**DSBA/MBAD 6211 Assignment 1**

Due: 11:59pm @ 10/07/2020

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.

![A close up of text on a white background

Description automatically generated]()

![Text

Description automatically generated]()

* 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.

I rejected CONTACT\_DATE, LEVEL\_YEAR, telecq, and sex.

I decided on CONTACT\_ DATE because it seemed irrelevant to the dataset altogether. I rejected LEVEL\_YEAR because it had the same value for all the rows in the dataset. Telecq was rejected because of the number of missing values. It had over 70000 out of a whole dataset of 90000. I also rejected sex because it should not matter what sex you are to be accepted into university.

I also decided to reject TOTAL\_CONTACTS, SELF\_INIT\_CNTCTS, TRAVEL\_INIT\_CNTCTS, SOLICITED\_CNTCTS, REFERRAL\_CNTCTS due to their large vif from the multicollinearity function.![Text

Description automatically generated]()

I also rejected ethnicity because it was not statistically significant.

![Table

Description automatically generated]()

* 1. Which variable is your target variable?

I chose Enroll.

* 1. Do you need to change any measurement levels of your existing variables (e.g., binary, numeric, factor)? Why?

I changed all the variables from int to factor.

1. Explain whether variable imputation and transformation are needed for the regression model. If so, please explain which variables have been imputed, transformed and how.

dfRC$Enroll <- factor(dfRC$Enroll)

dfRC$sex <- factor(dfRC$sex)

dfRC$premiere<-factor(dfRC$premiere)

dfRC$stucell<-factor(dfRC$stucell)

dfRC$TRAVEL\_INIT\_CNTCTS<-factor(dfRC$TRAVEL\_INIT\_CNTCTS)

dfRC$SOLICITED\_CNTCTS<-factor(dfRC$SOLICITED\_CNTCTS)

dfRC$REFERRAL\_CNTCTS<-factor(dfRC$REFERRAL\_CNTCTS)

dfRC$CAMPUS\_VISIT<-factor(dfRC$CAMPUS\_VISIT)

I transformed all these variables due to them being ints. The variable sex was transformed but later not used. Enroll, premiere, and stucell all only had two options therefore they had to be transformed.

TRAVEL\_INIT\_CNTCTS, SOLICITED\_CNTCTS, REFERRAL\_CNTCTS, and CAMPUS\_VISIT were transformed as well because they only had a few different options compared to others. TRAVEL\_INIT\_CNTCTS, SOLICITED\_CNTCTS, and REFERRAL\_CNTCTS were later deleted because they were not significant to the dataset.

dfRC$ETHNICITY <- with(dfRC,impute(ETHNICITY,max))

dfRC$TERRITORY <- with(dfRC,impute(TERRITORY,max))

dfRC$satscore <- with(dfRC,impute(satscore,mean))

dfRC$avg\_income <- with(dfRC,impute(avg\_income,mean))

dfRC$distance <- with(dfRC,impute(distance,mean))

Then, I imputed some variables I thought were fitting towards the dataset. They had a lot of missing values so the ones that were categorical I chose to use maximum values, while variables that were numerical were given mean values.

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

Initial regression:

![A screenshot of text

Description automatically generated]()

Chart

Description automatically generated



Final Regression:

![Table

Description automatically generated]()

Chart

Description automatically generated

Chart, histogram

Description automatically generated

![Text

Description automatically generated]()

![A close up of text on a white background

Description automatically generated]()

![Text, letter

Description automatically generated]()

* 1. Tree plot for the decision tree model

Initial tree model:

![A close up of a receipt

Description automatically generated]()

  
Diagram

Description automatically generated

Chart

Description automatically generated

Final Tree model:

![A screenshot of a cell phone

Description automatically generated]()

![Text

Description automatically generated]()

Chart

Description automatically generatedDiagram

Description automatically generated

1. Which model will you choose? Why? Please provide support for your answer.

I believe the linear regression model is the best model out of the two. The linear regression model has a higher area under the curve in both instances which proves to be more accurate. The linear regression model both had AUC over .9 while the tree model both were below.

1. Please explain and summarize your major findings to the director of the Office of Enrollment Management.

Director of the Office of Enrollment Management,

I have done what you asked. I have found major findings in the dataset you have provided me to help you decide on the new enrollment process for our university. There were many different variables that showed to be irrelevant and others had to many missing values. First, I dropped the variables you recommended, and then ran a multicollinearity function on the remaining variables. This led to some major findings. I decided to drop some variables including CONTACT\_DATE, LEVEL\_YEAR, telecq, and sex because I found them to be irrelevant to the enrollment process. We found that TRAVEL\_INIT\_CNTCTS, SOLICITED\_CNTCTS, REFERRAL\_CNTCTS, and CAMPUS\_VISIT all had a high vif showing that these variables are not needed to be used to determine who gets enrolled. Later, I decided to run a baseline model to determine the p values of each variable. I found that ethnicity had high p values making it not statically efficient to the dataset.

Next, the model of linear regression was more precise then the tree model due to the values of each area under the curve. The area under the curve for the linear regression model was .9683 while the tree model was a mere .877.

Lastly, the tree model created two different decision trees based on its own findings. The first decision tree depicts that the SELF\_INIT variable is one of the most important variables and then HSCRAT, PREMIERE, and HSCRAT and stucell, however; due to the p value SELF\_INIT had to be deleted. The second decision tree depicts premiere being a top variable following satscore, HSCRAT and stucell.

These are just my opinions based on my findings. Please let me know if you have any questions.

Yours truly,

Ryan Carriere

1. Copy and paste your R codes at the end of the documents.

install.packages("caret")

install.packages("e1071")

install.packages("car")

install.packages("pROC")

install.packages("dplyr")

library(caret)

library(e1071)

library(car)

library(pROC)

library(dplyr)

setwd('C:/Users/grays/Desktop/ConsumerAnalytics')

dfRC <- read.csv("inq2015.csv",na.strings=c("NA",""))

summary(dfRC)

str(dfRC)

cols.dont.want <- c("ACADEMIC\_INTEREST\_1","ACADEMIC\_INTEREST\_2","IRSCHOOL","CONTACT\_CODE1","CONTACT\_DATE1")

dfRC <- dfRC[, ! names(dfRC) %in% cols.dont.want, drop = F]

#drop additional unwanted variables

#additional one that could be deleted is ethnicity

cols.dont.want <- c("CONTACT\_DATE","LEVEL\_YEAR","telecq","sex")

dfRC <- dfRC[, ! names(dfRC) %in% cols.dont.want, drop = F]

#comment out, before running the vif so it shows the high vif

cols.dont.want <- c("TOTAL\_CONTACTS","SELF\_INIT\_CNTCTS","TRAVEL\_INIT\_CNTCTS","SOLICITED\_CNTCTS","REFERRAL\_CNTCTS")

dfRC <- dfRC[, ! names(dfRC) %in% cols.dont.want, drop = F]

#comment out before running everything, p value too high

cols.dont.want <- c("ETHNICITY")

dfRC <- dfRC[, ! names(dfRC) %in% cols.dont.want, drop = F]

#variables in which vif was high

#transform variables from int to factor

dfRC$Enroll <- factor(dfRC$Enroll)

#dfRC$sex <- factor(dfRC$sex)

dfRC$premiere<-factor(dfRC$premiere)

dfRC$stucell<-factor(dfRC$stucell)

#dfRC$TRAVEL\_INIT\_CNTCTS<-factor(dfRC$TRAVEL\_INIT\_CNTCTS)

#dfRC$SOLICITED\_CNTCTS<-factor(dfRC$SOLICITED\_CNTCTS)

#dfRC$REFERRAL\_CNTCTS<-factor(dfRC$REFERRAL\_CNTCTS)

dfRC$CAMPUS\_VISIT<-factor(dfRC$CAMPUS\_VISIT)

str(dfRC)

summary(dfRC,na.strings=c("NA",""))

vif(glm(formula=Enroll ~ . , family = binomial(link='logit'),data = dfRC))

install.packages("Hmisc")

library(Hmisc)

#commented out, deleted due to pvalue

#dfRC$ETHNICITY <- with(dfRC,impute(ETHNICITY,max))

dfRC$TERRITORY <- with(dfRC,impute(TERRITORY,max))

dfRC$satscore <- with(dfRC,impute(satscore,mean))

dfRC$avg\_income <- with(dfRC,impute(avg\_income,mean))

dfRC$distance <- with(dfRC,impute(distance,mean))

summary(dfRC,na.strings=c("NA",""))

str(dfRC)

set.seed(60)

trainIndex <- createDataPartition(dfRC$Enroll,

p=0.7,

list=FALSE,

times=1)

dfRC.train <- dfRC[trainIndex,]

dfRC.valid <-dfRC[-trainIndex,]

#Run baseline model

baseline.model <- train(Enroll~.,

data=dfRC.train,

method='glm',

family='binomial',

na.action=na.pass)

summary(baseline.model)

#Evaluation model performance using the validation dataset

#Criteria 1: the confusion matrix

predictionRC <- predict(baseline.model,newdata=dfRC.valid)

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

confusionMatrix(predictionRC,dfRC.valid.nonmissing$Enroll)

predRC.probabilities <- predict(baseline.model,newdata=dfRC.valid,type='prob')

regressionRC.ROC <- roc(predictor=predRC.probabilities$`1`,

response=dfRC.valid.nonmissing$Enroll,

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

plot(regressionRC.ROC)

regressionRC.ROC$auc

# User-defined functions to calculate cumulative lift & gains

lift <- function(depvar, predcol, groups=10) {

if(is.factor(depvar)) depvar <- as.integer(as.character(depvar))

helper <- data.frame(cbind(depvar, predcol))

helper <- helper[order(-helper$predcol),]

helper[,"bucket"] = ntile(-helper[,"predcol"], groups)

gaintable = helper %>% group\_by(bucket) %>%

summarise\_at(vars(depvar), funs(total = n(),

totalresp=sum(., na.rm = TRUE))) %>%

mutate(Cumresp = cumsum(totalresp),

Gain=Cumresp/sum(totalresp)\*100,

Cumlift=Gain/(bucket\*(100/groups)))

return(gaintable)

}

dtRC=lift(dfRC.valid.nonmissing$Enroll,predRC.probabilities$`1`,groups=10)

print(dtRC)

plot(dtRC$bucket, dtRC$Cumlift, type="l", ylab="Cumulative lift", xlab="Bucket")

#DECISION TREE MODEL

install.packages('rpart')

install.packages('rpart.plot')

library(rpart)

library(rpart.plot)

# Build a decision tree model

treeRC.model <- train(Enroll~.,

data=dfRC.train,

method="rpart",

na.action=na.pass)

treeRC.model

prp(treeRC.model$finalModel,type=2,extra=106)

#Criteria 1: the confusion matrix

predictionRC <- predict(treeRC.model,newdata=dfRC.valid,na.action = na.pass)

confusionMatrix(predictionRC,dfRC.valid$Enroll)

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

treeRC.probabilities <- predict(treeRC.model,newdata=dfRC.valid,type='prob',na.action=na.pass)

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

response=dfRC.valid$Enroll,

levels=levels(dfRC.valid$Enroll))

plot(treeRC.ROC)

treeRC.ROC$auc