Northeastern University
College of Professional Studies

ALY6015-GLM and Logistic Regression

Instructor: Prof. Ji-Young Yun

Date: 08 Feb 2022

By: Jeseeka Shah

Introduction

The assignment has question which must be performed using college dataset. Using different header files for performing the logistic regression , confusion matrix, calculating the accuracy, precision, recall , and specificity. The assignment involves the EDA on the data using different methods and plots and summary. The EDA helps in getting the information about the data and understanding the parameter are considered based on the descriptive analysis. The data given is then distributed as the train and the test data which is used for testing purpose. Here, after model creation, the next step will be checking the false positive and negativity which will be the based on how I tell the result of the analysis.

Analysis

After reading the data and performing the exploratory data analysis. Firstly, stating with the taking the summary of data. I get all the parameter such as mean, median, and different quartile values.

```
> summary(College)
 Private        Apps           Accept          Enroll         Top10perc       Top25perc
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
           Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0
           Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
           3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
           Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
  F.Undergrad     P.Undergrad         Outstate       Room.Board       Books
 Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0
 1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0
 Median : 1707   Median :  353.0   Median : 9990   Median :4200   Median : 500.0
 Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4
 3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0
 Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0
    Personal         PhD            Terminal       S.F.Ratio      perc.alumni
 Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00
 1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00
 Median :1200   Median : 75.00   Median : 82.0   Median :13.60   Median :21.00
 Mean   :1341   Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74
 3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00
 Max.   :6800   Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00
     Expend         Grad.Rate
 Min.   : 3186   Min.   : 10.00
 1st Qu.: 6751   1st Qu.: 53.00
 Median : 8377   Median : 65.00
 Mean   : 9660   Mean   : 65.46
 3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :56233   Max.   :118.00
```

Figure 1 : Showing the description of the dataset

I have considered the dependent variable as Private feature of the College dataset. I have removed the relation of other feature with respect to the categorical variable Private. Figure 2 is about the Expend vs. Private, when it's a private the expend is high as compared to non-private college. Figure 3 shows the accept vs enrol based on the college being private or not. The trend is that it's a linear graph in which accepted admission normally do enrol in the college. College those are non-private have a greater number of enrolment than that of private college based on the acceptancy rate as shown in the figure 3.

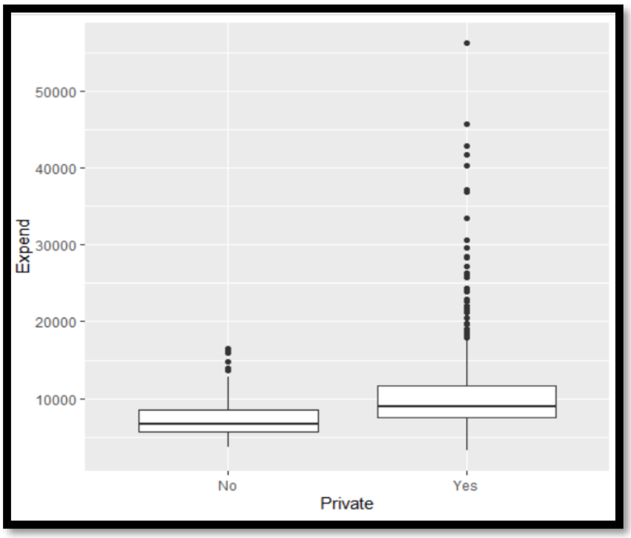

Figure 2: Private vs. Expend

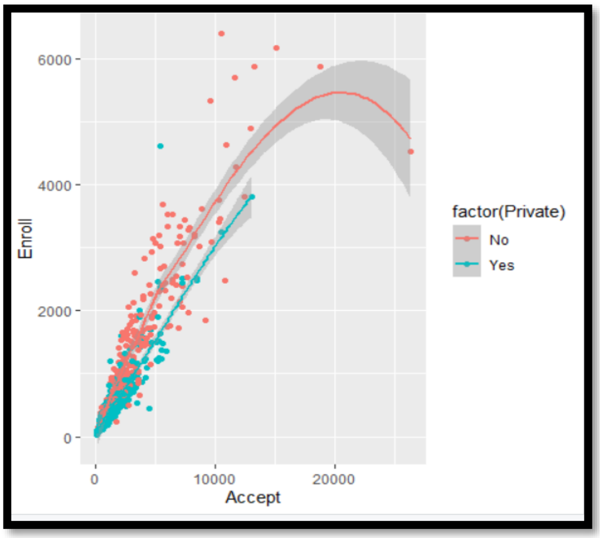

Figure 3: Accept Vs. Enroll based on the factor Private

The boxplot below shows the pass and fail of the undergrad students in the private colleges as per the figure 4. And the figure 5 shows the student to faculty ratio in non-private and private college. Its clear that student to faculty ratio is better in non-private college as compared to that of private.

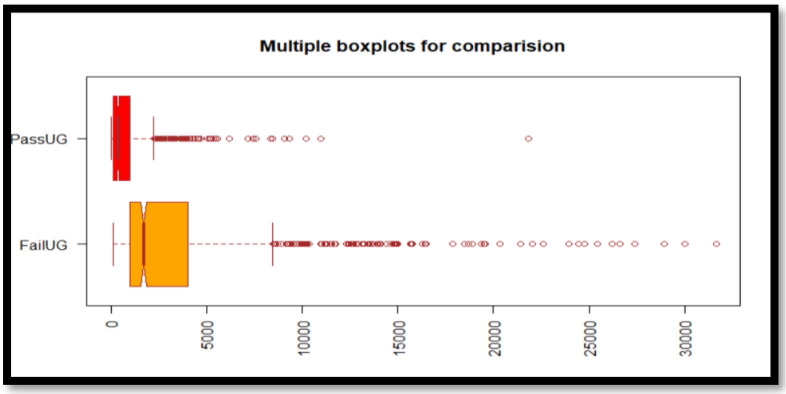

Figure 4: Boxplot that shows the passed and failed undergrad student

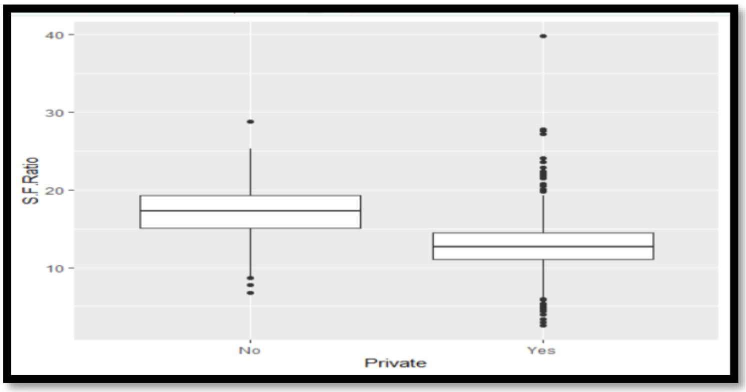

Figure 5: Student and Faculty ratio of the colleges

Taking the correlation matrix of all the numeric values to check the relation as shown in the figure 6. The scale shows which is correlated to one-another. The one tending towards 1 is highly correlated and this is depicted based on the colour encoding on the scale. Looking at this I have considered different combination to check the model creation.

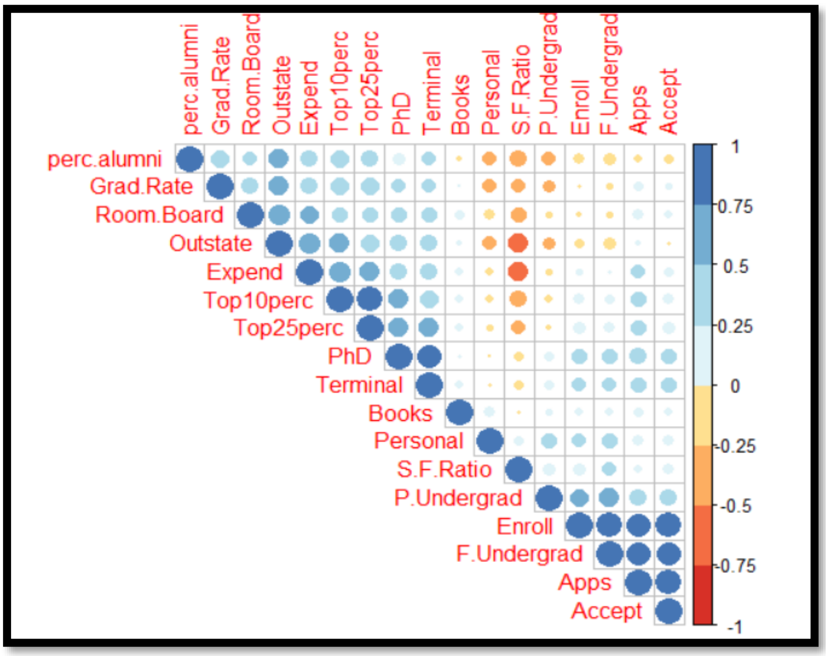

Figure 6: showing the correlation matrix of the college data set

Data distribution into train and test

Using the **createDataPartition** to create the partition of the college dataset based on the 70-30 ratio respectively as train and test dataset. This function belongs to the caret library when split the data as required. Figure 8 shows after distribution of the data into 70-30 the train data has around 281 observation and 546 observations out of 18 variables.

```
#Create the data partitionn
trainCollege1<-createDataPartition(College$Accept, p = 0.70, list = FALSE)
trainCollege1

College_train <- College[ trainCollege1,]
College_test <-College[-trainCollege1,]
```

Figure 7: splitting the data into train and test dataset

Data	
College_test	231 obs. of 18 variables
College_train	546 obs. of 18 variables

Figure 8: shows the distribution of train and test

Applying the logistic regression

To decide the features that needs to be considered for creating a model for checking public or private college. I have taken into consideration all the values that are shown in the model 1 in figure 9. Here , the factors that are needed to be check for a model to be fit are AIC and area under the curve. AUC shows the accuracy and that should be the more or close to 100. On the other hand, the AIC values should be less.

```
> model2 <- glm(Private~Apps + Enroll+F.Undergrad+P.Undergrad+Outstate+Books+Persona
+PhD+Expend,
+                data = College_train,family = binomial(link = "logit") )
> summary(model2)

Call:
glm(formula = Private ~ Apps + Enroll + F.Undergrad + P.Undergrad + 
    Outstate + Books + Personal + PhD + Expend, family = binomial(link = "logit"), 
    data = College_train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.9441  -0.0373   0.0324   0.1364   4.6301  

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.1651847  1.4161544  -0.117  0.90714    
Apps        -0.0004818  0.0001723  -2.797  0.00516 ** 
Enroll       0.0018942  0.0008690   2.180  0.02928 *  
F.Undergrad -0.0004623  0.0001596  -2.897  0.00377 ** 
P.Undergrad -0.0001435  0.0001734  -0.828  0.40783    
Outstate     0.0010107  0.0001393   7.256 4.00e-13 ***
Books        0.0013330  0.0014970   0.890  0.37324    
Personal    -0.0006367  0.0003133  -2.032  0.04215 *  
PhD         -0.1072038  0.0221354  -4.843 1.28e-06 ***
Expend       0.0003161  0.0001222   2.586  0.00970 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 643.92  on 545  degrees of freedom
Residual deviance: 169.57  on 536  degrees of freedom
AIC: 189.57

Number of Fisher Scoring iterations: 8
```

Figure 9 : Showing the regression mode

```
> ## get the coefficients of the model
> coef(model2)
  (Intercept)          Apps        Enroll   F.Undergrad   P.Undergrad      Outstate 
-1.921854e+00 -4.857349e-04  1.825397e-03 -4.915465e-04 -1.426308e-05  7.420917e-04 
        Books      Personal           PhD        Expend 
 2.558966e-03 -6.005083e-04 -6.754556e-02  3.527928e-04 
> ## display the regression coefficients (odds)
> exp(coef(model2))
(Intercept)        Apps      Enroll F.Undergrad P.Undergrad    Outstate       Books 
  0.1463354   0.9995144   1.0018271   0.9995086   0.9999857   1.0007424   1.0025622 
   Personal         PhD      Expend 
  0.9993997   0.9346851   1.0003529 
```

Figure 10: The coefficients and odds of the model

The figure 10 shows, the predictor factors predicting the result variable . The coefficients of the model that are not in better format and the next step was to take the log odds.
The probabilities of private rise by a factor of 0.9995 for every 1 unit rise in the number of applications received, according to various readings of the coefficients. For every 1 unit increase in the number of new students enrolled, the probability of private increase by a factor of 1.0027.

Now, taking the new model based on the train dataset using predict function as shown below. Here, the dependent variable is taken when the college is Private. First, I have created the model and the used coef() and then calculated the odds regression variables just like above.

```
> #Forward
> model_forward <- glm(formula = Private ~ F.Undergrad + P.Undergrad + Outstate + Grad.Rate + PhD
 +Outstate+ Apps + Accept + Expend + Enroll, data = College_train, family = binomial(link = "logi
t"))
> model_forward

Call:  glm(formula = Private ~ F.Undergrad + P.Undergrad + Outstate +
    Grad.Rate + PhD + Outstate + Apps + Accept + Expend + Enroll,
    family = binomial(link = "logit"), data = College_train)

Coefficients:
(Intercept)  F.Undergrad  P.Undergrad     Outstate    Grad.Rate          PhD         Apps
 -2.383e+00   -6.206e-04    5.399e-05    6.648e-04    2.395e-02   -6.698e-02   -5.356e-04
     Accept       Expend       Enroll
  1.964e-04    3.793e-04    1.929e-03

Degrees of Freedom: 545 Total (i.e. Null);  536 Residual
Null Deviance:      640
Residual Deviance: 186.3        AIC: 206.3
```

Figure 11: Model created with relevant variables

```
> test_mat=confusionMatrix(pre_forward.min, College_train$Private, positive='Yes')
> test_mat
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  130  12
       Yes  19 385

               Accuracy : 0.9432
                 95% CI : (0.9204, 0.9611)
    No Information Rate : 0.7271
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8548

 Mcnemar's Test P-Value : 0.2812

            Sensitivity : 0.9698
            Specificity : 0.8725
         Pos Pred Value : 0.9530
         Neg Pred Value : 0.9155
             Prevalence : 0.7271
         Detection Rate : 0.7051
   Detection Prevalence : 0.7399
      Balanced Accuracy : 0.9211

       'Positive' Class : Yes
```

Figure 12 : Showing the confusion matrix

The figure 12 shows that the confusion matrix for the train dataset has an accuracy of 94 percent where 512 cases were predicted correctly. The false positive is 19 , and false negative is 12 . The most impacting parameter are false negative. But in our case the impact is less as the numbers are small. The true positive is 385 and true negative are 130.

```
> auc = auc(ROC1)
> auc
Area under the curve: 0.9796
> auc = auc(ROC1)
> auc
Area under the curve: 0.9796
> libra
```

Figure 13 : area under the curve

The confusion matrix for showing the model with accuracy of 94.32 meaning that model is performing that the unseen data so accurately. The sensitivity is 96.98 % and specificity is 87.25% as shown in the figure 12. The precision is 95.30 %, indicating that the model predicted the actual positive cases 95.30 % accurately. The accuracy for the area under the curve is 97 percent.

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No   57   6
       Yes   6 162

               Accuracy : 0.9481
                 95% CI : (0.911, 0.9729)
    No Information Rate : 0.7273
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.869

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9643
            Specificity : 0.9048
         Pos Pred Value : 0.9643
         Neg Pred Value : 0.9048
             Prevalence : 0.7273
         Detection Rate : 0.7013
   Detection Prevalence : 0.7273
      Balanced Accuracy : 0.9345

       'Positive' Class : Yes
```

```
> auc
Area under the curve: 0.9795
```

Figure 14: test confusion matrix

Figure 15 : Area under the curve for test

Looking at the accuracy of the test data which is 94.81 percent which is close to that of train data. Even the sensitivity is almost close with 96.43 percent and specificity is 90.48 percent. Even here the accuracy for the area under the curve is close to the area under the curve accuracy for the train data part.

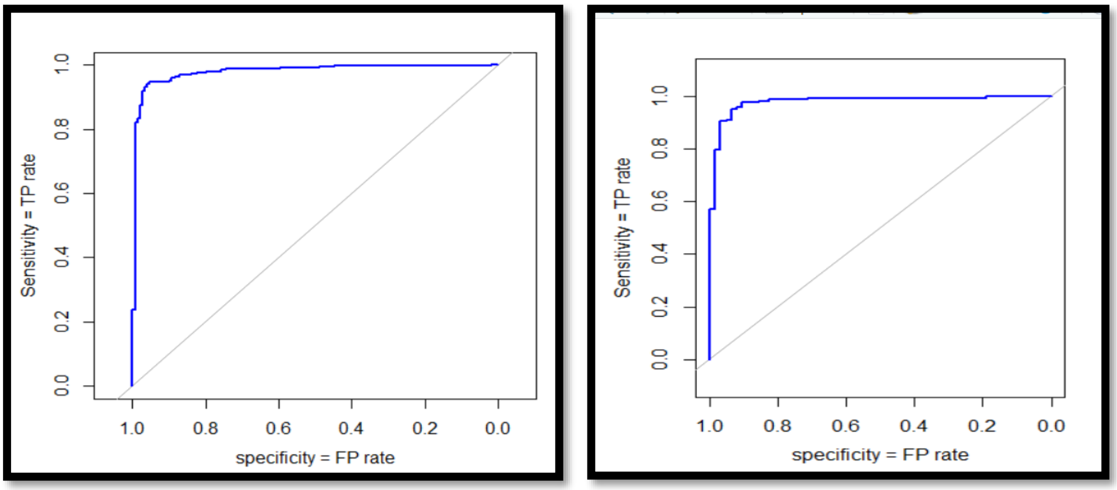

Figure 16 :showing graphically the curve of both train and test respectively

Conclusions

I have made three model and almost all gave the similar result but the model that I have explained in the report is the one with low AIC values compared to other models. From the analysis answered the logistic regression model can be used to predict that college public or private , the result proved that model fits perfectly with accuracy of 94 percent on the test set. Area under the curve shows 97 percent of good generalization based on the test data which the model had not seen. The confusion matrix and other metrics proves with parameter that the model is accurate enough even when new dataset is used for identifying the college type. The model predicts well with positive, and negatives based on the confusion matrix generated. Therefore, this model is effective in predicting the university type.

References

Excillio. (2022). Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures - Exsilio Blog. Retrieved 2 February 2022, from https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Steward, K. (2019). Sensitivity vs. Specificity. Retrieved 2 February 2022, from https://www.technologynetworks.com/analysis/articles/sensitivity-vs-specificity-318222

Appendix

```
#Header Files
install.packages("caret")
install.packages("ISLR")
library('caret')
library(ISLR)
library(Hmisc)
library(corrplot)
library(RColorBrewer)
library(pROC)

install.packages("dplyr")                         # Install dplyr
library("dplyr")

#Reading the data
attach(College)
#College
#Performing the descriptive analytics
describe(College)
summary(College)
is.null(College)

 ###################################
#Creating the plot to show the trend that the Accept and Enroll for the college dataset
qplot(Accept, Enroll, data = College, color = factor(Private),
      geom=c("point", "smooth"))
## Private and Expend ratio and target variable
College %>% ggplot(aes(Private, Expend))+
  geom_boxplot()

## Boxplot showing F.Undergrad and pass rate  for undergrad
boxplot(College$F.Undergrad,College$P.Undergrad,
        main = "Multiple boxplots for comparision",
        at = c(1,2),
        names = c("FailUG", "PassUG"),
        las = 2,
        col = c("orange","red"),
        border = "brown",
        horizontal = TRUE,
        notch = TRUE
)
# Private college vs. Public college
College %>% ggplot(aes(Private, S.F.Ratio))+
  geom_boxplot()

# making the correlation plot
num_cols <- unlist(lapply(College, is.numeric))         # Identify numeric columns
num_cols
#Saving the numeric values into a variables
data_num2 <- select_if(College, is.numeric)             # Subset numeric columns with dplyr
data_num2

#Correaltion Matrix
```

```
#Numerical
M <-cor(data_num2)
corrplot(M, type="upper", order="hclust",
        col=brewer.pal(n=8, name="RdYlBu"))
##################################
#Splitting the data into 70-30 train-test
#Create the data partitionn
trainCollege1<-createDataPartition(College$Accept, p = 0.70, list = FALSE)
trainCollege1

College_train <- College[ trainCollege1,]
College_test <-College[-trainCollege1,]
head(trainCollege1)
###############################################

model1 <- glm(Private~., data = College_train, family = binomial(link = "logit"))
summary(model1)


model2 <- glm(Private~Apps +
Enroll+F.Undergrad+P.Undergrad+Outstate+Books+Personal+PhD+Expend,
             data = College_train,family = binomial(link = "logit") )
summary(model2)

## get the coefficients of the model
coef(model2)

## display the regression coefficients (odds)
exp(coef(model2))
########################################
#Feature Selection Method
#Forward Selection Method

#Forward
model_forward <- glm(formula = Private ~ F.Undergrad + P.Undergrad + Outstate + Grad.Rate +
PhD + Apps + Accept + Expend + Enroll, data = College_train, family = binomial(link = "logit"))
model_forward
summary(model_forward)
colnames(College)

## get the coefficients of the model
coef(model_forward)

## display the regression coefficients (odds)
exp(coef(model_forward))

#### tRAIN DATA #1
pre_forward.train =predict(model_forward, newdata=College_train, type="response")
pre_forward.min<- as.factor(ifelse(pre_forward.train >=0.5,"Yes","No"))
pre_forward.min

 #Creating a Confusion Matrix
train_mat=confusionMatrix(pre_forward.min, College_train$Private, positive='Yes')
```

```
train_mat


#Roc  for the forward _ Model
ROC1 = roc(College_train$Private, pre_forward.train)
plot(ROC1, col = "blue", ylab = "Sensitivity = TP rate", xlab = 'specificity = FP rate')

auc = auc(ROC1)
auc
#################################



### tEST DATA  #1
pre_forward.test =predict(model_forward, newdata=College_test, type="response")
pre_forward.min_tst<- as.factor(ifelse(pre_forward.test >=0.5,"Yes","No"))
pre_forward.min_tst

#Creating a Confusion Matrix
test_mat=confusionMatrix(pre_forward.min_tst, College_test$Private, positive='Yes')
test_mat

############
#Roc  for the forward _ Model
ROC1 = roc(College_test$Private, pre_forward.test)
plot(ROC1, col = "blue", ylab = "Sensitivity = TP rate", xlab = 'specificity = FP rate')

auc = auc(ROC1)
auc


################### Model 2###############################################333
#Significant removed P.Undergrad , Grad.Rate. Apps
#Relevant for the model :- F.Undergrad ,

model_backward <- glm(formula = Private ~ F.Undergrad + PhD + Outstate + Expend , data =
College, family = binomial(link = "logit"))
model_backward

summary(model_backward)
# Getting Regression Coefficitents
coef(model_backward)


#Creating a dataset to check the impact on the probabilites
model_test_data



#Displaying the regression coefficient in r
exp(coef(model_backward))



#confusion matrix train data #2
```

```
Pre_backward_1=predict(model_backward, College_train, type="response")

pre.min_train<- as.factor(ifelse(Pre_backward_1 >=0.5,"Yes","No"))

train_mat_2=confusionMatrix(pre.min_train, College_train$Private, positive='Yes')
train_mat_2

#Roc
ROC1 = roc(College_train$Private,Pre_backward_1 )
plot(ROC1, col = "blue", ylab = "Sensitivity = TP rate", xlab = 'specificity = FP rate')

auc = auc(ROC1)
auc

###Test  confusion matrix test data #2

Pre_backward_test=predict(model_backward, College_test, type="response")

pre.min_test<- as.factor(ifelse(Pre_backward_test >=0.5,"Yes","No"))

test_mat_2=confusionMatrix(pre.min_test, College_test$Private, positive='Yes')
test_mat_2

#Roc
ROC1 = roc(College_test$Private,Pre_backward_test )
plot(ROC1, col = "blue", ylab = "Sensitivity = TP rate", xlab = 'specificity = FP rate')

auc = auc(ROC1)
auc

#Taking Final Model with same values and less variables
# Taking another model  with four parameter
#Significant removed P.Undergrad , Grad.Rate , removing Expend and adding prec.alumini
#Relevant for the model :- F.Undergrad ,

model_backward_2<- glm(formula = Private ~  PhD + F.Undergrad +  Outstate +perc.alumni +
Expend , data = College, family = binomial(link = "logit"))
model_backward_2

summary(model_backward_2)
# Getting Regression Coefficitents
coef(model_backward)


#Creating a dataset to check the impact on the probabilites
model_test_data


#Displaying the regression coefficient in r
exp(coef(model_backward_2))

################################ Train
#confusion matrix train #3
```

```
Pre_backward_2=predict(model_backward_2, College_train, type="response")

pre.min<- as.factor(ifelse(Pre >=0.5,"Yes","No"))

test_mat=confusionMatrix(pre.min, College_train$Private, positive='Yes')
test_mat

#Roc
ROC1 = roc(College_train$Private, Pre_backward_2)
plot(ROC1, col = "blue", ylab = "Sensitivity = TP rate", xlab = 'specificity = FP rate')

auc = auc(ROC1)
auc

###################################### Test

#confusion matrix test #3

Pre_backward_2_test=predict(model_backward_2, College_test, type="response")

pre.min_test<- as.factor(ifelse(Pre_backward_2_test >=0.5,"Yes","No"))

test_mat=confusionMatrix(pre.min_test, College_test$Private, positive='Yes')
test_mat

#Roc
ROC1 = roc(College_test$Private, Pre_backward_2_test)
plot(ROC1, col = "blue", ylab = "Sensitivity = TP rate", xlab = 'specificity = FP rate')

auc = auc(ROC1)
auc

# In this model the accuracy is too lower than other model.
```