Module 3 Assignment — GLM and Logistic Regression

Sourabh D. Khot (ID 002754952)

College of Professional Studies, Northeastern University

ALY 6015: Intermediate Analytics CRN 81176

Professor Behzad Ahmadi

May 2, 2022

Table of Contents

Introduction

This assignment aims to understand the concepts of logistic regression and implement the learnings hands-on. The dataset used in this assignment is college.csv from the ISLR package. The dataset contains details of colleges and classifies whether the college is private or non-private (public).

The question in the dataset is whether the type of college (private/non-private) can be predicted by building a model using other details and the accuracy of the model. This is a binary classification problem with supervised learning.

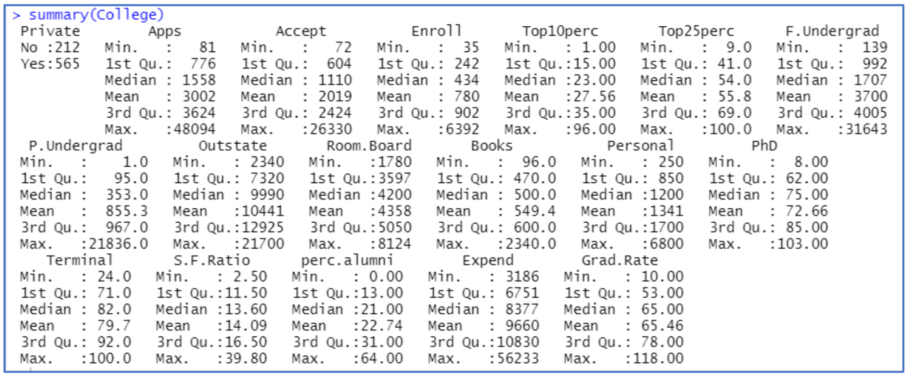
I will implement logistic regression to address the problem in the R language. The data will be explored using Exploratory Data Analysis techniques. Further, I will split the data into train and test sets, and basis the train set, I will fit the best model with an appropriate number of features. Lastly, I will perform diagnostics on the final model and report the model metrics.

Analysis

Exploratory Data Analysis

1. Import the dataset and perform Exploratory Data Analysis by using descriptive statistics and plots to describe the dataset.

I imported the ISLR library and attached the College dataset so I can easily access the dataset variables by their names. There are 777 records and 18 variables, of which one variable (‘Private’) is categorical, and the remaining 17 are numerical. The dataset contains various details of the 777 colleges related to the college, students, faculty, and alumni. ‘Private’ variable is the outcome and has yes/no values indicating whether it is a private college or not, while others are the input variables. A description of all the variables is given in the appendix.



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

Figure 1. Descriptive Statistics

It can be observed in Figure 1 that the outcome variable has a class bias. I will address this class bias by creating a training dataset having equal proportions of Private No’s and Yes’s.

In figure 2, I have selected four variables as per my understanding and checked their boxplot distribution against the target variable. Similarly, I have performed a linear regression of the number of enrollments against out-of-state tuition fees with a different legend for private and non-private universities.

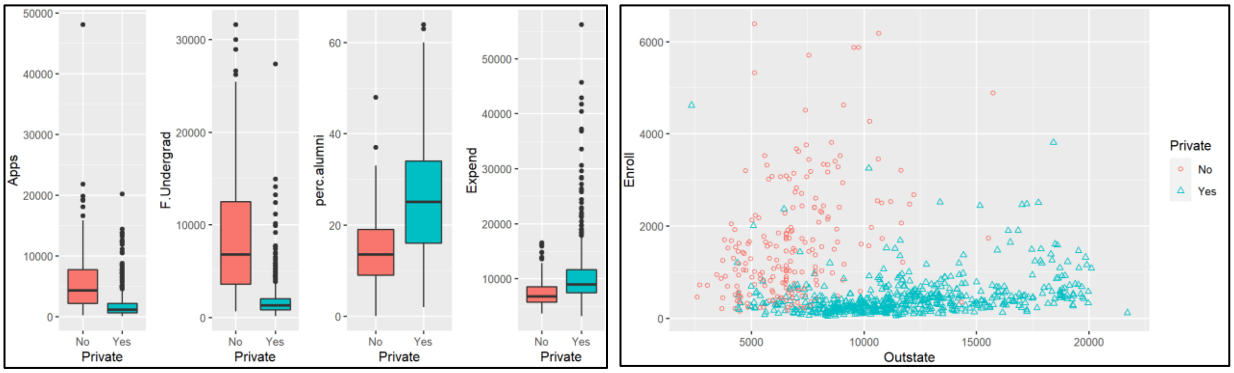


Figure 2. Exploratory Data Analysis with ‘Private’ as the outcome variable

From the boxplots, it can be inferred that ‘number of applications’ and ‘number of full-time undergraduate students’ have a much different distribution in private and non-private universities. From the scatterplot, the data points of private and non-private universities are well-separated in terms of the out-of-state tuition fees, while there is also some separation in terms of enrolment.

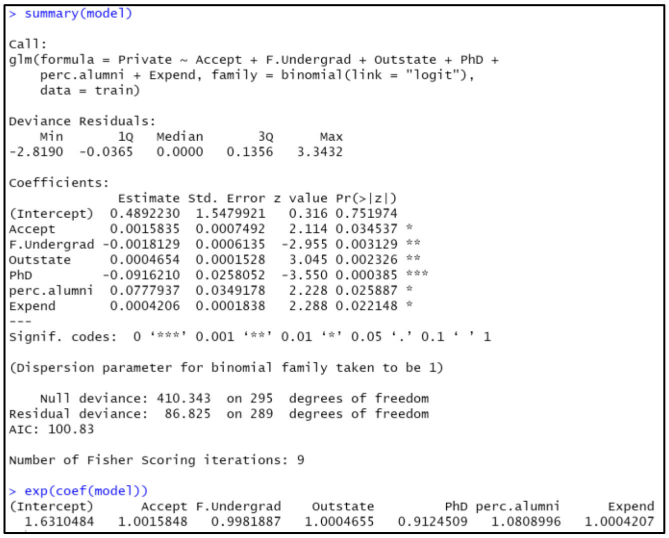
Logistic Regression

2. Split the data into a train and test set.

Because of the target variable’s class bias, I have split the data with equal proportions of private and non-private universities in the train set and kept the remaining in the test set. More specifically, I have taken 70% non-private and an equal number of private universities into the train set. This is implemented using the sample() function of R (full code in the appendix).

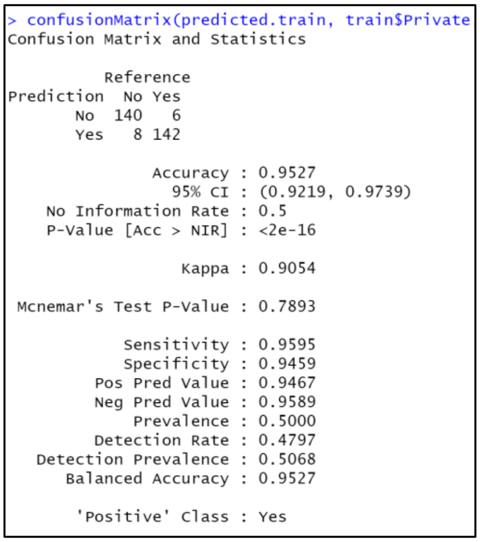
3. Use the glm() function in the ‘stats’ package to fit a logistic regression model to the training set using at least two predictors.

‘Private’ is the output variable, and the potential 17 input variables are all numerical. For feature selection among the input variables, I used the p-value from the Logit model. First, I fitted a logistic regression model with all input variables and iteratively removed the feature with the highest p-value to fit another model until I fitted a model with all features having p-value less than 0.05. The features finally selected are ‘number of applications accepted’, number of full-time undergrads, out-of-state tuition, percentage of faculty with PhDs, and expenditure per student. A summary of this final model is given in Figure 3.



```
> summary(model)

Call:
glm(formula = Private ~ Accept + F.Undergrad + Outstate + PhD +
    perc.alumni + Expend, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8190  -0.0365   0.0000   0.1356   3.3432

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.4892230  1.5479921   0.316 0.751974
Accept       0.0015835  0.0007492   2.114 0.034537 *
F.Undergrad -0.0018129  0.0006135  -2.955 0.003129 **
Outstate     0.0004654  0.0001528   3.045 0.002326 **
PhD         -0.0916210  0.0258052  -3.550 0.000385 ***
perc.alumni  0.0777937  0.0349178   2.228 0.025887 *
Expend       0.0004206  0.0001838   2.288 0.022148 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 410.343  on 295  degrees of freedom
Residual deviance:  86.825  on 289  degrees of freedom
AIC: 100.83

Number of Fisher Scoring iterations: 9

> exp(coef(model))
(Intercept)      Accept F.Undergrad    Outstate         PhD perc.alumni      Expend
  1.6310484   1.0015848   0.9981887   1.0004655   0.9124509   1.0808996   1.0004207
```

Figure 3. Summary of the Final Model



```
> confusionMatrix(predicted.train, train$Private
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  140   6
       Yes   8 142

               Accuracy : 0.9527
                 95% CI : (0.9219, 0.9739)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9054

 Mcnemar's Test P-Value : 0.7893

            Sensitivity : 0.9595
            Specificity : 0.9459
         Pos Pred Value : 0.9467
         Neg Pred Value : 0.9589
             Prevalence : 0.5000
         Detection Rate : 0.4797
   Detection Prevalence : 0.5068
      Balanced Accuracy : 0.9527

       'Positive' Class : Yes
```

Figure 4. Confusion Matrix for train

The exponentiated coefficients at the end of Figure 3 can be interpreted as e.g., the odds of college being ‘Private’ increases by a factor of 1.00158 with a unit increase in the ‘number of applications accepted’. Null and residual deviances have a desirably significant difference.

Model Diagnostics

4. **Create a confusion matrix and report the results for the train set. Which misclassifications are more damaging, False Positives or False Negatives?**

Confusion Matrix for the train set is plotted in Figure 4. Out of 148 actual/reference non-private colleges, the model accurately predicts 140 of them (94.59%), while out of 148 actual private colleges, the model predicts 142 of them (95.94%). I think the misclassification of a private college as a public college (False Negatives) is more damaging for this analysis as a private college has high out-of-state fees. This mistake can be costly compared to classifying a public college as private (False Positive). Hence, the precision of the model should be high.

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

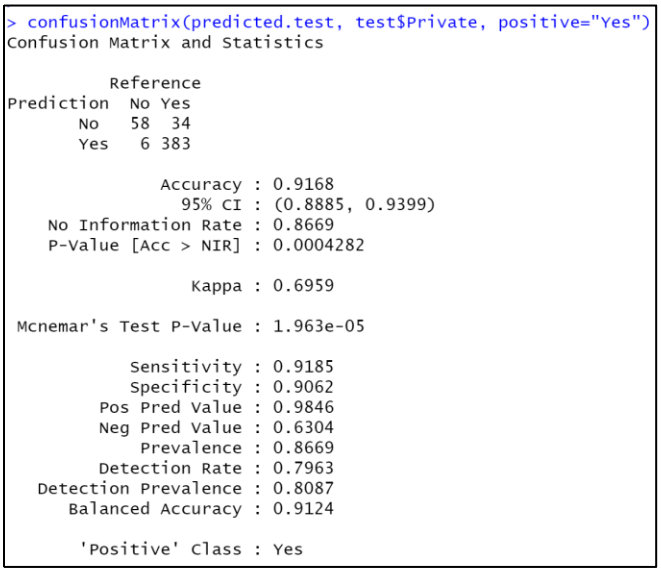
Accuracy [(TN + TP) / All] is 0.9527 for this model on the train set, which is very high, especially when the outcome variable class is balanced. It signifies among all cases the percentage of cases accurately classified as private and non-private.

Precision / Pos Pred Value [TP / (FP+TP)] is 0.9467, which is good and can be interpreted as a share of correct predictions when predicated private (low False Negatives).

Recall / Sensitivity [TP / (TP+FN)] is 0.9595, a high number to correctly identify private among all private universities.

Specificity [TN / (TN + FP)] is 0.9459 is good for identifying non-private among all non-private universities.

6. Create a confusion matrix and report the results of model for the test set.



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

          Reference
Prediction  No Yes
       No   58  34
       Yes   6 383

               Accuracy : 0.9168
                 95% CI : (0.8885, 0.9399)
    No Information Rate : 0.8669
    P-Value [Acc > NIR] : 0.0004282

                  Kappa : 0.6959

 Mcnemar's Test P-Value : 1.963e-05

            Sensitivity : 0.9185
            Specificity : 0.9062
         Pos Pred Value : 0.9846
         Neg Pred Value : 0.6304
             Prevalence : 0.8669
         Detection Rate : 0.7963
   Detection Prevalence : 0.8087
      Balanced Accuracy : 0.9124

       'Positive' Class : Yes
```

Figure 4. Confusion Matrix for test

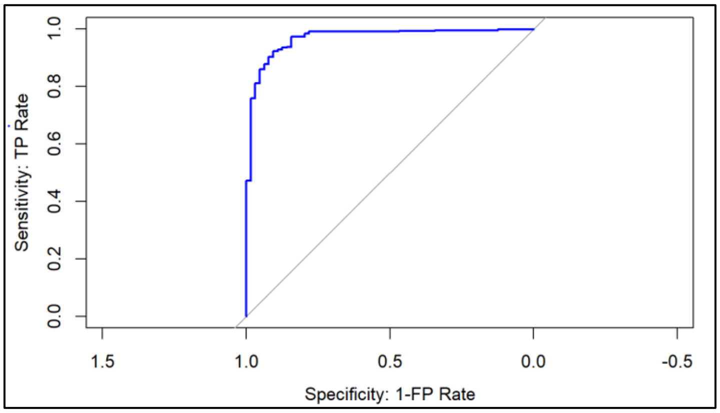
I have summarized the confusion matrix metric results with train and test data for easy comparison into Figure 5.

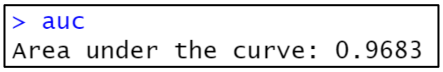
Metric	Description	Train Set	Test Set
Accuracy	Correct identification among all	0.9527	0.9168
Precision	Correct when identified as private	0.9467	0.9846
Recall	Correctly identified private among all private	0.9595	0.9185
Specificity	Correctly identified non-private among all non-private	0.9459	0.9062

Figure 5. Model Metrics for Train and Test Set

On applying the model to the test set, the precision increases to 0.9846 (lower False Negatives), which is excellent since False Negatives are more damaging. Other metrics are still above 0.9. Thus, the model is good at identifying private universities with low False Negatives.

7. Plot and interpret the ROC curve.





```
> auc
Area under the curve: 0.9683
```

Figure 6. Receiver Operator Characteristics (ROC) Curve and Area Under the Curve

It can be observed that the ROC curve is pretty close to the 'Specificity = 0' line near the vertex and then suddenly bends and is much flat towards 'Sensitivity = 0'. This indicates that the model is excellent in predicting whether a college is private or public.

8. Calculate and interpret the AUC.

As indicated in Figure 6, the area under the curve is 0.9683, which is a high number closer to 1, showing the model has high accuracy.

Conclusion

I have performed Exploratory Data Analysis. It can be observed that ‘number of applications’, ‘number of full-time undergraduate students’ and ‘out-of-state tuition fees’ have the highest impact on the outcome variable of whether the college is private.

As the predictor class was biased, I created a train set with equal proportions of private and non-private colleges. Using the p-value of the Logit model, I selected five features (‘number of applications accepted’, ‘number of full-time undergrads’, ‘out-of-state tuition’, ‘percentage of faculty with PhDs’, and ‘expenditure per student’) and fitted a final model with an AIC of 100.83.

The model can accurately predict whether a college is private or not using the above five features and has all metrics of accuracy, precision, recall, and specificity above 90% for both train and test sets. Classifying whether private university as public (False Negative) is more damaging, and is taken care of by the model as seen with high precision. The ROC curve is far from the diagonal, with a high area under the curve of 0.9683.

To answer the original question, I built a model to accurately predict the type of college (private/non-private) using five features with high accuracy and low damaging misclassification.

References

APA Style Table: APA.org. (n.d.). Retrieved from https://apastyle.apa.org/style-grammar-guidelines/tables-figures/tables

Bluman, A. G. (2018). *Elementary Statistics.* New York: McGraw Hill Education.

Canvas Module 3 Assignment. (n.d.). Retrieved from https://northeastern.instructure.com/courses/110053/assignments/1345430

Clear objects: StackOverFlow.com. (2013). Retrieved from https://stackoverflow.com/questions/11761992/how-do-i-clear-only-a-few-specific-objects-from-the-workspace

INTERPRET ODDS RATIOS: UCLA.edu. (n.d.). Retrieved from https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

Kelly, R. (2014, June 23). *Classification: R Studio*. Retrieved from https://rstudio-pubs-static.s3.amazonaws.com/21379_a214540189fb45f1915ba171c286a9d1.html#example-1-college-admissions

Prabhakaran, S. (n.d.). *Logistic Regression: r-statistics.co*. Retrieved from http://r-statistics.co/Logistic-Regression-With-R.html

ROC Curve: DisplayR.com. (n.d.). Retrieved from https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/

Understanding the components of a GLM: Statistical Odds & Ends Blog. (2019). Retrieved from https://statisticaloddsandends.wordpress.com/2019/10/31/understanding-the-components-of-a-generalized-linear-model-glm/

Appendix

Description of Variables

- Private: A factor with levels No and Yes indicating private or public university
- Apps: Number of applications received
- Accept: Number of applications accepted
- Enroll: Number of new students enrolled
- Top10perc: Pct. new students from top 10% of H.S. class
- Top25perc: Pct. new students from top 25% of H.S. class
- F.Undergrad: Number of fulltime undergraduates
- P.Undergrad: Number of parttime undergraduates
- Outstate: Out-of-state tuition
- Room.Board: Room and board costs
- Books: Estimated book costs
- Personal: Estimated personal spending
- PhD: Pct. of faculty with Ph.D.'s
- Terminal: Pct. of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Pct. alumni who donate
- Expend : Instructional expenditure per student
- Grad.Rate : Graduation rate

R Code

```
# Importing libraries
library(ISLR)       # for datasets
library(caret)      # for helper fns: partition dataset, confusion matrix, wrapper around fitted
models
library(ggplot2)    # graphing library
library(gridExtra)  # grids for plots
library(pROC)       # for ROC and AUC

# 1. EDA ####

attach(College)
str(College)
head(College)
summary(College)

table(Private)  # checking class bias

box1 <- qplot(x=Private, y=Apps, fill=Private, geom='boxplot') + guides(fill="none")
box2 <- qplot(x=Private, y=F.Undergrad, fill=Private, geom='boxplot') + guides(fill="none")
box3 <- qplot(x=Private, y=perc.alumni, fill=Private, geom='boxplot') + guides(fill="none")
box4 <- qplot(x=Private, y=Expend, fill=Private, geom='boxplot') + guides(fill="none")
grid.arrange(box1, box2, box3, box4, nrow=1)
rm(box1, box2, box3, box4)

qplot(x=Outstate, y=Enroll, color=Private, shape=Private, geom='point') +
scale_shape(solid=FALSE)

# 2. Splitting Dataset ####
## Splitting data without address class bias (code is commented since using different method)
# train_index <- createDataPartition(Private, p=0.70, list = FALSE)
# train <- College[trainIndex,]
# test <- College[-trainIndex,]
# rm(train_index)

# Splitting data keeping equal class in training
data_nos <- College[which(Private=="No"), ]
data_yess <- College[which(Private=="Yes"), ]

set.seed(952)
train_nos_index <- sample(1:nrow(data_nos), 0.7*nrow(data_nos))
train_yess_index <- sample(1:nrow(data_yess), 0.7*nrow(data_nos))
train_nos <- data_nos[train_nos_index, ]
train_yess <- data_yess[train_yess_index, ]
train <- rbind(train_nos, train_yess)
```

```
test_nos <- data_nos[-train_nos_index, ]
test_yess <- data_yess[-train_yess_index, ]
test <- rbind(test_nos, test_yess)
rm(data_nos, data_yess, train_nos_index, train_yess_index, train_nos, train_yess, test_nos, test_yess)

# 3. Logistic Regression ####

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

# Final model
model = glm(Private~Accept+F.Undergrad+Outstate+PhD+perc.alumni+Expend, data = train, family = binomial(link="logit"))
summary(model)
exp(coef(model))

# 4. Confusion Matrix for Train ####

prob.train <- predict(model, newdata=train, type="response")
predicted.train <- as.factor( ifelse(prob.train>=0.5, "Yes", "No") )
confusionMatrix(predicted.train, train$Private, positive="Yes")

# 5. Metrics of Model ####

# Accuracy [(TN + TP) / All]
# Precision [TP / (FP+TP)]
# Recall / Sensitivity [TP / (TP+FN)]
# Specificity [TN / (TN + FP)]

# 6. Confusion Matrix for Test ####

prob.test <- predict(model, newdata=test, type="response")
predicted.test <- as.factor( ifelse(prob.test>=0.5, "Yes", "No") )
confusionMatrix(predicted.test, test$Private, positive="Yes")

# 7. ROC Curve ####

ROC <- roc(test$Private, prob.test)
plot(ROC, col="blue", ylab = "Sensitivity: TP Rate", xlab = "Specificity: 1-FP Rate")

# 8. AUC ####

auc <- auc(ROC)
auc
```