GLM & Logistic Regression

Name : Shivani Vellanki

Date : March 19, 2023.

Introduction:

In this project, we're using R to analyze the 'College' dataset from the ISLR package. We use EDA, divide the dataset into train and test sets, fit a logistic regression model to both train and test sets using the glm() function, and then build a confusion matrix.

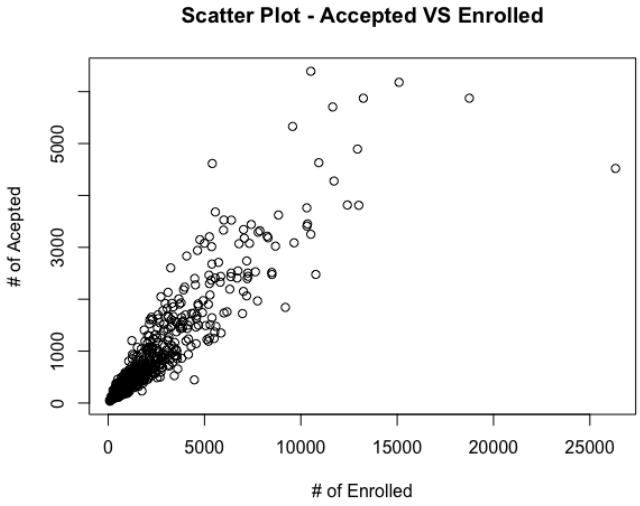
Analysis:

1. **Importing the Dataset into R and Exploratory Data Analysis:**

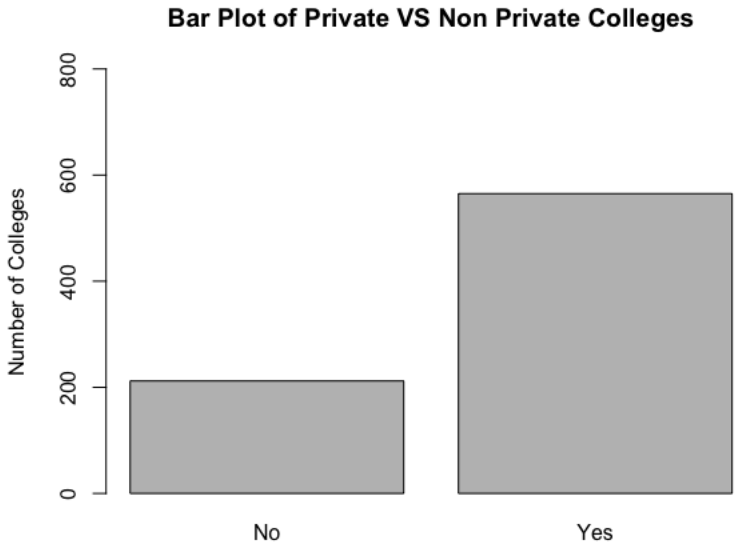
```
> College <- ISLR :: College
>
> summary(College)
 Private        Apps           Accept          Enroll       Top10perc       Top25perc
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   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           Personal
 Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
 1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
 Median : 1707   Median :  353.0   Median : 9990   Median :4200   Median : 500.0   Median :1200
 Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341
 3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
 Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800
      PhD            Terminal       S.F.Ratio      perc.alumni        Expend        Grad.Rate
 Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median : 75.00   Median : 82.0   Median :13.60   Median :21.00   Median : 8377   Median : 65.00
 Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
> psych::describe(College)
             vars   n     mean      sd median  trimmed     mad    min     max   range  skew kurtosis
Private*        1 777     1.73    0.45    2.0     1.78    0.00    1.0     2.0     1.0 -1.02    -0.96
Apps            2 777  3001.64 3870.20 1558.0  2193.01 1463.33   81.0 48094.0 48013.0  3.71    26.52
Accept          3 777  2018.80 2451.11 1110.0  1510.29 1008.17   72.0 26330.0 26258.0  3.40    18.75
Enroll          4 777   779.97  929.18  434.0   575.95  354.34   35.0  6392.0  6357.0  2.68     8.74
Top10perc       5 777    27.56   17.64   23.0    25.13   13.34    1.0    96.0    95.0  1.41     2.17
Top25perc       6 777    55.80   19.80   54.0    55.12   20.76    9.0   100.0    91.0  0.26    -0.57
F.Undergrad     7 777  3699.91 4850.42 1707.0  2574.88 1441.09  139.0 31643.0 31504.0  2.60     7.61
P.Undergrad     8 777   855.30 1522.43  353.0   536.36  449.23    1.0 21836.0 21835.0  5.67    54.52
Outstate        9 777 10440.67 4023.02 9990.0 10181.66 4121.63 2340.0 21700.0 19360.0  0.51    -0.43
Room.Board     10 777  4357.53 1096.70 4200.0  4301.70 1005.20 1780.0  8124.0  6344.0  0.48    -0.20
Books          11 777   549.38  165.11  500.0   535.22  148.26   96.0  2340.0  2244.0  3.47    28.06
Personal       12 777  1340.64  677.07 1200.0  1268.35  593.04  250.0  6800.0  6550.0  1.74     7.04
PhD            13 777    72.66   16.33   75.0    73.92   17.79    8.0   103.0    95.0 -0.77     0.54
Terminal       14 777    79.70   14.72   82.0    81.10   14.83   24.0   100.0    76.0 -0.81     0.22
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
perc.alumni    16 777    22.74   12.39   21.0    21.86   13.34    0.0    64.0    64.0  0.60    -0.11
Expend         17 777  9660.17 5221.77 8377.0  8823.70 2730.95 3186.0 56233.0 53047.0  3.45    18.59
Grad.Rate      18 777    65.46   17.18   65.0    65.60   17.79   10.0   118.0   108.0 -0.11    -0.22
```

- Checking for Correlation in the dataset:

```
> # Scatter Plot of Accept and Enroll
> plot(College$Accept, College$Enroll ,
+      main = 'Scatter Plot - Accepted VS Enrolled',
+      ylab = '# of Acepted',
+      xlab = '# of Enrolled')
```



- From the above graph we can see there is a strong positive correlation between Accepted Applications and Number of students Enrolled



- The above bar graph shows the number of colleges which are private and number of colleges which are not private.

2. Splitting dataset into Train and Test datasets:

```
> set.seed(123)
> trainIndex <- createDataPartition(College$Private, p=0.7, list = FALSE)
> train <- College[trainIndex,]
> test <- College[-trainIndex,]
```

- The dataset is split into train and test datasets using the createDataPartition() method in R.

3. Logistic Regression Model:

A logistic regression model using glm() function is used to fit the model. To compare, two models are being created. In contrast to the second model, which only contained the two independent variables outstate and s.f.ratio , the first model included every independent variable from the original dataset.

```
> RegModel1 <- glm(Private~.,data=train, family = binomial(link = logit))
> summary(RegModel1)

Call:
glm(formula = Private ~ ., family = binomial(link = logit), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.9395  -0.0155   0.0450   0.1528   2.7862

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.3837441  2.5287015  -0.152  0.87938
Apps        -0.0005226  0.0003089  -1.692  0.09070 .
Accept       0.0003383  0.0006281   0.539  0.59019
Enroll       0.0017904  0.0012618   1.419  0.15592
Top10perc    0.0127635  0.0370418   0.345  0.73042
Top25perc    0.0053642  0.0267867   0.200  0.84128
F.Undergrad -0.0007627  0.0002739  -2.785  0.00536 **
P.Undergrad  0.0002173  0.0002614   0.831  0.40573
Outstate     0.0007750  0.0001624   4.773 1.82e-06 ***
Room.Board  -0.0000698  0.0003345  -0.209  0.83472
Books        0.0024696  0.0018723   1.319  0.18716
Personal    -0.0003566  0.0003468  -1.028  0.30382
PhD         -0.0539319  0.0345821  -1.560  0.11887
Terminal    -0.0401208  0.0336091  -1.194  0.23258
S.F.Ratio   -0.0625279  0.0939568  -0.665  0.50573
perc.alumni  0.0354363  0.0281107   1.261  0.20745
Expend       0.0001801  0.0001780   1.012  0.31163
Grad.Rate    0.0271715  0.0177348   1.532  0.12550
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.40  on 544  degrees of freedom
Residual deviance: 145.98  on 527  degrees of freedom
AIC: 181.98

Number of Fisher Scoring iterations: 8
```

Logistic Regression Model 1

- The summary of model 1, which was derived from the independent variables in the initial dataset, is depicted in the above image. The p-values for each variable are highlighted in the summary. We can observe that the outstate has the lowest p-value. So, for the second model, we apply Outstate and s.f.ratio.

```
> RegModel2 <- glm(Private ~ Outstate + S.F.Ratio ,data=train, family = binomial(link = logit))
> summary(RegModel2)

Call:
glm(formula = Private ~ Outstate + S.F.Ratio, family = binomial(link = logit),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3602  -0.3883   0.2099   0.4777   2.4508

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.739e-01  8.886e-01  -0.983    0.325
Outstate     5.451e-04  6.229e-05   8.751  < 2e-16 ***
S.F.Ratio   -1.962e-01  4.018e-02  -4.883 1.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 639.40  on 544  degrees of freedom
Residual deviance: 370.49  on 542  degrees of freedom
AIC: 376.49

Number of Fisher Scoring iterations: 6
```

Logistic Regression Model 2

- As seen in the above figure, the second model's summary has a higher AIC value than the first model. This indicates that the model has eliminated the outliers..
- The Log Odd values of coefficients are as below.

```
> # Regression Coef (Log Odds)
> coef(RegModel2)
  (Intercept)      Outstate     S.F.Ratio
-0.8739307782  0.0005451172 -0.1961779751
```

- The Odds values of the coefficients are as below.

```
> # Regression Coef (Odds)
> exp(coef(RegModel2))
(Intercept)    Outstate   S.F.Ratio
  0.4173080   1.0005453   0.8218659
```

4. Confusion Matrix for Train Set:

```
> probabilities.train <- predict(RegModel2, newdata=train, type="response")
> predicted.classes.min <- as.factor(ifelse(probabilities.train>=0.5, "Yes", "No"))
> confusionMatrix(predicted.classes.min, train$Private, positive = "Yes")
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  110  35
       Yes  39 361
```

```
               Accuracy : 0.8642
                 95% CI : (0.8326, 0.8919)
    No Information Rate : 0.7266
    P-Value [Acc > NIR] : 8.76e-15

                  Kappa : 0.6554

 Mcnemar's Test P-Value : 0.7273

            Sensitivity : 0.9116
            Specificity : 0.7383
         Pos Pred Value : 0.9025
         Neg Pred Value : 0.7586
             Prevalence : 0.7266
         Detection Rate : 0.6624
   Detection Prevalence : 0.7339
      Balanced Accuracy : 0.8249

       'Positive' Class : Yes
```

- In the above figure, we can see the True Positive and True Negative values are 361 and 110. The False Positive and False Negative values are 35 and 39.
- The accuracy of the model is 86.42% which means this model is successful.
- False Negative is more damaging than False Positive for this analysis because from the outcome of this analysis, students may face problems in applying for colleges if the information is turned out to be false.

5. Report and interpret metrics for Accuracy, Precision, Recall, and Specificity:

- Precision (Pos Pred Value) equals 0.9025. This indicates that 90.25% of schools categorized as private are genuinely private institutions.
- Recall (Sensitivity): 0.9116, indicating that 91.16 % of all real private schools were accurately predicted as private schools.
- Specificity: 0.7383 is evaluated as 73.83 % of real public schools were properly predicted as public.
- Accuracy: The model's accuracy is 0.8642 which is 86.42% showing this is an effective model.

6. Confusion Matrix for Test Set;

```
> probabilities.test = predict(RegModel2, newdata = test, type="response")
> predicted.classes.min = as.factor(ifelse(probabilities.test>=0.5, "Yes", "No"))
> head(predicted.classes.min)
           Abilene Christian University                     Adelphi University
                                     No                                    Yes
                        Adrian College              Alaska Pacific University
                                    Yes                                    Yes
                     Albright College Allentown Coll. of St. Francis de Sales
                                    Yes                                    Yes
Levels: No Yes
```

- From the above result, we can see the True Positive value is 158 and True Negative value is 52 whereas the False Positive and False Negative vales are 11 and 11, which is a good indication of the model's effectiveness.

```
> confusionMatrix(predicted.classes.min, test$Private, positive = "Yes")
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No   42  16
       Yes  21 153

               Accuracy : 0.8405
                 95% CI : (0.7869, 0.8852)
    No Information Rate : 0.7284
    P-Value [Acc > NIR] : 3.808e-05

                  Kappa : 0.5866

 Mcnemar's Test P-Value : 0.5108

            Sensitivity : 0.9053
            Specificity : 0.6667
         Pos Pred Value : 0.8793
         Neg Pred Value : 0.7241
             Prevalence : 0.7284
         Detection Rate : 0.6595
   Detection Prevalence : 0.7500
      Balanced Accuracy : 0.7860

       'Positive' Class : Yes
```

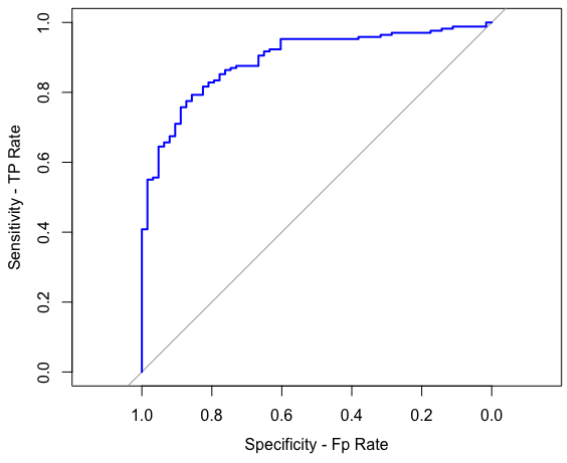
- Although if the metrics indicate that the model performed better on the train set than the test set, it is still a successful model since, when comparing two confusion matrices of the Train set and Test set, 84.05% percent correct predictions on a fresh dataset is a reasonable result.

7. ROC Curve:

```
# 7. ROC Curve

ROC1 = roc(test$Private, probabilities.test)

plot(ROC1, col="blue", ylab="Sensitivity - TP Rate", xlab= "Specificity - Fp Rate")
```



The balance between accuracy and precision is demonstrated by the receiver operating characteristic, or ROC. When the curve closely fits the left and top boundaries of the ROC space, the test is more accurate. The ROC curve is shown in the above picture to be toward the upper left corner of the space, indicating that this model is almost perfect.

8. **AUC:**

```
> AUC1 = auc(ROC1)
> AUC1
Area under the curve: 0.8936
```

How well a measure of separability may be detected is measured by the area under the curve (AUC) of a ROC curve. It shows how the model can distinguish between various types of data. The more AUC a model has, the more accurate it becomes. This model's AUC is 0.8936, which is a successful result.

Conclusion:

In order to identify whether a school is public or private, the above logistic regression model was developed. The regression model is simply confirmed by comparing the values in the matrix's four sections. True Positives and True Negatives should outweigh False Positives and False Negatives in a good model. Other metrics including precision, accuracy, sensitivity, and specificity can also be calculated using the matrix. Real-time performance evaluation of a model has also been done using AUC-ROC.

Reference:

Performing Logistic Regression Analysis Using R. (n.d.). https://sphweb.bumc.bu.edu/otlt/MPH-Modules/PH717-QuantCore/PH717-Module12-MultipleRegression/PH717-Module12-MultipleRegression8.html