**DSBA/MBAD 6211 Assignment 1**

Due: 11:59pm @ 07/12/2020

In the fall of 2014, the administration of a large private university requested that the Office of Enrollment Management and the Office of Institutional Research work together to identify prospective students who would most likely enroll as new freshmen in the Fall 2015 semester. Historically, inquiries numbered about 90,000+ students, and the university enrolled from 2400 to 2800 new freshmen each Fall semester. It was decided that inquiries for Fall 2014 would be used to build the model to help shape the Fall 2015 freshman class. The data set ***INQ2015*** was built over a period of a several months in consultation with Enrollment Management. Please carefully explore all variables and build a predictive model for better enrollment management. Please apply regression and decision tree models to analyze the data.

* **Variable and model naming requirements:**
  + Please include your ***name initials*** to the data frame names as well as model names in your R coding.
  + Please instance, in my coding, I would name the data frames as ***dfKZ, dfKZ.train***, and ***dfKZ.valid.*** I would also name the models as ***regressionKZ, treeKZ***, etc.

Please submit a Word document including:

1. A table showing the overall structure of the dataset, including variable names, data types, and whether the variables will be used in your analyses. Also, please answer questions c, d, e.
   1. The nominal variables **ACADEMIC\_INTEREST\_1**, **ACADEMIC\_INTEREST\_2**, and **IRSCHOOL** were rejected because they were replaced by the interval variables **INT1RAT**, **INT2RAT**, and **HSCRAT**, respectively. For example, academic interest codes 1 and 2 were replaced by the percentage of inquirers over the past five years who indicated those interest codes and then enrolled. The variable **IRSCHOOL** is the high school code of the student, and it was replaced by the percentage of inquirers from that high school over the last five years who enrolled.
   2. **CONTACT\_CODE1** and **CONTACT\_DATE1** are also rejected due to their irrelevance suggested by Enrollment Management.
   3. ***Should your model reject any other variables for your analyses?*** If so, please explain reasons for each additionally rejected variable.
      * + Rejected IRSCHOOL and LEVEL\_YEAR as these variable doesn’t add any value for dependent variable prediction. LEVEL\_YEAR has always constant value as - FR14 and IRSCHOOL is just a high school code which is not significant.
   4. Which variable is your target variable?
      * + Enroll is target variable
   5. Do you need to change any measurement levels of your existing variables? Why?
      * + Yes I changed measurement level for AVG\_INCOME and DISTANCE variable to avoid skew problem, and to distribute it normally.
2. Explain whether variable imputation and transformation are needed. If so, please explain which variables have been imputed, transformed and how.
   * + - Below are the imputation and transformation. Most of the variables are quantitative and continuous in nature but has missing data so all the mission data is replaced by mean of that same column.
       - For categorical and ordinal in nature variable missing values are replaced by max occurrence (MODE) of the same variable.
       - Also DISTANCE and AVG\_INCOME columns are log transformed to avoid skew.

dfNK1$satscore <-with(dfNK1,impute(satscore,mean)) # Quantitative - Continuous Variable

dfNK1$avg\_income <-with(dfNK1,impute(avg\_income,mean)) # Quantitative - Continuous Variable

dfNK1$distance <-with(dfNK1,impute(distance,mean)) # Quantitative - Continuous Variable

dfNK1$sex <-with(dfNK1,impute(sex,max)) # Categorical - Ordinal variable

dfNK1$telecq<-with(dfNK1,impute(telecq,max)) # Categorical - Ordinal variable

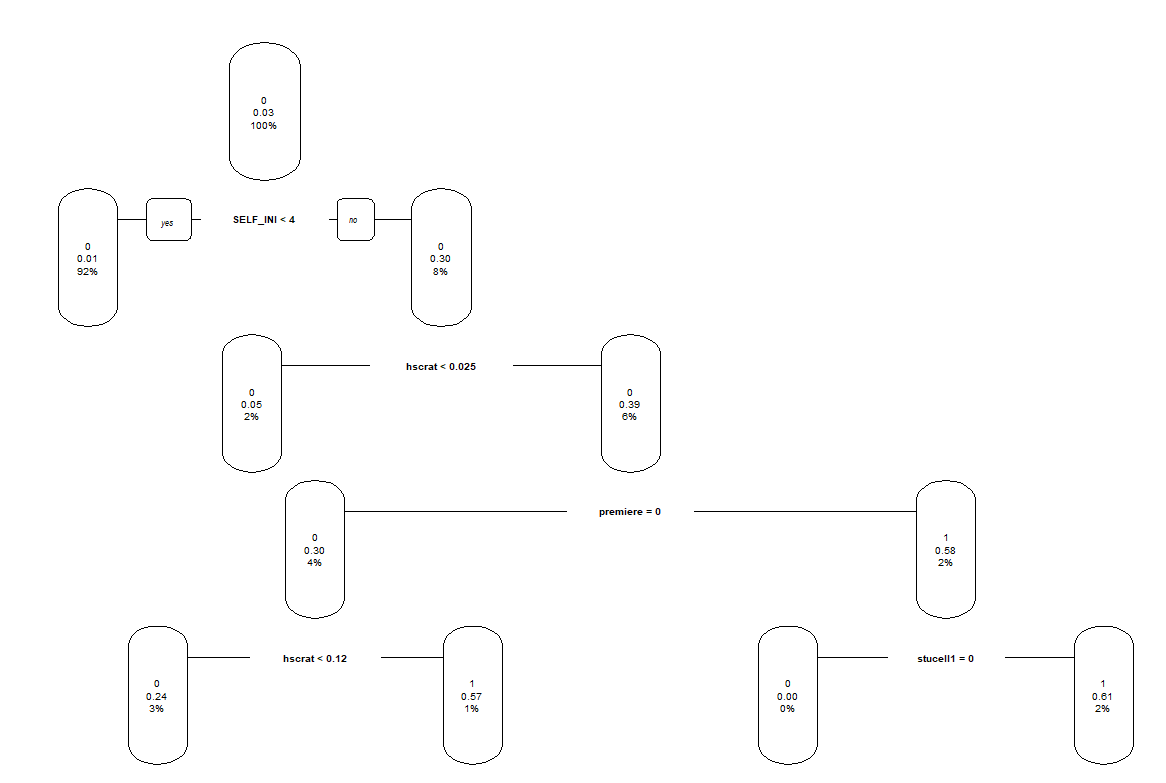
dfNK1$distance <- log10(dfNK1$distance+1) # Transforming distance to reduce skew and for proper distribution

dfNK1$avg\_income <- log10(dfNK1$avg\_income+1) # Transforming avg\_income to reduce skew and for proper distribution

1. Please provide the following results for each model:
   1. Model result summary for the regression model

|  |
| --- |
| > summary(regression.modelNK) # Summary of baseline model  Call:  NULL  Deviance Residuals:  Min 1Q Median 3Q Max  -5.6262 -0.1202 -0.0245 0.0000 3.2275  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -4.038e+01 1.259e+04 -0.003 0.997441  ETHNICITYB -1.322e-01 1.509e-01 -0.876 0.380888  ETHNICITYC -5.646e-02 1.153e-01 -0.490 0.624379  ETHNICITYH -7.995e-02 1.429e-01 -0.560 0.575677  ETHNICITYI -8.812e-01 4.248e-01 -2.074 0.038042 \*  ETHNICITYN -2.340e+00 2.379e-01 -9.837 < 2e-16 \*\*\*  ETHNICITYO -2.156e-01 2.263e-01 -0.953 0.340843  TERRITORY1 1.771e+01 1.259e+04 0.001 0.998878  TERRITORY2 1.767e+01 1.259e+04 0.001 0.998881  TERRITORY3 1.756e+01 1.259e+04 0.001 0.998887  TERRITORY4 1.750e+01 1.259e+04 0.001 0.998891  TERRITORY5 1.790e+01 1.259e+04 0.001 0.998866  TERRITORY6 1.737e+01 1.259e+04 0.001 0.998899  TERRITORY7 1.755e+01 1.259e+04 0.001 0.998888  TERRITORY8 1.724e+01 1.259e+04 0.001 0.998907  TERRITORYA 1.761e+01 1.259e+04 0.001 0.998884  TERRITORYN 1.687e+01 1.259e+04 0.001 0.998931  TOTAL\_CONTACTS -1.710e+01 1.519e+04 -0.001 0.999102  SELF\_INIT\_CNTCTS 1.746e+01 1.519e+04 0.001 0.999083  TRAVEL\_INIT\_CNTCTS 1.751e+01 1.519e+04 0.001 0.999080  SOLICITED\_CNTCTS 1.702e+01 1.519e+04 0.001 0.999106  REFERRAL\_CNTCTS 1.744e+01 1.519e+04 0.001 0.999084  CAMPUS\_VISIT 9.808e-01 8.828e-02 11.111 < 2e-16 \*\*\*  satscore 2.133e-03 2.552e-04 8.359 < 2e-16 \*\*\*  sex1 -2.521e-01 6.641e-02 -3.796 0.000147 \*\*\*  mailq2 -1.751e-01 1.146e-01 -1.529 0.126359  mailq3 -4.357e-01 1.629e-01 -2.674 0.007498 \*\*  mailq4 -1.621e+01 6.488e+03 -0.002 0.998007  mailq5 5.764e-01 9.885e-02 5.831 5.50e-09 \*\*\*  telecq2 -1.198e+00 1.296e-01 -9.248 < 2e-16 \*\*\*  telecq3 -1.860e+00 2.033e-01 -9.149 < 2e-16 \*\*\*  telecq4 -2.745e+00 2.482e-01 -11.063 < 2e-16 \*\*\*  premiere1 1.155e+00 8.838e-02 13.069 < 2e-16 \*\*\*  interest 3.288e-01 7.037e-02 4.672 2.99e-06 \*\*\*  stucell1 1.803e+01 1.760e+02 0.102 0.918401  init\_span -5.153e-02 4.892e-03 -10.533 < 2e-16 \*\*\*  int1rat 3.428e+00 1.009e+00 3.396 0.000683 \*\*\*  int2rat 6.573e+00 1.049e+00 6.265 3.73e-10 \*\*\*  hscrat 1.195e+01 4.596e-01 25.994 < 2e-16 \*\*\*  avg\_income -8.055e-02 2.131e-01 -0.378 0.705460  distance -2.966e-01 1.144e-01 -2.592 0.009554 \*\*  InstateY 1.938e-01 1.323e-01 1.465 0.143013  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 16842.0 on 50349 degrees of freedom  Residual deviance: 7104.3 on 50308 degrees of freedom  (13688 observations deleted due to missingness)  AIC: 7188.3  Number of Fisher Scoring iterations: 20  >  >  > #Evaluation model performance using the validation dataset  >  > #Criteria 1: the confusion matrix  > prediction <- predict(regression.modelNK,newdata=dfNK1.valid)  >  > #Need to remove missing values from the validation dataset for evaluation  > dfNK1.valid.nonmissing <- na.omit(dfNK1.valid)  > confusionMatrix(prediction,dfNK1.valid.nonmissing$Enroll)  Confusion Matrix and Statistics  Reference  Prediction 0 1  0 20368 470  1 195 388    Accuracy : 0.969  95% CI : (0.9665, 0.9712)  No Information Rate : 0.9599  P-Value [Acc > NIR] : 1.604e-12    Kappa : 0.5231    Mcnemar's Test P-Value : < 2.2e-16    Sensitivity : 0.9905  Specificity : 0.4522  Pos Pred Value : 0.9774  Neg Pred Value : 0.6655  Prevalence : 0.9599  Detection Rate : 0.9508  Detection Prevalence : 0.9728  Balanced Accuracy : 0.7214    'Positive' Class : 0    >  > #Criteria 2: the ROC curve and area under the curve  > pred.probabilities <- predict(regression.modelNK,newdata=dfNK1.valid,type='prob')  >  > regression.ROC <- roc(predictor=pred.probabilities$`1`,  + response=dfNK1.valid.nonmissing$Enroll ,  + levels=levels(dfNK1.valid.nonmissing$Enroll))  Setting direction: controls < cases  > plot(regression.ROC)  > regression.ROC$auc  Area under the curve: 0.9746 |
|  |
| |  | | --- | | > | |

* 1. Tree plot for the decision tree model



1. Which model will you choose? Why? Please provide support for your answer.
   * + - I will choose regression model. Based on the accuracy and ROC curve which are 97% and 97% resp. compared to Decision Tree where the accuracy is 97% but the ROC curve is 87% defines that regression model prediction performance is better.
2. Please explain and summarize your major findings to the director of the Office of Enrollment Management.
   * + - ETHNICITY type N, Campus visit, sex type 1(which is Male), Sat Score, Mail and Telecounciling qualifying score (MAILQ, TELECQ), One who attended campus recruitment event (PREMIERE=1) and academic interest are the major factors to consider while targeting the new enrollment.
       - INSTATE, Avg\_income and Distance though these looks significant factors but are not majorly impacted on the students decision of enrollment.
3. Attach your R codes at the end of the documents.

|  |  |
| --- | --- |
| **Name** | **Description** |
| **ACADEMIC\_INTEREST\_1** | Primary academic interest code |
| **ACADEMIC\_INTEREST\_2** | Secondary academic interest code |
| **CAMPUS\_VISIT** | Campus visit code |
| **CONTACT\_CODE1** | First contact code |
| **CONTACT\_DATE1** | First contact date |
| **ETHNICITY** | Ethnicity |
| **ENROLL** | 1=Enrolled F2014, 0=Not enrolled F2014 |
| **IRSCHOOL** | High school code |
| **INSTATE** | 1=In state, 0=Out of state |
| **LEVEL\_YEAR** | Student academic level |
| **REFERRAL\_CNTCTS** | Referral contact count |
| **SELF\_INIT\_CNTCTS** | Self-initiated contact count |
| **SOLICITED\_CNTCTS** | Solicited contact count |
| **TERRITORY** | Recruitment area |
| **TOTAL\_CONTACTS** | Total contact count |
| **TRAVEL\_INIT\_CNTCTS** | Travel initiated contact count |
| **AVG\_INCOME** | Commercial HH income estimate |
| **DISTANCE** | Distance from university |
| **HSCRAT** | 5-year high school enrollment rate |
| **INIT\_SPAN** | Time from first contact to enrollment date |
| **INT1RAT** | 5-year primary interest code rate |
| **INT2RAT** | 5-year secondary interest code rate |
| **INTEREST** | Number of indicated extracurricular interests |
| **MAILQ** | Mail qualifying score (1=very interested) |
| **PREMIERE** | 1=Attended campus recruitment event, 0=Did not |
| **SATSCORE** | SAT (original) score |
| **SEX** | Sex |
| **STUCELL** | 1=Have a cell phone, 0=Do not |
| **TELECQ** | Telecounciling qualifying score (1=very interested) |

#=====================================================================================================

################# REGRESSION MODEL ##################################################################

#=====================================================================================================

#install.packages('caret')

#install.packages('car')

#install.packages('e1071', dependencies=TRUE)

#install.packages('pROC')

#install.packages('dplyr')

#install.packages('Hmisc')

library(caret) # Load all required package

library(car)

library(pROC)

library(dplyr)

library(Hmisc)

getwd()

setwd("C:\\Users\\P2190101\\Desktop\\NK Personal\\NK Study\\UNCC\\6211 - Advanced BI\\Assignment 1") # Set working directory

getwd()

dfNK <- read.csv('inq2015.csv',na.strings = c('NA','')) # Read the csv file

summary(dfNK) # Summary of new dataset

str(dfNK) # Structure of new dataset

ExcludeList <- names(dfNK) %in% c("ACADEMIC\_INTEREST\_1", "ACADEMIC\_INTEREST\_2", "CONTACT\_DATE","CONTACT\_CODE1","IRSCHOOL","LEVEL\_YEAR") # Exclude variables which are not required.

dfNK1 <- dfNK[!ExcludeList] # New dataset with excluded list of columns

dfNK1$Enroll <- factor(dfNK1$Enroll) # Convert Int variables to factor

dfNK1$sex <- factor(dfNK1$sex) # Convert Int variables to factor

dfNK1$Instate <- factor(dfNK1$Instate) # Convert Int variables to factor

dfNK1$premiere <- factor(dfNK1$premiere) # Convert Int variables to factor

dfNK1$stucell <- factor(dfNK1$stucell) # Convert Int variables to factor

dfNK1$mailq <- factor(dfNK1$mailq) # Convert Int variables to factor

dfNK1$telecq <- factor(dfNK1$telecq) # Convert Int variables to factor

vif(glm(formula=Enroll~.,family = binomial(link='logit'),data=dfNK1)) # Using VIF to test Multicollinearity

# Impute columns

dfNK1$satscore <-with(dfNK1,impute(satscore,mean)) # Quantitative - Continuous Variable

dfNK1$avg\_income <-with(dfNK1,impute(avg\_income,mean)) # Quantitative - Continuous Variable

dfNK1$distance <-with(dfNK1,impute(distance,mean)) # Quantitative - Continuous Variable

dfNK1$sex <-with(dfNK1,impute(sex,max)) # Categorical - Ordinal variable

dfNK1$telecq<-with(dfNK1,impute(telecq,max)) # Categorical - Ordinal variable

dfNK1$distance <- log10(dfNK1$distance+1) # Transforming distance to reduce skew and for proper distribution

dfNK1$avg\_income <- log10(dfNK1$avg\_income+1) # Transforming avg\_income to reduce skew and for proper distribution

summary(dfNK1) # Summary of new dataset

str(dfNK1) # Structure of new dataset

head(dfNK1) # Top rows of new dataset

set.seed(101)

trainIndex <- createDataPartition(dfNK1$Enroll, # Create an index to partition the data for training and validation

p=0.7,

list=FALSE,

times=1)

dfNK1.train <- dfNK1[trainIndex,] # Create Training Data

dfNK1.valid <-dfNK1[-trainIndex,] # Create Validation Data

regression.modelNK <- train(Enroll~., # Create baseline model with the training Dataset

data=dfNK1.train,

method='glm',

family='binomial',

na.action=na.pass)

summary(regression.modelNK) # Summary of baseline model

#Evaluation model performance using the validation dataset

#Criteria 1: the confusion matrix

prediction <- predict(regression.modelNK,newdata=dfNK1.valid)

#Need to remove missing values from the validation dataset for evaluation

dfNK1.valid.nonmissing <- na.omit(dfNK1.valid)

confusionMatrix(prediction,dfNK1.valid.nonmissing$Enroll)

#Criteria 2: the ROC curve and area under the curve

pred.probabilities <- predict(regression.modelNK,newdata=dfNK1.valid,type='prob')

regression.ROC <- roc(predictor=pred.probabilities$`1`,

response=dfNK1.valid.nonmissing$Enroll ,

levels=levels(dfNK1.valid.nonmissing$Enroll))

plot(regression.ROC)

regression.ROC$auc

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | | #=====================================================================================================  > ################# REGRESSION MODEL ##################################################################  > #=====================================================================================================  > #install.packages('caret')  > #install.packages('car')  > #install.packages('e1071', dependencies=TRUE)  > #install.packages('pROC')  > #install.packages('dplyr')  > #install.packages('Hmisc')  >  > library(caret) # Load all required package  > library(car)  > library(pROC)  > library(dplyr)  > library(Hmisc)  >  >  > getwd()  [1] "C:/Users/P2190101/Desktop/NK Personal/NK Study/UNCC/6211 - Advanced BI/Assignment 1"  >  > setwd("C:\\Users\\P2190101\\Desktop\\NK Personal\\NK Study\\UNCC\\6211 - Advanced BI\\Assignment 1") # Set working directory  >  > getwd()  [1] "C:/Users/P2190101/Desktop/NK Personal/NK Study/UNCC/6211 - Advanced BI/Assignment 1"  >  > dfNK <- read.csv('inq2015.csv',na.strings = c('NA','')) # Read the csv file  >  > summary(dfNK) # Summary of new dataset  ETHNICITY TERRITORY ACADEMIC\_INTEREST\_1 ACADEMIC\_INTEREST\_2 Enroll CONTACT\_DATE TOTAL\_CONTACTS SELF\_INIT\_CNTCTS  Length:91482 Length:91482 Length:91482 Length:91482 Min. :0.00000 Length:91482 Min. : 1.000 Min. : 0.000  Class :character Class :character Class :character Class :character 1st Qu.:0.00000 Class :character 1st Qu.: 1.000 1st Qu.: 0.000  Mode :character Mode :character Mode :character Mode :character Median :0.00000 Mode :character Median : 2.000 Median : 1.000  Mean :0.03135 Mean : 2.166 Mean : 1.214  3rd Qu.:0.00000 3rd Qu.: 2.000 3rd Qu.: 1.000  Max. :1.00000 Max. :58.000 Max. :56.000    TRAVEL\_INIT\_CNTCTS SOLICITED\_CNTCTS REFERRAL\_CNTCTS CAMPUS\_VISIT CONTACT\_CODE1 LEVEL\_YEAR IRSCHOOL satscore  Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.00000 Length:91482 Length:91482 Length:91482 Min. : 420  1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00000 Class :character Class :character Class :character 1st Qu.: 940  Median :0.0000 Median :0.0000 Median :0.0000 Median :0.00000 Mode :character Mode :character Mode :character Median :1070  Mean :0.3648 Mean :0.5472 Mean :0.0399 Mean :0.03467 Mean :1073  3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:1210  Max. :8.0000 Max. :9.0000 Max. :5.0000 Max. :2.00000 Max. :1600  NA's :64479  sex mailq telecq premiere interest stucell init\_span int1rat  Min. :0.000 Min. :1.000 Min. :1.00 Min. :0.00000 Min. :0.00000 Min. :0.0000 Min. :-216.00 Min. :0.00000  1st Qu.:0.000 1st Qu.:3.000 1st Qu.:2.00 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.: 13.00 1st Qu.:0.01718  Median :1.000 Median :5.000 Median :2.00 Median :0.00000 Median :0.00000 Median :0.0000 Median : 19.00 Median :0.04211  Mean :0.632 Mean :4.084 Mean :2.78 Mean :0.02891 Mean :0.05364 Mean :0.4899 Mean : 19.69 Mean :0.03709  3rd Qu.:1.000 3rd Qu.:5.000 3rd Qu.:4.00 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:1.0000 3rd Qu.: 25.00 3rd Qu.:0.04927  Max. :1.000 Max. :5.000 Max. :4.00 Max. :1.00000 Max. :3.00000 Max. :1.0000 Max. : 228.00 Max. :1.00000  NA's :3833 NA's :70880  int2rat hscrat avg\_income distance Instate  Min. :0.00000 Min. :0.00000 Min. : 4940 Min. : 0.417 Length:91482  1st Qu.:0.02038 1st Qu.:0.00000 1st Qu.: 32045 1st Qu.: 114.471 Class :character  Median :0.05667 Median :0.03333 Median : 42324 Median : 183.547 Mode :character  Mean :0.04290 Mean :0.03765 Mean : 47315 Mean : 380.428  3rd Qu.:0.05667 3rd Qu.:0.04803 3rd Qu.: 57314 3rd Qu.: 539.433  Max. :1.00000 Max. :1.00000 Max. :200001 Max. :4798.899  NA's :20929 NA's :19468  > str(dfNK) # Structure of new dataset  'data.frame': 91482 obs. of 29 variables:  $ ETHNICITY : chr "C" "C" "B" NA ...  $ TERRITORY : chr "N" "N" "N" "N" ...  $ ACADEMIC\_INTEREST\_1: chr NA NA NA NA ...  $ ACADEMIC\_INTEREST\_2: chr NA NA NA NA ...  $ Enroll : int 0 0 0 0 0 0 0 0 0 0 ...  $ CONTACT\_DATE : chr "21MAY14" "15FEB13" "09JUN12" "25NOV11" ...  $ TOTAL\_CONTACTS : int 1 1 2 3 1 1 1 1 2 1 ...  $ SELF\_INIT\_CNTCTS : int 0 0 2 2 1 0 0 1 2 1 ...  $ TRAVEL\_INIT\_CNTCTS : int 0 0 0 1 0 1 0 0 0 0 ...  $ SOLICITED\_CNTCTS : int 1 1 0 0 0 0 1 0 0 0 ...  $ REFERRAL\_CNTCTS : int 0 0 0 0 0 0 0 0 0 0 ...  $ CAMPUS\_VISIT : int 0 0 0 0 0 0 0 0 0 1 ...  $ CONTACT\_CODE1 : chr NA "C01" "LMI" "LDZ" ...  $ LEVEL\_YEAR : chr "FR14" "FR14" "FR14" "FR14" ...  $ IRSCHOOL : chr NA NA NA NA ...  $ satscore : int NA NA NA NA NA NA NA NA NA NA ...  $ sex : int 1 0 0 1 0 1 1 NA 1 0 ...  $ mailq : int 5 5 5 5 5 2 5 5 5 4 ...  $ telecq : int NA NA NA NA NA 3 NA NA NA NA ...  $ premiere : int 0 0 0 0 0 0 0 0 1 0 ...  $ interest : int 0 0 0 0 0 0 0 0 0 0 ...  $ stucell : int 0 1 0 1 1 1 1 1 0 0 ...  $ init\_span : int 4 19 27 34 45 46 29 27 36 34 ...  $ int1rat : num 0.0172 0.0172 0.0172 0.0172 0.0172 ...  $ int2rat : num 0.0204 0.0204 0.0204 0.0204 0.0204 ...  $ hscrat : num 0.0377 0.0377 0.0377 0.0377 0.0377 ...  $ avg\_income : int NA NA NA NA NA NA NA NA NA NA ...  $ distance : num NA NA NA NA NA NA NA NA NA NA ...  $ Instate : chr "N" "N" "N" "N" ...  >  > ExcludeList <- names(dfNK) %in% c("ACADEMIC\_INTEREST\_1", "ACADEMIC\_INTEREST\_2", "CONTACT\_DATE","CONTACT\_CODE1","IRSCHOOL","LEVEL\_YEAR") # Exclude variables which are not required.  >  > dfNK1 <- dfNK[!ExcludeList] # New dataset with excluded list of columns  >  > dfNK1$Enroll <- factor(dfNK1$Enroll) # Convert Int variables to factor  > dfNK1$sex <- factor(dfNK1$sex) # Convert Int variables to factor  > dfNK1$Instate <- factor(dfNK1$Instate) # Convert Int variables to factor  > dfNK1$premiere <- factor(dfNK1$premiere) # Convert Int variables to factor  > dfNK1$stucell <- factor(dfNK1$stucell) # Convert Int variables to factor  > dfNK1$mailq <- factor(dfNK1$mailq) # Convert Int variables to factor  > dfNK1$telecq <- factor(dfNK1$telecq) # Convert Int variables to factor  >  >  > vif(glm(formula=Enroll~.,family = binomial(link='logit'),data=dfNK1)) # Using VIF to test Multicollinearity  GVIF Df GVIF^(1/(2\*Df))  ETHNICITY 1.424793e+00 6 1.029942e+00  TERRITORY 2.235856e+00 8 1.051575e+00  TOTAL\_CONTACTS 7.708492e+11 1 8.779802e+05  SELF\_INIT\_CNTCTS 5.192997e+11 1 7.206245e+05  TRAVEL\_INIT\_CNTCTS 5.064545e+10 1 2.250454e+05  SOLICITED\_CNTCTS 6.998184e+10 1 2.645408e+05  REFERRAL\_CNTCTS 9.634156e+09 1 9.815374e+04  CAMPUS\_VISIT 1.320906e+00 1 1.149307e+00  satscore 1.327940e+00 1 1.152363e+00  sex 1.053580e+00 1 1.026440e+00  mailq 1.579077e+00 4 1.058767e+00  telecq 1.068871e+00 3 1.011162e+00  premiere 1.493108e+00 1 1.221928e+00  interest 1.055865e+00 1 1.027553e+00  stucell 1.000347e+00 1 1.000173e+00  init\_span 1.387989e+00 1 1.178129e+00  int1rat 1.145232e+00 1 1.070155e+00  int2rat 1.181703e+00 1 1.087062e+00  hscrat 1.207191e+00 1 1.098722e+00  avg\_income 1.334105e+00 1 1.155034e+00  distance 2.717773e+00 1 1.648567e+00  Instate 2.679944e+00 1 1.637053e+00  >  > # Impute columns  > dfNK1$satscore <-with(dfNK1,impute(satscore,mean)) # Quantitative - Continuous Variable  > dfNK1$avg\_income <-with(dfNK1,impute(avg\_income,mean)) # Quantitative - Continuous Variable  > dfNK1$distance <-with(dfNK1,impute(distance,mean)) # Quantitative - Continuous Variable  > dfNK1$sex <-with(dfNK1,impute(sex,max)) # Categorical - Ordinal variable  > dfNK1$telecq<-with(dfNK1,impute(telecq,max)) # Categorical - Ordinal variable  > dfNK1$distance <- log10(dfNK1$distance+1) # Transforming distance for proper distribution  > dfNK1$avg\_income <- log10(dfNK1$avg\_income+1) # Transforming distance for proper distribution  >  > summary(dfNK1) # Summary of new dataset  64479 values imputed to 1072.993  3833 values imputed to 1  70880 values imputed to 2  20929 values imputed to 4.675011  19468 values imputed to 2.581412  ETHNICITY TERRITORY Enroll TOTAL\_CONTACTS SELF\_INIT\_CNTCTS TRAVEL\_INIT\_CNTCTS SOLICITED\_CNTCTS REFERRAL\_CNTCTS  Length:91482 Length:91482 0:88614 Min. : 1.000 Min. : 0.000 Min. :0.0000 Min. :0.0000 Min. :0.0000  Class :character Class :character 1: 2868 1st Qu.: 1.000 1st Qu.: 0.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000  Mode :character Mode :character Median : 2.000 Median : 1.000 Median :0.0000 Median :0.0000 Median :0.0000  Mean : 2.166 Mean : 1.214 Mean :0.3648 Mean :0.5472 Mean :0.0399  3rd Qu.: 2.000 3rd Qu.: 1.000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.0000  Max. :58.000 Max. :56.000 Max. :8.0000 Max. :9.0000 Max. :5.0000  CAMPUS\_VISIT satscore sex mailq telecq premiere interest stucell init\_span int1rat  Min. :0.00000 Min. : 420 0:32289 1: 7981 1: 763 0:88837 Min. :0.00000 0:46663 Min. :-216.00 Min. :0.00000  1st Qu.:0.00000 1st Qu.:1073 1:59193 2:11709 2:80687 1: 2645 1st Qu.:0.00000 1:44819 1st Qu.: 13.00 1st Qu.:0.01718  Median :0.00000 Median :1073 3: 8344 3: 3265 Median :0.00000 Median : 19.00 Median :0.04211  Mean :0.03467 Mean :1073 4: 24 4: 6767 Mean :0.05364 Mean : 19.69 Mean :0.03709  3rd Qu.:0.00000 3rd Qu.:1073 5:63424 3rd Qu.:0.00000 3rd Qu.: 25.00 3rd Qu.:0.04927  Max. :2.00000 Max. :1600 Max. :3.00000 Max. : 228.00 Max. :1.00000  int2rat hscrat avg\_income distance Instate  Min. :0.00000 Min. :0.00000 Min. :3.694 Min. :0.1514 N:34722  1st Qu.:0.02038 1st Qu.:0.00000 1st Qu.:4.541 1st Qu.:2.1696 Y:56760  Median :0.05667 Median :0.03333 Median :4.675 Median :2.5192  Mean :0.04290 Mean :0.03765 Mean :4.647 Mean :2.4228  3rd Qu.:0.05667 3rd Qu.:0.04803 3rd Qu.:4.716 3rd Qu.:2.5814  Max. :1.00000 Max. :1.00000 Max. :5.301 Max. :3.6812  > str(dfNK1) # Structure of new dataset  'data.frame': 91482 obs. of 23 variables:  $ ETHNICITY : chr "C" "C" "B" NA ...  $ TERRITORY : chr "N" "N" "N" "N" ...  $ Enroll : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  $ TOTAL\_CONTACTS : int 1 1 2 3 1 1 1 1 2 1 ...  $ SELF\_INIT\_CNTCTS : int 0 0 2 2 1 0 0 1 2 1 ...  $ TRAVEL\_INIT\_CNTCTS: int 0 0 0 1 0 1 0 0 0 0 ...  $ SOLICITED\_CNTCTS : int 1 1 0 0 0 0 1 0 0 0 ...  $ REFERRAL\_CNTCTS : int 0 0 0 0 0 0 0 0 0 0 ...  $ CAMPUS\_VISIT : int 0 0 0 0 0 0 0 0 0 1 ...  $ satscore : 'impute' num 1073 1073 1073 1073 1073 ...  ..- attr(\*, "imputed")= int [1:64479] 1 2 3 4 5 6 7 8 9 10 ...  $ sex : Factor w/ 2 levels "0","1": 2 1 1 2 1 2 2 2 2 1 ...  ..- attr(\*, "imputed")= int [1:3833] 8 20 29 46 82 87 90 99 103 108 ...  $ mailq : Factor w/ 5 levels "1","2","3","4",..: 5 5 5 5 5 2 5 5 5 4 ...  $ telecq : Factor w/ 4 levels "1","2","3","4": 2 2 2 2 2 3 2 2 2 2 ...  ..- attr(\*, "imputed")= int [1:70880] 1 2 3 4 5 7 8 9 10 11 ...  $ premiere : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 ...  $ interest : int 0 0 0 0 0 0 0 0 0 0 ...  $ stucell : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 2 2 1 1 ...  $ init\_span : int 4 19 27 34 45 46 29 27 36 34 ...  $ int1rat : num 0.0172 0.0172 0.0172 0.0172 0.0172 ...  $ int2rat : num 0.0204 0.0204 0.0204 0.0204 0.0204 ...  $ hscrat : num 0.0377 0.0377 0.0377 0.0377 0.0377 ...  $ avg\_income : 'impute' num 4.68 4.68 4.68 4.68 4.68 ...  ..- attr(\*, "imputed")= int [1:20929] 1 2 3 4 5 6 7 8 9 10 ...  $ distance : 'impute' num 2.58 2.58 2.58 2.58 2.58 ...  ..- attr(\*, "imputed")= int [1:19468] 1 2 3 4 5 6 7 8 9 10 ...  $ Instate : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 1 1 ...  > head(dfNK1) # Top rows of new dataset  ETHNICITY TERRITORY Enroll TOTAL\_CONTACTS SELF\_INIT\_CNTCTS TRAVEL\_INIT\_CNTCTS SOLICITED\_CNTCTS REFERRAL\_CNTCTS CAMPUS\_VISIT satscore sex mailq  1 C N 0 1 0 0 1 0 0 1072.993 1 5  2 C N 0 1 0 0 1 0 0 1072.993 0 5  3 B N 0 2 2 0 0 0 0 1072.993 0 5  4 <NA> N 0 3 2 1 0 0 0 1072.993 1 5  5 <NA> N 0 1 1 0 0 0 0 1072.993 0 5  6 N N 0 1 0 1 0 0 0 1072.993 1 2  telecq premiere interest stucell init\_span int1rat int2rat hscrat avg\_income distance Instate  1 2 0 0 0 4 0.01718275 0.02038 0.03765163 4.675011 2.581412 N  2 2 0 0 1 19 0.01718275 0.02038 0.03765163 4.675011 2.581412 N  3 2 0 0 0 27 0.01718275 0.02038 0.03765163 4.675011 2.581412 N  4 2 0 0 1 34 0.01718275 0.02038 0.03765163 4.675011 2.581412 N  5 2 0 0 1 45 0.01718275 0.02038 0.03765163 4.675011 2.581412 N  6 3 0 0 1 46 0.01718275 0.02038 0.03765163 4.675011 2.581412 N  >  > set.seed(101)  >  > trainIndex <- createDataPartition(dfNK1$Enroll, # Create an index to partition the data for training and validation  + p=0.7,  + list=FALSE,  + times=1)  > dfNK1.train <- dfNK1[trainIndex,] # Create Training Data  > dfNK1.valid <-dfNK1[-trainIndex,] # Create Validation Data  >  >  > regression.modelNK <- train(Enroll~., # Create baseline model with the training Dataset  + data=dfNK1.train,  + method='glm',  + family='binomial',  + na.action=na.pass)  There were 26 warnings (use warnings() to see them)  >  > summary(regression.modelNK) # Summary of baseline model  Call:  NULL  Deviance Residuals:  Min 1Q Median 3Q Max  -5.6262 -0.1202 -0.0245 0.0000 3.2275  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -4.038e+01 1.259e+04 -0.003 0.997441  ETHNICITYB -1.322e-01 1.509e-01 -0.876 0.380888  ETHNICITYC -5.646e-02 1.153e-01 -0.490 0.624379  ETHNICITYH -7.995e-02 1.429e-01 -0.560 0.575677  ETHNICITYI -8.812e-01 4.248e-01 -2.074 0.038042 \*  ETHNICITYN -2.340e+00 2.379e-01 -9.837 < 2e-16 \*\*\*  ETHNICITYO -2.156e-01 2.263e-01 -0.953 0.340843  TERRITORY1 1.771e+01 1.259e+04 0.001 0.998878  TERRITORY2 1.767e+01 1.259e+04 0.001 0.998881  TERRITORY3 1.756e+01 1.259e+04 0.001 0.998887  TERRITORY4 1.750e+01 1.259e+04 0.001 0.998891  TERRITORY5 1.790e+01 1.259e+04 0.001 0.998866  TERRITORY6 1.737e+01 1.259e+04 0.001 0.998899  TERRITORY7 1.755e+01 1.259e+04 0.001 0.998888  TERRITORY8 1.724e+01 1.259e+04 0.001 0.998907  TERRITORYA 1.761e+01 1.259e+04 0.001 0.998884  TERRITORYN 1.687e+01 1.259e+04 0.001 0.998931  TOTAL\_CONTACTS -1.710e+01 1.519e+04 -0.001 0.999102  SELF\_INIT\_CNTCTS 1.746e+01 1.519e+04 0.001 0.999083  TRAVEL\_INIT\_CNTCTS 1.751e+01 1.519e+04 0.001 0.999080  SOLICITED\_CNTCTS 1.702e+01 1.519e+04 0.001 0.999106  REFERRAL\_CNTCTS 1.744e+01 1.519e+04 0.001 0.999084  CAMPUS\_VISIT 9.808e-01 8.828e-02 11.111 < 2e-16 \*\*\*  satscore 2.133e-03 2.552e-04 8.359 < 2e-16 \*\*\*  sex1 -2.521e-01 6.641e-02 -3.796 0.000147 \*\*\*  mailq2 -1.751e-01 1.146e-01 -1.529 0.126359  mailq3 -4.357e-01 1.629e-01 -2.674 0.007498 \*\*  mailq4 -1.621e+01 6.488e+03 -0.002 0.998007  mailq5 5.764e-01 9.885e-02 5.831 5.50e-09 \*\*\*  telecq2 -1.198e+00 1.296e-01 -9.248 < 2e-16 \*\*\*  telecq3 -1.860e+00 2.033e-01 -9.149 < 2e-16 \*\*\*  telecq4 -2.745e+00 2.482e-01 -11.063 < 2e-16 \*\*\*  premiere1 1.155e+00 8.838e-02 13.069 < 2e-16 \*\*\*  interest 3.288e-01 7.037e-02 4.672 2.99e-06 \*\*\*  stucell1 1.803e+01 1.760e+02 0.102 0.918401  init\_span -5.153e-02 4.892e-03 -10.533 < 2e-16 \*\*\*  int1rat 3.428e+00 1.009e+00 3.396 0.000683 \*\*\*  int2rat 6.573e+00 1.049e+00 6.265 3.73e-10 \*\*\*  hscrat 1.195e+01 4.596e-01 25.994 < 2e-16 \*\*\*  avg\_income -8.055e-02 2.131e-01 -0.378 0.705460  distance -2.966e-01 1.144e-01 -2.592 0.009554 \*\*  InstateY 1.938e-01 1.323e-01 1.465 0.143013  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 16842.0 on 50349 degrees of freedom  Residual deviance: 7104.3 on 50308 degrees of freedom  (13688 observations deleted due to missingness)  AIC: 7188.3  Number of Fisher Scoring iterations: 20  >  >  > #Evaluation model performance using the validation dataset  >  > #Criteria 1: the confusion matrix  > prediction <- predict(regression.modelNK,newdata=dfNK1.valid)  >  > #Need to remove missing values from the validation dataset for evaluation  > dfNK1.valid.nonmissing <- na.omit(dfNK1.valid)  > confusionMatrix(prediction,dfNK1.valid.nonmissing$Enroll)  Confusion Matrix and Statistics  Reference  Prediction 0 1  0 20368 470  1 195 388    Accuracy : 0.969  95% CI : (0.9665, 0.9712)  No Information Rate : 0.9599  P-Value [Acc > NIR] : 1.604e-12    Kappa : 0.5231    Mcnemar's Test P-Value : < 2.2e-16    Sensitivity : 0.9905  Specificity : 0.4522  Pos Pred Value : 0.9774  Neg Pred Value : 0.6655  Prevalence : 0.9599  Detection Rate : 0.9508  Detection Prevalence : 0.9728  Balanced Accuracy : 0.7214    'Positive' Class : 0    >  > #Criteria 2: the ROC curve and area under the curve  > pred.probabilities <- predict(regression.modelNK,newdata=dfNK1.valid,type='prob')  >  > regression.ROC <- roc(predictor=pred.probabilities$`1`,  + response=dfNK1.valid.nonmissing$Enroll ,  + levels=levels(dfNK1.valid.nonmissing$Enroll))  Setting direction: controls < cases  > plot(regression.ROC)  > regression.ROC$auc  Area under the curve: 0.9746 | |  | | |  | | --- | | > | | |
|  |
| |  | | --- | | > | |

#=====================================================================================================

################# DECISION TREE #####################################################################

#=====================================================================================================

#install.packages('caret')

#install.packages('car')

#install.packages('e1071', dependencies=TRUE)

#install.packages('pROC')

#install.packages('dplyr')

#install.packages('Hmisc')

library(caret) # Load all required package

library(car)

library(pROC)

library(dplyr)

library(Hmisc)

library(rpart)

library(rpart.plot)

getwd()

setwd("C:\\Users\\P2190101\\Desktop\\NK Personal\\NK Study\\UNCC\\6211 - Advanced BI\\Assignment 1") # Set working directory

getwd()

dfNK <- read.csv('inq2015.csv',na.strings = c('NA','')) # Read the csv file

summary(dfNK) # Summary of new dataset

str(dfNK) # Structure of new dataset

ExcludeList <- names(dfNK) %in% c("ACADEMIC\_INTEREST\_1", "ACADEMIC\_INTEREST\_2", "CONTACT\_DATE","CONTACT\_CODE1","IRSCHOOL","LEVEL\_YEAR") # Exclude variables which are not required.

dfNK1 <- dfNK[!ExcludeList] # New dataset with excluded list of columns

dfNK1$Enroll <- factor(dfNK1$Enroll) # Convert Int variables to factor

dfNK1$sex <- factor(dfNK1$sex) # Convert Int variables to factor

dfNK1$Instate <- factor(dfNK1$Instate) # Convert Int variables to factor

dfNK1$premiere <- factor(dfNK1$premiere) # Convert Int variables to factor

dfNK1$stucell <- factor(dfNK1$stucell) # Convert Int variables to factor

dfNK1$mailq <- factor(dfNK1$mailq) # Convert Int variables to factor

dfNK1$telecq <- factor(dfNK1$telecq) # Convert Int variables to factor

set.seed(101)

trainIndex <- createDataPartition(dfNK1$Enroll, # Create an index to partition the data for training and validation

p=0.7,

list=FALSE,

times=1)

dfNK1.train <- dfNK1[trainIndex,] # Create Training Data

dfNK1.valid <-dfNK1[-trainIndex,] # Create Validation Data

tree.modelNK <- train(Enroll~.,

data=dfNK1.train,

method='rpart',

na.action=na.pass)

tree.modelNK

prp(tree.modelNK$finalModel,type=2,extra=106)

prediction.tree <- predict(tree.modelNK,

newdata = dfNK1.valid,

na.action=na.pass)

confusionMatrix(prediction.tree,dfNK1.valid$Enroll)

tree.probabliities <- predict(tree.modelNK,

newdata=dfNK1.valid,

type='prob',

na.action=na.pass)

tree.ROC <-roc(predictor=tree.probabliities$`1`,

response = dfNK1.valid$Enroll,

levels=levels(dfNK1.valid$Enroll))

plot(tree.ROC)

tree.ROC$auc

> #=====================================================================================================

> ################# DECISION TREE #####################################################################

> #=====================================================================================================

>

> #install.packages('caret')

> #install.packages('car')

> #install.packages('e1071', dependencies=TRUE)

> #install.packages('pROC')

> #install.packages('dplyr')

> #install.packages('Hmisc')

>

> library(caret) # Load all required package

> library(car)

> library(pROC)

> library(dplyr)

> library(Hmisc)

> library(rpart)

> library(rpart.plot)

>

>

> getwd()

[1] "C:/Users/P2190101/Desktop/NK Personal/NK Study/UNCC/6211 - Advanced BI/Assignment 1"

>

> setwd("C:\\Users\\P2190101\\Desktop\\NK Personal\\NK Study\\UNCC\\6211 - Advanced BI\\Assignment 1") # Set working directory

>

> getwd()

[1] "C:/Users/P2190101/Desktop/NK Personal/NK Study/UNCC/6211 - Advanced BI/Assignment 1"

>

> dfNK <- read.csv('inq2015.csv',na.strings = c('NA','')) # Read the csv file

>

> summary(dfNK) # Summary of new dataset

ETHNICITY TERRITORY ACADEMIC\_INTEREST\_1 ACADEMIC\_INTEREST\_2 Enroll CONTACT\_DATE TOTAL\_CONTACTS SELF\_INIT\_CNTCTS

Length:91482 Length:91482 Length:91482 Length:91482 Min. :0.00000 Length:91482 Min. : 1.000 Min. : 0.000

Class :character Class :character Class :character Class :character 1st Qu.:0.00000 Class :character 1st Qu.: 1.000 1st Qu.: 0.000

Mode :character Mode :character Mode :character Mode :character Median :0.00000 Mode :character Median : 2.000 Median : 1.000

Mean :0.03135 Mean : 2.166 Mean : 1.214

3rd Qu.:0.00000 3rd Qu.: 2.000 3rd Qu.: 1.000

Max. :1.00000 Max. :58.000 Max. :56.000

TRAVEL\_INIT\_CNTCTS SOLICITED\_CNTCTS REFERRAL\_CNTCTS CAMPUS\_VISIT CONTACT\_CODE1 LEVEL\_YEAR IRSCHOOL satscore

Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.00000 Length:91482 Length:91482 Length:91482 Min. : 420

1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00000 Class :character Class :character Class :character 1st Qu.: 940

Median :0.0000 Median :0.0000 Median :0.0000 Median :0.00000 Mode :character Mode :character Mode :character Median :1070

Mean :0.3648 Mean :0.5472 Mean :0.0399 Mean :0.03467 Mean :1073

3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:1210

Max. :8.0000 Max. :9.0000 Max. :5.0000 Max. :2.00000 Max. :1600

NA's :64479

sex mailq telecq premiere interest stucell init\_span int1rat

Min. :0.000 Min. :1.000 Min. :1.00 Min. :0.00000 Min. :0.00000 Min. :0.0000 Min. :-216.00 Min. :0.00000

1st Qu.:0.000 1st Qu.:3.000 1st Qu.:2.00 1st Qu.:0.00000 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.: 13.00 1st Qu.:0.01718

Median :1.000 Median :5.000 Median :2.00 Median :0.00000 Median :0.00000 Median :0.0000 Median : 19.00 Median :0.04211

Mean :0.632 Mean :4.084 Mean :2.78 Mean :0.02891 Mean :0.05364 Mean :0.4899 Mean : 19.69 Mean :0.03709

3rd Qu.:1.000 3rd Qu.:5.000 3rd Qu.:4.00 3rd Qu.:0.00000 3rd Qu.:0.00000 3rd Qu.:1.0000 3rd Qu.: 25.00 3rd Qu.:0.04927

Max. :1.000 Max. :5.000 Max. :4.00 Max. :1.00000 Max. :3.00000 Max. :1.0000 Max. : 228.00 Max. :1.00000

NA's :3833 NA's :70880

int2rat hscrat avg\_income distance Instate

Min. :0.00000 Min. :0.00000 Min. : 4940 Min. : 0.417 Length:91482

1st Qu.:0.02038 1st Qu.:0.00000 1st Qu.: 32045 1st Qu.: 114.471 Class :character

Median :0.05667 Median :0.03333 Median : 42324 Median : 183.547 Mode :character

Mean :0.04290 Mean :0.03765 Mean : 47315 Mean : 380.428

3rd Qu.:0.05667 3rd Qu.:0.04803 3rd Qu.: 57314 3rd Qu.: 539.433

Max. :1.00000 Max. :1.00000 Max. :200001 Max. :4798.899

NA's :20929 NA's :19468

> str(dfNK) # Structure of new dataset

'data.frame': 91482 obs. of 29 variables:

$ ETHNICITY : chr "C" "C" "B" NA ...

$ TERRITORY : chr "N" "N" "N" "N" ...

$ ACADEMIC\_INTEREST\_1: chr NA NA NA NA ...

$ ACADEMIC\_INTEREST\_2: chr NA NA NA NA ...

$ Enroll : int 0 0 0 0 0 0 0 0 0 0 ...

$ CONTACT\_DATE : chr "21MAY14" "15FEB13" "09JUN12" "25NOV11" ...

$ TOTAL\_CONTACTS : int 1 1 2 3 1 1 1 1 2 1 ...

$ SELF\_INIT\_CNTCTS : int 0 0 2 2 1 0 0 1 2 1 ...

$ TRAVEL\_INIT\_CNTCTS : int 0 0 0 1 0 1 0 0 0 0 ...

$ SOLICITED\_CNTCTS : int 1 1 0 0 0 0 1 0 0 0 ...

$ REFERRAL\_CNTCTS : int 0 0 0 0 0 0 0 0 0 0 ...

$ CAMPUS\_VISIT : int 0 0 0 0 0 0 0 0 0 1 ...

$ CONTACT\_CODE1 : chr NA "C01" "LMI" "LDZ" ...

$ LEVEL\_YEAR : chr "FR14" "FR14" "FR14" "FR14" ...

$ IRSCHOOL : chr NA NA NA NA ...

$ satscore : int NA NA NA NA NA NA NA NA NA NA ...

$ sex : int 1 0 0 1 0 1 1 NA 1 0 ...

$ mailq : int 5 5 5 5 5 2 5 5 5 4 ...

$ telecq : int NA NA NA NA NA 3 NA NA NA NA ...

$ premiere : int 0 0 0 0 0 0 0 0 1 0 ...

$ interest : int 0 0 0 0 0 0 0 0 0 0 ...

$ stucell : int 0 1 0 1 1 1 1 1 0 0 ...

$ init\_span : int 4 19 27 34 45 46 29 27 36 34 ...

$ int1rat : num 0.0172 0.0172 0.0172 0.0172 0.0172 ...

$ int2rat : num 0.0204 0.0204 0.0204 0.0204 0.0204 ...

$ hscrat : num 0.0377 0.0377 0.0377 0.0377 0.0377 ...

$ avg\_income : int NA NA NA NA NA NA NA NA NA NA ...

$ distance : num NA NA NA NA NA NA NA NA NA NA ...

$ Instate : chr "N" "N" "N" "N" ...

>

> ExcludeList <- names(dfNK) %in% c("ACADEMIC\_INTEREST\_1", "ACADEMIC\_INTEREST\_2", "CONTACT\_DATE","CONTACT\_CODE1","IRSCHOOL","LEVEL\_YEAR") # Exclude variables which are not required.

>

> dfNK1 <- dfNK[!ExcludeList] # New dataset with excluded list of columns

>

> dfNK1$Enroll <- factor(dfNK1$Enroll) # Convert Int variables to factor

> dfNK1$sex <- factor(dfNK1$sex) # Convert Int variables to factor

> dfNK1$Instate <- factor(dfNK1$Instate) # Convert Int variables to factor

> dfNK1$premiere <- factor(dfNK1$premiere) # Convert Int variables to factor

> dfNK1$stucell <- factor(dfNK1$stucell) # Convert Int variables to factor

> dfNK1$mailq <- factor(dfNK1$mailq) # Convert Int variables to factor

> dfNK1$telecq <- factor(dfNK1$telecq) # Convert Int variables to factor

>

> set.seed(101)

>

> trainIndex <- createDataPartition(dfNK1$Enroll, # Create an index to partition the data for training and validation

+ p=0.7,

+ list=FALSE,

+ times=1)

> dfNK1.train <- dfNK1[trainIndex,] # Create Training Data

> dfNK1.valid <-dfNK1[-trainIndex,] # Create Validation Data

>

> tree.modelNK <- train(Enroll~.,

+ data=dfNK1.train,

+ method='rpart',

+ na.action=na.pass)

> tree.modelNK

CART

64038 samples

22 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 64038, 64038, 64038, 64038, 64038, 64038, ...

Resampling results across tuning parameters:

cp Accuracy Kappa

0.01543825 0.9729814 0.4703344

0.03187251 0.9713940 0.3632696

0.03203851 0.9713940 0.3632696

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was cp = 0.01543825.

>

> prp(tree.modelNK$finalModel,type=2,extra=106)

>

> prediction.tree <- predict(tree.modelNK,

+ newdata = dfNK1.valid,

+ na.action=na.pass)

>

> confusionMatrix(prediction.tree,dfNK1.valid$Enroll)

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 26290 421

1 294 439

Accuracy : 0.9739

95% CI : (0.972, 0.9758)

No Information Rate : 0.9687

P-Value [Acc > NIR] : 1.323e-07

Kappa : 0.5378

Mcnemar's Test P-Value : 2.451e-06

Sensitivity : 0.9889

Specificity : 0.5105

Pos Pred Value : 0.9842

Neg Pred Value : 0.5989

Prevalence : 0.9687

Detection Rate : 0.9580

Detection Prevalence : 0.9733

Balanced Accuracy : 0.7497

'Positive' Class : 0

>

>

>

> tree.probabliities <- predict(tree.modelNK,

+ newdata=dfNK1.valid,

+ type='prob',

+ na.action=na.pass)

> tree.ROC <-roc(predictor=tree.probabliities$`1`,

+ response = dfNK1.valid$Enroll,

+ levels=levels(dfNK1.valid$Enroll))

Setting direction: controls < cases

> plot(tree.ROC)

>

> tree.ROC$auc

Area under the curve: 0.8714