**Module 3 Assignment**

Dhairyav Jatin Shah

College of Professional Studies, Northeastern University

Master of Professional Studies in Analytics, ALY6015

Professor Roy Wada

Due Date: 11/29/202

**Module 2 Assignment**

**Introduction**

In the week 3 assignment, we have a data set of statistics from the 1995 issue of US News and World Report for a wide number of US colleges. The ISLR R package contains the College data set. On this data set, we are going to perform exploratory data analysis to understand the data and further build a prediction model to predict if a given school is public or private using logistic regression. Hence, we can say that our target variable for this analysis is a categorical variable “Private” from the data set. In addition to this, there are 777 observations and 17 more variables that, depending on their reliance on the target variable, can be utilized as predictor variables.

**Analysis**

**Exploratory Data Analysis**

The initial step to start data analysis is to install and import the required libraries in R Studio. Next, we start by importing the data into R. Below is the descriptive statistics table of the entire population.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table No. 1 Descriptive Statistics of the Population | | | | | | |
|  | n | mean | min | sd | median | max |
| Apps | 777 | 3001.638 | 81 | 3870.201 | 1558 | 48094 |
| Accept | 777 | 2018.804 | 72 | 2451.114 | 1110 | 26330 |
| Enroll | 777 | 779.973 | 35 | 929.1762 | 434 | 6392 |
| Top10perc | 777 | 27.55856 | 1 | 17.64036 | 23 | 96 |
| Top25perc | 777 | 55.79665 | 9 | 19.80478 | 54 | 100 |
| F.Undergrad | 777 | 3699.907 | 139 | 4850.421 | 1707 | 31643 |
| P.Undergrad | 777 | 855.2986 | 1 | 1522.432 | 353 | 21836 |
| Outstate | 777 | 10440.67 | 2340 | 4023.016 | 9990 | 21700 |
| Room.Board | 777 | 4357.526 | 1780 | 1096.696 | 4200 | 8124 |
| Books | 777 | 549.381 | 96 | 165.1054 | 500 | 2340 |
| Personal | 777 | 1340.642 | 250 | 677.0715 | 1200 | 6800 |
| PhD | 777 | 72.66023 | 8 | 16.32815 | 75 | 103 |
| Terminal | 777 | 79.7027 | 24 | 14.72236 | 82 | 100 |
| S.F.Ratio | 777 | 14.0897 | 2.5 | 3.958349 | 13.6 | 39.8 |
| perc.alumni | 777 | 22.74389 | 0 | 12.3918 | 21 | 64 |
| Expend | 777 | 9660.171 | 3186 | 5221.768 | 8377 | 56233 |
| Grad.Rate | 777 | 65.46332 | 10 | 17.17771 | 65 | 118 |

From the above table, we can see that the first column represents the number of rows present in the data set looking at this we can see that there are no null values present in any of the variables. “Private” is a categorical variable hence we have decided not to include it in the descriptive statistics table. Furthermore, we split this table into 2 parts where the college is private and public to derive the statistics table using which we can derive more insights into the data set.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table No. 2 Descriptive Statistics of Public and Private Colleges | | | | | | | | | | |
| Private | | | | | | Public | | | | |
|  | mean | min | sd | median | max | mean | min | sd | median | max |
| Apps | 1977.929 | 81 | 2443.341 | 1133 | 20192 | 5729.92 | 233 | 5370.675 | 4307 | 48094 |
| Accept | 1305.703 | 72 | 1369.549 | 859 | 13007 | 3919.288 | 233 | 3477.266 | 2929.5 | 26330 |
| Enroll | 456.9451 | 35 | 457.5291 | 328 | 4615 | 1640.873 | 153 | 1261.592 | 1337.5 | 6392 |
| Top10perc | 29.33097 | 1 | 17.85139 | 25 | 96 | 22.83491 | 1 | 16.18044 | 19 | 95 |
| Top25perc | 56.95752 | 9 | 19.58836 | 55 | 100 | 52.70283 | 12 | 20.09106 | 51 | 100 |
| F.Undergrad | 1872.168 | 139 | 2110.662 | 1274 | 27378 | 8571.005 | 633 | 6467.696 | 6785.5 | 31643 |
| P.Undergrad | 433.9664 | 1 | 722.3705 | 207 | 10221 | 1978.189 | 9 | 2321.035 | 1375 | 21836 |
| Outstate | 11801.69 | 2340 | 3707.471 | 11200 | 21700 | 6813.41 | 2580 | 2145.248 | 6609 | 15732 |
| Room.Board | 4586.143 | 2370 | 1089.698 | 4400 | 8124 | 3748.241 | 1780 | 858.1399 | 3708 | 6540 |
| Books | 547.5062 | 250 | 174.9323 | 500 | 2340 | 554.3774 | 96 | 135.7299 | 550 | 1125 |
| Personal | 1214.441 | 250 | 632.8796 | 1100 | 6800 | 1676.981 | 400 | 677.5157 | 1649 | 4288 |
| PhD | 71.09381 | 8 | 17.35089 | 73 | 100 | 76.83491 | 33 | 12.31753 | 78.5 | 103 |
| Terminal | 78.53451 | 24 | 15.45025 | 81 | 100 | 82.81604 | 33 | 12.06967 | 86 | 100 |
| S.F.Ratio | 12.94549 | 2.5 | 3.518573 | 12.7 | 39.8 | 17.13915 | 6.7 | 3.418049 | 17.25 | 28.8 |
| perc.alumni | 25.89027 | 2 | 12.40075 | 25 | 64 | 14.35849 | 0 | 7.518935 | 13.5 | 48 |
| Expend | 10486.35 | 3186 | 5682.577 | 8954 | 56233 | 7458.316 | 3605 | 2695.542 | 6716.5 | 16527 |
| Grad.Rate | 68.99823 | 15 | 16.74946 | 69 | 118 | 56.04245 | 10 | 14.58341 | 55 | 100 |

The average number of applications for public schools is significantly higher than for private schools, as shown in table no. 2. Public schools receive an average of more applicants and enrollees than private schools. For further information, we are going to plot some visualizations to get a better understanding of colleges and other variables.

**Visualization**

In this part of the analysis, we will plot various graphs to understand the relationships between the target variable “Private” and other predictor variables. We start by plotting a bar chart to understand the number of public and private colleges in the data set.

Chart, bar chart, waterfall chart

Description automatically generated From figure 1, we can clearly see that there are almost twice the number of private colleges that public colleges in the given data set. There are more than 500 private colleges whereas there is only 200+ public colleges.

Chart, box and whisker chart

Description automatically generatedNext, we plot the number of applications made by students in each college. Here in figure 2, we can see that even if the number of public colleges are half than that of private but the number of applications in public colleges is greater than private.

Chart, bar chart, waterfall chart

Description automatically generated In Figures 3 & 4, we have plotted a box plot to study accepted and enrolled students. In figure 3, we can see that number of students accepted in public schools is greater than in private schools as well as in figure 4 we can see that the number of students enrolled in public schools is much greater than the private schools. 95th percentile in figure 4 is more than 4000 students in public schools whereas the 95th percentile is just 1000 students in private schools.

In figures 5 & 6, we have plotted a bar chart studying the full-time and part-time students enrolled in public and private schools. Clearly, we can see that both part-time and full-time students are enrolled more in public than private schools.

In figures 7,8,9 & 10, we have plotted the average cost per student for out-of-state tuition, instructional expense, cost of books, and personal expenses. From fig. 7, we see that the average cost of out-state tuition for private college students is almost double that of public college students. The average out-of-state tuition for a private college student cost $11,801.69 whereas for public college students costs only $6,813.41. From fig. 8, the average cost of the instructional expense for a private college student is $10,486.35 whereas for a public college student is only $7,458.32. From fig. 9, the average cost of books for a private college student is $547.51 and for a public college student is $554.38. From fig. 10, the average cost of personal expenses for a private college student is much less than that of a public college student. We can say that the overall cost of a private college student is higher than that of a private college student.

Chart, waterfall chart

Description automatically generatedChart, treemap chart

Description automatically generated In figures 11,12,13 & 14, we study the faculties, student-to-faculty ratio, and graduation rate in public and private schools. Private colleges have more professors with Ph.D. as well as a terminal degree, this can be the case since the number of private schools are also high compared to public schools. Also, the student-to-faculty ratio is less in private schools with a ratio of 12.95:1. We can also see that the graduation rate of students in private schools is higher than that of public schools with a graduation rate of 56.04% and private schools with a graduation rate of 69%. The three variables faculty with Ph.D., terminal degree, and student-faculty ratio are not very helpful for now as the number of public and private colleges are not equal. Hence, they can give biased analysis towards private colleges.

**Prediction Using Logistic Regression**

We start the prediction by splitting the main data set into 2 different subsets, “Train” and “Test”. The train data set contains 70% of the main data set and the test data set contains the remaining 30% of the data set.Once we have split the data into 2 parts. We run logistic regression on the train of the data set using the glm() function in R. Where we must mention the family= “Binomial”, which states we want to run the model using logistic regression. We run the model with all the predictor variables and choose the variables that are correlated with the target variable to get accurate results.

The below table shows the output we get once run the model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table no.3: Logistic Regression Model on Training Data. | | | | |
|  | Estimate | Std.Error | Z value | p-value(>|z|) |
| Intercept | -2.004e-02 | 1.008 | -0.020 | 0.984131 |
| Applications | -4.802e-04 | 7.901e-05 | -6.078 | 1.22e-09 \*\*\* |
| Outstate Tuition | 8.068e-04 | 8.756e-05 | 9.214 | < 2e-16 \*\*\* |
| Faculty with Ph.D | -8.035e-02 | 1.712e-02 | -4.693 | 2.69e-06 \*\*\* |
| Alumni (%) | 7.481e-02 | 2.106e-02 | 3.553 | 0.000382 \*\*\* |

From table 3, we see that all 4 variables used to predict if the college is public or private have very low p-value hence all variables used to predict are significant in the model.

**Confusion Matrix of Training Data Set.**

The below table displays the confusion matrix of the logistic model fit on the training data set.

|  |  |  |
| --- | --- | --- |
| Table no.4: Confusion Matrix of the training data set | | |
|  | Reference | |
| Prediction | **No** | **Yes** |
| **No** | 126 | 20 |
| **Yes** | 22 | 376 |

**Interpretation:**

The above confusion matrix is obtained from the training data set. The True Negatives are 126, True Positives are 376, False Positives are 22 and False Negatives are 20.

False Positives – If the predicted college is Private and if it is false

False Negatives – If the predicted college is not Private and if it is false.

In our case, the predictions for false negatives are more damaging to the analysis. As if the college is not a private college and it is false it will affect the prediction.

**Accuracy, Precision, Recall, and Specificity of the Training Data set.**

**Metrics of the training model:**

|  |  |
| --- | --- |
| Table no. 5 Metrics of the Model | |
| **Metrics** | **Values** |
| Accuracy | 92.28% |
| Precision | 94.47% |
| Recall | 94.94% |
| Specificity | 85.14% |

**Interpretation:**

From the above metrics, we have observed that the accuracy of the model is 92.28%, which is good. The Precision is derived from (TP/(TP+FP)) which is 94.47%, Recall is derived from (TP/(TP+FN)) which is 94.94%, and the specificity is 85.14%. Considering the obtained values, we can conclude that the model is accurate enough to predict whether a college is private or public.

**Confusion Matrix of Test Data Set.**

The below table displays the confusion matrix of the logistic model run on the test data set.

|  |  |  |
| --- | --- | --- |
| Table no.6: Confusion Matrix of the test data set | | |
|  | Reference | |
| Prediction | **No** | **Yes** |
| **No** | 56 | 6 |
| **Yes** | 8 | 163 |

**Interpretation:**

The above confusion matrix is obtained from the training data set. The True Negatives are 126, True Positives are 163, False Positives are 8 and False Negatives are 6.

**Accuracy, Precision, Recall, and Specificity of the Training Data set.**

**Metrics of the training model:**

|  |  |
| --- | --- |
| Table no. 7 Metrics of the Test Data | |
| **Metrics** | **Values** |
| Accuracy | 93.99% |
| Precision | 95.32% |
| Recall | 96.45% |
| Specificity | 87.50% |

**Interpretation:**

From the above metrics, we observed the following metrics when we ran the model on the test data set the accuracy of the model is 93.99%, which is increased from the training data set. The Precision is derived from (TP/(TP+FP)) which is 95.32%, Recall is derived from (TP/(TP+FN)) which is 96.45%, and the specificity is 87.50%. Looking at the above metrics we can conclude that the model is predicting accurately.

**Plotting and interpreting ROC curve and calculating AOC.**

Chart

Description automatically generatedAn illustration of a classification model's performance at each classification threshold is called a receiver operating characteristic curve (ROC Curve).

**Interpretation:**

The above plot represents the ROC curve. It is derived by plotting Specificity (False Positive Rate) vs Sensitivity (True Positive Rate). The ROC curve should be closer to 1. Hence, from the above graph, we can see that the peak is closer to 1 and it is representing an accurate Model.

**Calculating the Area Under the Curve:**

The area under the ROC curve is 98.5%. Since the value of AOC is higher, we can say that the model is predicting Private colleges as Private and Public Colleges as Public.

**Conclusion:**

We have found that there are more private colleges than public institutions according to our analysis of the college data collection. Public colleges, however, receive more applications, are accepted by more students, and have a higher enrolment rate. A student in a private institution will spend more overall than a student in a public college. We achieved an accuracy of 92% after doing logistic regression on the training data set, and we achieved an accuracy of 93.99% after using this model on the test data set. The curve's area is 98.5%, making it a good model.

**Reference**

1. *R Dataset / Package ISLR / College* <https://r-data.pmagunia.com/dataset/r-dataset-package-islr-college>
2. Zach(Apr 2021) *How to Create a Confusion Matrix in R (Step-by-Step)* <https://www.statology.org/confusion-matrix-in-r/>
3. DataTechNotes(June 2019) *How to create a ROC curve in R* <https://www.datatechnotes.com/2019/03/how-to-create-roc-curve-in-r.html#:~:text=ROC%20curve%20is%20a%20metric,multi%2Dclass%20classification%20accuracy%20checking>.
4. Narkhede.S(June 2018) *Understanding AUC - ROC Curve* <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>

**Appendix**

# Week 3 Assignment

# Dhairyav Shah

# Importing Libraries

library(ISLR)

library(psych)

library(dplyr)

library(ggplot2)

library(ggpubr)

library(caTools)

library(caret)

library(regclass)

library(pROC)

library(stargazer)

# Importing the data set

df<-as.data.frame(College)

# Exploratory Data analysis

str(df)

df1 <- describe(df)

df2 <- df1 %>% select(n, mean, min, sd, median, max) # you still need to round decimals (see below)

write.csv(df2,"Summary\_Popluation.csv")

# Sub-setting into Public and Private

df\_pub<-subset(df,Private=="No")

df\_priv<-subset(df,Private=="Yes")

df3 <- describe(df\_pub)

df4 <- df3 %>% select(n, mean, min, sd, median, max) # you still need to round decimals (see below)

write.csv(df4,"Summary\_Pub.csv")

df5 <- describe(df\_priv)

df6 <- df5 %>% select(n, mean, min, sd, median, max) # you still need to round decimals (see below)

write.csv(df6,"Summary\_Priv.csv")

# Visualization

#Bar plot Public and Private colleges

g1<-ggplot(data = df,aes(fill = Private)) +

geom\_bar(mapping = aes(x=Private)) +

ggtitle("Fig 1:Number of Colleges") +

scale\_y\_continuous(name = "Number of Colleges") +

scale\_x\_discrete(name = "Private College") +

scale\_fill\_manual(values = c("steelblue","Red")) +

theme\_light()

guide = guide\_legend(title = "Private Type")

df\_app<- df %>%

group\_by(Private) %>%

summarise(Applications=sum(Apps))

g2<-ggplot(data = df\_app, aes(x= Private, y=Applications, fill= Private))+

geom\_bar(stat = "identity")+

ggtitle("Fig 2: Number of Applications") +

scale\_y\_continuous(name = "Number of Colleges") +

scale\_x\_discrete(name = "Private College") +

scale\_fill\_manual(values = c("steelblue","Red")) +

geom\_text(aes(label=Applications), vjust=2, color="black", size=3.5)+

theme\_light()

ggarrange(g1,g2,

ncol = 2, nrow = 1)

# Box plots to study the Accepted and Enrolled

# Applications

p1<-ggplot(data = df, aes(Private,Accept,Apps,fill = Private))+

geom\_boxplot(color = "black", fill = "green")+

ggtitle("Fig 3: Number of Applications Accepted")+

theme\_light()

#Enrolled

p2<-ggplot(data = df, aes(Private,Enroll,Apps,fill = Private))+

geom\_boxplot(color = "black", fill = "green")+

ggtitle("Fig 4: Number of Students Enrolled")+

theme\_light()

ggarrange(p1,p2,

ncol = 2, nrow = 1)

# Number of students enrolled in part-time and full-time in public and private university

df\_time<- df %>%

group\_by(Private) %>%

summarise(full\_time=sum(F.Undergrad),

part\_time=sum(P.Undergrad))

p3<-ggplot(data = df\_time, aes(x= Private, y=full\_time, fill= Private))+

geom\_bar(stat = "identity")+

ggtitle("Fig 5: Number of Full-time Students") +

scale\_y\_continuous(name = "Number of Colleges") +

scale\_x\_discrete(name = "Private College") +

scale\_fill\_manual(values = c("steelblue","Red")) +

geom\_text(aes(label=full\_time), vjust=2, color="black", size=3.5)+

theme(legend.position = "none")+

theme\_light()

p4<-ggplot(data = df\_time, aes(x= Private, y=part\_time, fill= Private))+

geom\_bar(stat = "identity")+

ggtitle("Fig 6: Number of Part-time Students") +

scale\_y\_continuous(name = "Number of Colleges") +

scale\_x\_discrete(name = "Private College") +

scale\_fill\_manual(values = c("steelblue","Red")) +

geom\_text(aes(label=part\_time), vjust=2, color="black", size=3.5)+

theme\_light()

ggarrange(p3,p4,

ncol = 2, nrow = 1)

df\_expenses<- df %>%

group\_by(Private) %>%

summarise(out\_state=round(mean(Outstate),2),

expense=round(mean(Expend),2),

books=round(mean(Books),2),

personal=round(mean(Personal),2))

p5<-ggplot(data = df\_expenses, aes(x= Private, y=out\_state, fill= Private))+

geom\_bar(stat = "identity")+

ggtitle("Fig 7: Average Cost of Out of State Tution") +

scale\_y\_continuous(name = "Number of Colleges") +

scale\_x\_discrete(name = "Private College") +

scale\_fill\_manual(values = c("steelblue","Red")) +

geom\_text(aes(label=out\_state), vjust=2, color="black", size=3.5)+

theme(legend.position = "none")+

theme\_light()

p6<-ggplot(data = df\_expenses, aes(x= Private, y=expense, fill= Private))+

geom\_bar(stat = "identity")+

ggtitle("Fig 8: Average Cost of Instructional Expense per Student") +

scale\_y\_continuous(name = "Number of Colleges") +

scale\_x\_discrete(name = "Private College") +

scale\_fill\_manual(values = c("steelblue","Red")) +

geom\_text(aes(label=expense), vjust=2, color="black", size=3.5)+

theme\_light()

p7<-ggplot(data = df\_expenses, aes(x= Private, y=books, fill= Private))+

geom\_bar(stat = "identity")+

ggtitle("Fig 9: Average Cost for books") +

scale\_y\_continuous(name = "Number of Colleges") +

scale\_x\_discrete(name = "Private College") +

scale\_fill\_manual(values = c("steelblue","Red")) +

geom\_text(aes(label=books), vjust=2, color="black", size=3.5)+

theme(legend.position = "none")+

theme\_light()

p8<-ggplot(data = df\_expenses, aes(x= Private, y=personal, fill= Private))+

geom\_bar(stat = "identity")+

ggtitle("Fig 10: Average Cost of Personal Expenses") +

scale\_y\_continuous(name = "Number of Colleges") +

scale\_x\_discrete(name = "Private College") +

scale\_fill\_manual(values = c("steelblue","Red")) +

geom\_text(aes(label=personal), vjust=2, color="black", size=3.5)+

theme(legend.position = "none")+

theme\_light()

ggarrange(p5,p6,p7,p8,

ncol = 2, nrow = 2)

df\_misc<- df %>%

group\_by(Private) %>%

summarise(PHD=round(sum(PhD),2),

terminal=round(sum(Terminal),2),

s.f=round(mean(S.F.Ratio),2),

grad.rate=round(mean(Grad.Rate),2))

p9<-ggplot(data = df\_misc, aes(x= Private, y=PHD, fill= Private))+

geom\_bar(stat = "identity")+

ggtitle("Fig 11: Number of Faculty with PHD") +

scale\_y\_continuous(name = "Number of Colleges") +

scale\_x\_discrete(name = "Private College") +

scale\_fill\_manual(values = c("steelblue","Red")) +

geom\_text(aes(label=PHD), vjust=2, color="black", size=3.5)+

theme(legend.position = "none")+

theme\_light()

p10<-ggplot(data = df\_misc, aes(x= Private, y=terminal, fill= Private))+

geom\_bar(stat = "identity")+

ggtitle("Fig 12: Number of Faculty with Terminal Degree ") +

scale\_y\_continuous(name = "Number of Colleges") +

scale\_x\_discrete(name = "Private College") +

scale\_fill\_manual(values = c("steelblue","Red")) +

geom\_text(aes(label=terminal), vjust=2, color="black", size=3.5)+

theme\_light()

p11<-ggplot(data = df\_misc, aes(x= Private, y=s.f, fill= Private))+

geom\_bar(stat = "identity")+

ggtitle("Fig 13: Average Student Faculty Ratio") +

scale\_y\_continuous(name = "Number of Colleges") +

scale\_x\_discrete(name = "Private College") +

scale\_fill\_manual(values = c("steelblue","Red")) +

geom\_text(aes(label=s.f), vjust=2, color="black", size=3.5)+

theme(legend.position = "none")+

theme\_light()

p12<-ggplot(data = df\_misc, aes(x= Private, y=grad.rate, fill= Private))+

geom\_bar(stat = "identity")+

ggtitle("Fig 14: Average Graduation Rate") +

scale\_y\_continuous(name = "Number of Colleges") +

scale\_x\_discrete(name = "Private College") +

scale\_fill\_manual(values = c("steelblue","Red")) +

geom\_text(aes(label=grad.rate), vjust=2, color="black", size=3.5)+

theme(legend.position = "none")+

theme\_light()

ggarrange(p9,p10,p11,p12,

ncol = 2, nrow = 2)

# Creating Test and Train Data set

set.seed(1234)

sample <- sample.split(df$Private, SplitRatio = 0.7)

train <- subset(df, sample == TRUE)

test <- subset(df, sample == FALSE)

# Logistic Regression Analysis

m<- glm(Private~.,data = train,family = "binomial")

summary(m)

m1<- glm(Private~ Apps+Outstate+PhD+perc.alumni,data = train,family = "binomial")

summary(m1)

#Creating a Confusion Matrix for the train data set

p.train<-predict(m1,newdata= train,type = "response")

pred.prob<-as.factor(ifelse(p.train>=0.5,"Yes","No"))

confusionMatrix(pred.prob,as.factor(train$Private),positive = "Yes")

#Creating a Confusion Matrix for the test data set

p.test<-predict(m1,newdata= test,type = "response")

pred.prob\_test<-as.factor(ifelse(p.test>=0.5,"Yes","No"))

confusionMatrix(pred.prob\_test,as.factor(test$Private),positive = "Yes")

# Plotting the ROC curve

roc\_curve<-roc(test$Private,p.test)

plot(roc\_curve,col = "red", main = "ROC Chart", ylab = "Sensitivity", xlab = "Specificity")

# Calculating AUC

auc\_curve<-auc(roc\_curve)

auc\_curve