INTRODUCTION

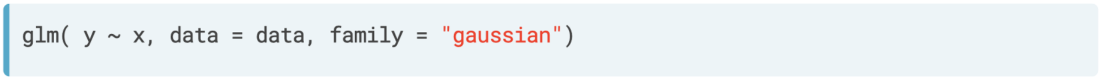
This report includes a diagnosis that shows how to use RStudio to fit, analyses, and assess a regression model. This assignment module's learning outcomes are as follows:

- Develop more powerful data interpretation models.
- Answer strategic and operational problems with sophisticated generalized linear approaches.
- Create a complex dataset for analysis.
- To enhance predicted outcomes, use multivariable and logistic regression methods.

What exactly is the glm function?

Generalized linear model (GLM) is a generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution like Gaussian distribution (Datacamp., 2020).

GLMs are fit with function glm (). Like linear models (lm () s), glm () s have formulas and data as inputs, but also have a family input.



```
glm( y ~ x, data = data, family = "gaussian")
```

Generalized linear models can have non-normal errors or distributions. However, there are limitations to the possible distributions. For example, you can use Poisson family for count data, or you can use binomial family for binomial data.

What is Logistic Regression?

Logistic regression is an instance of classification technique that you can use to predict a qualitative response. More specifically, logistic regression models the probability that gender belongs to a particular category.

That means that, if you are trying to do gender classification, where the response gender falls into one of the two categories, male or female, you'll use logistic regression models to estimate the probability that gender belongs to a particular category (Datacamp., 2018).

For example, the probability of gender given long hair can be written as:

Pr(gender = female |long hair)Pr(gender = female |long hair)

The values of Pr(gender=female| long hair)Pr(gender=female| long hair) (abbreviated as p(longhair)p(longhair)) will range between 0 and 1. Then, for any given value of longhair, a prediction can be made for gender.

Given XX as the explanatory variable and YY as the response variable, how should you then model the relationship between p(X)=Pr(Y=1|X)p(X)=Pr(Y=1|X) and XX? The linear regression model represents these probabilities as:

$p(X)=\beta 0 + \beta 1X$

What is ROC & Confusion Matrix?

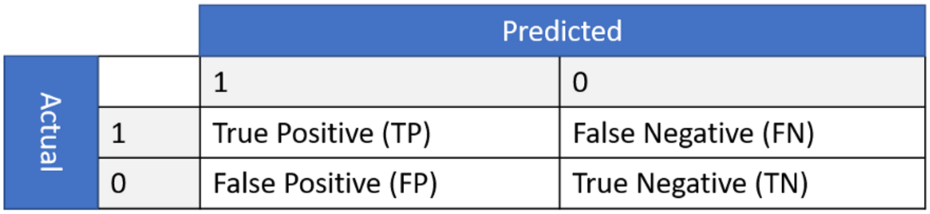
The world is facing a unique crisis these days and we all are stuck in a never seen before lockdown. As all of us are utilizing this time in many productive ways, I thought of creating some blogs of data concepts I know, not only to share it with the community but also to develop a deeper understanding of the concept as I write it down (Deshpande R., 2021).

The first one is here about the most loved evaluation metric — The ROC curve.

ROC (Receiver Operating Characteristic) Curve is a way to visualize the performance of a binary classifier.

Understanding the confusion matrix

In order to understand AUC/ROC curve, it is important to understand the confusion matrix first.



		Predicted	
		1	0
Actual	1	True Positive (TP)	False Negative (FN)
	0	False Positive (FP)	True Negative (TN)

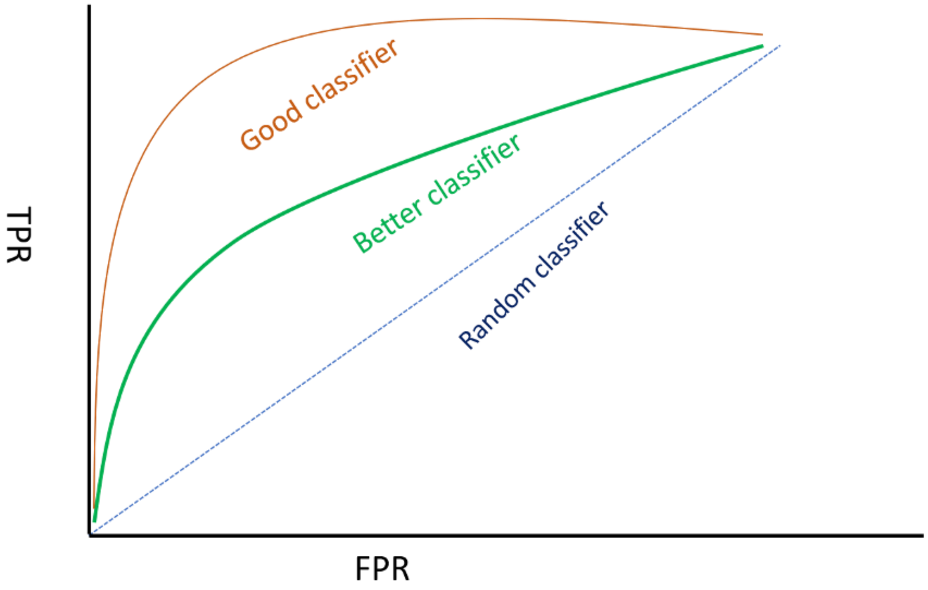
TPR = TP/(TP+FN)

FPR = FP/(TN+FP)

TPR or True Positive Rate answers the question — When the actual classification is positive, how often does the classifier predict positive?

FPR or False Positive Rate answers the question — When the actual classification is negative, how often does the classifier incorrectly predict positive?

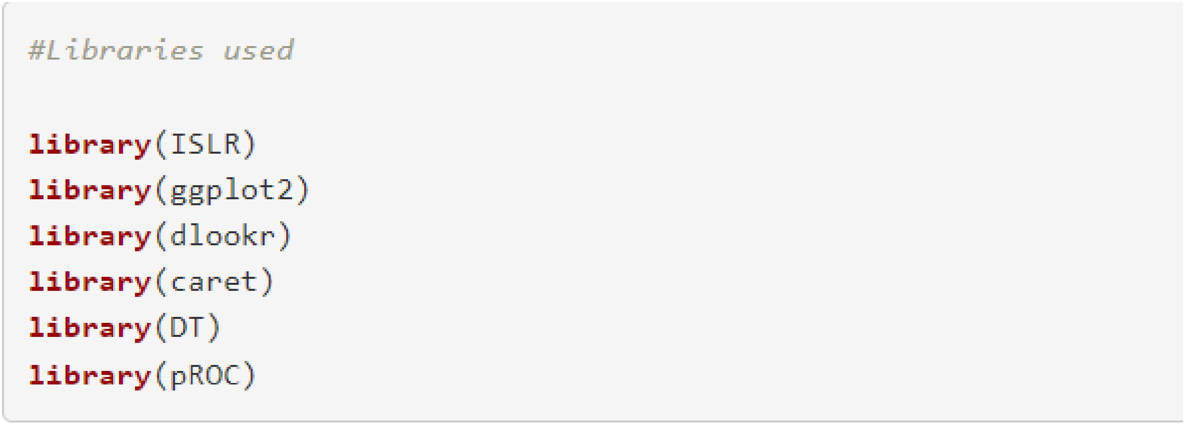
Now, if I plot this data on a graph, I will get a ROC curve.
The ROC curve is the graph plotted with TPR on y-axis and FPR on x-axis for all possible threshold. Both TPR and FPR vary from 0 to 1.



Therefore, a good classifier will have an arc/ curve and will be further away from the random classifier line.
To quantify a good classifier from a bad one using a ROC curve, is done by AUC (Area under Curve). From the graph it is quite clear that a good classifier will have AUC higher than a bad classifier as the area under curve will be higher for the former.

ANALYSIS SECTION

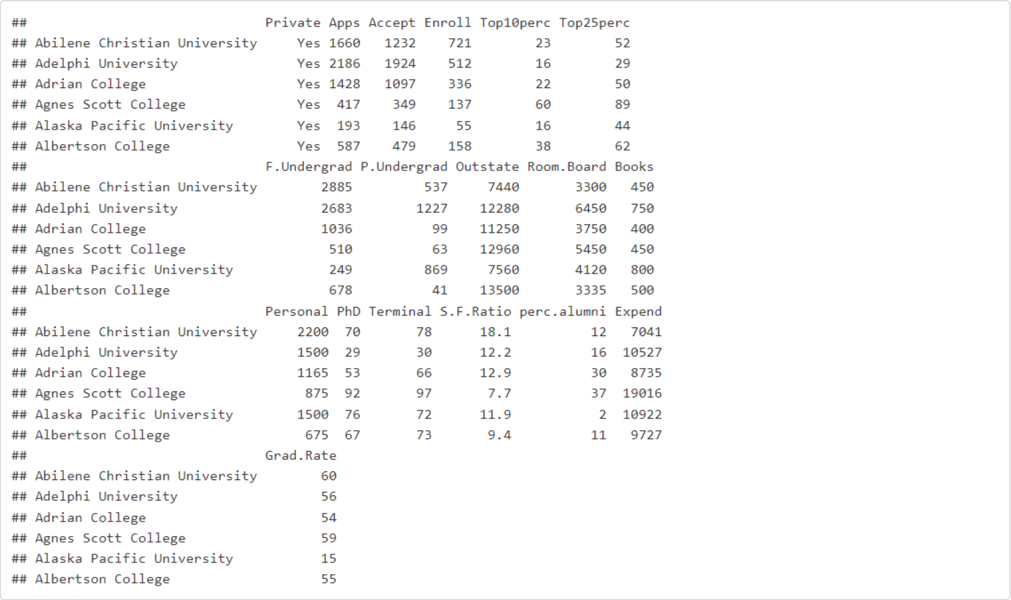
Libraries Used



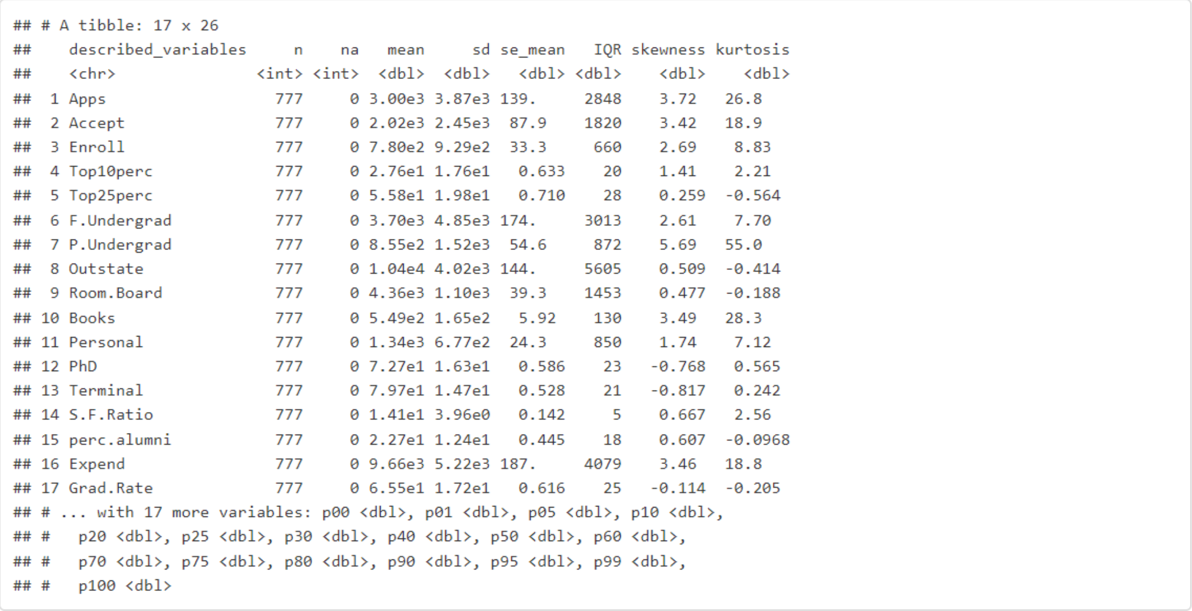
```
#Libraries used

library(ISLR)
library(ggplot2)
library(dlookr)
library(caret)
library(DT)
library(pROC)
```

First task

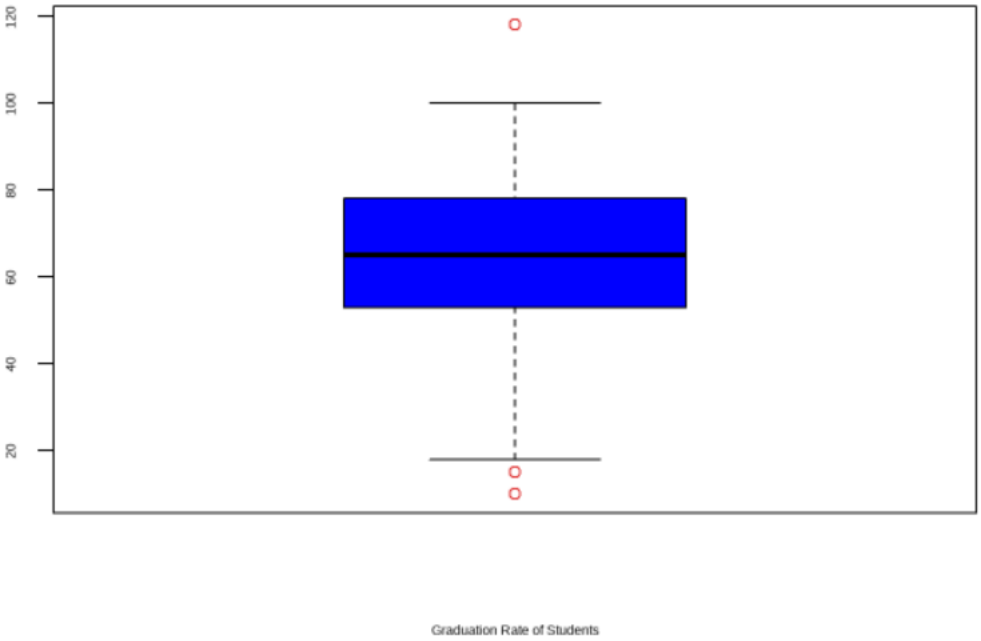
Import the dataset and conduct exploratory data analysis, describing the dataset using descriptive statistics and graphs.

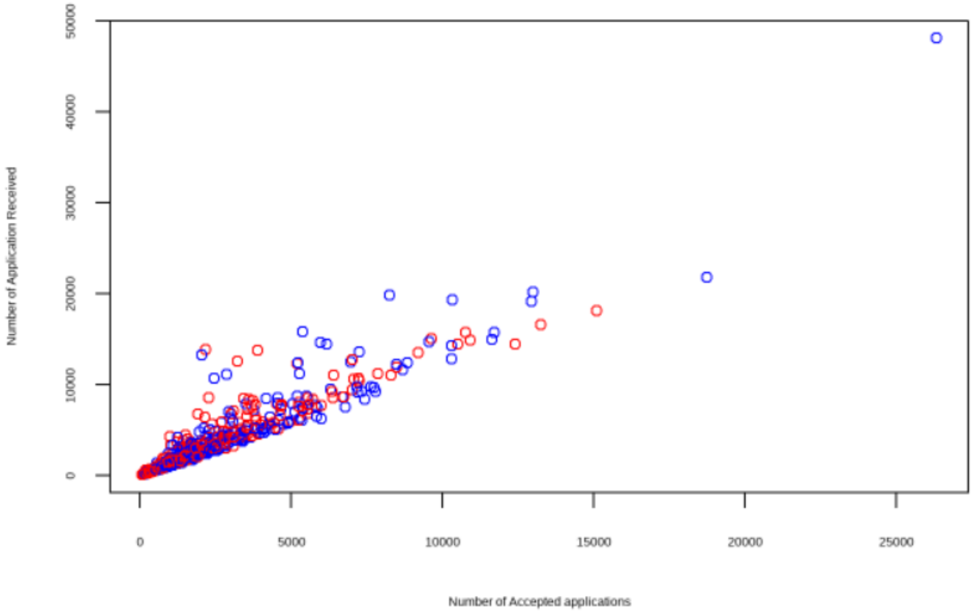


```
##                              Private Apps Accept Enroll Top10perc Top25perc
## Abilene Christian University     Yes 1660   1232    721        23        52
## Adelphi University               Yes 2186   1924    512        16        29
## Adrian College                   Yes 1428   1097    336        22        50
## Agnes Scott College              Yes  417    349    137        60        89
## Alaska Pacific University        Yes  193    146     55        16        44
## Albertson College                Yes  587    479    158        38        62
##                              F.Undergrad P.Undergrad Outstate Room.Board Books
## Abilene Christian University        2885         537     7440       3300   450
## Adelphi University                  2683        1227    12280       6450   750
## Adrian College                      1036          99    11250       3750   400
## Agnes Scott College                  510          63    12960       5450   450
## Alaska Pacific University            249         869     7560       4120   800
## Albertson College                    678          41    13500       3335   500
##                              Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1          12   7041
## Adelphi University               1500  29       30      12.2          16  10527
## Adrian College                   1165  53       66      12.9          30   8735
## Agnes Scott College               875  92       97       7.7          37  19016
## Alaska Pacific University        1500  76       72      11.9           2  10922
## Albertson College                 675  67       73       9.4          11   9727
##                              Grad.Rate
## Abilene Christian University        60
## Adelphi University                  56
## Adrian College                      54
## Agnes Scott College                 59
## Alaska Pacific University           15
## Albertson College                   55
```



```
## # A tibble: 17 x 26
##    described_variables     n    na   mean     sd se_mean   IQR skewness kurtosis
##    <chr>               <int> <int>  <dbl>  <dbl>   <dbl> <dbl>    <dbl>    <dbl>
##  1 Apps                  777     0 3.00e3 3.87e3 139.     2848    3.72   26.8   
##  2 Accept                777     0 2.02e3 2.45e3  87.9    1820    3.42   18.9   
##  3 Enroll                777     0 7.80e2 9.29e2  33.3     660    2.69    8.83  
##  4 Top10perc             777     0 2.76e1 1.76e1   0.633    20    1.41    2.21  
##  5 Top25perc             777     0 5.58e1 1.98e1   0.710    28    0.259  -0.564 
##  6 F.Undergrad           777     0 3.70e3 4.85e3 174.     3013    2.61    7.70  
##  7 P.Undergrad           777     0 8.55e2 1.52e3  54.6     872    5.69   55.0   
##  8 Outstate              777     0 1.04e4 4.02e3 144.     5605    0.509  -0.414 
##  9 Room.Board            777     0 4.36e3 1.10e3  39.3    1453    0.477  -0.188 
## 10 Books                 777     0 5.49e2 1.65e2   5.92    130    3.49   28.3   
## 11 Personal              777     0 1.34e3 6.77e2  24.3     850    1.74    7.12  
## 12 PhD                   777     0 7.27e1 1.63e1   0.586    23   -0.768   0.565 
## 13 Terminal              777     0 7.97e1 1.47e1   0.528    21   -0.817   0.242 
## 14 S.F.Ratio             777     0 1.41e1 3.96e0   0.142     5    0.667   2.56  
## 15 perc.alumni           777     0 2.27e1 1.24e1   0.445    18    0.607  -0.0968
## 16 Expend                777     0 9.66e3 5.22e3 187.     4079    3.46   18.8   
## 17 Grad.Rate             777     0 6.55e1 1.72e1   0.616    25   -0.114  -0.205 
## # ... with 17 more variables: p00 <dbl>, p01 <dbl>, p05 <dbl>, p10 <dbl>,
## #   p20 <dbl>, p25 <dbl>, p30 <dbl>, p40 <dbl>, p50 <dbl>, p60 <dbl>,
## #   p70 <dbl>, p75 <dbl>, p80 <dbl>, p90 <dbl>, p95 <dbl>, p99 <dbl>,
## #   p100 <dbl>
```

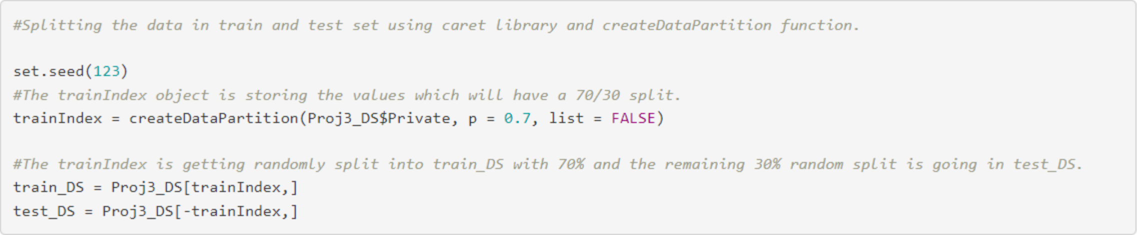




The head and dlookr::describe functions were used to do exploratory and descriptive analysis on the imported College data, respectively. Then, for the graduation rate of the students, a boxplot was constructed, which shows that there aren't many outliers, indicating that our data set is correct. The number of approved candidates was then plotted against the number of applications in a scatterplot.

Second task

Divide the data into two sets: one for training and one for testing.

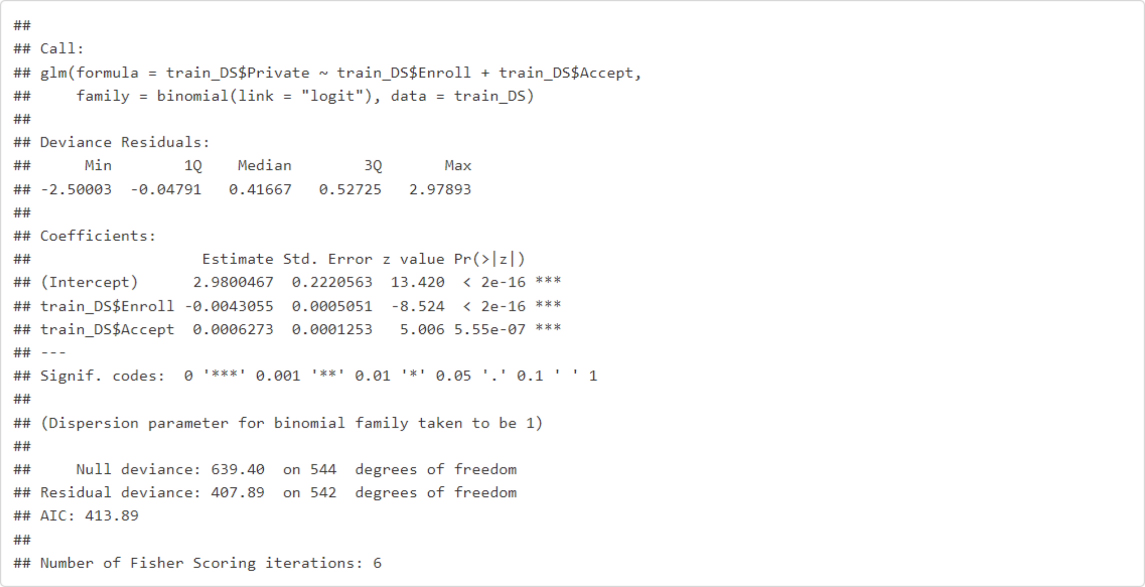


```
#Splitting the data in train and test set using caret library and createDataPartition function.

set.seed(123)
#The trainIndex object is storing the values which will have a 70/30 split.
trainIndex = createDataPartition(Proj3_DS$Private, p = 0.7, list = FALSE)

#The trainIndex is getting randomly split into train_DS with 70% and the remaining 30% random split is going in test_DS.
train_DS = Proj3_DS[trainIndex,]
test_DS = Proj3_DS[-trainIndex,]
```

The dataset was split into train and test using the caret library and createDataPartition function. The trainIndex is used to store the values where the data was splitted randomly in a 70/30 split.

Third task

Fit a logistic regression model to the training set with at least two predictors using the glm() method.

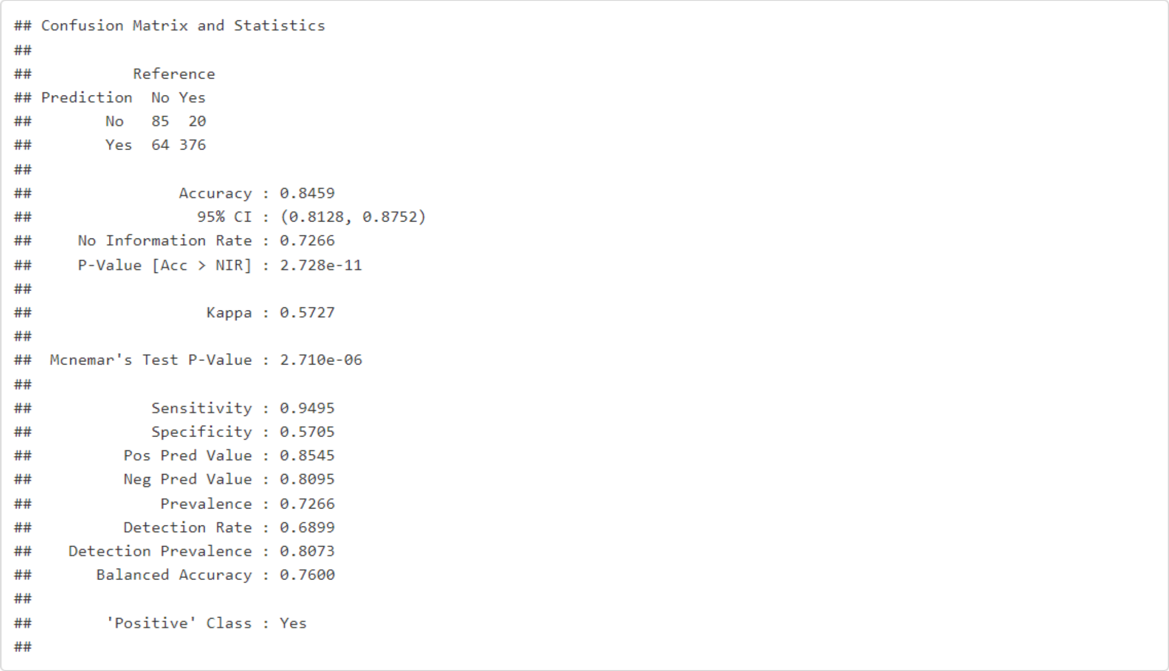


```
##
## Call:
## glm(formula = train_DS$Private ~ train_DS$Enroll + train_DS$Accept,
##     family = binomial(link = "logit"), data = train_DS)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.50003  -0.04791   0.41667   0.52725   2.97893
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.9800467  0.2220563  13.420  < 2e-16 ***
## train_DS$Enroll -0.0043055  0.0005051  -8.524  < 2e-16 ***
## train_DS$Accept  0.0006273  0.0001253   5.006 5.55e-07 ***
## ---
## 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: 407.89  on 542  degrees of freedom
## AIC: 413.89
##
## Number of Fisher Scoring iterations: 6
```

Using the glm () function to fit a logistic regression model, the training set was created using two predictors for train data set. Then, to check for better model we look at the AIC value where the lowest AIC values means it’s a preferred model.

Fourth task

Make a confusion matrix and submit your model's results for the train set. The confusion matrix should be interpreted and discussed. False Positives or False Negatives: which is more damaging to the analysis?

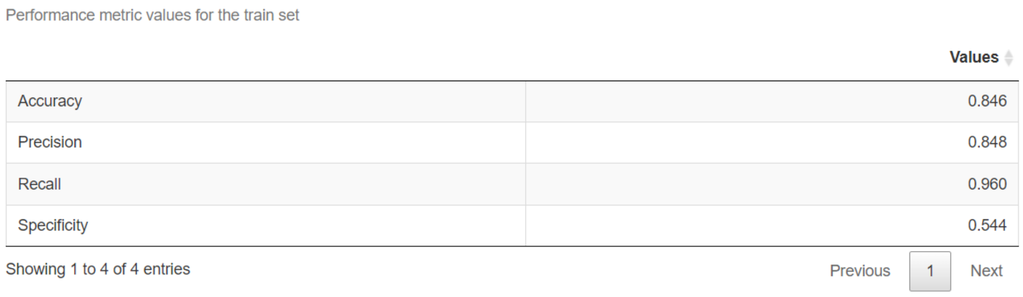


```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No   85  20
##        Yes  64 376
##
##                Accuracy : 0.8459
##                  95% CI : (0.8128, 0.8752)
##     No Information Rate : 0.7266
##     P-Value [Acc > NIR] : 2.728e-11
##
##                   Kappa : 0.5727
##
##  Mcnemar's Test P-Value : 2.710e-06
##
##             Sensitivity : 0.9495
##             Specificity : 0.5705
##          Pos Pred Value : 0.8545
##          Neg Pred Value : 0.8095
##              Prevalence : 0.7266
##          Detection Rate : 0.6899
##    Detection Prevalence : 0.8073
##       Balanced Accuracy : 0.7600
##
##        'Positive' Class : Yes
##
```

Then, using the train data set, predictions are created by comparing predicted values to actual values using a confusion matrix. The number of false positives in this case is 64. When we anticipate yes but the actual response is no, we call it a false positive. The number of false negatives in this case is 20. When we forecast no, but the actual answer is yes, we have a false negative. The true positive number is 376, whereas the true negative number is 85. This model has a precision of 0.8459. To improve our model, we can reduce the sensitivity or modify the probability value from 0.5 to something lower. Our prevalence value is 0.7266, which implies that if we take 100 values, our model will predict 72 of them to be yes and the remaining 28 to be no. False Negative or False Positive both are damaging but that depends on our model and question that is being answered. In some case, False Negative can be more damaging than False Positive or vice versa.

Fifth task

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



Performance metric values for the train set

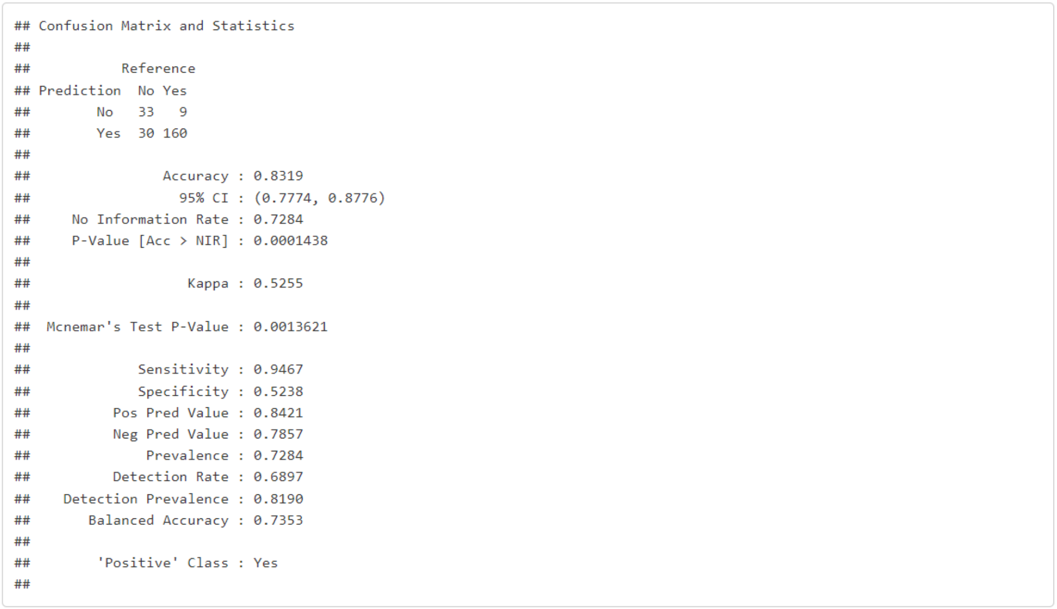
	Values
Accuracy	0.846
Precision	0.848
Recall	0.960
Specificity	0.544

Showing 1 to 4 of 4 entries

Based on the train predicted values acquired in the previous task using confusion Matrix, the Accuracy, Precision, Recall, and Specificity values were computed and displayed by using the DT library table.

Sixth task

Make a confusion matrix and report your model's results for the test set.

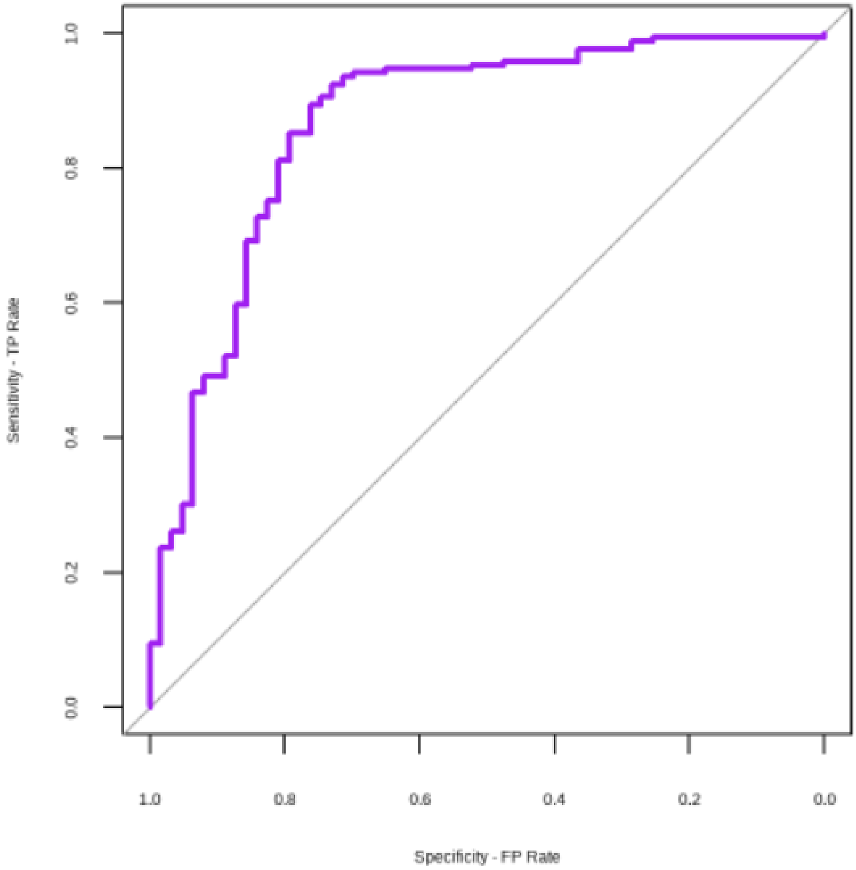


```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No   33   9
##        Yes  30 160
##
##                Accuracy : 0.8319
##                  95% CI : (0.7774, 0.8776)
##     No Information Rate : 0.7284
##     P-Value [Acc > NIR] : 0.0001438
##
##                   Kappa : 0.5255
##
##  Mcnemar's Test P-Value : 0.0013621
##
##             Sensitivity : 0.9467
##             Specificity : 0.5238
##          Pos Pred Value : 0.8421
##          Neg Pred Value : 0.7857
##              Prevalence : 0.7284
##          Detection Rate : 0.6897
##    Detection Prevalence : 0.8190
##       Balanced Accuracy : 0.7353
##
##        'Positive' Class : Yes
##
```

The training set was built using two predictors for test data and the glm () method to design a logistic regression model. The number of false positives in this case is 30. When we anticipate yes but the actual response is no, we call it a false positive. The number of false negatives in this case is 9. When we forecast no, but the actual answer is yes, we have a false negative. The number of true positives is 160, whereas the number of true negatives is 33. This model has a precision of 0.8534. To improve our model, we can reduce the sensitivity or modify the probability value from 0.5 to something lower. Our prevalence value is 0.7284, which suggests that out of 100 possible values, our model will forecast 73 as yes and the remaining 27 as no.

Seventh task

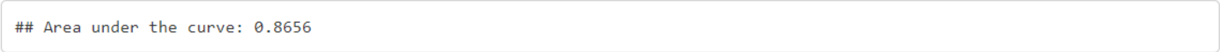
Plot and interpret the ROC curve. Please discuss what will happen to Precision and Recall if raising the classification threshold.



The ROC graph was plotted using the roc function from the pROC library for our model. Then, the par function was used to remove the padding from our ROC curve. We use pty which is the plot type to s which is the square to remove the padding. Ideally, we want the purple line to go straight to the top and then right horizontally. That is a model that is capable of perfect predictions but it never happens. But what we do not want is this purple line hugging the gray diagonal line. Because that will be the model that will make no correct predictions The diagonal line shows where the TP Rate is the same as the FP Rate. So, the ROC curve is a pretty good way to understand how your model is performing over specified thresholds. Consider a classification model that separates email into two categories: "spam" or "not spam." If you raise the classification threshold, what will happen to precision? In general, raising the classification threshold reduces false positives, thus raising precision. Now, if you raise the classification threshold, what will happen to recall? Raising our classification threshold will cause the number of true positives to decrease or stay the same and will cause the number of false negatives to increase or stay the same. Thus, recall will either stay constant or decrease.

Eighth task

Calculate the AUC and analyze the results.



```
## Area under the curve: 0.8656
```

AUC, or Area under the curve, is another approach we may employ in addition to ROC. As a result, AUC stands for the area under the ROC curve. We can see that the result is 0.9082, indicating that the area under the purple line and between the gray horizontal line is 0.9082.

CONCLUSION

Finally, I learned about several libraries such as caret, ISLR, and pROC. Worked with college data and performed complex analysis on it, such as extracting the glm value and then randomly splitting the data set into two parts, the train and test dataset, and working on them by predicting their values using the confusion matrix and interpreting their Accuracy, Precision, Recall, and Specificity values for both the train and test data sets. Finally, I learnt what ROC and AUC are and how to use the ROC and AUC functions from the pROC package to build them in R.

BIBLIOGRAPHY

1) GLM in R: Generalized Linear Model Tutorial. (2020). DataCamp Community. https://www.datacamp.com/community/tutorials/generalized-linear-models

2) Logistic Regression in R Tutorial. (2018). DataCamp Community. https://www.datacamp.com/community/tutorials/logistic-regression-R#logistic-regression

3) Deshpande, R. (2021, December 14). ROC Curve and AUC in Machine learning and R pROC Package. Medium. https://medium.com/swlh/roc-curve-and-auc-detailed-understanding-and-r-proc-package-86d1430a3191