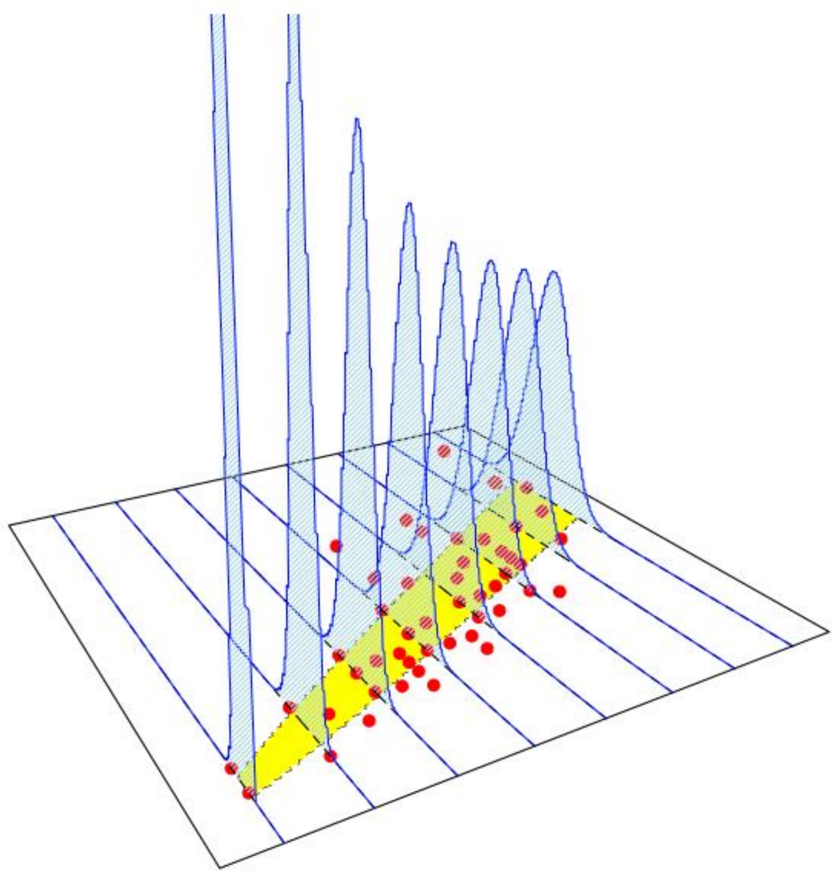
Sean McLean

ALY 6015

Module 3 Assignment

GLM and Logistic Regression



Introduction

The focus of the assignment is to analyze a dataset on colleges across the country and to predict whether a school is private or not private. This question is explored and answered by using logistic regression methods and advanced modeling techniques with a college dataset. This problem can be solved because the response variable being looked at is categorical and only has two possible outcomes. Three predictor variables will be used from the dataset and will be run through a series of logistical regression tests and will be essential for modeling the probability of a categorical outcome based on those predictor variables.

Analysis

Importing the necessary libraries including the ISLR library where the college dataset is located.

```
install.packages("ISLR")
library(ISLR)
install.packages("pROC")
library(pROC)
install.packages("caret")
library(caret)
library(ggplot2)
library(gridExtra)
attach(College)
```

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

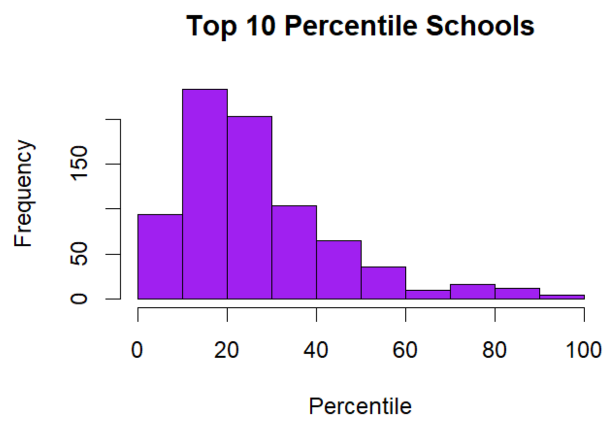
From the summary statistics there are 18 columns and 777 rows of data in the college dataset. The 'Private' variable is the only categorical variable in the dataset, with the other variables being all numerical. At the beginning of the summary, it indicates that there are 212 schools that are not private and 565 schools that are private. The information provided in the summary statistics for each column include minimum and maximum, median, mean, and first and third quartiles.

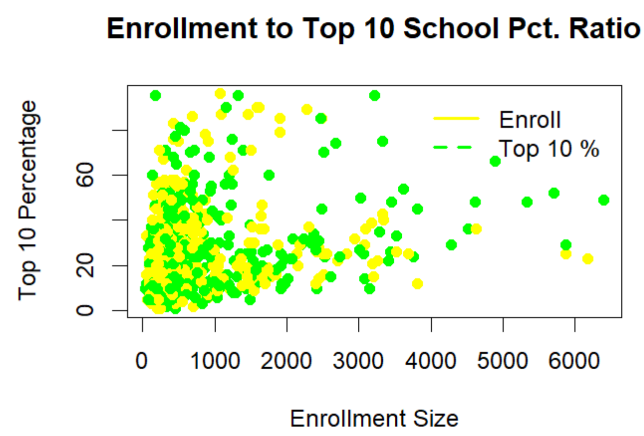
```
2), lwd = 2, bty = "n", inset = c(0.02, 0.02))
> summary(College)
 Private        Apps           Accept          Enroll        Top10perc       Top25perc    
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0  
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 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  
```

Several plots were created to show the correlations between the predicting variables, including a histogram, boxplots, and quick plots.

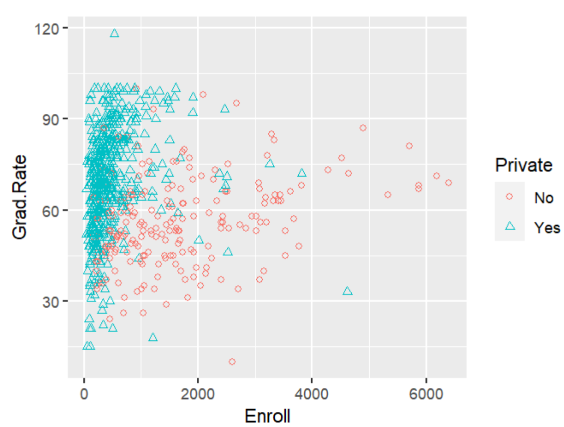
```r
hist(College$Top10perc, main = "Top 10 Percentile Schools", col = "purple", xlab = "Percentile")

plot(College$Enroll, College$Top10perc, main = "Enrollment to Top 10 School Pct. Ratio",
     col = c("yellow", "green"), pch = 19, xlab = "Enrollment Size", ylab = "Top 10 Percentage")
legend("topright", legend = c("Enroll", "Top 10 %"), col = c("yellow", "green"), lty = c(1, 2), lwd = 2, bty = "n", inset = c(0.02, 0.02))

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

x <- qplot(x=Private, y=Top10perc, fill=Private, geom = "boxplot") + guides(fill = FALSE)
y <- qplot(x=Private, y=Enroll, fill=Private, geom = "boxplot") + guides(fill = FALSE)
z <- qplot(x=Private, y=Grad.Rate, fill=Private, geom = "boxplot") + guides(fill = FALSE)
grid.arrange(x,y,z, nrow = 1)
```

The histogram on the left lays out the top 10 percentile schools in the dataset and is right skewered with a majority of the schools having scores below 30. The scatterplot on the right compares the top 10 percentile schools with enrollment size. It shows a lot of outliers but the majority of the plots of clumped together where enrollment size is smaller and where the schools have low percentile scores.

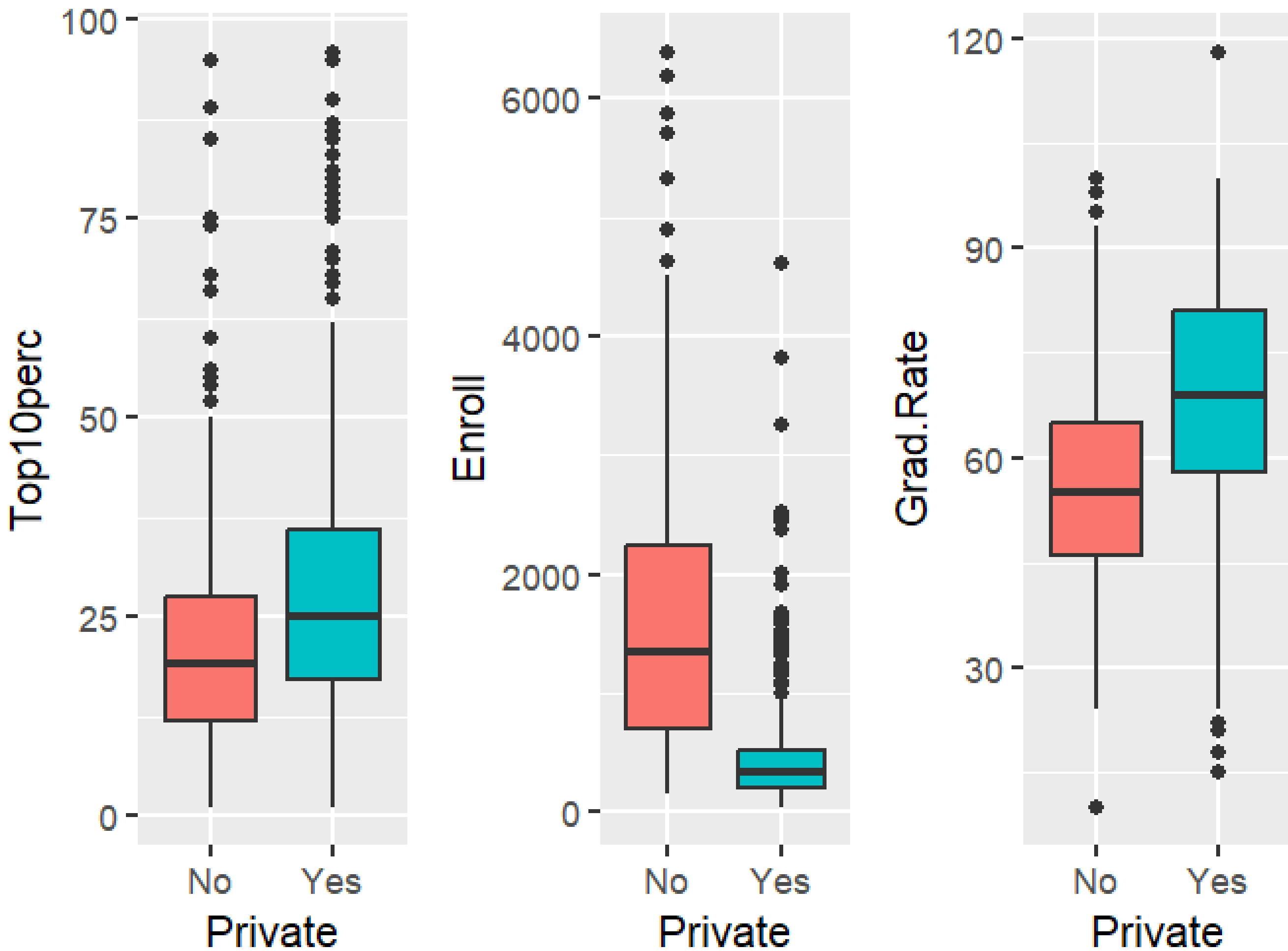




The quick plot compares the enrollment size of private and public schools with graduation rates. An analysis of the plot shows that the private schools are smaller in enrollment size than public schools and have higher graduation rates than public schools.



A comparison of the three boxplots shows all the correlations between the response variable and the three independent variables that were chosen. A common theme between the three boxplots for the private schools in the dataset are that they are more likely to be higher up in the top 10 percentile, have a smaller enrollment size than schools that are not private, and they have higher graduation rates than public schools. One noticeable observation is that the private school's enrollment sizes are quite small in range but a lot of outliers that stretch well beyond the whiskers of the boxplot.



Question 2: Split the data into a train and test set – refer to the Feature_Selection_R.pdf document for information on how to split a dataset.

The dataset is divided into a training set and test set, with 70 percent of the data being allocated to the training model and the remaining 30 percent being used for evaluation of the model's performance. After the data has been split up the head of the two sets are run, with the training set showing the top 100 results contained.

```r
#Question 2: Split the data into a train and test set - refer to the Feature_Selection_R.pdf document for
#information on how to split a dataset.

# load data
data("College")
head(College)

# Create Train and Test set - random sample (70/30 split)
trainIndex <- sort(sample(x = nrow(College), size = nrow(College) * 0.7))
sample_train <- College[trainIndex,]
sample_test <- College[-trainIndex,]

# Create Train and Test set - maintain % of event rate (70/30 split)
library(caret)
set.seed(123)
trainIndex <- createDataPartition(College$Private, p = 0.7, list = FALSE, times = 1)
caret_train <- College[ trainIndex,]
caret_test <- College[-trainIndex,]

head(caret_train, n = 100)
head(caret_test)
```

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

The glm() function is used to establish a logistic regression model between the 'Private' response variable and the three predictors. The enrollment variable shows a small decrease of –0.0032 with the possibility of the university being private for every additional student that enrolls there. The graduation rate and top 10 percentile variables show an increase for every unit increase in the model. The p-values of all three explanatory variables are smaller than 0.05 which suggests that there a strong possibility that the response variable is being impacted by those factors.

```
> model2 <- glm(Private ~ Enroll + Grad.Rate + Top10perc, data = caret_train, family = binomial(link = "logit"))
> summary(model2)

Call:
glm(formula = Private ~ Enroll + Grad.Rate + Top10perc, family = binomial(link = "logit"),
    data = caret_train)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.5870860  0.5651930  -2.808  0.00498 ** 
Enroll      -0.0032562  0.0003317  -9.816  < 2e-16 ***
Grad.Rate    0.0668055  0.0110449   6.049 1.46e-09 ***
Top10perc    0.0357326  0.0116282   3.073  0.00212 ** 
---
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: 321.84  on 541  degrees of freedom
AIC: 329.84

Number of Fisher Scoring iterations: 6
```

Question 4: Create a confusion matrix and report the results of your model for the train set. Interpret and discuss the confusion matrix. Which misclassifications are more damaging for the analysis, False Positives or False Negatives?

The evaluation of the confusion matrix concluded that there was 379 True Positives, 97 True Negatives, 52 False Positives, and 17 False Negatives. These numbers indicate how each instance in the training set was predicted. The 52 instances that were false positives were due to being incorrectly predicted as yes when they are actually no, known as a Type I Error. The 17 instances that were false negatives were due to being incorrectly predicted no when they were actually a yes which is known as a Type II Error. The Type I Error is a more damaging misclassification I believe for students enrolling or applying to what they think is a private school when it is not, perhaps taking away the prestige factor of the university. A Type II Error could be an issue as well with the cost of the school, believing that a school is not private and it is, so the cost could potentially be higher.

```
> #Model Accuracy
> confusionMatrix(predicted.classes.min, caret_train$Private, positive = 'Yes')
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No   97  17
       Yes  52 379

               Accuracy : 0.8734
                 95% CI : (0.8425, 0.9001)
    No Information Rate : 0.7266
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6561

 Mcnemar's Test P-Value : 4.256e-05

            Sensitivity : 0.9571
            Specificity : 0.6510
         Pos Pred Value : 0.8794
         Neg Pred Value : 0.8509
             Prevalence : 0.7266
         Detection Rate : 0.6954
   Detection Prevalence : 0.7908
      Balanced Accuracy : 0.8040

       'Positive' Class : Yes
```

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

Calculating the formula for accuracy of the matrix, 0.8734 or 87 percent of the predicted instances were correct. The positive predicted value or the precision was at 0.8794, meaning that 87 percent of the instances that were predicted positive were actually positive. The sensitivity value in the matrix or the recall was 0.9571, indicating that 95 percent of the actual positive occurrences were correctly predicted. And in the specificity metric, the value of 0.6510 means that 65 percent of the actual negative instances were correctly predicted which isn't very strong compared to the other tests conducted in the matrix.

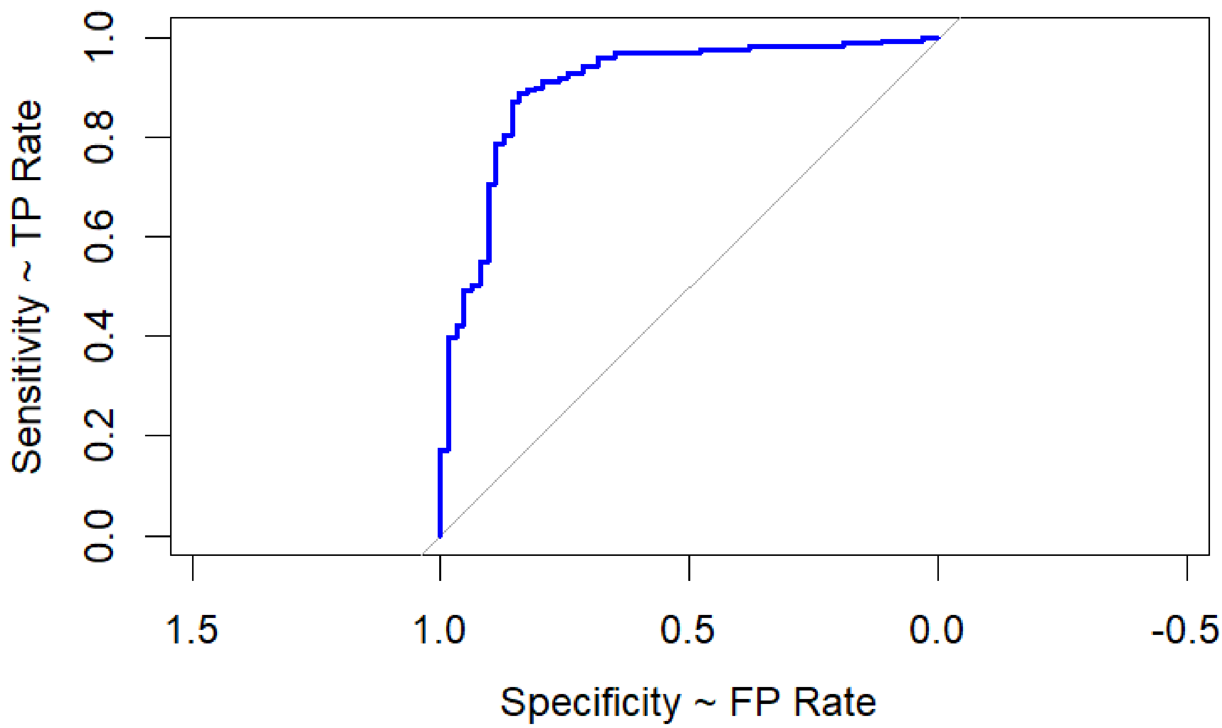
Question 6: Create a confusion matrix and report the results of your model for the test set.

The confusion matrix model that was made for the test set shows that there were 157 True Positives, 45 True Negatives, 18 False Positives, and 12 False Negatives. The overall accuracy of 87 percent the matrix was nearly identical to the confusion matrix that was used for the training set. The values for precision, recall, and specificity were all fairly close to the previous confusion matrix for the training set.

```
> #Model Accuracy
> confusionMatrix(predicted.classes.min, caret_test$Private, positive = 'Yes')
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No   45  12
       Yes  18 157

               Accuracy : 0.8707
                 95% CI : (0.8206, 0.911)
    No Information Rate : 0.7284
    P-Value [Acc > NIR] : 1.326e-07

                  Kappa : 0.6631

 Mcnemar's Test P-Value : 0.3613

            Sensitivity : 0.9290
            Specificity : 0.7143
         Pos Pred Value : 0.8971
         Neg Pred Value : 0.7895
             Prevalence : 0.7284
         Detection Rate : 0.6767
   Detection Prevalence : 0.7543
      Balanced Accuracy : 0.8216

       'Positive' Class : Yes
```

#Question 7: Plot and interpret the ROC curve

The model looks to be performing well, with the curve well above the random line in the plot. The curve rises steeply and indicates that the model has a large positive rate and a low false positive rate. The curve also nearly reaches the top-left corner which means the model's performance will be better with regards to sensitivity and specificity.



Question 8: Calculate and interpret the AUC.

The area under the ROC curve is 0.9049, and the closer the score is to one, the better the overall performance of the model will be. Because it is close to one it has a high positive rate and will also mean that the false positive rate will be low.

Conclusion/Interpretations

Using logistic regression modeling on a dataset provided me with the opportunity to create training and test sets and then a confusion matrix for the data. The relationship between the response variable and the independent variables indicated that there was enough evidence in forecasting whether a college was private or not. The confusion matrix was effective in its high accuracy rates and also had very few Type I and Type II errors. This is a great and efficient tool for using logistic regression modeling on a dataset and I think is especially useful when using larger datasets that include categorical variables.

References

Chat GPT. (2023, December 10th). Default (GPT 3.5). < https://chat.openai.com/>

Kabacoff, R. (2015). *R in Action.* 2nd Edition, Manning Publisher.

Prabhakaran, S. (2016-17). R-statistics.co.
< https://r-statistics.co/Logistic-Regression-With-R.html>

Appendix

```r
install.packages("ISLR")
library(ISLR)
install.packages("pROC")
library(pROC)
install.packages("caret")
library(caret)
library(ggplot2)
library(gridExtra)
attach(College)

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

summary(College)
View(College)
dim(College)
head(College)
tail(College)
str(College)
class(College)
range(College$Grad.Rate)

hist(College$Top10perc, main = "Top 10 Percentile Schools", col = "purple", xlab = "Percentile")

plot(College$Enroll, College$Top10perc, main = "Enrollment to Top 10 School Pct. Ratio",
     col = c("yellow", "green"), pch = 19, xlab = "Enrollment Size", ylab = "Top 10 Percentage")
legend("topright", legend = c("Enroll", "Top 10 %"), col = c("yellow", "green"), lty = c(1, 2), lwd = 2, bty = "n", inset = c(0.02, 0.02))

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

x <- qplot(x=Private, y=Top10perc, fill=Private, geom = "boxplot") + guides(fill = FALSE)
y <- qplot(x=Private, y=Enroll, fill=Private, geom = "boxplot") + guides(fill = FALSE)
z <- qplot(x=Private, y=Grad.Rate, fill=Private, geom = "boxplot") + guides(fill = FALSE)
grid.arrange(x,y,z, nrow = 1)

#Question 2: Split the data into a train and test set - refer to the Feature_Selection_R.pdf document for
#information on how to split a dataset.

# load data
data("College")
head(College)

# Create Train and Test set - random sample (70/30 split)
trainIndex <- sort(sample(x = nrow(College), size = nrow(College) * 0.7))
sample_train <- College[trainIndex,]
sample_test <- College[-trainIndex,]

# Create Train and Test set - maintain % of event rate (70/30 split)
library(caret)
set.seed(123)
trainIndex <- createDataPartition(College$Private, p = 0.7, list = FALSE, times = 1)
caret_train <- College[ trainIndex,]
caret_test <- College[-trainIndex,]

head(caret_train, n = 100)
head(caret_test)

dim(caret_train)
dim(caret_test)

#Question 3: Use the glm() function in the 'stats' package to fit a logistic regression model to the
#training set using at least two predictors.
```

```r

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

model2 <- glm(Private ~ Enroll + Grad.Rate + Top10perc, data = caret_train, family = binomial(link = "logit"))
summary(model2)

#Display regression coefficients (log-odds)
coef(model2)

#Display rergression coefficents (odds)
exp(coef(model2))

#View min, mean and max values for pedigree
summary(College$Grad.Rate)

#Question 4: Create a confusion matrix and report the results of your model for the train set. Interpret
#and discuss the confusion matrix. Which misclassifications are more damaging for the
#analysis, False Positives or False Negatives?

#Make predictions on the test data using lambda.min
probabilities.train <- predict(model2, newdata = caret_train, type = 'response')
predicted.classes.min <- as.factor(ifelse(probabilities.train >= 0.5, "Yes", "No"))

#Model Accuracy
confusionMatrix(predicted.classes.min, caret_train$Private, positive = 'Yes')

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

#Question 6: Create a confusion matrix and report the results of your model for the test set.

#Test set predictions
probabilities.test <- predict(model2, newdata = caret_test, type = 'response')
predicted.classes.min <- as.factor(ifelse(probabilities.test >= 0.5, "Yes", "No"))

#Model Accuracy
confusionMatrix(predicted.classes.min, caret_test$Private, positive = 'Yes')

#Question 7: Plot and interpret the ROC curve
ROC1 <- roc(caret_test$Private, probabilities.test)

plot(ROC1, col = "blue", ylab = "Sensitivity ~ TP Rate", xlab = "Specificity ~ FP Rate")

#Question 8: Calculate and interpret the AUC.

#Calculate the area under the ROC curve
auc <- auc(ROC1)
auc

```