**Homework\_3 Data Cleaning Part**

Sai Hruthik Reddy Varakantham-Dorababu Chintala

2022-04-05

## **R Markdown**

*#Retention Modeling At Scholastic Travel Company*

*#Before constructing our decision tree models and random forest related charts we first need to clean our data. The data can be cleaned for numerical data by replacing the NA values with the mean value.*

*#Whereas in case of categorical variables we can replace the NA values with the most occuring values in the column*

*#But sometimes they can also lead to outliers for which in the future we would need to use other packages*

*#First loading our file into a variable to create a dataframe and creating Backup\_xlsx to keep the raw data in place*

setwd("~/Desktop/Data\_Mining") *#Setting our working directory*

pacman::p\_load("pacman","rio","tidyverse","rpart","rpart.plot","readxl") *#Loading our packages*

Assignment\_3<-read\_xlsx("Homework-3.xlsx") *#Storing our raw data into The Assignment\_3 variable*

Backup\_xlsx <- Assignment\_3

Backup\_xlsx2<-Assignment\_3

The data cleaning can be done on the Assignment\_3 file

*#Data Cleaning*

*#Replace NA values with mean and converting into factor since grade is easier to work with when it is in factor format*

*#So removing NA values for From.Grade*

library("zoo")

##

## Attaching package: 'zoo'

## The following objects are masked from 'package:base':

##

## as.Date, as.Date.numeric

e1<- as.numeric(Assignment\_3$From.Grade)

## Warning: NAs introduced by coercion

e1<-na.aggregate(e1)

Assignment\_3$From.Grade<- as.factor(e1)

*#view(Assignment\_3)*

*#So removing NA values for To.Grade*

library("zoo")

e1<- as.numeric(Assignment\_3$To.Grade)

## Warning: NAs introduced by coercion

e1<-na.aggregate(e1)

Assignment\_3$To.Grade<- as.factor(e1)

*#view(Assignment\_3)*

*#There are no outliers in Group State*

sum(is.na(Assignment\_3$Group.State))

## [1] 0

table(Assignment\_3$Group.State)

##

## AB AK AL AR AZ

## 1 5 21 10 53

## Bermuda CA Cayman Islands CO CT

## 1 718 1 89 15

## 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

Assignment\_3$Group.State<-as.factor(Assignment\_3$Group.State)

*#There are no outliers in Is.Non.Annual.*

sum(is.na(Assignment\_3$Is.Non.Annual.))

## [1] 0

table(Assignment\_3$Is.Non.Annual.)

##

## 0 1

## 2021 368

*#There are no outliers in Days.*

sum(is.na(Assignment\_3$Days))

## [1] 0

table(Assignment\_3$Days)

##

## 1 2 3 4 5 6 7 8 9 10 11 12

## 77 84 269 621 907 264 111 34 6 5 10 1

*#There are no outliers in Travel.Type*

sum(is.na(Assignment\_3$Travel.Type))

## [1] 0

table(Assignment\_3$Travel.Type)

##

## A B N T

## 2014 367 2 6

*#There are no outliers in Travel.Type*

sum(is.na(Assignment\_3$Special.Pay))

## [1] 2

table(Assignment\_3$Special.Pay)

##

## CP FR NA SA

## 70 293 1917 107

a3<-as.character(Assignment\_3$Special.Pay)

e3<-replace\_na(a3,"NA")

Assignment\_3$Special.Pay<-e3

Assignment\_3$Special.Pay<-as.factor(Assignment\_3$Special.Pay)

*#view(Assignment\_3)*

*# For posix format*

Assignment\_3$Departure.Date <- as.Date(Assignment\_3$Departure.Date)

Assignment\_3$Return.Date <- as.Date(Assignment\_3$Return.Date)

Assignment\_3$Deposit.Date <- as.Date(Assignment\_3$Deposit.Date)

Assignment\_3$Early.RPL <- as.numeric(Assignment\_3$Early.RPL)

## Warning: NAs introduced by coercion

Assignment\_3$Early.RPL <- as.Date(Assignment\_3$Early.RPL, origin = "1899-12-30")

Assignment\_3$Latest.RPL <- as.numeric(Assignment\_3$Latest.RPL)

## Warning: NAs introduced by coercion

Assignment\_3$Latest.RPL <- as.Date(Assignment\_3$Latest.RPL, origin = "1899-12-30")

Assignment\_3$Initial.System.Date <- as.numeric(Assignment\_3$Initial.System.Date)

## Warning: NAs introduced by coercion

Assignment\_3$Initial.System.Date <- as.Date(Assignment\_3$Initial.System.Date, origin = "1899-12-30")

Assignment\_3$FirstMeeting <- as.numeric(Assignment\_3$FirstMeeting)

## Warning: NAs introduced by coercion

Assignment\_3$FirstMeeting <- as.Date(Assignment\_3$FirstMeeting, origin = "1899-12-30")

Assignment\_3$LastMeeting <- as.numeric(Assignment\_3$LastMeeting)

## Warning: NAs introduced by coercion

Assignment\_3$LastMeeting <- as.Date(Assignment\_3$LastMeeting, origin = "1899-12-30")

Assignment\_3 <- Assignment\_3 %>%

mutate(Column = pmin(Assignment\_3$Departure.Date,Assignment\_3$Return.Date,Assignment\_3$Deposit.Date,Assignment\_3$Early.RPL, Assignment\_3$Latest.RPL))

Assignment\_3$Initial.System.Date <- as.Date(ifelse(is.na(Assignment\_3$Initial.System.Date),Assignment\_3$Column,Assignment\_3$Initial.System.Date), origin = "1970-01-01" )

Assignment\_3 <- select(Assignment\_3, -Column)

*#There are no outliers in Tuition*

sum(is.na(Assignment\_3$Tuition))

## [1] 0

table(Assignment\_3$Tuition)

##

## 79 100 119 143 149 156 163 168 172 199 243 249 298 299 349 358

## 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1

## 359 369 374 375 376 380 382 383 384 385 386 389 390 392 393 394

## 2 3 2 3 1 1 1 3 1 3 3 1 1 3 3 1

## 395 396 398 399 400 402 406 408 409 410 415 417 419 420 424 425

## 3 2 4 11 2 1 1 3 6 1 2 1 5 2 2 2

## 428 429 435 438 439 440 441 443 445 449 453 457 459 460 461 463

## 2 4 1 1 7 1 1 1 1 10 1 1 2 2 1 1

## 469 470 475 479 480 488 489 496 499 515 516 525 537 539 551 552

## 1 1 1 2 1 1 3 1 4 1 1 1 1 2 1 1

## 554 555 557 558 559 566 567 570 573 574 575 578 579 583 584 585

## 1 1 2 1 2 2 1 1 2 1 2 2 1 2 4 1

## 587 588 589 592 593 595 596 598 599 607 609 610 612 613 614 2867

## 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1

## 2870 2894 2899 2907 2948 2949 2953 2976 2989 2990 3027 3050 3074 3085 3089 3098

## 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1

## 3110 3114 3129 3224 3349 3379 3439 3479 3499 3628 3799 3884 4199 4200

## 1 1 1 1 1 1 1 1 1 1 1 7 1 1

*#There are no outliers in FRP.active*

sum(is.na(Assignment\_3$FRP.Active))

## [1] 0

table(Assignment\_3$FRP.Active)

##

## 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

## 124 44 85 82 115 109 100 109 99 90 85 84 70 98 69 68 61 44 58 64

## 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39

## 46 48 32 38 30 39 29 31 23 25 25 26 22 28 25 16 11 21 15 8

## 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59

## 13 10 11 12 13 6 7 5 9 2 10 7 5 3 6 7 3 4 5 4

## 61 62 64 65 66 67 68 70 71 72 73 74 75 78 80 81 83 85 86 88

## 2 2 2 1 1 5 2 1 2 3 1 2 1 1 1 2 3 1 1 3

## 93 94 97 101 104 108 115 116 124 139 149 160 257

## 1 2 1 1 1 1 1 1 1 1 1 1 1

*#There are no outliers in FRP.Cancelled*

sum(is.na(Assignment\_3$FRP.Cancelled))

## [1] 0

table(Assignment\_3$FRP.Cancelled)

##

## 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

## 477 397 374 314 234 146 112 80 56 49 32 33 17 22 10 10 4 5 3 3

## 20 22 23 26 27 28 30 32 45

## 2 1 1 1 2 1 1 1 1

*#There are no outliers in FRP.Takeup percent*

sum(is.na(Assignment\_3$FRP.Take.up.percent.))

## [1] 0

table(Assignment\_3$FRP.Take.up.percent.)

##

## 0 0.013 0.02 0.028 0.029 0.031 0.046 0.053 0.059 0.067 0.07 0.071 0.077

## 124 1 1 2 1 1 1 1 2 1 1 1 3

## 0.083 0.086 0.095 0.1 0.111 0.118 0.119 0.121 0.125 0.128 0.132 0.143 0.158

## 2 1 1 4 2 2 1 2 5 1 1 3 2

## 0.16 0.167 0.17 0.171 0.174 0.177 0.179 0.182 0.188 0.192 0.194 0.198 0.2

## 1 8 1 3 1 1 1 2 3 1 2 1 13

## 0.208 0.212 0.213 0.216 0.217 0.222 0.227 0.231 0.233 0.235 0.237 0.238 0.239

## 1 2 1 1 1 4 2 3 1 1 1 1 3

## 0.889 0.893 0.894 0.898 0.9 0.903 0.905 0.908 0.909 0.917 0.923 0.929 0.933

## 5 1 1 2 9 1 1 1 8 6 8 5 1

## 0.938 0.941 0.944 0.947 0.952 0.96 0.966 1

## 1 1 1 1 1 1 1 72

*#There are no outliers in Cancelled.Pax*

sum(is.na(Assignment\_3$Cancelled.Pax))

## [1] 0

table(Assignment\_3$Cancelled.Pax)

##

## 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

## 242 316 300 294 247 223 173 135 97 81 49 46 41 22 20 15 17 16 8 10

## 20 21 22 23 24 25 26 29 32 33 34 37 38 39

## 9 5 5 2 1 2 3 2 1 1 1 2 2 1

*#There are no outliers in Total.Discount.pax*

sum(is.na(Assignment\_3$Total.Discount.Pax))

## [1] 0

table(Assignment\_3$Total.Discount.Pax)

##

## 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

## 15 795 595 345 238 127 82 59 37 21 20 12 8 7 4 5 2 2 2 6

## 20 22 26 27 29 47

## 1 2 1 1 1 1

*#There are outliers in Poverty.Code*

sum(is.na(Assignment\_3$Poverty.Code))

## [1] 599

table(Assignment\_3$Poverty.Code)

##

## 0 A B C D E

## 4 265 961 507 36 17

a10<-as.character(Assignment\_3$Poverty.Code)

e10<-replace\_na(a10,"B")

Assignment\_3$Poverty.Code<-as.factor(e10)

*#view(Assignment\_3)*

*#There are no outliers in Region*

sum(is.na(Assignment\_3$Region))

## [1] 0

table(Assignment\_3$Region)

##

## Dallas Houston Northern California Other

## 163 145 275 1165

## Pacific Northwest Southern California

## 198 443

Assignment\_3$Region <- as.factor(Assignment\_3$Region)

*#There are outliers in CRM.Segment*

sum(is.na(Assignment\_3$CRM.Segment))

## [1] 0

table(Assignment\_3$CRM.Segment)

##

## 1 10 11 2 3 4 5 6 7 8 9 NA

## 77 914 13 47 11 228 788 94 111 93 9 4

*#Assignment\_3$CRM.Segment <- Backup\_xlsx3$CRM.Segment*

a22<-as.character(Assignment\_3$CRM.Segment)

e22<-replace\_na(a22,"10")

Assignment\_3$CRM.Segment<-as.factor(e22)

*#view(Assignment\_3)*

*#There are no outliers in School.Type*

sum(is.na(Assignment\_3$School.Type))

## [1] 0

table(Assignment\_3$School.Type)

##

## Catholic CHD Private non-Christian

## 163 257 151

## PUBLIC

## 1818

Assignment\_3$Region <- as.factor(Assignment\_3$School.Type)

*#There are no outliers in Parent.Meeting.Flag*

sum(is.na(Assignment\_3$Parent.Meeting.Flag))

## [1] 0

table(Assignment\_3$Parent.Meeting.Flag)

##

## 0 1

## 337 2052

*#There are no outliers in MDR.LowGrade*

sum(is.na(Assignment\_3$MDR.Low.Grade))

## [1] 68

table(Assignment\_3$MDR.Low.Grade)

##

## 1 10 2 3 4 5 6 7 8 9 K PK

## 8 3 2 12 17 96 888 348 14 104 428 401

Assignment\_3$MDR.Low.Grade<-as.factor(Assignment\_3$MDR.Low.Grade)

*#There are outliers in Parent.Meeting.Flag*

sum(is.na(Assignment\_3$MDR.High.Grade))

## [1] 0

table(Assignment\_3$MDR.High.Grade)

##

## 1 10 11 12 2 3 4 5 6 7 8 9 NA

## 2 3 3 358 1 1 6 99 110 25 1659 54 68

a80<-as.character(Assignment\_3$MDR.High.Grade)

e80<-replace\_na(a80,"8")

Assignment\_3$MDR.High.Grade<-as.factor(e80)

*#view(Assignment\_3)*

*#There are outliers in School.Type*

sum(is.na(Assignment\_3$School.Type))

## [1] 0

table(Assignment\_3$School.Type)

##

## Catholic CHD Private non-Christian

## 163 257 151

## PUBLIC

## 1818

Assignment\_3$School.Type<-as.factor(Assignment\_3$School.Type)

*#There are outliers in Total.school.Enrolment*

sum(is.na(Assignment\_3$Total.School.Enrollment))

## [1] 91

table(Assignment\_3$Total.School.Enrollment)

##

## 19 36 50 52 56 57 60 65 70 75 80 82 85 86 90 92

## 1 1 2 2 1 1 1 1 4 2 1 1 2 1 1 2

## 95 98 100 104 105 106 107 108 109 110 111 112 114 115 118 119

## 2 1 5 1 5 2 2 1 1 2 1 1 1 1 1 1

## 120 122 125 126 129 130 131 133 135 138 140 142 143 145 146 147

## 7 1 4 1 1 2 1 1 2 1 3 1 1 3 1 1

## 148 149 150 151 155 158 159 160 162 163 164 165 167 168 170 172

## 1 1 11 1 4 1 1 2 2 1 5 8 1 2 6 4

## 175 179 180 181 182 183 184 185 188 189 190 192 193 195 197 198

## 10 1 10 1 2 1 3 2 1 1 4 2 1 2 1 2

1

## 2000 2050 2087 2098 2120 2127 2159 2165 2168 2169 2175 2200 2235 2300 2351 2375

## 3 1 1 1 1 1 1 1 1 2 1 1 1 2 1 1

## 2393 2441 2520 2700 2765 2778 2850 3000 3100 3200 3600 3700 3990

## 1 1 1 1 1 1 1 2 1 2 2 1 1

e48<- as.numeric(Assignment\_3$Total.School.Enrollment)

e48<-na.aggregate(e48)

Assignment\_3$Total.School.Enrollment<-e48

*#There are outliers in EZ.Pay.Take.Up.Rate*

sum(is.na(Assignment\_3$EZ.Pay.Take.Up.Rate))

## [1] 0

table(Assignment\_3$EZ.Pay.Take.Up.Rate)

##

## 0 0.008 0.011 0.012 0.016 0.019 0.02 0.021 0.022 0.024 0.025 0.026 0.027

## 341 1 1 1 1 2 2 1 2 1 1 3 2

## 0.028 0.029 0.03 0.031 0.032 0.033 0.036 0.037 0.039 0.04 0.041 0.042 0.043

## 1 3 3 1 1 1 1 3 5 2 1 4 1

## 0.423 0.424 0.426 0.429 0.433 0.436 0.438 0.44 0.441 0.444 0.446 0.448 0.45

## 1 3 1 14 2 1 6 1 2 11 1 1 2

## 0.455 0.457 0.458 0.462 0.464 0.467 0.471 0.474 0.475 0.476 0.478 0.479 0.48

## 6 1 2 4 1 1 3 2 1 3 1 1 2

## 0.482 0.483 0.484 0.486 0.5 0.524 0.526 0.533 0.535 0.539 0.546 0.552 0.556

## 1 2 2 1 43 1 3 2 1 4 4 1 2

## 0.565 0.571 0.583 0.591 0.594 0.6 0.625 0.643 0.667 0.692 0.714 0.722 0.75

## 1 6 1 1 1 10 3 1 10 1 2 2 4

## 0.769 0.773 0.786 0.8 1 1.205 1.75

## 1 1 1 2 4 1 1

*#There are outliers in School.Sponsor*

sum(is.na(Assignment\_3$School.Sponsor))

## [1] 0

table(Assignment\_3$School.Sponsor)

##

## 0 1

## 2136 253

*#There are no outliers in SPR.Product.Type*

sum(is.na(Assignment\_3$SPR.Product.Type))

## [1] 0

table(Assignment\_3$SPR.Product.Type)

##

## CA History Costa Rica East Coast IL History International

## 175 46 2005 5 15

## Science

## 143

Assignment\_3$SPR.Product.Type<-as.factor(Assignment\_3$SPR.Product.Type)

*#There are no outliers in SPR.New.Existing*

sum(is.na(Assignment\_3$SPR.New.Existing))

## [1] 0

table(Assignment\_3$SPR.New.Existing)

##

## EXISTING NEW

## 1607 782

Assignment\_3$SPR.New.Existing<-as.factor(Assignment\_3$SPR.New.Existing)

*#There are no outliers in FPP*

sum(is.na(Assignment\_3$FPP))

## [1] 0

table(Assignment\_3$FPP)

##

## 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

## 9 36 41 67 44 74 78 60 65 58 77 61 65 51 66 55 65 51 45 48

## 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41

## 40 50 38 40 44 38 31 39 33 34 42 38 35 33 29 23 34 28 24 20

## 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61

## 32 23 22 34 17 22 19 16 20 18 16 13 9 10 5 9 8 7 8 5

## 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81

## 12 6 5 10 15 5 8 11 8 11 16 2 4 8 5 2 7 2 3 5

## 82 83 84 85 86 87 88 89 90 91 92 93 94 95 97 98 100 101 102 103

## 2 4 5 5 4 4 1 3 1 4 4 2 1 2 2 2 2 2 1 2

## 104 105 106 107 108 109 110 112 115 118 119 120 123 125 126 131 132 135 136 137

## 4 2 2 2 4 1 2 1 2 1 4 2 2 2 1 1 3 2 1 1

## 142 145 156 159 162 165 166 167 174 177 181 183 185 190 191 196 197 199 204 205

## 1 2 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

## 212 222 230 243 257 286

## 1 1 1 1 1 1

*#There are no outliers in Total.Pax*

sum(is.na(Assignment\_3$Total.Pax))

## [1] 0

table(Assignment\_3$Total.Pax)

##

## 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21

## 1 9 36 37 65 41 76 67 62 65 56 70 64 56 53 47 57 56 58 51

## 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41

## 38 47 34 46 48 39 29 29 31 49 31 35 27 42 40 25 28 26 17 34

## 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61

## 26 23 24 23 22 27 23 25 19 26 15 17 18 15 16 14 9 12 8 4

## 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81

## 8 9 2 6 10 8 6 8 7 5 9 10 7 7 8 14 7 6 9 4

## 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101

## 3 4 3 6 6 2 5 5 2 1 4 2 4 3 4 3 3 2 2 1

## 102 104 105 106 107 108 109 110 112 113 114 115 116 117 118 120 122 123 126 127

## 3 2 1 2 2 2 2 3 2 2 2 3 2 3 2 1 1 1 3 1

## 129 130 131 132 135 136 137 139 140 141 143 148 150 151 154 159 163 168 175 177

## 2 1 1 1 1 1 1 2 1 2 2 1 1 1 1 1 1 1 3 1

## 182 185 186 192 194 197 202 205 210 218 220 225 239 241 250 251 262 276 313

## 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1

*#There are no outliers in SPR.Group.Revenue*

sum(is.na(Assignment\_3$SPR.Group.Revenue))

## [1] 0

table(Assignment\_3$SPR.Group.Revenue)

##

## 79 100 119 143 149 156 163 168 172 199 243 249 298 299 349 358

## 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 1

## 359 369 374 375 376 380 382 383 384 385 386 389 390 392 393 394

## 2 3 2 3 1 1 1 3 1 3 3 1 1 3 3 1

## 395 396 398 399 400 402 406 408 409 410 415 417 419 420 424 425

## 3 2 4 11 2 1 1 3 6 1 2 1 5 2 2 2

## 428 429 435 438 439 440 441 443 445 449 453 457 459 460 461 463

1

## 2870 2894 2899 2907 2948 2949 2953 2976 2989 2990 3027 3050 3074 3085 3089 3098

## 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 1

## 3110 3114 3129 3224 3349 3379 3439 3479 3499 3628 3799 3884 4199 4200

## 1 1 1 1 1 1 1 1 1 1 1 7 1 1

*#There are no outliers in Number of meetings with parents*

sum(is.na(Assignment\_3$NumberOfMeetingswithParents))

## [1] 0

table(Assignment\_3$NumberOfMeetingswithParents)

##

## 0 1 2

## 337 1471 581

Assignment\_3$NumberOfMeetingswithParents<-as.factor(Assignment\_3$NumberOfMeetingswithParents)

*#There are no outliers in Number of meetings with parents*

sum(is.na(Assignment\_3$NumberOfMeetingswithParents))

## [1] 0

table(Assignment\_3$NumberOfMeetingswithParents)

##

## 0 1 2

## 337 1471 581

*#There are no outliers in DifferenceTraveltoFirstMeeting*

sum(is.na(Assignment\_3$DifferenceTraveltoFirstMeeting))

## [1] 0

table(Assignment\_3$DifferenceTraveltoFirstMeeting)

##

## -188 -204 116 124 128 130 131 132 133 134 135 136 137 138 139 14

## 1 1 2 1 1 2 1 1 1 1 1 1 3 1 2 1

## 140 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156

## 1 2 6 1 5 4 1 3 1 4 3 1 1 3 2 2

2

## 401 402 403 404 405 406 407 408 409 410 411 412 413 414 416 417

## 3 2 6 4 2 4 4 3 2 3 3 8 1 4 1 3

## 418 419 420 421 422 423 424 425 426 427 428 431 432 433 434 436

## 3 5 1 1 1 1 6 3 4 1 3 3 2 2 1 2

## 437 438 439 440 444 445 448 451 452 453 455 456 458 459 460 461

## 1 4 2 3 1 1 2 2 3 2 2 3 1 1 1 1

## 466 468 474 475 477 480 483 485 505 530 569 598 604 623 63 651

## 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1

## 73 749 78 80 82 91 NA

## 1 1 1 1 2 1 337

Assignment\_3$DifferenceTraveltoFirstMeeting<-as.factor(Assignment\_3$DifferenceTraveltoFirstMeeting)

*#There are no outliers in Different Travel to Last Meeting*

sum(is.na(Assignment\_3$Difference Travels to Last Meeting))

## [1] 0

table(Assignment\_3$Difference Travels to Last Meeting)

##

## -17 -188 -204 -4 109 116 120 121 124 127 128 130 131 132 133 134

## 1 1 1 1 1 2 1 1 1 1 1 2 2 2 1 2

## 135 136 137 138 139 14 140 142 143 144 145 146 147 148 149 15

## 2 1 6 1 3 2 2 2 5 1 5 3 1 6 1 1

## 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165

## 6 5 2 2 7 2 2 5 5 3 5 3 6 3 5 5

## 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 7

## 288 289 290 291 292 293 294 296 298 299 30 300 302 303 306 307

## 1 3 1 2 2 2 1 1 2 2 1 3 1 3 1 1

## 308 314 315 316 318 32 320 322 324 325 327 328 333 336 342 345

## 1 1 1 1 1 1 1 1 1 1 1 1 2 1 3 1

## 347 348 352 354 356 358 367 368 369 376 377 38 381 383 384 385

## 1 1 1 2 1 1 1 1 1 1 2 1 1 2 1 1

## 389 391 392 395 398 402 403 410 412 419 425 432 434 437 438 455

## 2 1 1 1 1 2 1 2 1 2 1 1 1 1 1 1

## 456 530 63 651 73 749 78 80 82 9 91 NA

## 1 1 1 1 1 1 1 1 2 1 1 337

Assignment\_3$DifferenceTraveltoLastMeeting<-as.factor(Assignment\_3$Difference Travels to Last Meeting)

*#There are no outliers in SchoolgradetypeLow*

sum(is.na(Assignment\_3$SchoolGradeTypeLow))

## [1] 0

table(Assignment\_3$SchoolGradeTypeLow)

##

## Elementary High Middle Undefined

## 259 141 1862 127

Assignment\_3$SchoolGradeTypeLow<-as.factor(Assignment\_3$SchoolGradeTypeLow)

*#There are no outliers in SchoolgradetypeLow*

sum(is.na(Assignment\_3$SchoolGradeTypeHigh))

## [1] 0

table(Assignment\_3$SchoolGradeTypeHigh)

##

## Elementary High Middle Undefined

## 196 265 1778 150

Assignment\_3$SchoolGradeTypeHigh<-as.factor(Assignment\_3$SchoolGradeTypeHigh)

Assignment\_3$SchoolGradeType<-as.factor(Assignment\_3$SchoolGradeType)

Assignment\_3$SchoolGradeType<-as.factor(Assignment\_3$MajorProgramCode)

Assignment\_3$DepartureMonth<-as.factor(Assignment\_3$DepartureMonth)

Assignment\_3$GroupGradeTypeLow<-as.factor(Assignment\_3$GroupGradeTypeLow)

Assignment\_3$GroupGradeTypeHigh<-as.factor(Assignment\_3$GroupGradeTypeHigh)

Assignment\_3$GroupGradeType<-as.factor(Assignment\_3$GroupGradeType)

Assignment\_3$MajorProgramCode<-as.factor(Assignment\_3$MajorProgramCode)

Assignment\_3$SingleGradeTripFlag<-as.factor(Assignment\_3$SingleGradeTripFlag)

Assignment\_3$GroupGradeType<-as.factor(Assignment\_3$GroupGradeType)

*#There are no outliers in FPP.TO.PAX*

sum(is.na(Assignment\_3$FPP.to.PAX))

## [1] 0

table(Assignment\_3$FPP.to.PAX)

##

## 0.6 0.625 0.666667 0.692308 0.705882 0.714286 0.727273 0.75

## 2 1 13 1 1 4 2 44

## 0.769231 0.775862 0.777778 0.783784 0.785714 0.789474 0.8 0.809524

## 1 1 5 1 1 1 51 1

## 0.8125 0.812749 0.813559 0.818182 0.823529 0.825 0.826087 0.829268

## 2 1 1 12 6 1 2 2

## 0.833333 0.835616 0.836735 0.837209 0.83871 0.839286 0.839623 0.84

## 83 1 1 3 4 1 1 3

## 0.842105 0.843137 0.84375 0.844444 0.846154 0.849057 0.849315 0.85

## 8 1 2 2 19 1 1 3

## 0.851064 0.851852 0.852273 0.852459 0.852941 0.853659 0.854167 0.854545

## 2 3 1 1 2 2 2 1

## 0.857143 0.858974 0.862069 0.862745 0.863636 0.864865 0.865979 0.866667

## 73 2 3 1 6 3 1 19

## 0.868421 0.869159 0.869565 0.87013 0.870968 0.872093 0.872727 0.875

## 2 1 12 1 7 2 1 108

## 0.877193 0.877551 0.878049 0.878788 0.879668 0.88 0.880952 0.881356

## 1 2 6 5 1 14 3 1

## 0.882353 0.883117 0.883333 0.883721 0.884444 0.884615 0.884892 0.885057

## 23 1 3 3 1 16 1 1

## 0.885417 0.885714 0.886364 0.8875 0.888889 0.89 0.890411 0.890909

## 1 6 7 2 115 1 1 1

## 0.891304 0.891429 0.891892 0.892308 0.892683 0.892857 0.893617 0.893939

## 3 1 7 1 1 11 8 1

## 0.894309 0.894737 0.895522 0.895833 0.896104 0.896552 0.896907 0.897297

## 1 44 3 2 2 15 1 1

## 0.897436 0.897959 0.898551 0.898734 0.899083 0.899281 0.9 0.901639

## 9 5 2 2 1 1 114 1

## 0.901961 0.902174 0.902439 0.902778 0.903226 0.903846 0.90411 0.904255

## 7 1 8 1 13 5 2 2

## 0.904762 0.905405 0.90566 0.905983 0.90625 0.906977 0.907407 0.907692

## 37 1 2 1 13 9 6 1

## 0.907895 0.908397 0.908571 0.909091 0.909524 0.910112 0.910448 0.910714

## 1 1 1 100 1 3 1 3

## 0.911111 0.911392 0.911504 0.911765 0.91195 0.912162 0.912281 0.912409

## 8 3 1 11 1 1 6 1

## 0.9125 0.912698 0.913043 0.913333 0.91358 0.913738 0.913793 0.914286

## 1 2 36 1 1 1 3 23

## 0.91453 0.914634 0.914894 0.915254 0.915493 0.915663 0.915789 0.916084

## 1 1 9 4 3 2 2 1

## 0.916667 0.917582 0.917647 0.918182 0.918367 0.918919 0.919192 0.919355

## 85 1 1 1 9 6 2 3

## 0.92 0.920354 0.920635 0.921053 0.921348 0.921569 0.922078 0.922481

## 34 1 1 10 1 7 1 1

## 0.923077 0.923913 0.924242 0.924528 0.924731 0.925 0.925373 0.925926

## 91 1 1 5 1 6 2 25

## 0.926471 0.926606 0.926829 0.927083 0.927273 0.927481 0.927536 0.928571

## 2 1 14 1 5 1 1 75

## 0.92887 0.929412 0.929825 0.93 0.930233 0.930693 0.931034 0.931159

## 1 1 2 1 5 1 14 1

## 0.931373 0.931507 0.931818 0.932203 0.932432 0.932584 0.932692 0.93299

## 1 3 5 5 1 1 2 1

## 0.933333 0.934211 0.934426 0.934783 0.935065 0.935484 0.93617 0.936508

## 65 3 3 9 4 26 7 5

## 0.936709 0.937008 0.9375 0.938144 0.938272 0.938462 0.938776 0.939086

## 1 1 49 1 1 2 7 1

## 0.93913 0.939394 0.939759 0.94 0.940299 0.940397 0.941176 0.941558

## 2 21 1 5 1 1 41 1

## 0.94186 0.942029 0.942308 0.942857 0.943396 0.943662 0.944056 0.944444

## 1 3 4 17 7 1 1 41

## 0.945055 0.945205 0.945455 0.945946 0.946429 0.946667 0.947368 0.947826

## 1 1 3 10 4 1 24 1

## 0.948052 0.948276 0.948718 0.949153 0.95 0.95122 0.951613 0.951807

## 1 2 5 2 23 4 1 1

## 0.952381 0.952941 0.953488 0.954023 0.954128 0.954545 0.956522 0.957055

## 20 1 4 1 1 16 9 1

## 0.957447 0.958333 0.959184 0.959459 0.96 0.960784 0.961538 0.962264

## 4 9 3 1 9 2 13 3

## 0.962617 0.962963 0.963636 0.964286 0.964539 0.964912 0.965116 0.965517

## 1 11 2 3 1 2 1 2

## 0.965909 0.966102 0.967742 0.968421 0.96875 0.969697 0.970149 0.970588

## 1 1 5 1 3 3 1 7

## 0.971429 0.971831 0.972222 0.972603 0.972727 0.972973 0.973684 0.974359

## 4 1 4 1 1 2 3 3

## 0.975 0.975248 0.97561 0.976744 0.977273 0.977778 0.978261 0.980769

## 1 1 1 2 2 1 2 1

## 0.984848 1

## 1 15

*#There are no outliers in Num.of.non\_FPP*

sum(is.na(Assignment\_3$Num.of.Non\_FPP.PAX))

## [1] 0

table(Assignment\_3$Num.of.Non\_FPP.PAX)

##

## 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

## 15 795 595 345 238 127 82 59 37 21 20 12 8 7 4 5 2 2 2 6

## 20 22 26 27 29 47

## 1 2 1 1 1 1

*#There are no outliers in SchoolSizeIndicator*

sum(is.na(Assignment\_3$SchoolSizeIndicator))

## [1] 91

table(Assignment\_3$SchoolSizeIndicator)

##

## L M-L S S-M

## 597 594 507 600

a220<-as.character(Assignment\_3$SchoolSizeIndicator)

e220<-replace\_na(a220,"S-M")

Assignment\_3$SchoolSizeIndicator<-as.factor(e220)

*#view(Assignment\_3)*

Assignment\_3$FPP.to.School.enrollment<-NULL

Assignment\_3$ID<-NULL

Assignment\_3$Program.Code<-as.factor(Assignment\_3$Program.Code)

Assignment\_3$Travel.Type<-as.factor(Assignment\_3$Travel.Type)

## **R Markdown**

Exploratory Data Analysis(EDA)

*#Exploratory DataAnalysis Part*

*#After replacing all the columns having NA values we remove some unimportant variables from the dataset*

*#We do the chisquare tests on the variables we take the predictor variable and the targe variable and plot against each other.*

*#Sample Example*

chisq.test(Assignment\_3$Tuition,Assignment\_3$Retained.in.2012.,correct = F)

## Warning in chisq.test(Assignment\_3$Tuition, Assignment\_3$Retained.in.2012., :

## Chi-squared approximation may be incorrect

##

## Pearson's Chi-squared test

##

## data: Assignment\_3$Tuition and Assignment\_3$Retained.in.2012.

## X-squared = 1271.5, df = 1229, p-value = 0.1944

*# Chi Square Test for FRP.Cancelled*

chisq.test(Assignment\_3$FRP.Cancelled,Assignment\_3$Retained.in.2012.,correct = F)

## Warning in chisq.test(Assignment\_3$FRP.Cancelled,

## Assignment\_3$Retained.in.2012., : Chi-squared approximation may be incorrect

##

## Pearson's Chi-squared test

##

## data: Assignment\_3$FRP.Cancelled and Assignment\_3$Retained.in.2012.

## X-squared = 33.176, df = 28, p-value = 0.2293

chisq.test(Assignment\_3$Parent.Meeting.Flag,Assignment\_3$Retained.in.2012.,correct = F)

##

## Pearson's Chi-squared test

##

## data: Assignment\_3$Parent.Meeting.Flag and Assignment\_3$Retained.in.2012.

## X-squared = 1.0022, df = 1, p-value = 0.3168

chisq.test(Assignment\_3$Total.School.Enrollment,Assignment\_3$Retained.in.2012.,correct = F)

## Warning in chisq.test(Assignment\_3$Total.School.Enrollment,

## Assignment\_3$Retained.in.2012., : Chi-squared approximation may be incorrect

##

## Pearson's Chi-squared test

##

## data: Assignment\_3$Total.School.Enrollment and Assignment\_3$Retained.in.2012.

## X-squared = 968.43, df = 893, p-value = 0.03984

chisq.test(Assignment\_3$FRP.Cancelled,Assignment\_3$Retained.in.2012.,correct = F)

## Warning in chisq.test(Assignment\_3$FRP.Cancelled,

## Assignment\_3$Retained.in.2012., : Chi-squared approximation may be incorrect

##

## Pearson's Chi-squared test

##

## data: Assignment\_3$FRP.Cancelled and Assignment\_3$Retained.in.2012.

## X-squared = 33.176, df = 28, p-value = 0.2293

*#We did chi square test for all the variables and found these variables to be removed since their P-value is >0.05 which establishes null hypothesis therefore we remove them/*

*#After the chi square test is done we remove the values giving p-value is more pertaining that we accept null hypothesis for this case and therefore we conclude that there is very less/ no relation virtually.*

*#For most of the variables Chi-SQuare values we calculated many were in powers of e therefore concluding that p value is so less than 0.05 meaning that we could reject null hypothesis good enough to predict that there is a relationship between the variables.*

Assignment\_3 <- select(Assignment\_3,-c('Tuition','Early.RPL','Parent.Meeting.Flag','Total.School.Enrollment','SPR.Group.Revenue','FirstMeeting','LastMeeting','SchoolGradeType','Group.State','NumberOfMeetingswithParents','DifferenceTraveltoFirstMeeting','Difference Travels to Last Meeting'))

*#EDA(Continued)*

*#More of them are in June month are retained*

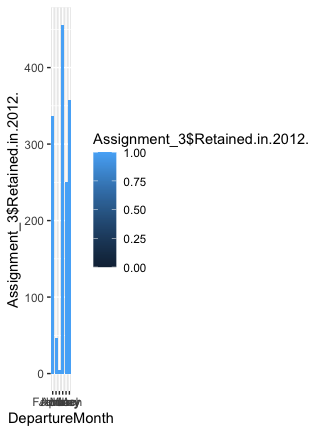
Barplot\_1<- ggplot(data=Assignment\_3,aes(x=DepartureMonth,y=Assignment\_3$Retained.in.2012.,fill=Assignment\_3$Retained.in.2012.))+geom\_bar(stat = "identity")

Barplot\_1

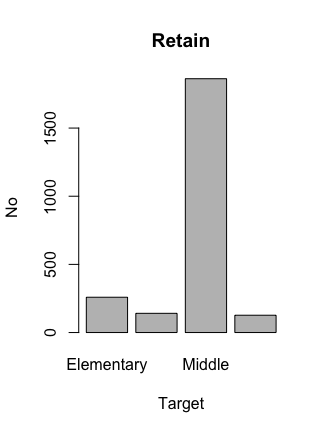
## Warning: Use of `Assignment\_3$Retained.in.2012.` is discouraged. Use `Retained.in.2012.` instead.

## Use of `Assignment\_3$Retained.in.2012.` is discouraged. Use `Retained.in.2012.` instead.

*#Most of them retained are in middle school*

**

plot(Assignment\_3[,42],main="Retain",ylab ="No",xlab="Target")



*#hist(Assignment\_3$Tuition,xlab = "Tuition",col = "red")*

*#More number of people pay an tuition of over 2000*

Decison Tree Construction

*#Decision Tree Construction*

*#The number of terminal nodes while pruning decreases or the R-software package tries to decrease n so we need to increase the value of cp to avoid overfitting*

*#We also need to choose the cp values that give us the minimum error*

set.seed(110)

indx<-sample(2,nrow(Assignment\_3),replace=TRUE,prob = c(0.7,0.3))

train<-Assignment\_3[indx==1,]

nrow(train)

## [1] 1675

test<-Assignment\_3[indx==2,]

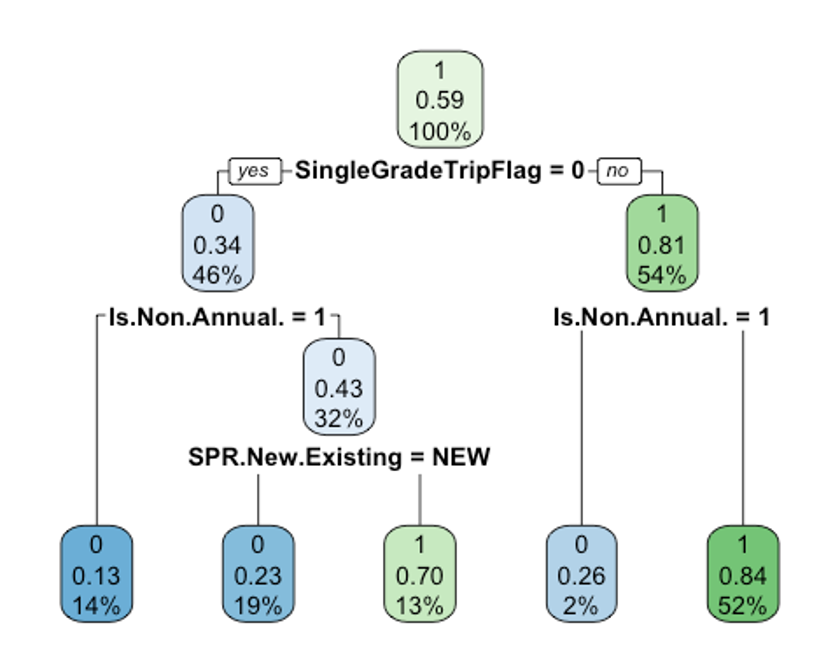
nrow(test)

## [1] 714

myFormula<-Retained.in.2012.~.

myTree<-rpart(myFormula, data = train, method = "class")

rpart.plot(myTree)



predict\_train<-predict(myTree,data=train,type="class")

mean\_train<-mean(train$Retained.in.2012.!=predict\_train)

mean\_train

## [1] 0.1922388

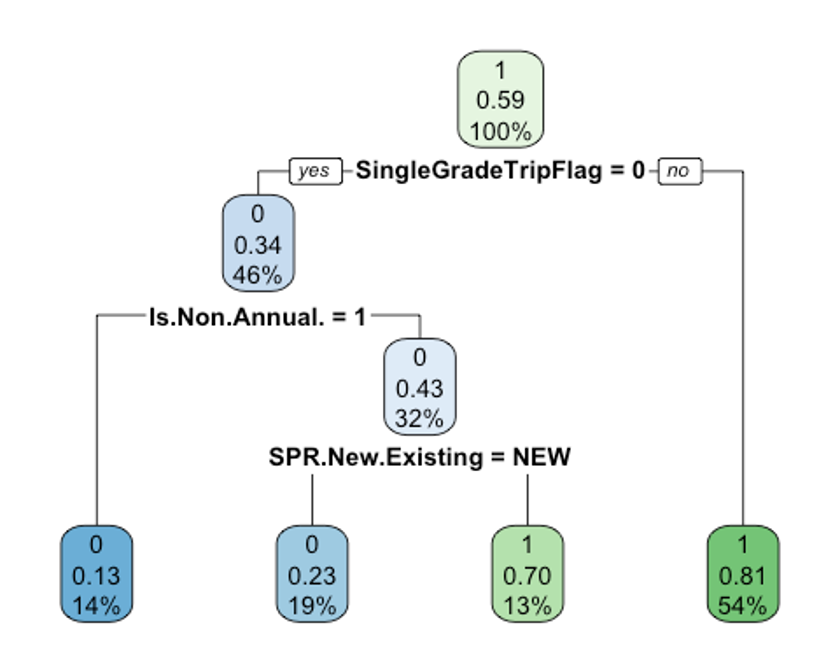
*#predict\_test<-predict(myTree, newdata=test,type="class")*

*#mean\_test<-mean(test$Retained.in.2012.!=predict\_test)*

*#mean\_test*

myTreeHPT<-rpart(myFormula, data=train,method = "class",parms = list(split="information"), control = rpart.control(minsplit = 15,minbucket=10,cp=0.01))

rpart.plot(myTreeHPT)



library(randomForest)

## randomForest 4.7-1

## Type rfNews() to see new features/changes/bug fixes.

##

## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':

##

## combine

## The following object is masked from 'package:ggplot2':

##

## margin

Target <- Assignment\_3$Retained.in.2012.

rf <- randomForest(Target~ ., data =Assignment\_3, mtry = 3,

importance = TRUE, na.action = na.omit)

## Warning in randomForest.default(m, y, ...): The response has five or fewer

## unique values. Are you sure you want to do regression?

print(rf)

##

## Call:

## randomForest(formula = Target ~ ., data = Assignment\_3, mtry = 3, importance = TRUE, na.action = na.omit)

## Type of random forest: regression

## Number of trees: 500

## No. of variables tried at each split: 3

##

## Mean of squared residuals: 0.02939162

## % Var explained: 87.67

attributes(rf)

## $names

## [1] "call" "type" "predicted" "mse"

## [5] "rsq" "oob.times" "importance" "importanceSD"

## [9] "localImportance" "proximity" "ntree" "mtry"

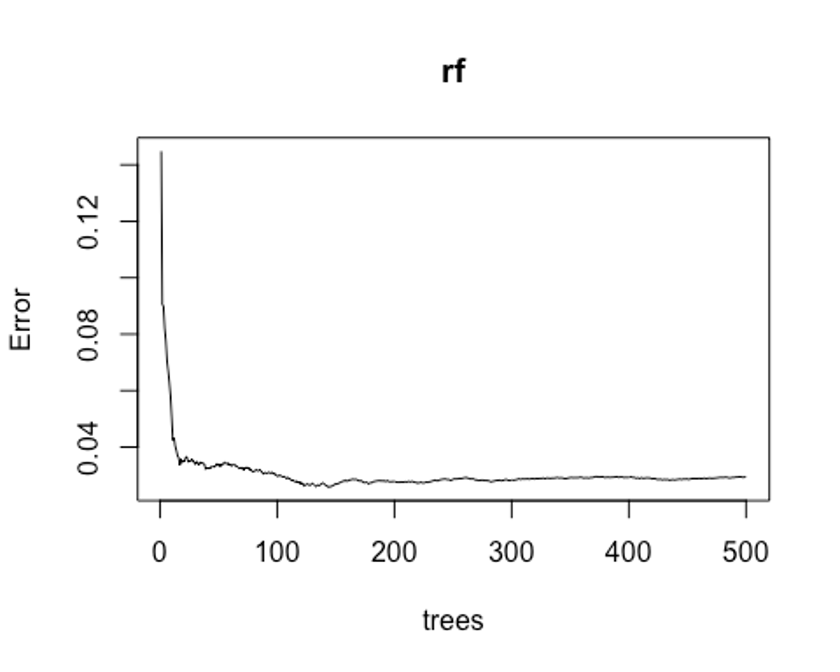
## [13] "forest" "coefs" "y" "test"

## [17] "inbag" "terms" "na.action"

##

## $class

## [1] "randomForest.formula" "randomForest"



plot(rf)

head(rf$predicted)

## 1 2 3 5 6 7

## 0.9220854 0.8114134 0.9590502 0.3480489 0.6188887 0.3796595

*#df$Num.of.Non\_FPP.PAX[is.na(df$SchoolSizeIndicator)]<-as.character(df$FPP.to.PAX[is.na(df$SchoolSizeIndicator)])*

ind<-sample(2,nrow(train),replace = T,prob=c(0.7,0.3))

Trainrf<-train[ind==1,]

Validation<-train[ind==2,]

pr.err<-c()

rf$predicted

## 1 2 3 5 6 7

## 0.922085386 0.811413409 0.959050179 0.348048942 0.618888698 0.379659498

## 9 10 11 12 13 14

## 0.846794872 0.983306480 0.951831502 0.965861316 0.902461193 0.981595577

## 16 17 18 19 20 21

## 0.954690117 0.957212476 0.790810811 0.991165756 0.894067797 0.988989899

## 22 24 25 26 27 28

## 0.721910112 0.987306631 0.885397086 0.992982456 0.873963964 0.929729730

## 29 30 31 32 33 34

## 0.939714286 0.967352941 0.966889483 0.981395349 0.765065913 0.836931491

## 35 36 37 38 39 40

## 0.975227687 0.970879121 0.996542553 0.767796610 0.901714286 0.982954545

## 42 43 44 45 46 47

## 0.979549550 0.339199255 0.887243948 0.830683761 0.961881571 0.261623780

## 48 49 50 51 52 53

## 0.962612613 0.859186074 0.313162303 0.947895623 0.923180077 0.974365079

## 54 55 56 57 58 59

## 0.961315789 0.918354978 0.750113982 0.687673611 0.936648250 0.985029940

## 60 61 62 63 64 65

## 0.892090395 0.274656357 0.997129187 0.214853195 0.179087049 0.967249931

## 66 67 68 69 70 71

## 0.228888889 0.896254682 0.986559140 0.953629032 0.957655502 0.907541899

## 72 73 74 75 76 77

## 0.959420290 0.854182864 0.916387097 0.971976840 0.842817014 0.912176166

## 78 79 80 81 82 83

## 0.853910615 0.958894646 0.971445020 0.974250441 0.168353175 0.904943639

## 84 85 87 89 90 91

## 0.930587121 0.833142680 0.988062678 0.270043996 0.756510417 0.857711443

## 92 93 94 95 96 97

## 0.840600775 0.871557971 0.132240437 0.975468165 0.939814815 0.358958333

## 98 99 100 101 102 103

## 0.263569519 0.966859345 0.861080074 0.872517730 0.811603376 0.298724490

## 104 105 106 107 108 109

## 0.923325563 0.947735507 0.832587065 0.156134969 0.125567190 0.833709981

## 110 111 112 113 114 115

## 0.867517007 0.116577540 0.895333333 0.898847518 0.960845778 0.926799242

## 116 117 118 119 120 121

```

Lift Chart

perf <- performance(pred, "lift","rpp")

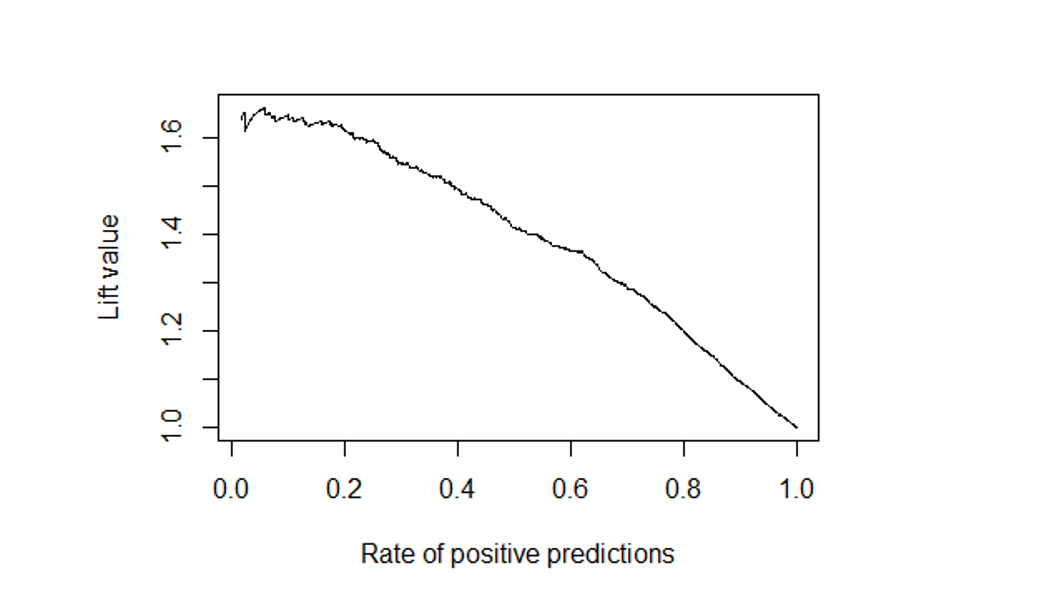
perf

## A performance instance

## 'Rate of positive predictions' vs. 'Lift value' (alpha: 'Cutoff')

## with 1255 data points

plot(perf)



#Gain Chart

per <- performance(pred, "tpr","rpp")

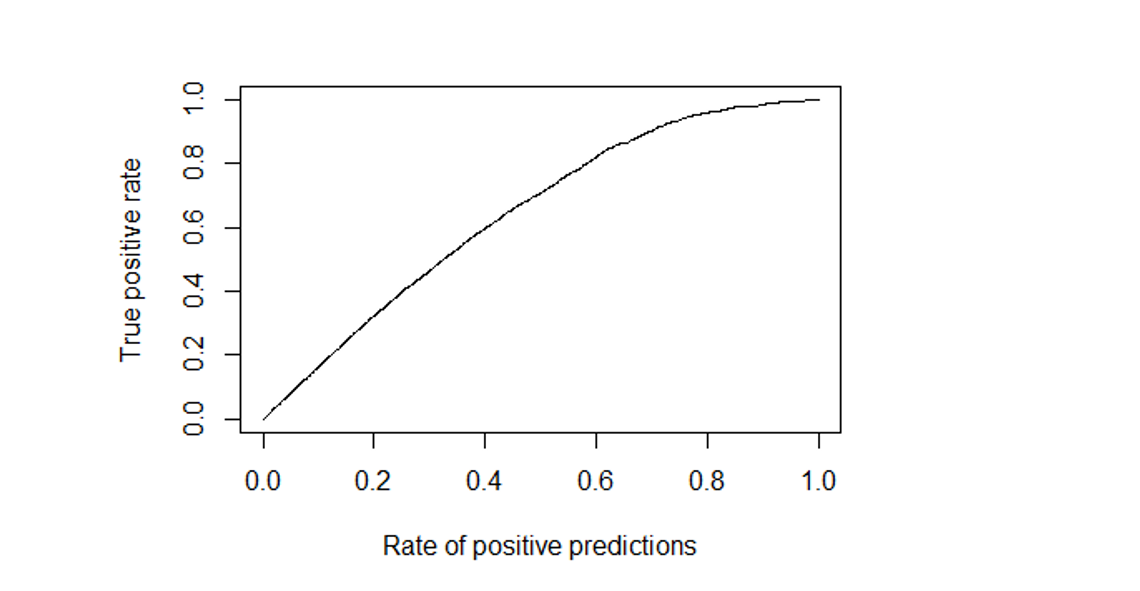
per

## A performance instance

## 'Rate of positive predictions' vs. 'True positive rate' (alpha: 'Cutoff')

## with 1255 data points

plot(per)



#Response Chart

perfo <- performance(pred, "ppv","rpp")

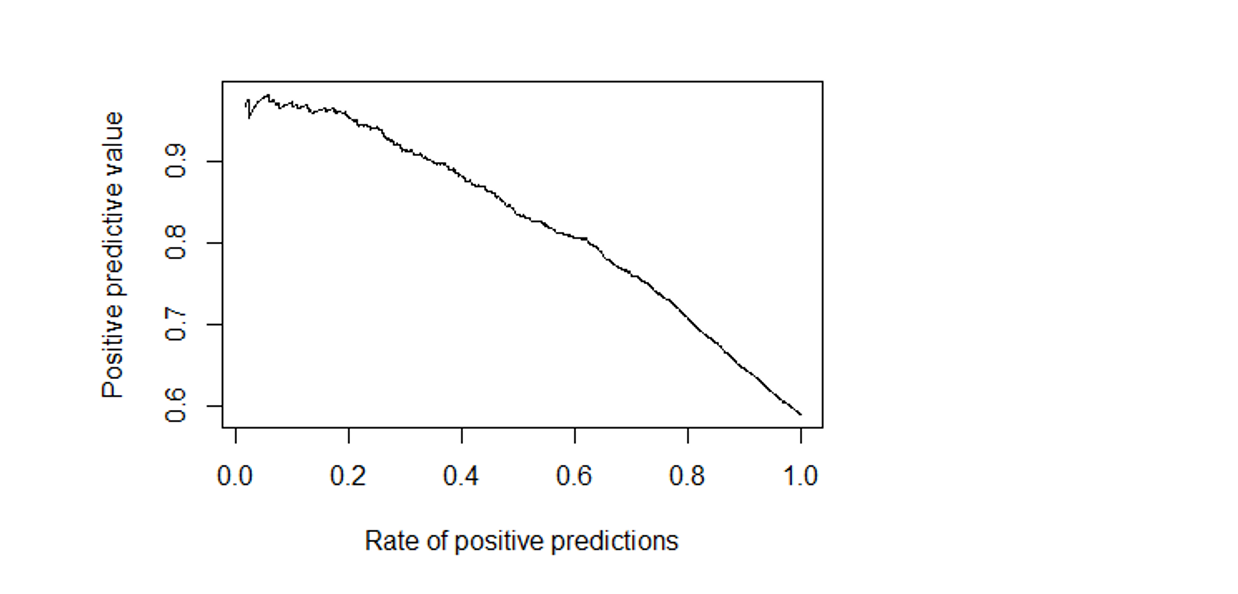
perfo

## A performance instance

## 'Rate of positive predictions' vs. 'Positive predictive value' (alpha: 'Cutoff')

## with 1255 data points

plot(perfo)



#ROC Curve

perform <- performance(pred, "tpr","fpr")

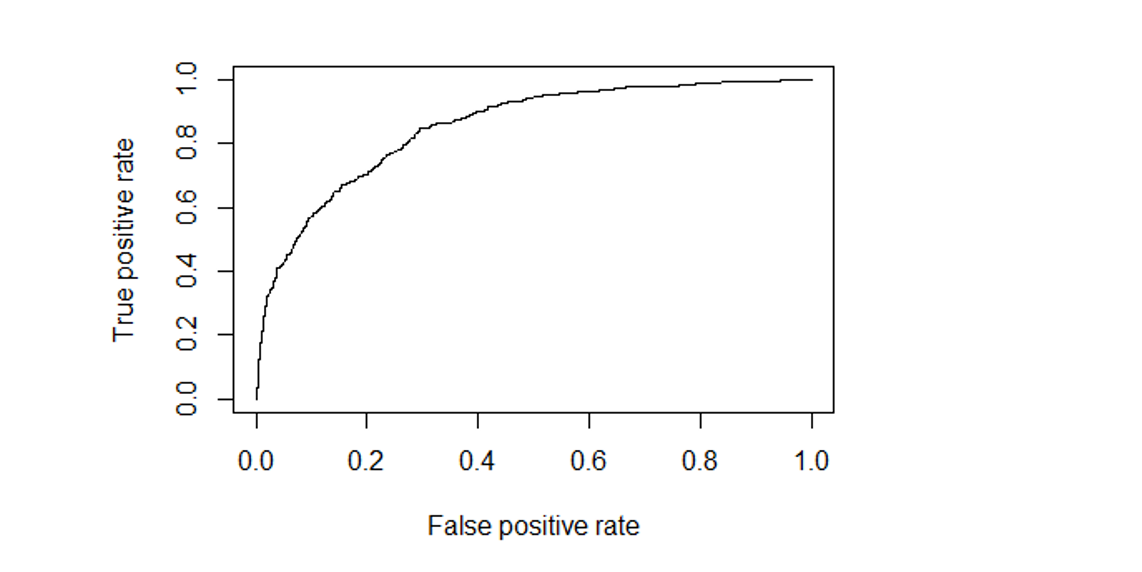
perform

## A performance instance

## 'False positive rate' vs. 'True positive rate' (alpha: 'Cutoff')

## with 1255 data points

plot(perform)



#Cross Validation tells us how good our model is good

Assignment\_3<-Assignment\_3

colnames(Assignment\_3)[42]<-"Some"

library(ISLR)

data("Assignment\_3")

attach(Assignment\_3)

High <- Assignment\_3$Some

Data<-data.frame(Assignment\_3,High)

#colnames(Assignment\_3)[42]<- "Target"

Data <- Data[sample(nrow(Data)),]

k<-10

method <- 1

folds <- cut(seq(1,nrow(Data)), breaks = k,labels = F)

mod.err<- matrix(-1,k,method,dimnames = list(paste0("Fold",1:k,c("Logistic Regression"))))

for(i in 1:k)

{

test \_indexes <- which(folds==i,arr.ind=T)

Test2 <- Data[text \_indexes,]

Train2 <- Data[-testindexes,]

LogitModel <- glm(Some~.,data=Train2, family = "binomial")

pred <- predict(LogitModel,newdata = Test2,type="response")

pred\_class <- as.factor(ifelse(pred>=0.5,"YES","NO"))

mod.err[i] <- mean(Test2$Some != pred\_class)

}

mean(model.err)

**Conclusion:**

According to the analysis on various measures, Random Forest outperforms Decision Tree. As a result, we propose utilizing the Random Forest model to forecast values.