

*Toronto*

**GLM and Logistic Regression**

Soni Manana

a *College of Professional Studies****,*** *Master of Professional Studies in Analytics.*

***Subject****: ALY6015* ***NUID****: 002982645*

Under the guidance of

**Dr. Prof. Alex Maizlish**

**Introduction**

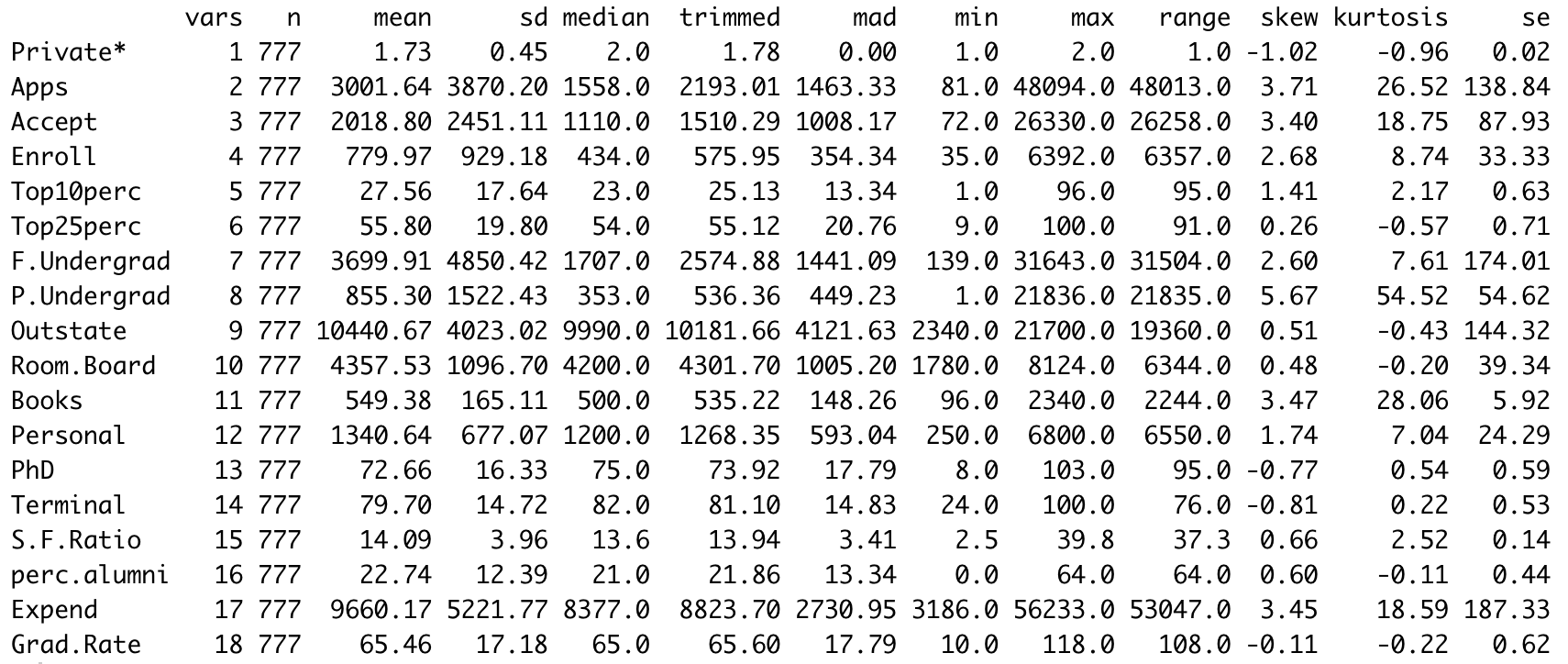
Data was provided from library named ISLR and dataset has data regarding college of US. Data of private and community colleges of United states at 1995 has been provided. We can answer many research question for that time such as how many applications were there, how many accepted, who enrolled from that. In addition, graduation rate and student/faculty ration have given in data so for financial and educational answers can be provided from this data. Purpose of this assignment is to learn GLM and Logistic Regression by implementation in R and interpretation of models. Here we will answer the question if a university or college is private or not by using logistic regression.

**Logistic Regression**

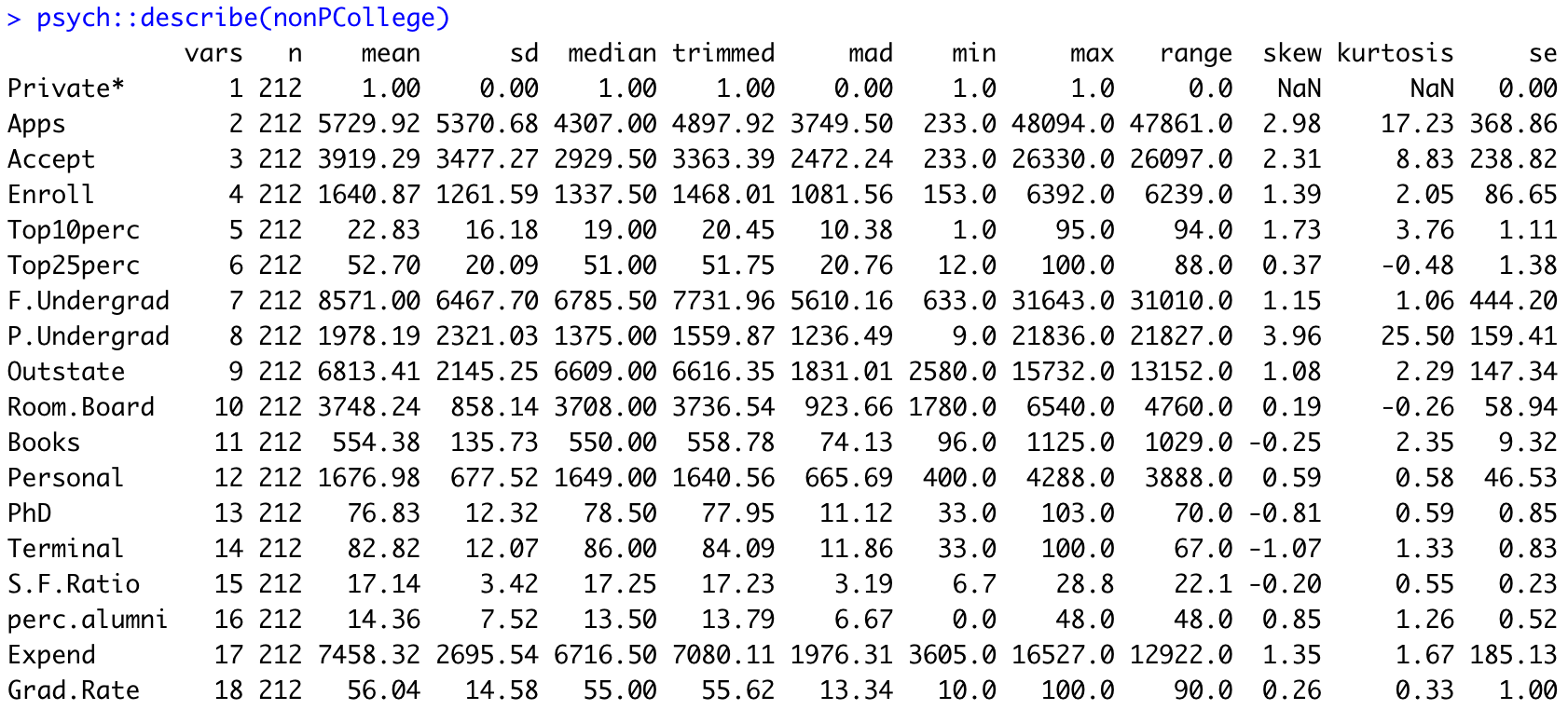
* A Logistic regression is a type of regression model we use when the response variable is binary or Boolean.
* For example, if someone asks a question and answer is either yes or no and by using data and logistic regression we can answer with evidence.
* We will find if there is any impact of Top 10% student

**Descriptive Analysis**

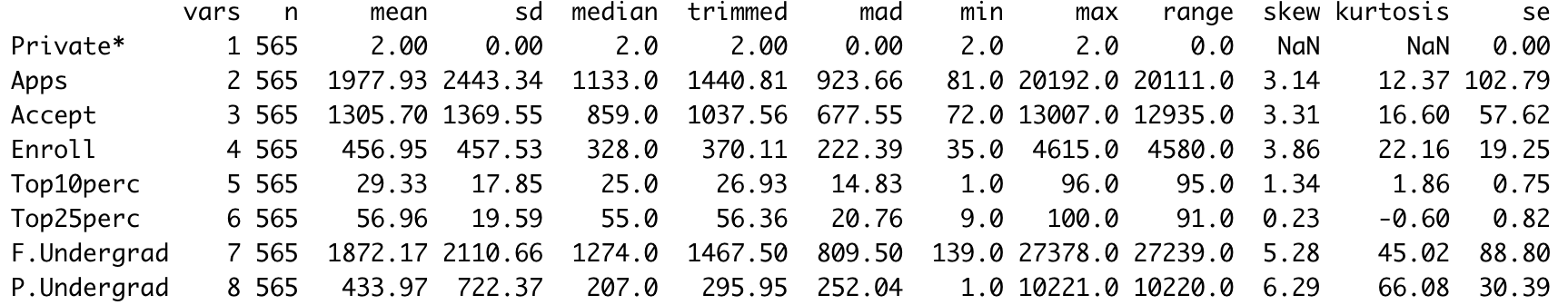
* Data has 777 samples and 18 variables.
* Skewness is varying from 5 to -1 which illustrates distribution variations.
* Mean applications to colleges is 3001 and standard deviation is 3870.



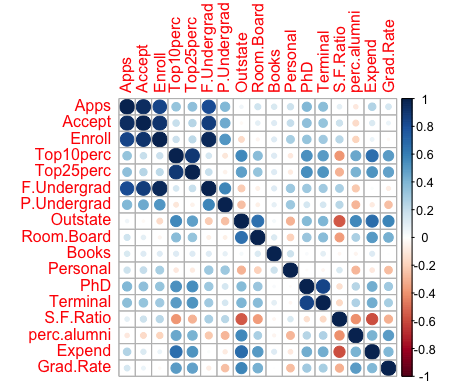
* To go in deep, I took subset of private and non-private found visible differences. I’ve mentioned below

****

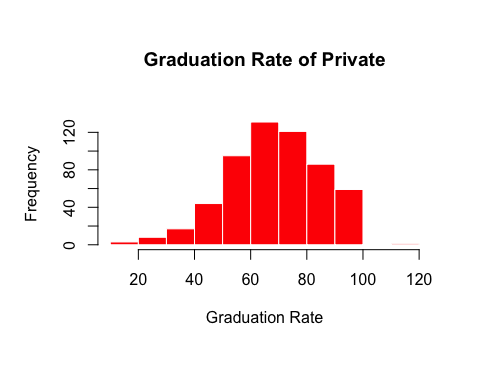
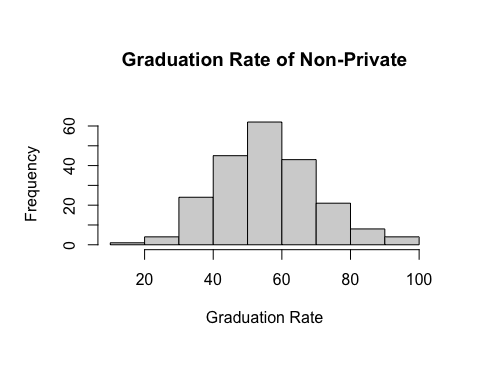
* As we can see describe of non-private college mean application rate is 5729 with SD of 5370 and mean fulltime student is 8571 whereas for private colleges have one third of non-private college in application as we can say private college could have more fees than non-private.

****

* Correlation plot shows us relations between the variable for regression problems.
* Correlation ease the process for feature modelling for models.
* From this we can say, Application, Acceptance and Enrolment have strong relation like they are growing with other.

****

* From these two histograms, we can note that passing chances is 10% more in private colleges.
* Most of colleges have set their graduation rate more than 50% and For non-private distribution is almost normal.



**Train-Test Split from Dataset**

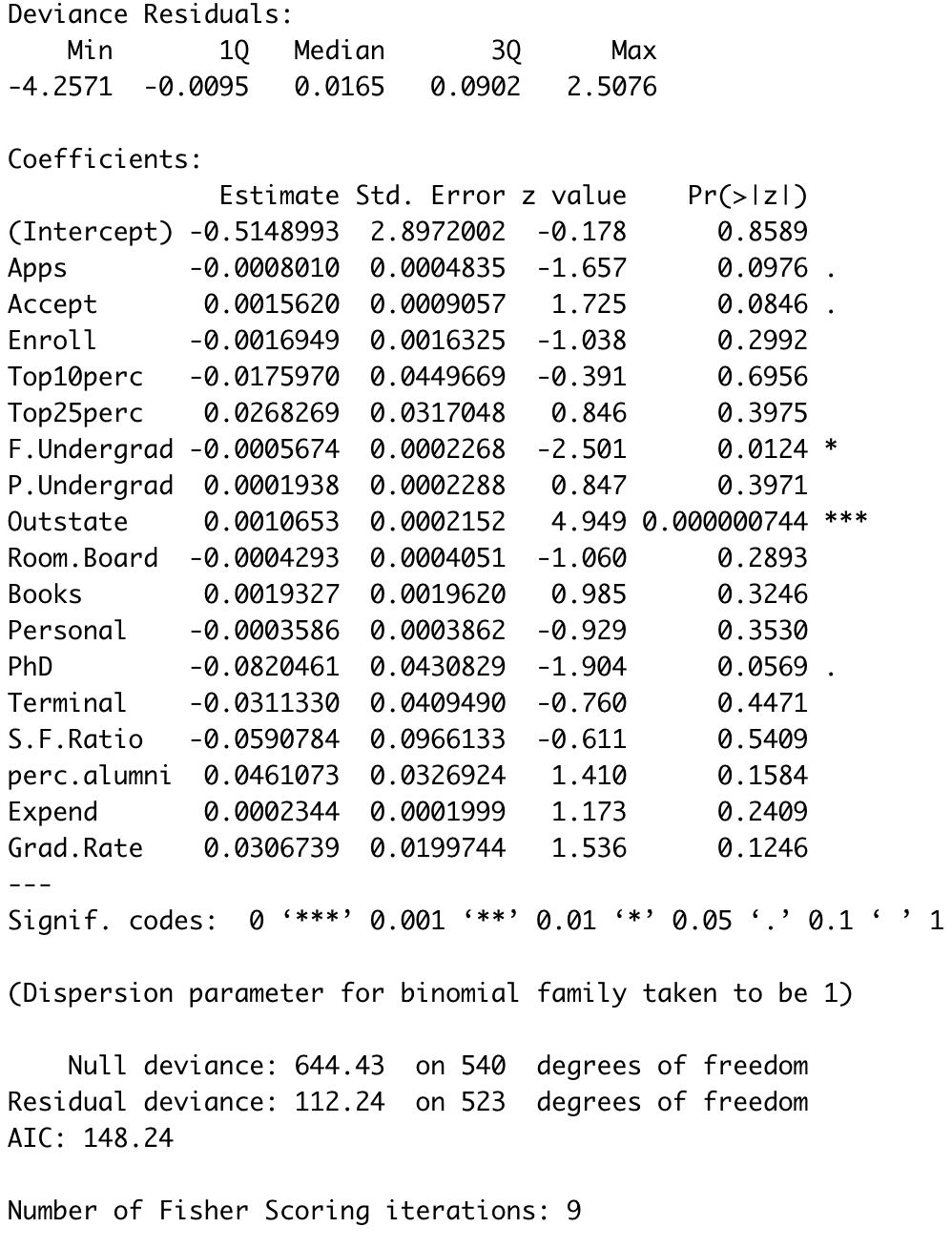
After splitting train-test data from College with the 70% threshold. So test will have 30% of data.

**Set.Seed()** is used for setting random number generation for random row distribution.

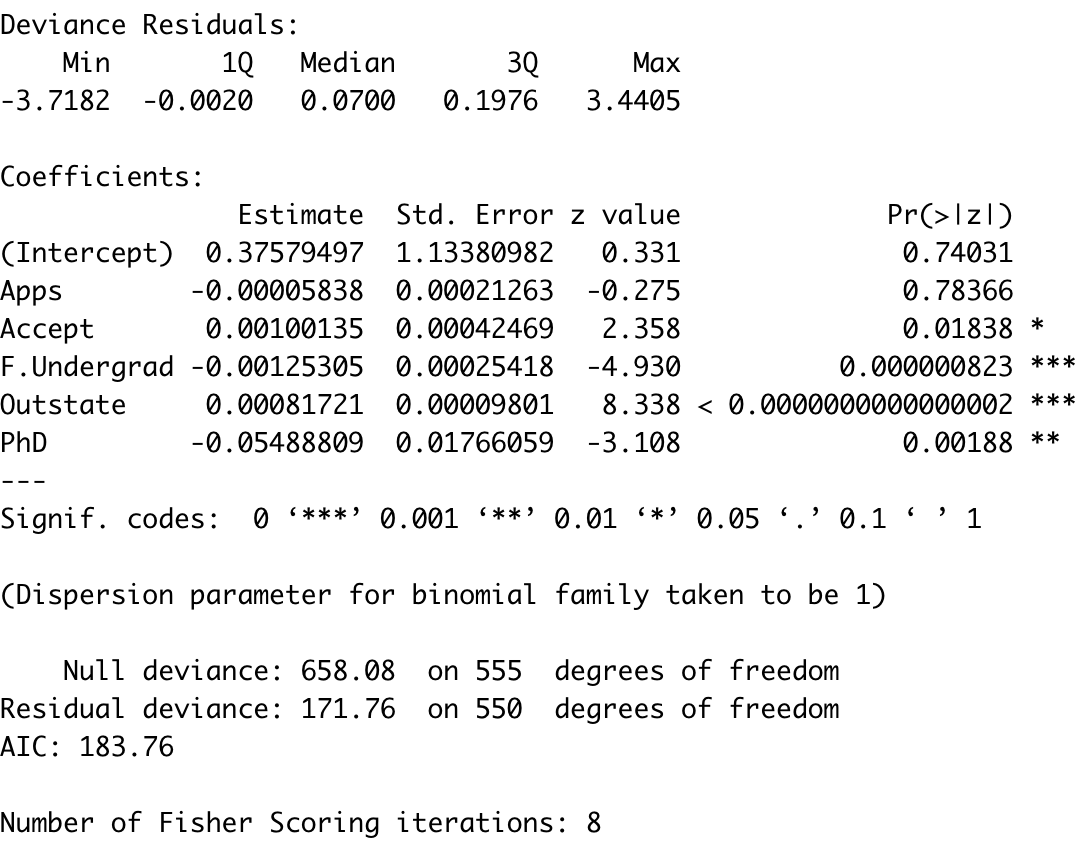
By using, **Sample(Data,prob=c(0.7,0.3))** we can split between train, test. And probability is argument where we can pass the percentage for setting the threshold.

**Logistic Regression and GLM**

* As we can see very few are scoring significance like around 0.05 and Most of them except number of Fulltime undergrads are signifies.
* We use **summary()** to get the performance measures of models.

****

* From this we can see that Fulltime undergrads, Outstate, PhDs and Accept variables are not scored to significance.

****

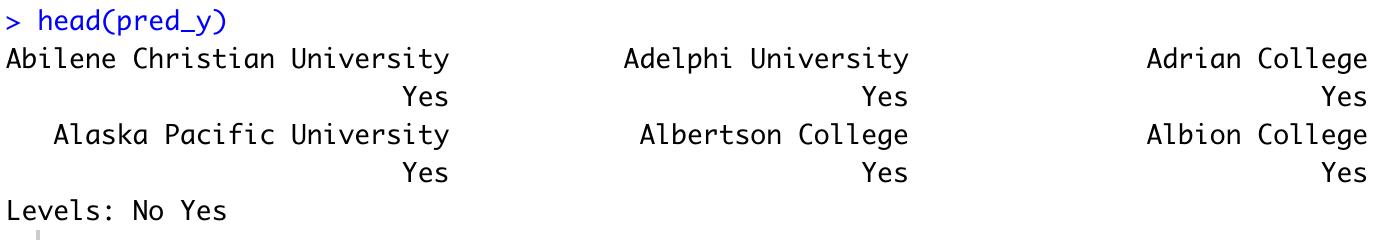
* It also provides deviances like from the mean known as null and residual for the model with predictors.

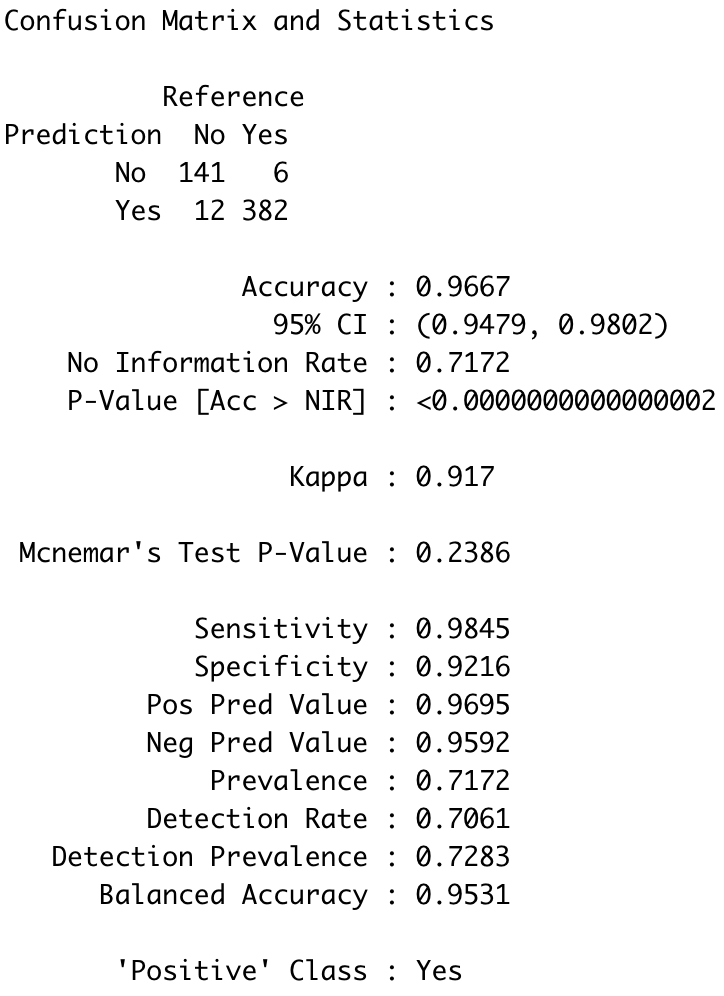
**Prediction for Test dataset**

We will use **Predict()** function to make prediction will have to feed model and data to make prediction.

After that, Initiated a variable for ‘Yes’ and ‘No’ for the response.

confusionMatrix() will be used to measure the accuracy of this model.

As we can see accuracy of this model is 0.9667 which means 96% accurate. We can calculate using method of True Positive, True Negative, False Negative and False Positive. But R does that for us.

****

method of True Positive, True Negative, False Negative and False Positive. But R does that for us. We will check this value by calculating on our own to test.

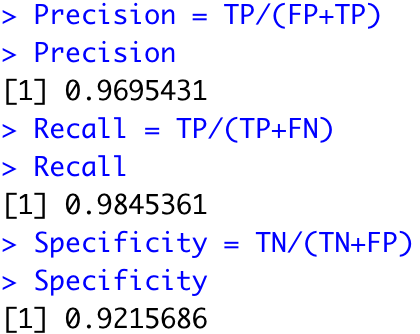
**Recall, Precision and Specificity**

Precision and Sensitivity are same which is known as Positive predictive value.

Recall is value of false negative

There are some threshold of these values if they are less than that we have to reject the model and work for the other variables.

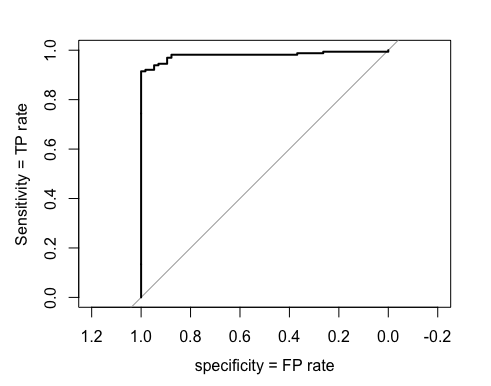
Specificity provides number for correctly identified among colleges.

****

**ROC and AUC**

AUC is the are under the curve of ROC and ROC is receiver operating characteristic curve which is to measure performance of the logistic regression or classification models.

For our model AUC value is 0.98%

****

****

**Conclusion**

From this we can learn how to make a logistic regression model for binary solutions. How to calculate miscalculation (False Negatives and Positives) by model.

Finding the accuracy and Performance by finding confusion matrix, ROC and AUC.

For the dataset, From the Applications, Accept, Fulltime Grads, Outstates and PhD are the variables who leading the model to 98%. As I mentioned, correlation matrix plays important role just like EDA and feature modelling.

Accuracy and AUC is just higher than 90-95% are consider as overfitting model however for qualified data it is possible.

**References**

Machine Learning Crash Course, Google *2022.*Classification: ROC Curce and AUC *Source:*https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

JournalDev, *2022. Confusion matrix in R Source:* https://www.journaldev.com/46732/confusion-matrix-in-r

**Appendix**

#install.packages('tidyverse')

#install.packages('ISLR')

#install.packages('psych')

#install.packages('ggridges')

#install.packages("simex")

#install.packages("InformationValue")

#install.packages("caret")

#install.packages('Hmisc')

#### INSTALL IF NOT #####

library(tidyverse)

library(ISLR)

library(ggridges)

library(simex)

library(InformationValue)

library(caret)

library(Hmisc)

data(College)

psych::describe(College)

hist(College$F.Undergrad,

col = 'red',border='white', main = paste("Number of applications received"))

df = subset(College)

library(corrplot)

corrplot(cor(df[2:18]))

nonPCollege = College %>% filter(College$Private == 'No')

PCollege = College %>% filter(College$Private == 'Yes')

psych::describe(nonPCollege)

psych::describe(PCollege)

hist(nonPCollege$Grad.Rate,main="Graduation Rate of Non-Private",xlab='Graduation Rate')

hist(PCollege$Grad.Rate,main="Graduation Rate of Private",xlab='Graduation Rate',col='red',border='white')

lmdl = lm(df$Enroll~df$Accept)

plot(df$Enroll~df$Accept ,main=" Enroll ~ Accept Students",xlab='Accepted Students',ylab='Enrolled Student')

abline(a=lmdl$coefficients[1],b=lmdl$coefficients[2])

summary(lmdl)

#02 train-test Spliting the Data

set.seed(120)

?set.seed

mIndex = sample(2, nrow(College),

replace = T,

prob = c(0.7,0.3))

train\_x = College[mIndex == 1,]

test\_x = College[mIndex == 2,]

head(train\_x)

set.seed(123)

mIndex = sample(2, nrow(College),

replace = T,

prob = c(0.7,0.3))

train\_x = College[mIndex == 1,]

test\_x = College[mIndex == 2,]

head(train\_x)

#03

LR\_model <- glm(Private ~ Apps + Accept + F.Undergrad + Outstate + PhD,

data = train\_x,

family = binomial(link = "logit"))

summary(LR\_model)

#04

install.packages("e1071")

library(e1071)

# Confusion Matrix

prob.train\_x = predict(LR\_model,

newdata = train\_x,

type = "response")

cm\_data = as.factor(ifelse

(prob.train\_x >= 0.5,

"Yes", "No"))

confusionMatrix(cm\_data,

train\_x$Private,

positive = 'Yes')

#install.packages("misclassGLM")

library(misclassGLM)

#05

TP = 383 # True +ve

TN = 127 # True -ve

FN = 16 # False -ve

FP = 19 # False +ve

# Predicted Accuracy

Accuracy = (TN + TP)/(TN+FP+FN+TP)

Accuracy

# Actual Accuracy

212/(212+565)

Precision = TP/(FP+TP)

Precision

Recall = TP/(TP+FN)

Recall

Specificity = TN/(TN+FP)

Specificity

#06 Create a confusion matrix and report the results of your model for the test set.

prob.test\_x = predict(LR\_model,

newdata = test\_x,

type = "response")

prob.test\_x

cm\_data = as.factor(ifelse

(prob.test\_x >= 0.5,

"Yes", "No"))

cm\_data

head(cm\_data)

confusionMatrix(cm\_data,

test\_x$Private,

positive = "Yes")

#07 Plot and interpret the ROC curve.

library(pROC)

ROC = roc (test\_x$Private, prob.test\_x)

X = plot(ROC,

col = "black",

ylab = "Sensitivity = TP rate",

xlab = 'specificity = FP rate')

#08 Calculate and interpret the AUC.

AUC = auc(ROC)

cat("ROC area under the curve = ", AUC)