Northeastern University

Toronto

GLM and Logistic Regression

Soni Manan[a]

[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.

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Private*	1	777	1.73	0.45	2.0	1.78	0.00	1.0	2.0	1.0	-1.02	-0.96	0.02
Apps	2	777	3001.64	3870.20	1558.0	2193.01	1463.33	81.0	48094.0	48013.0	3.71	26.52	138.84
Accept	3	777	2018.80	2451.11	1110.0	1510.29	1008.17	72.0	26330.0	26258.0	3.40	18.75	87.93
Enroll	4	777	779.97	929.18	434.0	575.95	354.34	35.0	6392.0	6357.0	2.68	8.74	33.33
Top10perc	5	777	27.56	17.64	23.0	25.13	13.34	1.0	96.0	95.0	1.41	2.17	0.63
Top25perc	6	777	55.80	19.80	54.0	55.12	20.76	9.0	100.0	91.0	0.26	-0.57	0.71
F.Undergrad	7	777	3699.91	4850.42	1707.0	2574.88	1441.09	139.0	31643.0	31504.0	2.60	7.61	174.01
P.Undergrad	8	777	855.30	1522.43	353.0	536.36	449.23	1.0	21836.0	21835.0	5.67	54.52	54.62
Outstate	9	777	10440.67	4023.02	9990.0	10181.66	4121.63	2340.0	21700.0	19360.0	0.51	-0.43	144.32
Room.Board	10	777	4357.53	1096.70	4200.0	4301.70	1005.20	1780.0	8124.0	6344.0	0.48	-0.20	39.34
Books	11	777	549.38	165.11	500.0	535.22	148.26	96.0	2340.0	2244.0	3.47	28.06	5.92
Personal	12	777	1340.64	677.07	1200.0	1268.35	593.04	250.0	6800.0	6550.0	1.74	7.04	24.29
PhD	13	777	72.66	16.33	75.0	73.92	17.79	8.0	103.0	95.0	-0.77	0.54	0.59
Terminal	14	777	79.70	14.72	82.0	81.10	14.83	24.0	100.0	76.0	-0.81	0.22	0.53
S.F.Ratio	15	777	14.09	3.96	13.6	13.94	3.41	2.5	39.8	37.3	0.66	2.52	0.14
perc.alumni	16	777	22.74	12.39	21.0	21.86	13.34	0.0	64.0	64.0	0.60	-0.11	0.44
Expend	17	777	9660.17	5221.77	8377.0	8823.70	2730.95	3186.0	56233.0	53047.0	3.45	18.59	187.33
Grad.Rate	18	777	65.46	17.18	65.0	65.60	17.79	10.0	118.0	108.0	-0.11	-0.22	0.62

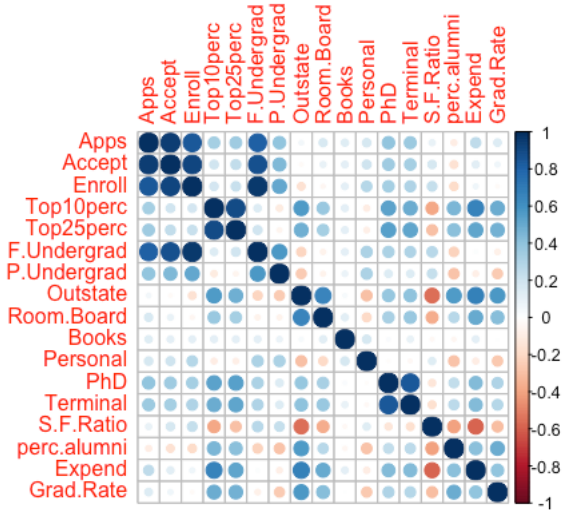
- To go in deep, I took subset of private and non-private found visible differences. I've mentioned below

```
> psych::describe(nonPCollege)
             vars   n    mean      sd  median trimmed     mad    min      max    range  skew kurtosis     se
Private*        1 212    1.00    0.00    1.00    1.00    0.00    1.0     1.00      0.0   NaN      NaN   0.00
Apps            2 212 5729.92 5370.68 4307.00 4897.92 3749.50  233.0 48094.0 47861.0  2.98    17.23 368.86
Accept          3 212 3919.29 3477.27 2929.50 3363.39 2472.24  233.0 26330.0 26097.0  2.31     8.83 238.82
Enroll          4 212 1640.87 1261.59 1337.50 1468.01 1081.56  153.0  6392.0  6239.0  1.39     2.05  86.65
Top10perc       5 212   22.83   16.18   19.00   20.45   10.38    1.0    95.0    94.0  1.73     3.76   1.11
Top25perc       6 212   52.70   20.09   51.00   51.75   20.76   12.0   100.0    88.0  0.37    -0.48   1.38
F.Undergrad     7 212 8571.00 6467.70 6785.50 7731.96 5610.16  633.0 31643.0 31010.0  1.15     1.06 444.20
P.Undergrad     8 212 1978.19 2321.03 1375.00 1559.87 1236.49    9.0 21836.0 21827.0  3.96    25.50 159.41
Outstate        9 212 6813.41 2145.25 6609.00 6616.35 1831.01 2580.0 15732.0 13152.0  1.08     2.29 147.34
Room.Board     10 212 3748.24  858.14 3708.00 3736.54  923.66 1780.0  6540.0  4760.0  0.19    -0.26  58.94
Books          11 212  554.38  135.73  550.00  558.78   74.13   96.0  1125.0  1029.0 -0.25     2.35   9.32
Personal       12 212 1676.98  677.52 1649.00 1640.56  665.69  400.0  4288.0  3888.0  0.59     0.58  46.53
PhD            13 212   76.83   12.32   78.50   77.95   11.12   33.0   103.0    70.0 -0.81     0.59   0.85
Terminal       14 212   82.82   12.07   86.00   84.09   11.86   33.0   100.0    67.0 -1.07     1.33   0.83
S.F.Ratio      15 212   17.14    3.42   17.25   17.23    3.19    6.7    28.8    22.1 -0.20     0.55   0.23
perc.alumni    16 212   14.36    7.52   13.50   13.79    6.67    0.0    48.0    48.0  0.85     1.26   0.52
Expend         17 212 7458.32 2695.54 6716.50 7080.11 1976.31 3605.0 16527.0 12922.0  1.35     1.67 185.13
Grad.Rate      18 212   56.04   14.58   55.00   55.62   13.34   10.0   100.0    90.0  0.26     0.33   1.00
```

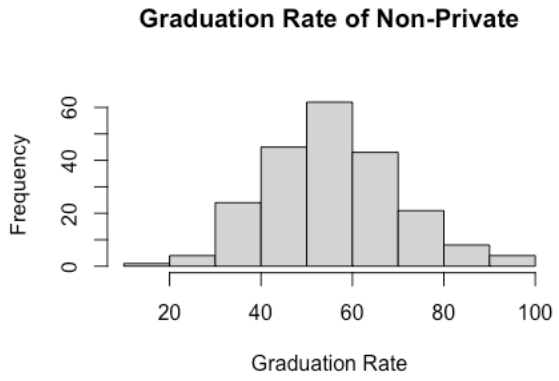
- As we can see describe of non-private college mean application rate is 5729 with SD of 5370 and mean fulltime student is 8571 whereas for private colleges have one third of non-private college in application as we can say private college could have more fees than non-private.

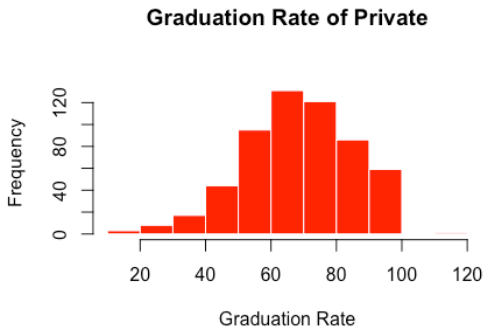
```
             vars   n    mean      sd  median trimmed     mad    min      max    range  skew kurtosis     se
Private*        1 565    2.00    0.00     2.0    2.00    0.00    2.0     2.0      0.0   NaN      NaN   0.00
Apps            2 565 1977.93 2443.34  1133.0 1440.81  923.66   81.0 20192.0 20111.0  3.14    12.37 102.79
Accept          3 565 1305.70 1369.55   859.0 1037.56  677.55   72.0 13007.0 12935.0  3.31    16.60  57.62
Enroll          4 565  456.95  457.53   328.0  370.11  222.39   35.0  4615.0  4580.0  3.86    22.16  19.25
Top10perc       5 565   29.33   17.85    25.0   26.93   14.83    1.0    96.0    95.0  1.34     1.86   0.75
Top25perc       6 565   56.96   19.59    55.0   56.36   20.76    9.0   100.0    91.0  0.23    -0.60   0.82
F.Undergrad     7 565 1872.17 2110.66  1274.0 1467.50  809.50  139.0 27378.0 27239.0  5.28    45.02  88.80
P.Undergrad     8 565  433.97  722.37   207.0  295.95  252.04    1.0 10221.0 10220.0  6.29    66.08  30.39
```

- 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.

```
Coefficients:
              Estimate Std. Error z value    Pr(>|z|)
(Intercept) -0.5148993  2.8972002  -0.178      0.8589
Apps        -0.0008010  0.0004835  -1.657      0.0976 .
Accept       0.0015620  0.0009057   1.725      0.0846 .
Enroll      -0.0016949  0.0016325  -1.038      0.2992
Top10perc   -0.0175970  0.0449669  -0.391      0.6956
Top25perc    0.0268269  0.0317048   0.846      0.3975
F.Undergrad -0.0005674  0.0002268  -2.501      0.0124 *
P.Undergrad  0.0001938  0.0002288   0.847      0.3971
Outstate     0.0010653  0.0002152   4.949 0.000000744 ***
Room.Board  -0.0004293  0.0004051  -1.060      0.2893
Books        0.0019327  0.0019620   0.985      0.3246
Personal    -0.0003586  0.0003862  -0.929      0.3530
PhD         -0.0820461  0.0430829  -1.904      0.0569 .
Terminal    -0.0311330  0.0409490  -0.760      0.4471
S.F.Ratio   -0.0590784  0.0966133  -0.611      0.5409
perc.alumni  0.0461073  0.0326924   1.410      0.1584
Expend       0.0002344  0.0001999   1.173      0.2409
Grad.Rate    0.0306739  0.0199744   1.536      0.1246
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 644.43  on 540  degrees of freedom
Residual deviance: 112.24  on 523  degrees of freedom
AIC: 148.24
```

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

```
Coefficients:
               Estimate  Std. Error z value             Pr(>|z|)
(Intercept)  0.37579497  1.13380982   0.331              0.74031
Apps        -0.00005838  0.00021263  -0.275              0.78366
Accept       0.00100135  0.00042469   2.358              0.01838 *
F.Undergrad -0.00125305  0.00025418  -4.930          0.000000823 ***
Outstate     0.00081721  0.00009801   8.338 < 0.0000000000000002 ***
PhD         -0.05488809  0.01766059  -3.108              0.00188 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 658.08  on 555  degrees of freedom
Residual deviance: 171.76  on 550  degrees of freedom
AIC: 183.76

Number of Fisher Scoring iterations: 8
```

- 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.

```
> head(pred_y)
Abilene Christian University           Adelphi University               Adrian College
                         Yes                          Yes                          Yes
   Alaska Pacific University            Albertson College               Albion College
                         Yes                          Yes                          Yes
Levels: No Yes
```

```
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  141   6
       Yes  12 382

               Accuracy : 0.9667
                 95% CI : (0.9479, 0.9802)
    No Information Rate : 0.7172
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.917

 Mcnemar's Test P-Value : 0.2386

            Sensitivity : 0.9845
            Specificity : 0.9216
         Pos Pred Value : 0.9695
         Neg Pred Value : 0.9592
             Prevalence : 0.7172
         Detection Rate : 0.7061
   Detection Prevalence : 0.7283
      Balanced Accuracy : 0.9531

       'Positive' Class : Yes
```

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.

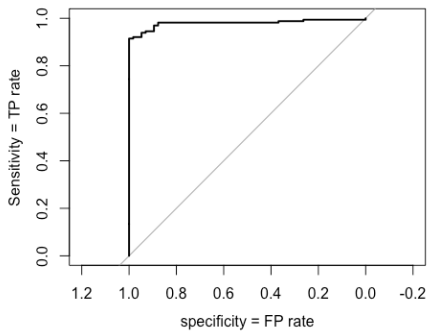
Specificity provides number for correctly identified among colleges.

```
> Precision = TP/(FP+TP)
> Precision
[1] 0.9695431
> Recall = TP/(TP+FN)
> Recall
[1] 0.9845361
> Specificity = TN/(TN+FP)
> Specificity
[1] 0.9215686
```

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%




```
ROC area under the curve =  0.9799957
```

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)
```