+ SchoolGradeType                           7   2076.3 2144.2
<none>                                          2129.3 2144.4
+ GroupGradeTypeLow                         5   2092.2 2145.0
+ School.Sponsor                            1   2123.5 2146.1
+ Special.Pay_surrogate                     1   2123.8 2146.5
+ FPP.to.PAX                                1   2123.8 2146.5
+ Tuition                                   1   2124.8 2147.4
+ SPR.Group.Revenue                         1   2124.8 2147.4
+ DifferenceTraveltoLastMeeting             1   2125.4 2148.0
+ FRP.Cancelled                             1   2125.5 2148.2
+ FPP.to.School.enrollment                  1   2125.8 2148.4
+ Deposit.Date                              1   2126.4 2149.1
+ Cancelled.Pax                             1   2126.7 2149.3
+ Early.RPL                                 1   2127.2 2149.8
+ Early.RPL_surrogate                       1   2127.2 2149.8
+ Latest.RPL_surrogate                      1   2127.4 2150.1
+ Latest.RPL                                1   2127.6 2150.2
+ CRM.Segment_surrogate                     1   2128.3 2150.9
+ Total.School.Enrollment_surrogate         1   2129.1 2151.8
+ FPP.to.School.enrollment_surrogate        1   2129.1 2151.8
+ EZ.Pay.Take.Up.Rate                       1   2129.1 2151.8
+ Initial.System.Date_surrogate             1   2129.2 2151.8
+ FRP.Take.up.percent.                      1   2129.2 2151.8
+ MDR.High.Grade_surrogate                  1   2129.2 2151.8
+ Initial.System.Date                       1   2129.2 2151.8
+ Parent.Meeting.Flag                       1   2129.3 2152.0
+ FirstMeeting_surrogate                    1   2129.3 2152.0
+ LastMeeting_surrogate                     1   2129.3 2152.0
+ DifferenceTraveltoFirstMeeting_surrogate  1   2129.3 2152.0
+ DifferenceTraveltoLastMeeting_surrogate   1   2129.3 2152.0
+ LastMeeting                               1   2129.3 2152.0
+ FirstMeeting                              1   2129.3 2152.0
+ Travel.Type                               2   2122.8 2152.9
+ MajorProgramCode                          3   2116.7 2154.4
+ NumberOfMeetingswithParents               2   2124.5 2154.7
+ MDR.High.Grade                            7   2087.9 2155.8
+ Special.Pay                               4   2112.9 2158.1
+ School.Type                               3   2125.2 2162.9
+ CRM.Segment                              10   2072.7 2163.2
+ From.Grade                               10   2073.9 2164.4
+ Poverty.Code                              6   2105.8 2166.2
+ To.Grade                                 10   2078.3 2168.8
+ SPR.Product.Type                          5   2117.6 2170.4
+ Region                                    5   2117.9 2170.7
+ GroupGradeType                           11   2079.5 2177.6
+ MDR.Low.Grade                            10   2090.1 2180.7
+ Days                                      9   2108.0 2191.0
+ Program.Code                             15   2100.3 2228.5
+ Income.Level                             19   2090.3 2248.8
+ Group.State                              35   2070.2 2349.3

Step:  AIC=2019.25
Retained.in.2012. ~ SingleGradeTripFlag + SPR.New.Existing

                                           Df Deviance    AIC
+ Is.Non.Annual.                            1   1797.5 1827.7
+ Total.Discount.Pax                        1   1949.7 1979.8
+ Num.of.Non_FPP.PAX                        1   1949.7 1979.8
+ Total.Pax                                 1   1952.3 1982.5
+ FPP                                       1   1953.9 1984.1
+ FRP.Active                                1   1959.0 1989.2
+ To.Grade_surrogate                        1   1959.5 1989.7
+ From.Grade_surrogate                      1   1965.6 1995.7
+ SchoolGradeTypeHigh                       3   1952.1 1997.4
+ SchoolGradeTypeLow                        3   1952.4 1997.6
+ Total.School.Enrollment                   1   1984.5 2014.6
+ SchoolSizeIndicator                       4   1963.4 2016.2
+ DifferenceTraveltoFirstMeeting            1   1986.8 2016.9
+ Tuition                                   1   1988.0 2018.1
+ SPR.Group.Revenue                         1   1988.0 2018.1
+ Departure.Date                            1   1988.5 2018.7
+ Return.Date                               1   1988.6 2018.8
<none>                                          1996.6 2019.2
+ Cancelled.Pax                             1   1990.3 2020.5
+ SchoolGradeType                           7   1945.8 2021.2
+ Early.RPL_surrogate                       1   1991.4 2021.6
+ Early.RPL                                 1   1991.4 2021.6
+ Travel.Type                               2   1984.3 2022.0
+ FRP.Cancelled                             1   1992.2 2022.3
+ DifferenceTraveltoLastMeeting             1   1992.5 2022.7
+ FPP.to.PAX                                1   1994.1 2024.3
+ EZ.Pay.Take.Up.Rate                       1   1995.2 2025.3
+ Initial.System.Date                       1   1995.6 2025.8
+ GroupGradeTypeHigh                        3   1980.8 2026.0
+ Initial.System.Date_surrogate             1   1996.0 2026.1
+ MDR.High.Grade_surrogate                  1   1996.1 2026.3
+ Latest.RPL_surrogate                      1   1996.1 2026.3
+ School.Sponsor                            1   1996.2 2026.3
+ Latest.RPL                                1   1996.2 2026.4
+ FPP.to.School.enrollment                  1   1996.2 2026.4
+ Special.Pay_surrogate                       1   1996.3 2026.5
+ Total.School.Enrollment_surrogate           1   1996.4 2026.6
+ FPP.to.School.enrollment_surrogate          1   1996.4 2026.6
+ CRM.Segment_surrogate                       1   1996.5 2026.7
+ Deposit.Date                                1   1996.6 2026.7
+ FRP.Take.up.percent.                        1   1996.6 2026.8
+ Parent.Meeting.Flag                         1   1996.6 2026.8
+ FirstMeeting_surrogate                      1   1996.6 2026.8
+ LastMeeting_surrogate                       1   1996.6 2026.8
+ DifferenceTraveltoFirstMeeting_surrogate    1   1996.6 2026.8
+ DifferenceTraveltoLastMeeting_surrogate     1   1996.6 2026.8
+ LastMeeting                                 1   1996.6 2026.8
+ FirstMeeting                                1   1996.6 2026.8
+ DepartureMonth                              5   1967.5 2027.8
+ GroupGradeTypeLow                           5   1967.6 2028.0
+ NumberOfMeetingswithParents                 2   1993.5 2031.2
+ MajorProgramCode                            3   1987.5 2032.7
+ MDR.High.Grade                              7   1961.4 2036.8
+ Special.Pay                                 4   1984.9 2037.7
+ School.Type                                 3   1993.4 2038.7
+ From.Grade                                 10   1941.4 2039.5
+ To.Grade                                   10   1945.4 2043.4
+ Region                                      5   1985.8 2046.2
+ SPR.Product.Type                            5   1987.8 2048.1
+ Poverty.Code                                6   1981.4 2049.3
+ CRM.Segment                                10   1962.8 2060.8
+ MDR.Low.Grade                              10   1964.4 2062.4
+ GroupGradeType                             11   1958.5 2064.1
+ Days                                        9   1978.9 2069.4
+ Program.Code                               15   1976.2 2112.0
+ Income.Level                               19   1966.9 2132.9
+ Group.State                                35   1939.6 2226.3

Step:  AIC=1827.69
Retained.in.2012. ~ SingleGradeTripFlag + SPR.New.Existing + 
    Is.Non.Annual.

                                           Df Deviance    AIC
+ FRP.Active                                  1   1743.4 1781.1
+ Total.Pax                                   1   1744.1 1781.8
+ FPP                                         1   1745.6 1783.3
+ Total.Discount.Pax                          1   1746.1 1783.8
+ Num.of.Non_FPP.PAX                          1   1746.1 1783.8
+ To.Grade_surrogate                          1   1781.5 1819.2
+ From.Grade_surrogate                        1   1783.6 1821.3
+ SchoolGradeTypeLow                          3   1769.6 1822.4
+ Return.Date                                 1   1786.6 1824.3
+ FPP.to.PAX                                  1   1786.6 1824.3
+ Departure.Date                              1   1786.6 1824.3
+ Tuition                                     1   1786.9 1824.6
+ SPR.Group.Revenue                           1   1786.9 1824.6
+ SchoolGradeTypeHigh                         3   1772.0 1824.8
<none>                                            1797.5 1827.7
+ Cancelled.Pax                               1   1790.4 1828.1
+ FPP.to.School.enrollment                    1   1790.6 1828.3
+ FRP.Cancelled                               1   1791.0 1828.7
+ DifferenceTraveltoFirstMeeting              1   1793.5 1831.2
+ Total.School.Enrollment                     1   1794.9 1832.7
+ School.Sponsor                              1   1795.5 1833.2
+ Special.Pay_surrogate                       1   1795.8 1833.5
+ DifferenceTraveltoLastMeeting               1   1795.8 1833.5
+ Early.RPL_surrogate                         1   1796.3 1834.0
+ Early.RPL                                   1   1796.3 1834.0
+ Deposit.Date                                1   1796.3 1834.0
+ Travel.Type                                 2   1789.0 1834.2
+ Initial.System.Date                         1   1796.7 1834.4
+ Initial.System.Date_surrogate               1   1796.8 1834.6
+ GroupGradeTypeHigh                          3   1781.8 1834.6
+ MDR.High.Grade_surrogate                    1   1797.0 1834.7
+ Total.School.Enrollment_surrogate           1   1797.1 1834.8
+ FPP.to.School.enrollment_surrogate          1   1797.1 1834.8
+ FRP.Take.up.percent.                        1   1797.2 1834.9
+ CRM.Segment_surrogate                       1   1797.3 1835.0
+ FirstMeeting                                1   1797.3 1835.0
+ LastMeeting                                 1   1797.3 1835.0
+ Parent.Meeting.Flag                         1   1797.3 1835.0
+ FirstMeeting_surrogate                      1   1797.3 1835.0
+ LastMeeting_surrogate                       1   1797.3 1835.0
+ DifferenceTraveltoFirstMeeting_surrogate    1   1797.3 1835.0
+ DifferenceTraveltoLastMeeting_surrogate     1   1797.3 1835.0
+ Latest.RPL_surrogate                        1   1797.3 1835.1
+ Latest.RPL                                  1   1797.4 1835.1
+ EZ.Pay.Take.Up.Rate                         1   1797.5 1835.2
+ MajorProgramCode                            3   1782.7 1835.5
+ DepartureMonth                              5   1768.2 1836.1
+ SchoolSizeIndicator                         4   1779.4 1839.8
+ NumberOfMeetingswithParents                 2   1794.8 1840.0
+ GroupGradeTypeLow                           5   1776.4 1844.3
+ School.Type                                 3   1792.9 1845.7
+ Special.Pay                                 4   1785.4 1845.7
+ MDR.High.Grade                              7   1764.2 1847.2
+ SchoolGradeType                             7   1766.0 1849.0
+ SPR.Product.Type                            5   1785.2 1853.1
+ Region                                    5   1786.5 1854.4
+ Poverty.Code                              6   1780.4 1855.8
+ From.Grade                               10   1757.7 1863.3
+ CRM.Segment                              10   1759.4 1865.0
+ To.Grade                                 10   1764.2 1869.8
+ Days                                      9   1777.7 1875.7
+ MDR.Low.Grade                            10   1774.1 1879.7
+ GroupGradeType                           11   1769.1 1882.2
+ Program.Code                             15   1768.8 1912.2
+ Income.Level                             19   1769.5 1943.0
+ Group.State                              35   1743.4 2037.6

Step:  AIC=1781.14
Retained.in.2012. ~ SingleGradeTripFlag + SPR.New.Existing + 
    Is.Non.Annual. + FRP.Active

                                           Df Deviance    AIC
+ To.Grade_surrogate                        1   1731.8 1777.1
+ Total.Discount.Pax                        1   1733.5 1778.7
+ Num.of.Non_FPP.PAX                        1   1733.5 1778.7
+ From.Grade_surrogate                      1   1733.7 1778.9
<none>                                          1743.4 1781.1
+ Total.Pax                                 1   1736.5 1781.8
+ FPP                                       1   1737.3 1782.5
+ Return.Date                               1   1737.4 1782.6
+ Departure.Date                            1   1737.4 1782.7
+ FRP.Take.up.percent.                      1   1739.1 1784.3
+ SchoolGradeTypeLow                        3   1724.1 1784.4
+ DifferenceTraveltoFirstMeeting            1   1739.3 1784.6
+ Special.Pay_surrogate                     1   1739.7 1784.9
+ Tuition                                   1   1740.5 1785.8
+ SPR.Group.Revenue                         1   1740.5 1785.8
+ Early.RPL_surrogate                       1   1741.1 1786.4
+ Early.RPL                                 1   1741.1 1786.4
+ SchoolGradeTypeHigh                       3   1726.2 1786.5
+ DifferenceTraveltoLastMeeting             1   1741.4 1786.7
+ School.Sponsor                            1   1741.5 1786.8
+ FPP.to.PAX                                1   1741.7 1787.0
+ Total.School.Enrollment                   1   1742.1 1787.3
+ Latest.RPL_surrogate                      1   1742.1 1787.4
+ EZ.Pay.Take.Up.Rate                       1   1742.2 1787.4
+ Latest.RPL                                1   1742.2 1787.5
+ GroupGradeTypeHigh                        3   1727.3 1787.7
+ Initial.System.Date                       1   1742.5 1787.8
+ MDR.High.Grade_surrogate                  1   1742.5 1787.8
+ Initial.System.Date_surrogate             1   1742.7 1788.0
+ Deposit.Date                              1   1742.9 1788.1
+ FRP.Cancelled                             1   1742.9 1788.2
+ Total.School.Enrollment_surrogate         1   1743.2 1788.4
+ FPP.to.School.enrollment_surrogate        1   1743.2 1788.4
+ CRM.Segment_surrogate                     1   1743.3 1788.6
+ Parent.Meeting.Flag                       1   1743.4 1788.6
+ FirstMeeting_surrogate                    1   1743.4 1788.6
+ LastMeeting_surrogate                     1   1743.4 1788.6
+ DifferenceTraveltoFirstMeeting_surrogate  1   1743.4 1788.6
+ DifferenceTraveltoLastMeeting_surrogate   1   1743.4 1788.6
+ LastMeeting                               1   1743.4 1788.7
+ FirstMeeting                              1   1743.4 1788.7
+ FPP.to.School.enrollment                  1   1743.4 1788.7
+ Cancelled.Pax                             1   1743.4 1788.7
+ Special.Pay                               4   1724.5 1792.4
+ MajorProgramCode                          3   1733.0 1793.3
+ Travel.Type                               2   1741.7 1794.5
+ NumberOfMeetingswithParents               2   1742.1 1794.9
+ School.Type                               3   1737.1 1797.5
+ SchoolSizeIndicator                       4   1730.9 1798.8
+ DepartureMonth                            5   1724.2 1799.7
+ GroupGradeTypeLow                         5   1728.2 1803.7
+ MDR.High.Grade                            7   1713.3 1803.8
+ Region                                    5   1734.9 1810.4
+ SchoolGradeType                           7   1721.1 1811.6
+ SPR.Product.Type                          5   1736.2 1811.6
+ Poverty.Code                              6   1731.0 1814.0
+ CRM.Segment                              10   1714.0 1827.1
+ From.Grade                               10   1715.3 1828.5
+ To.Grade                                 10   1720.1 1833.3
+ Days                                      9   1729.9 1835.5
+ MDR.Low.Grade                            10   1725.0 1838.2
+ GroupGradeType                           11   1719.0 1839.7
+ Program.Code                             15   1719.8 1870.7
+ Income.Level                             19   1717.5 1898.6
+ Group.State                              35   1690.5 1992.3

Step:  AIC=1777.07
Retained.in.2012. ~ SingleGradeTripFlag + SPR.New.Existing + 
    Is.Non.Annual. + FRP.Active + To.Grade_surrogate

                                           Df Deviance    AIC
+ Total.Discount.Pax                        1   1723.6 1776.4
+ Num.of.Non_FPP.PAX                        1   1723.6 1776.4
<none>                                          1731.8 1777.1
+ Return.Date                               1   1725.3 1778.2
+ Departure.Date                            1   1725.4 1778.2
+ Total.Pax                                 1   1727.0 1779.8
+ FPP                                       1   1727.7 1780.5
+ Special.Pay_surrogate                     1   1728.5 1781.3
+ DifferenceTraveltoFirstMeeting            1   1729.0 1781.9
+ FRP.Take.up.percent.                      1   1729.4 1782.2
+ Tuition                                   1   1729.8 1782.6
+ SPR.Group.Revenue                         1   1729.8 1782.6
+ School.Sponsor                            1   1730.1 1782.9
+ Early.RPL_surrogate                       1   1730.3 1783.1
+ Early.RPL                                 1   1730.3 1783.1
+ Deposit.Date                              1   1730.5 1783.3
+ FPP.to.PAX                                1   1730.5 1783.3
+ DifferenceTraveltoLastMeeting             1   1730.6 1783.4
+ Total.School.Enrollment                   1   1730.6 1783.4
+ GroupGradeTypeHigh                        3   1715.6 1783.5
+ Initial.System.Date                       1   1731.0 1783.8
+ Total.School.Enrollment_surrogate         1   1731.0 1783.8
+ FPP.to.School.enrollment_surrogate        1   1731.0 1783.8
+ EZ.Pay.Take.Up.Rate                       1   1731.1 1783.9
+ Initial.System.Date_surrogate             1   1731.2 1784.0
+ Latest.RPL_surrogate                      1   1731.4 1784.2
+ Latest.RPL                                1   1731.4 1784.2
+ FRP.Cancelled                             1   1731.5 1784.3
+ MDR.High.Grade_surrogate                  1   1731.5 1784.3
+ CRM.Segment_surrogate                     1   1731.5 1784.3
+ Cancelled.Pax                             1   1731.8 1784.6
+ FPP.to.School.enrollment                  1   1731.8 1784.6
+ From.Grade_surrogate                      1   1731.8 1784.6
+ FirstMeeting                              1   1731.8 1784.6
+ LastMeeting                               1   1731.8 1784.6
+ Parent.Meeting.Flag                       1   1731.8 1784.6
+ FirstMeeting_surrogate                    1   1731.8 1784.6
+ LastMeeting_surrogate                     1   1731.8 1784.6
+ DifferenceTraveltoFirstMeeting_surrogate  1   1731.8 1784.6
+ DifferenceTraveltoLastMeeting_surrogate   1   1731.8 1784.6
+ MajorProgramCode                          3   1718.7 1786.5
+ SchoolGradeTypeHigh                       2   1726.2 1786.5
+ SchoolGradeTypeLow                        3   1722.5 1790.4
+ Special.Pay                               4   1715.4 1790.8
+ NumberOfMeetingswithParents               2   1730.9 1791.3
+ Travel.Type                               2   1731.0 1791.3
+ School.Type                               3   1726.0 1793.9
+ SchoolSizeIndicator                       4   1720.2 1795.7
+ DepartureMonth                            5   1713.8 1796.8
+ GroupGradeTypeLow                         5   1716.0 1799.0
+ MDR.High.Grade                            7   1702.2 1800.3
+ SPR.Product.Type                          5   1722.3 1805.3
+ Region                                    5   1723.6 1806.5
+ Poverty.Code                              6   1721.6 1812.1
+ SchoolGradeType                           7   1720.0 1818.0
+ CRM.Segment                              10   1698.0 1818.7
+ Days                                      9   1716.8 1830.0
+ MDR.Low.Grade                            10   1712.5 1833.2
+ To.Grade                                  9   1720.1 1833.3
+ From.Grade                               10   1714.0 1834.7
+ GroupGradeType                           11   1707.4 1835.7
+ Program.Code                             15   1705.4 1863.8
+ Income.Level                             19   1708.1 1896.7
+ Group.State                              35   1681.9 1991.2

Step:  AIC=1776.43
Retained.in.2012. ~ SingleGradeTripFlag + SPR.New.Existing + 
    Is.Non.Annual. + FRP.Active + To.Grade_surrogate + Total.Discount.Pax

                                           Df Deviance    AIC
<none>                                          1723.6 1776.4
+ Return.Date                               1   1717.9 1778.3
+ Departure.Date                            1   1718.0 1778.3
+ FPP.to.PAX                                1   1719.3 1779.6
+ DifferenceTraveltoFirstMeeting            1   1721.5 1781.8
+ GroupGradeTypeHigh                        3   1706.5 1782.0
+ School.Sponsor                            1   1722.0 1782.4
+ Special.Pay_surrogate                     1   1722.2 1782.6
+ Deposit.Date                              1   1722.4 1782.7
+ Early.RPL_surrogate                       1   1722.4 1782.7
+ Early.RPL                                 1   1722.4 1782.7
+ Tuition                                   1   1722.5 1782.8
+ SPR.Group.Revenue                         1   1722.5 1782.8
+ Total.School.Enrollment                   1   1722.7 1783.0
+ DifferenceTraveltoLastMeeting             1   1722.7 1783.1
+ Initial.System.Date                       1   1722.8 1783.1
+ Total.School.Enrollment_surrogate         1   1722.8 1783.2
+ FPP.to.School.enrollment_surrogate        1   1722.8 1783.2
+ Initial.System.Date_surrogate             1   1722.9 1783.3
+ FPP.to.School.enrollment                  1   1723.0 1783.3
+ FRP.Cancelled                             1   1723.3 1783.6
+ CRM.Segment_surrogate                     1   1723.3 1783.6
+ MDR.High.Grade_surrogate                  1   1723.3 1783.7
+ Latest.RPL_surrogate                      1   1723.4 1783.8
+ Latest.RPL                                1   1723.4 1783.8
+ EZ.Pay.Take.Up.Rate                       1   1723.5 1783.8
+ From.Grade_surrogate                      1   1723.5 1783.9
+ FRP.Take.up.percent.                      1   1723.5 1783.9
+ FirstMeeting                              1   1723.6 1783.9
+ LastMeeting                               1   1723.6 1783.9
+ Parent.Meeting.Flag                       1   1723.6 1783.9
+ FirstMeeting_surrogate                    1   1723.6 1783.9
+ LastMeeting_surrogate                     1   1723.6 1783.9
+ DifferenceTraveltoFirstMeeting_surrogate  1   1723.6 1783.9
+ DifferenceTraveltoLastMeeting_surrogate   1   1723.6 1783.9
+ Cancelled.Pax                             1   1723.6 1784.0
+ FPP                                       1   1723.6 1784.0
+ Total.Pax                                 1   1723.6 1784.0
+ SchoolGradeTypeHigh                       2   1718.1 1786.0
+ MajorProgramCode                          3   1710.8 1786.2
+ SchoolGradeTypeLow                        3   1713.4 1788.8
+ NumberOfMeetingswithParents               2   1722.8 1790.7
+ Travel.Type                               2   1723.2 1791.1
+ School.Type                               3   1718.1 1793.5
+ Special.Pay                               4   1711.5 1794.5
+ SchoolSizeIndicator                       4   1713.0 1796.0
+ DepartureMonth                            5   1706.7 1797.2
+ GroupGradeTypeLow                         5   1707.9 1798.5
+ MDR.High.Grade                            7   1694.7 1800.3
+ SPR.Product.Type                          5   1714.3 1804.8
+ Region                                    5   1714.8 1805.3
+ Poverty.Code                              6   1713.7 1811.7
+ CRM.Segment                              10   1687.7 1815.9
+ SchoolGradeType                           7   1711.0 1816.6
+ Days                                      9   1709.4 1830.1
+ To.Grade                                  9   1711.5 1832.2
+ MDR.Low.Grade                            10   1704.3 1832.5
+ From.Grade                               10   1705.6 1833.9
+ GroupGradeType                           11   1699.0 1834.7
+ Program.Code                             15   1696.8 1862.8
+ Income.Level                             19   1700.0 1896.1
+ Group.State                              35   1674.6 1991.4
```

The algorithm stops once none of the 1-step expanded models lead to a lower BIC.

This is the selected model.

In [ ]:

```
summary(lgfit.selected)
```

```
Call:
glm(formula = Retained.in.2012. ~ SingleGradeTripFlag + SPR.New.Existing + 
    Is.Non.Annual. + FRP.Active + To.Grade_surrogate + Total.Discount.Pax, 
    family = "binomial", data = df.train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.8150  -0.7108   0.3982   0.6079   2.7149  

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)           0.100495   0.141541   0.710 0.477699    
SingleGradeTripFlag1  1.220935   0.130267   9.373  < 2e-16 ***
SPR.New.ExistingNEW  -1.597414   0.129210 -12.363  < 2e-16 ***
Is.Non.Annual.1      -2.427700   0.194144 -12.505  < 2e-16 ***
FRP.Active            0.023528   0.006669   3.528 0.000419 ***
To.Grade_surrogate1   0.738475   0.235902   3.130 0.001745 ** 
Total.Discount.Pax    0.108888   0.039687   2.744 0.006077 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2531.2  on 1888  degrees of freedom
Residual deviance: 1723.6  on 1882  degrees of freedom
AIC: 1737.6

Number of Fisher Scoring iterations: 5
```

You can predict probabilities from this model using the following.

In [ ]:

```
phat.lgfit.selected <- predict(lgfit.selected,
                               newdata = df.test,
                               type = "response")
```

You will use these probabilities later.

While we are investigating variable selection in logistic regression models, let us also use a more modern approach to variable selection. We will use the lasso.

If you have not seen this in BUS 41100 Applied regression course, do not worry. We will provide more details in the Week 5. You can also check Section 6.2.2 of the ISLR book.

I provide the code to fit a lasso logistic regression model. We find coefficients β that minimize the deviance loss plus the penalty: [-2\cdot\sum_{i=1}^n \log p(y_i, x*i; |beta) + |lambda |sum*{j=1}^p |\beta_j|.] Here, λ is the user chosen penalty that controls the flexibility of the fit.

First, we need to create a model matrix that will be used as an input to the package.

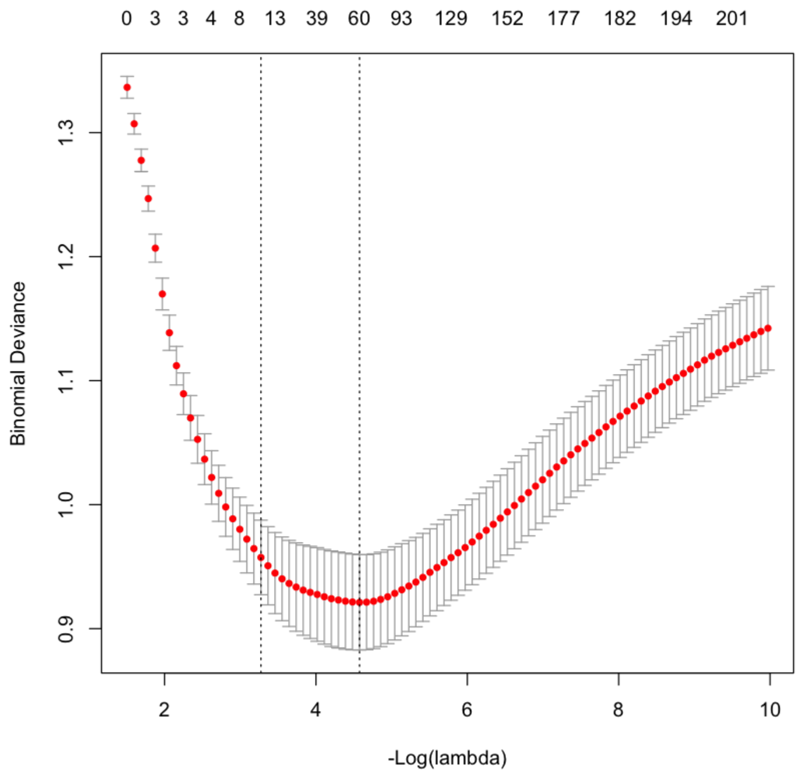
```
In [ ]: X <- model.matrix(formula(lgfit.all), STCdata_A)
        #need to subtract the intercept
        X <- X[,-1]

        X.train = X[ inTrain, ]
        X.test = X[ -inTrain, ]
```

Next, we run 5-fold cross-validation.

```
In [ ]: cv.l1.lgfit <- cv.glmnet(
          x        = X.train,
          y        = df.train$Retained.in.2012.,
          family   = "binomial",
          alpha    = 1,    #alpha=0 gives ridge regression
          nfolds   = 5)
```

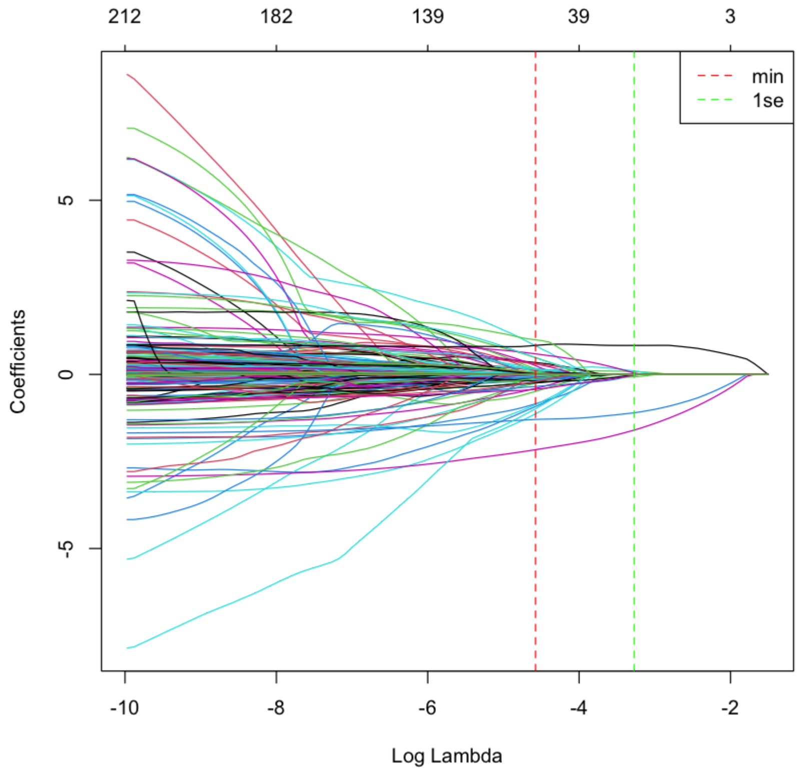
We can plot the cross-validation curve, which shows us an estimate of out-of-sample deviance as a function of the tuning parameter λ. The x-axis represents to $-\log(\lambda)$. Therefore, on the left we have large values of λ and on the right we have small values of λ. At the top, you can see the number variables that were selected into the model. The two vertical dashed lines correspond to λ values that minimize the cross-validation error and the largest value of lambda such that error is within 1 standard error of the minimum.

```
In [ ]: plot(cv.l1.lgfit, sign.lambda=-1)
```



Let us know plot the fitted coefficients as a function of λ. Note that `cv.l1.lgfit$glmnet.fit` corresponds to a fitted glmnet object for the full data.

```
In [ ]: glmnet.fit <- cv.l1.lgfit$glmnet.fit
        plot(glmnet.fit, xvar = "lambda")
        abline(v = log(cv.l1.lgfit$lambda.min), lty=2, col="red")
        abline(v = log(cv.l1.lgfit$lambda.1se), lty=2, col="green")
        legend("topright", legend=c("min", "1se"), lty=2, col=c("red", "green"))
```



For our predictive model, we will use 1 standard error λ. Below you can see the variables that are selected by the lasso.

In [ ]:

```
betas <- coef(cv.l1.lgfit, s = "lambda.1se")
model.1se <- which(betas[2:length(betas)]!=0)
colnames(X[,model.1se])
```

```
'From.Grade8' · 'Is.Non.Annual.1' · 'FRP.Active' · 'Total.Discount.Pax' · 'CRM.Segment8' · 'MDR.High.Grade8' · 'Income.LevelP' ·
'SPR.New.ExistingNEW' · 'Total.Pax' · 'SchoolGradeTypeHighHigh' · 'DepartureMonthJune' · 'SingleGradeTripFlag1' · 'SchoolSizeIndicatorS'
```

We now use our model to predict probabilities on the test set.

In [ ]:

```
phat.l1.lgfit <- predict(glmnet.fit,
                         newx = X.test,
                         s = cv.l1.lgfit$lambda.1se,
                         type = "response")
```

Questions

How well does logistic regression do?

1. Create a confusion matrix for two logistic regression models build above. Use probabilities `phat.lgfit.selected` and `phat.l1.lgfit` to do so.

   To solve this question, you need to make a major decision. What should the cutoff or "threshold" for the probability be, above which you will label a customer as being classified as "retained?" In our case, the data is slightly unbalanced---about 60.72% of data points are in Class 1. For very unbalanced data, we would first need to balance it (over- or under-sample). In this case, the benefits of balancing are unclear, hence one can implement the average probability of being retained as a cutoff.

   Predict classification using 0.6072 threshold.

   What can we see from the confusion matrices?

In [ ]:

```
threshold <- mean(phat.lgfit.selected)
```

In [ ]:

```
get_confusion_matrix = function(y, phat, thr=0.5){
  yhat = as.factor(ifelse(phat > thr, 1, 0)) # 1 of greater than thr, 0 o.w.
  confusionMatrix(yhat, y)
}
```

In [ ]:

```
get_confusion_matrix(df.test$Retained.in.2012., phat.lgfit.selected, threshold)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 147  66
         1  49 238

               Accuracy : 0.77
                 95% CI : (0.7306, 0.8062)
    No Information Rate : 0.608
    P-Value [Acc > NIR] : 1.062e-14

                  Kappa : 0.5248

 Mcnemar's Test P-Value : 0.1357

            Sensitivity : 0.7500
            Specificity : 0.7829
         Pos Pred Value : 0.6901
         Neg Pred Value : 0.8293
             Prevalence : 0.3920
         Detection Rate : 0.2940
   Detection Prevalence : 0.4260
      Balanced Accuracy : 0.7664

       'Positive' Class : 0
```

In [ ]:

```
get_confusion_matrix(df.test$Retained.in.2012., phat.l1.lgfit, threshold)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 149  60
         1  47 244

               Accuracy : 0.786
                 95% CI : (0.7474, 0.8212)
    No Information Rate : 0.608
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5563

 Mcnemar's Test P-Value : 0.246

            Sensitivity : 0.7602
            Specificity : 0.8026
         Pos Pred Value : 0.7129
         Neg Pred Value : 0.8385
             Prevalence : 0.3920
         Detection Rate : 0.2980
   Detection Prevalence : 0.4180
      Balanced Accuracy : 0.7814

       'Positive' Class : 0
```

From the confusion matrices, we can see that the lasso model does a better job, though not by much. It is observed that the lasso model has both less false positives and false negatives, increasing the accuracy by .01% which we consider not to be a significant improvement.

1. Plot ROC curves for the two classifiers and report the area under the curve.

   Note that the AUC of an error-free classifier would be 100%, and an AUC of a random guess would be 50%. For values in-between, we can think of AUC as follows:

   - 90%+ = excellent,
   - 80–90% = very good,
   - 70–80% = good,
   - 60–70% = so-so, and
   - below 60% = not much value.

In [ ]:

```
library(ROCR)
```

In [ ]:

```
# Create a list with the 2 phat vectors
phat_list = list()
phat_list$lgfit =  matrix(phat.lgfit.selected,  ncol = 1)
phat_list$lasso =  matrix(phat.l1.lgfit,  ncol = 1)
nmethod <- length(phat_list)
```

In [ ]:

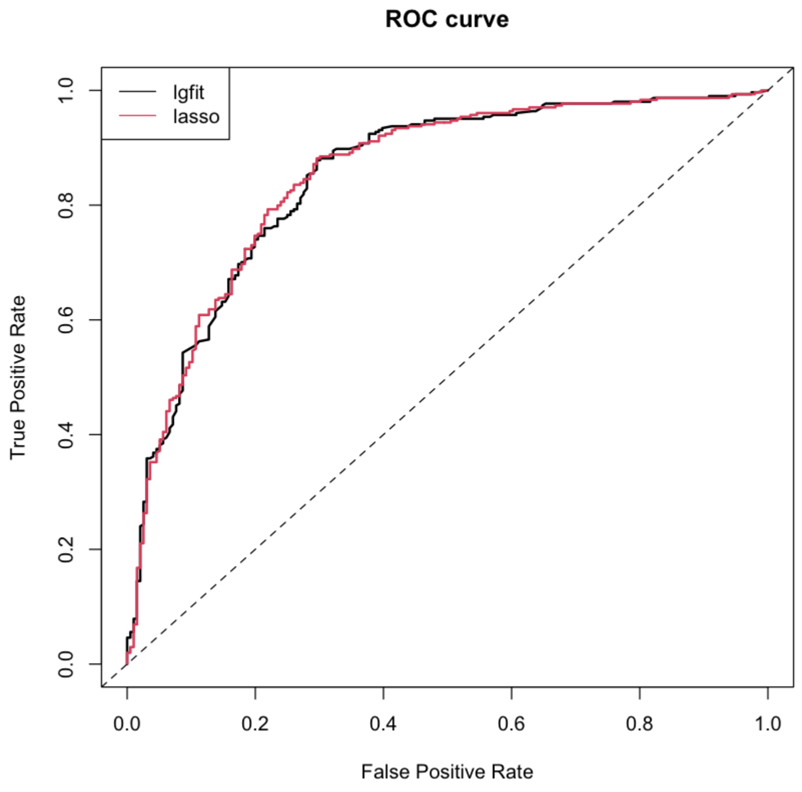
```
#' @param y: should be 0/1
#' @param phat: probabilities obtained by our algorithm
#' @param wht: shrinks probabilities in phat towards .5
#' this helps avoid numerical problems --- don't use log(0)!
#' @return deviance loss
get_deviance = function(y,phat,wht=1e-7) {
  if(is.factor(y)) y = as.numeric(y)-1
  phat = (1-wht)*phat + wht*.5
  py = ifelse(y==1, phat, 1-phat)
  return(-2*sum(log(py)))
}
```

In [ ]:

```
phat_best = matrix(0.0,nrow(df.test),nmethod) #pick off best from each method
colnames(phat_best) = names(phat_list)

for(i in 1:nmethod) {
  nrun = ncol(phat_list[[i]])
  lvec = rep(0,nrun)
  for(j in 1:nrun) lvec[j] = get_deviance(df.test$Retained.in.2012.,phat_list[[i]][,j])
    imin = which.min(lvec)
  phat_best[,i] = phat_list[[i]][,imin]
}
```

In [ ]:

```
for(i in 1:ncol(phat_best)) {
  pred = prediction(phat_best[,i], df.test$Retained.in.2012.)
  perf = performance(pred, measure = "tpr", x.measure = "fpr")

  if (i == 1) {
    plot(perf, col=1, lwd=2,
    main= 'ROC curve',
    xlab='False Positive Rate',
    ylab='True Positive Rate')
  }

  else {
    plot(perf, add=T, col=i, lwd=2)
  }
}
abline(0, 1, lty=2)
legend("topleft",legend=names(phat_list),col=1:nmethod,lty=rep(1,nmethod))
```



In [ ]:

```
for(i in 1:ncol(phat_best)) {
    pred = prediction(phat_best[,i], df.test$Retained.in.2012.)
    perf = performance(pred, measure = "auc")
    print(paste0("AUC ", names(phat_list)[i], " :: ", perf@y.values[[1]]))
}
```

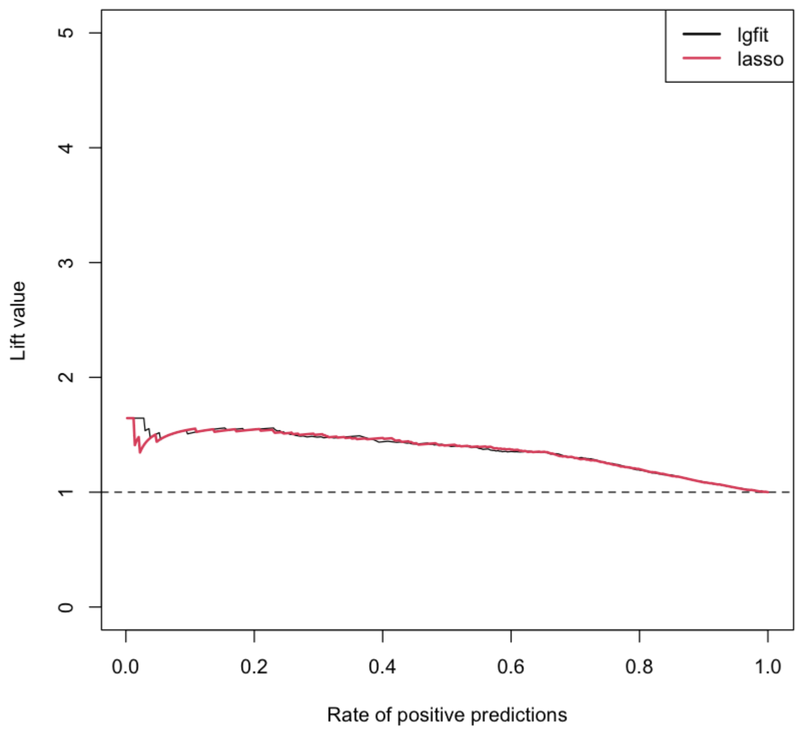
```
[1] "AUC lgfit :: 0.849321965628357"
[1] "AUC lasso :: 0.85170515574651"
```

The AUC (Area Under the Curve) values of 0.849 and 0.852 represent the performance of two different models, "lgfit" and "lasso", in a binary classification task. An AUC of 0.8 is considered a good model performance, and an AUC value close to 1 indicates a perfect classifier. The higher the AUC value, the better the model is at distinguishing between the positive and negative class. In this case, the "lasso" model has a slightly better performance with an AUC of 0.8517 compared to the "lgfit" model with an AUC of 0.8493.

1. Plot lift curves for the two classifiers.

In [ ]:

```
pred = prediction(phat_best[,1], df.test$Retained.in.2012.)
perf = performance(pred, measure="lift", x.measure="rpp", lwd=2)
plot(perf, col=1, ylim=c(0,5))
abline(h=1, lty=2)

for(i in 2:ncol(phat_best)) {
  pred = prediction(phat_best[,i], df.test$Retained.in.2012.)
  perf = performance(pred, measure="lift", x.measure="rpp")
  plot(perf, add=T, col=i, lwd=2)
}
legend("topright", legend=names(phat_list),col=1:nmethod, lty=rep(1,nmethod), lwd=2)
```



We can observe from the lift curves that they are very similar.

1. Create the profit curve (the amount of net profit vs the number of groups targeted for promotion) for the two classifiers. Suppose that the benefit of retaining a group is $100, whilethecostofapromotionis40$.

   How many groups should be targeted to maximize the profit?

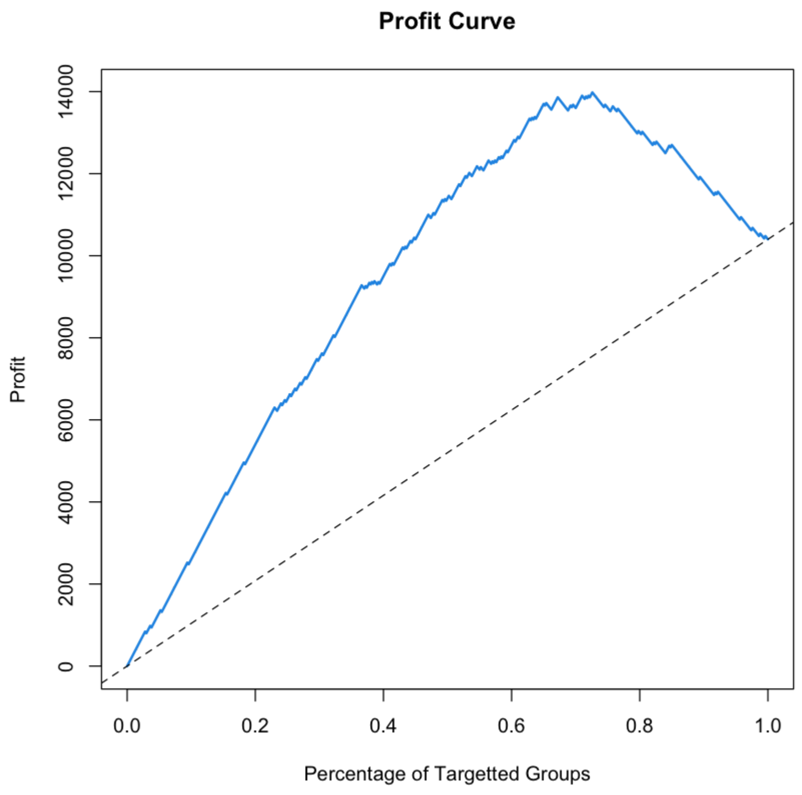
   How would this number change as the ratio between the benefit and cost changes?

In [ ]:

```
# Function to plot a profit curve
#
# Inputs:
#  - benefitTP(FN/FP/TN): the net benefit for a true positive (false negative,...)
#      which is positive for a gain, and negative for a loss
#  - y: vector of true labels, which has to be labeled as "0" and "1"
#  - phat: vector of predicted probabilities
# Outputs:
 #    the function returns the profit curve

ProfitCurve <- function(benefitTP, benefitFN, benefitFP, benefitTN, y, phat){
    if(length(y) != length(phat)) stop("Length of y and phat not identical")
    if(length(levels(y))!=2 | levels(y)[1]!="0" | levels(y)[2]!="1") stop("y should be a vector of factors, only with
    n <- length(y)
    df <- data.frame(y, phat)
# Order phat so that we can pick the k highest groups for promotion
    df <- df[order(df[,2], decreasing = T),]
    TP <- 0; FP <- 0; FN <- table(y)[2]; TN <- table(y)[1]
# Initializing the x and y coordinates of the plot
    ratio.vec <- seq(0,n)/n
    profit.vec <- rep(0,n+1)
    profit.vec[1] <- FN * benefitFN + TN * benefitTN
    for(k in 1:n){
        # k is the number of groups classified as "YES"
        # In every round, we are picking one more group for promotion.
        # If this group was ratained (positive), then in this round, it is classified
        # as a "YES" instead of "NO" before. The confusion matrix is updated each round
        # with one more TP, and one less FN. It's similar when the group was not ratained.
        if(df[k,1]=="1"){TP <- TP + 1; FN <- FN - 1}
        else{FP <- FP + 1; TN <- TN - 1}
        # print(paste(TP, FP, TP-FP, benefitTP, benefitFP))
        profit.vec[k+1] <- TP*benefitTP + FP*benefitFP + FN*benefitFN + TN*benefitTN
        }

    # Get a matrix with profit and ratio
    profit.mat <- cbind(ratio.vec, profit.vec)

    plt <- plot(ratio.vec, profit.vec, type="l", lwd=2, col=4, main="Profit Curve",
            xlab="Percentage of Targetted Groups", ylab="Profit")
    abline(b=(profit.vec[n+1]-profit.vec[1]), a=profit.vec[1], lty=2) #Random guess
    return(profit.mat )
    }
```

In [ ]:

```
curve_1 <- ProfitCurve(60,0,-40,0,df.test$Retained.in.2012.,phat_best[,1])
```



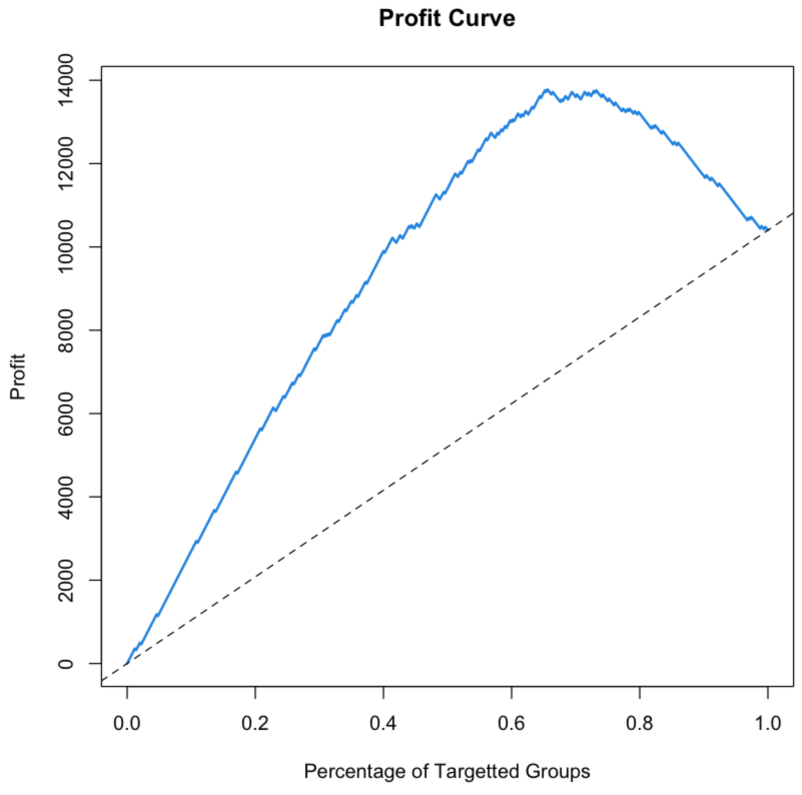
In [ ]:

```
# Get the maximum profit and the corresponding ratio with the correspinding row number
max_profit <- max(curve_1[,2])
max_ratio <- curve_1[which.max(curve_1[,2]),1]
max_row <- which.max(curve_1[,2])
print(paste("Maximum profit is", max_profit, "with ratio", max_ratio, "and the number of groups", max_row))
```

```
[1] "Maximum profit is 13980 with ratio 0.726 and the number of groups 364"
```

In [ ]:

```
curve_2 <- ProfitCurve(60,0,-40,0,df.test$Retained.in.2012.,phat_best[,2])
```



In [ ]:

```
# Get the maximum profit and the corresponding ratio with the correspinding row number
max_profit <- max(curve_2[,2])
max_ratio <- curve_2[which.max(curve_2[,2]),1]
max_row <- which.max(curve_2[,2])
print(paste("Maximum profit is", max_profit, "with ratio", max_ratio, "and the number of groups", max_row))
```

```
[1] "Maximum profit is 13780 with ratio 0.656 and the number of groups 329"
```

Let's suppose that the cost increases, we would expect the ratio of targeted groups to reduce as well. This is because the cost of the promotion is now higher than the benefit of retaining a group. The profit curve is a function of the ratio between the benefit and cost, and as the ratio increases, the number of groups targeted for promotion increases.

1. Develop a decision tree, random forest, and a boosting model using the training data.

   Report ROC, AUC, lift, and profit curves for these models.

   How do these methods compare to the logistic regression models?

In [ ]:

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

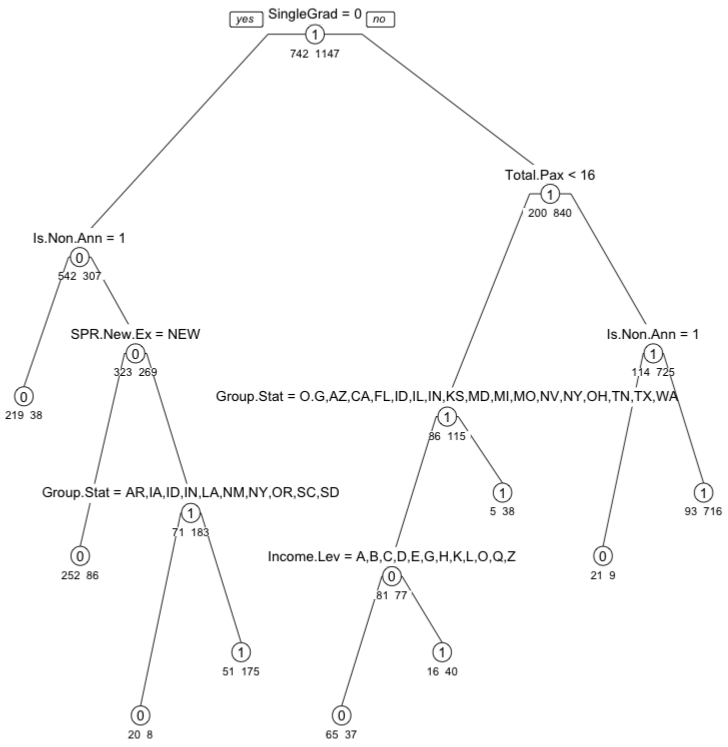
Decision Tree

In [ ]:

```
default.ct <- rpart(Retained.in.2012. ~ ., data = df.train, method = "class")
```

In [ ]:

```
prp(default.ct, type = 1, extra = 1, under = TRUE, split.font = 1, varlen = -10)
```



In [ ]:

```
deeper.ct <- rpart(Retained.in.2012. ~ ., data = df.train, method = "class", cp = 0, minsplit = 1)
```

In [ ]:

```
length(deeper.ct$frame$var[deeper.ct$frame$var == "<leaf>"])
```

229

In [ ]:

```
default.ct.point.pred.train <- predict(default.ct, df.train, type = "class")
deeper.ct.point.pred.train <- predict(deeper.ct, df.train, type = "class")
cm.default.train <- confusionMatrix(default.ct.point.pred.train, df.train$Retained.in.2012.)
cm.deeper.train <- confusionMatrix(deeper.ct.point.pred.train, df.train$Retained.in.2012.)
print(cm.default.train)
print(cm.deeper.train)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 577 178
         1 165 969

               Accuracy : 0.8184
                 95% CI : (0.8003, 0.8356)
    No Information Rate : 0.6072
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6205

 Mcnemar's Test P-Value : 0.517

            Sensitivity : 0.7776
            Specificity : 0.8448
         Pos Pred Value : 0.7642
         Neg Pred Value : 0.8545
             Prevalence : 0.3928
         Detection Rate : 0.3055
   Detection Prevalence : 0.3997
      Balanced Accuracy : 0.8112

       'Positive' Class : 0

Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  742    0
         1    0 1147

               Accuracy : 1
                 95% CI : (0.998, 1)
    No Information Rate : 0.6072
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

            Sensitivity : 1.0000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 1.0000
             Prevalence : 0.3928
         Detection Rate : 0.3928
   Detection Prevalence : 0.3928
      Balanced Accuracy : 1.0000

       'Positive' Class : 0
```

In [ ]:

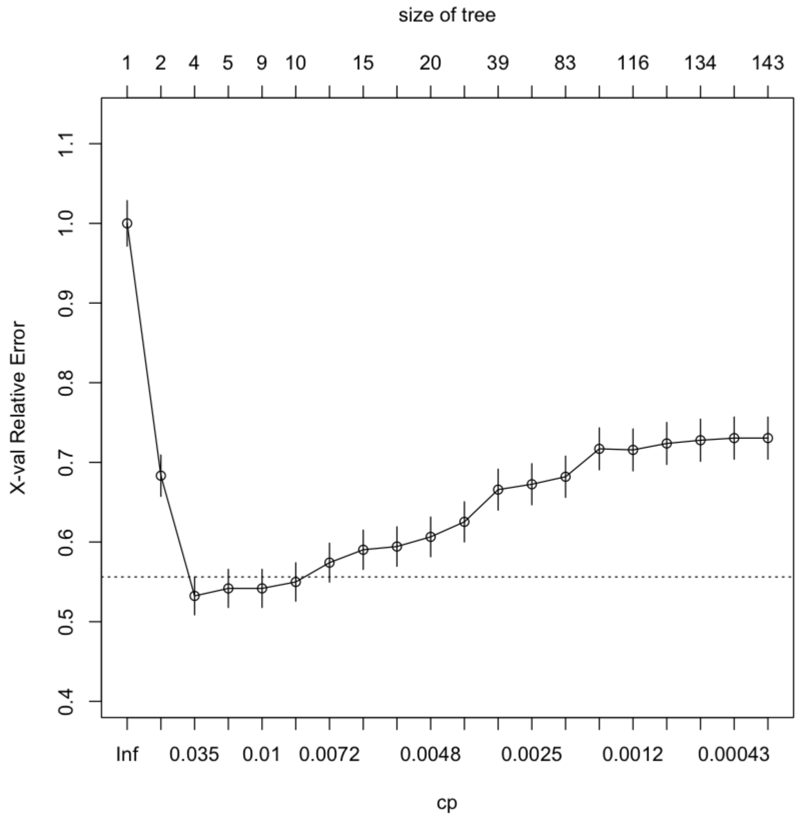
```r
default.ct.point.pred.valid <- predict(default.ct, df.test, type = "class")
deeper.ct.point.pred.valid <- predict(deeper.ct, df.test, type = "class")
cm.default.valid <- confusionMatrix(default.ct.point.pred.valid, df.test$Retained.in.2012.)
cm.deeper.valid <- confusionMatrix(deeper.ct.point.pred.valid, df.test$Retained.in.2012.)
print(cm.default.valid)
print(cm.deeper.valid)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 140  56
         1  56 248

               Accuracy : 0.776
                 95% CI : (0.7369, 0.8118)
    No Information Rate : 0.608
    P-Value [Acc > NIR] : 9.848e-16

                  Kappa : 0.5301

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.7143
            Specificity : 0.8158
         Pos Pred Value : 0.7143
         Neg Pred Value : 0.8158
             Prevalence : 0.3920
         Detection Rate : 0.2800
   Detection Prevalence : 0.3920
      Balanced Accuracy : 0.7650

       'Positive' Class : 0

Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 128  81
         1  68 223

               Accuracy : 0.702
                 95% CI : (0.6598, 0.7418)
    No Information Rate : 0.608
    P-Value [Acc > NIR] : 7.392e-06

                  Kappa : 0.3821

 Mcnemar's Test P-Value : 0.3256

            Sensitivity : 0.6531
            Specificity : 0.7336
         Pos Pred Value : 0.6124
         Neg Pred Value : 0.7663
             Prevalence : 0.3920
         Detection Rate : 0.2560
   Detection Prevalence : 0.4180
      Balanced Accuracy : 0.6933

       'Positive' Class : 0
```

In [ ]:

```r
cv.ct <- rpart(Retained.in.2012. ~ ., data = df.train, method = "class",
            cp = 0.00001, minsplit = 5, xval = 5)
printcp(cv.ct)
plotcp(cv.ct)
```

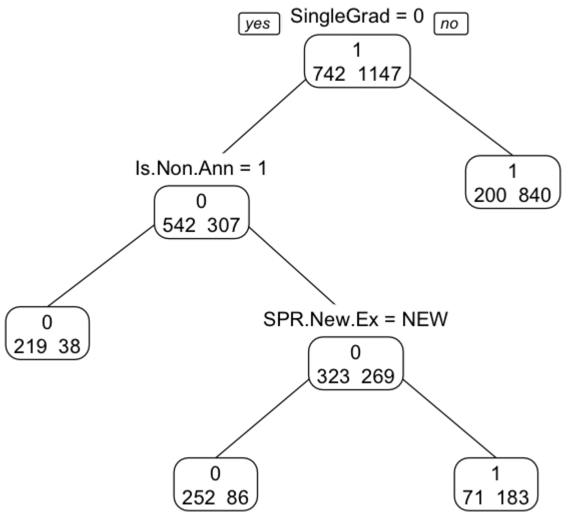
```
Classification tree:
rpart(formula = Retained.in.2012. ~ ., data = df.train, method = "class", 
    cp = 1e-05, minsplit = 5, xval = 5)

Variables actually used in tree construction:
 [1] CRM.Segment                    Days                          
 [3] Departure.Date                 Deposit.Date                  
 [5] DifferenceTraveltoFirstMeeting DifferenceTraveltoLastMeeting 
 [7] EZ.Pay.Take.Up.Rate            Early.RPL                     
 [9] FPP                            FPP.to.PAX                    
[11] FPP.to.School.enrollment       FRP.Active                    
[13] FRP.Cancelled                  FRP.Take.up.percent.          
[15] FirstMeeting                   From.Grade                    
[17] Group.State                    GroupGradeType                
[19] Income.Level                   Initial.System.Date           
[21] Is.Non.Annual.                 LastMeeting                   
[23] Latest.RPL                     MDR.High.Grade                
[25] MDR.Low.Grade                  Poverty.Code                  
[27] Program.Code                   Region                        
[29] SPR.New.Existing               School.Sponsor                
[31] School.Type                    SchoolGradeType               
[33] SchoolSizeIndicator            SingleGradeTripFlag           
[35] Special.Pay                    To.Grade                      
[37] Total.Pax                      Total.School.Enrollment       
[39] Tuition                       

Root node error: 742/1889 = 0.3928

n= 1889 

           CP nsplit rel error  xerror     xstd
1  0.31671159      0  1.000000 1.00000 0.028606
2  0.07547170      1  0.683288 0.68329 0.025956
3  0.01617251      3  0.532345 0.53235 0.023821
4  0.01347709      4  0.516173 0.54178 0.023974
5  0.00808625      8  0.462264 0.54178 0.023974
6  0.00763702      9  0.454178 0.54987 0.024104
7  0.00673854     13  0.420485 0.57412 0.024480
8  0.00539084     14  0.413747 0.59030 0.024720
9  0.00494160     16  0.402965 0.59434 0.024779
10 0.00471698     19  0.388140 0.60647 0.024953
11 0.00404313     21  0.378706 0.62534 0.025214
12 0.00269542     38  0.308625 0.66577 0.025741
13 0.00224618     71  0.207547 0.67251 0.025825
14 0.00202156     82  0.176550 0.68194 0.025940
15 0.00134771     92  0.150943 0.71698 0.026347
16 0.00112309    115  0.119946 0.71563 0.026332
17 0.00089847    121  0.113208 0.72372 0.026421
18 0.00067385    133  0.102426 0.72776 0.026466
19 0.00026954    137  0.099730 0.73046 0.026495
20 0.00001000    142  0.098383 0.73046 0.026495
```



In [ ]:

```
pruned.ct <- prune(cv.ct, cp = cv.ct$cptable[which.min(cv.ct$cptable[,"xerror"]),"CP"])
length(pruned.ct$frame$var[pruned.ct$frame$var == "<leaf>"])
prp(pruned.ct, type = 1, extra = 1, split.font = 1, varlen = -10)
```

```
4
```



In [ ]:

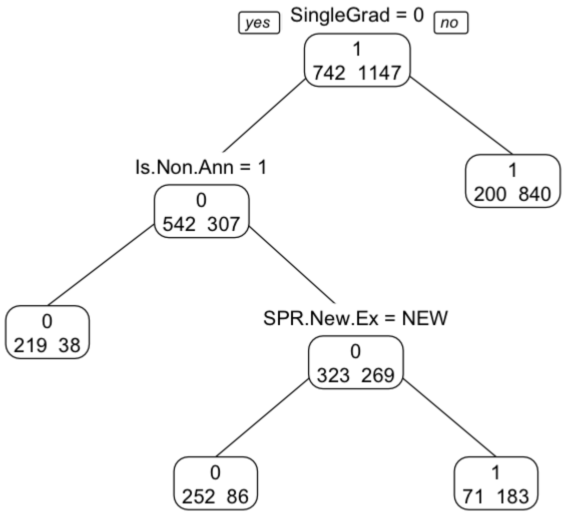
```
# this is the cp parameter with smallest cv-errror
index_cp_min = which.min(cv.ct$cptable[,"xerror"])

# one standard deviation rule
# need to find first cp value for which the xerror is below horizontal line on the plot
(val_h = cv.ct$cptable[index_cp_min, "xerror"] + cv.ct$cptable[index_cp_min, "xstd"])
(index_cp_std = Position(function(x) x < val_h, cv.ct$cptable[, "xerror"]))
(cp_std = cv.ct$cptable[ index_cp_std, "CP" ])
```

```
0.556165682910288
3
0.0161725067385445
```

In [ ]:

```
pruned.ct <- prune(cv.ct, cp = cp_std)
length(pruned.ct$frame$var[pruned.ct$frame$var == "<leaf>"])
prp(pruned.ct, type = 1, extra = 1, split.font = 1, varlen = -10)
```

```
4
```



In [ ]:

```
phat.tree <- predict(pruned.ct, df.test, type = "prob")
# Drop the first column, which is the probability of "NO"
phat.tree <- phat.tree[,2]
phat.tree <- data.frame(phat.tree)
# Add to phat_list the phat tree as a matrix 500 x 1
phat_list <- cbind(phat_list, phat.tree)
```

Random Forest

In [ ]:

```
# Run a random forest model
p = ncol(df.train) - 1

grid_rf = expand.grid(
  mtry = c(p, ceiling(sqrt(p))),
  node_size = c(5, 10, 20)
)

for (i in 1:nrow(grid_rf)) {
  rf = ranger(Retained.in.2012. ~ ., data = df.train, mtry = grid_rf$mtry[i],
              min.node.size = grid_rf$node_size[i], probability = TRUE)
  phat.rf <- predict(rf, df.test)$predictions
  phat.rf <- data.frame(phat.rf)
}

# Select the best model
phat.rf <- predict(rf, df.test)$predictions
phat.rf <- data.frame(phat.rf)
# Add to phat_list the phat rf as a matrix 500 x 1 only X1
phat_list <- cbind(phat_list, phat.rf[,1])
```

In [ ]:

```
phat_best = matrix(0.0,nrow(df.test),nmethod) #pick off best from each method
colnames(phat_best) = names(phat_list)

for(i in 1:nmethod) {
  nrun = ncol(phat_list[[i]])
  lvec = rep(0,nrun)
  for(j in 1:nrun) lvec[j] = get_deviance(df.test$Retained.in.2012.,phat_list[[i]][,j])
    imin = which.min(lvec)
  phat_best[,i] = phat_list[[i]][,imin]
}
```

```
Error in dimnames(x) <- dn: length of 'dimnames' [2] not equal to array extent
Traceback:

1. `colnames<-`(`*tmp*`, value = c("lgfit", "lasso", "phat.tree",
 . "phat.rf[, 1]"))
```