|  |  |  |  |
| --- | --- | --- | --- |
| DNSC 6313-10 Statistics for Analytics II | Lab 1 | GWlogoBlue | Student ID**:**  **G****27279521** |

**(This assignment is to be completed individually. While you may interact with others to gain mastery of the generic course content, all work associated with this assignment must be strictly and exclusively yours, including creating the R code and determining the correct answers/completing the template; the only collaboration permitted is with the instructor.)**

FlixIt Inc. purchases unlimited licenses to movie content that is then streamed on-demand to FlixIt subscribers. Subscribers pay a flat monthly fee, and are provided with unlimited access to FlixIt content. Last year, FlixIt implemented a “Recruit A Friend” (RAF) initiative. Under this initiative, any current FlixIt subscriber who recruits someone who purchases an annual FlixIt subscription is given a one-month rebate. FlixIt now wishes to determine the characteristics of subscribers who have participated in this initiative. Data collected from a random sample of FlixIt subscribers (contained in the file FlixIt\_2022.txt, which includes a header record) include age of the subscriber (Age: an integer), region of the country in which the subscriber resides (Region: 1=*north*, 2=*south*, 3=*east*, 4=*west*), and whether or not the subscriber participated in the RAF initiative (Partic: 0=*no*, 1=*yes*). Based on these data, and using a Logistic Regression framework, please use R (and ***only*** R) to complete the following template. ***Unless otherwise noted, all results should be accurate to at least 4 decimal places.*** Use an  level of .05. ***TO SIMPLIFY YOUR TASK, PLEASE USE AGE AS THE ONLY INDEPENDENT VARIABLE IN ANSWERING ALL OF THE FOLLOWING QUESTIONS***; please ignore all other independent variables. Unless otherwise specified, assume a threshold value of 0.5:

1. In the population, can we be reasonably certain that Age predicts Participation? (Yes/No) Yes  
   Why/why not? Because participation is categorical (binary) and it would be the dependent variable for the logistic regression.
2. How strong is the relationship (McFadden’s R2)? 0.3336328, Once the McFadden's pseudo R-squared value is between 0.2 to 0.4, I think it is the great fit and seems good relationship
3. What odds ratio is associated with Age? (provide a numerical value) exp(-12.96833 + 0.27695 \* Age) odd ratio would be 0.27695  
   Please provide an interpretation of this value: Once the value of odd ratio of Age is 0.27695. It means that it is the value that multiplied when the independent variable (age) is increased by 1 unit.
4. If you applied this model to 100 people, how many of them would you expect to be correctly classified? 82 Why? (134+29) / (134+29+15+22) \* 100 = 81.5 rounded up is 82.
5. If you had 100 people known to be Participants, how many of them would you expect to be ***incorrectly*** classified as non-participants by your logistic model? 43.13725 = 43 Why? FN / (FN + TP) is the equation for actual participant and predicted incorrectly as non-participant.
6. If you had 100 people known to be non-Participants, how many of them would you expect to be ***incorrectly*** classified as participants by your logistic model? 10.06711 = 10 Why? FP / (FP + TN) is the equation for actual non-participants and predicted incorrectly participant.\
7. If your model classifies 100 people as Participants, how many of those people would you expect to be ***incorrectly*** classified? 34.09091
8. If your model classifies 100 people as non-Participants, how many of those people would you expect to be ***incorrectly*** classified? 14.10256
9. Based on your model, what are the odds that a 30-year-old participates? 0.009468072 How did you derive this number? logit <- predict(flix.logit, newdata = data.frame(Age = c(30,40)), type = "link") // odds <- exp(logit) // odds or exp(-12.96833 + 0.27695 \* 30)
10. Based on your model, what are the odds that a 40-year-old participates? 0.151022
11. Based on the odds, how much more likely is it that a 40-year old participates compared to a 30-year-old? 40 years old is 15.95089 more likely compared to 30 years old.
12. What is the probability that a 30-year-old participates? 0.00937967 How did you derive this number? the equation of prob is (odds / (1+ odds)). So, odds of 30 years old is 0.009468481 so (0.009468481 / (1 + 0.009468481))
13. How much area is there under the ROC? (provide a numerical value) 0.8772  
    Please provide an interpretation of this value: it is the range between 0.50 to 1.00 and we say it is good model when the value is higher than 0.80. So when we plot the ROC.curve, it will shows the curved line on the plot section. Also, there is one straight line on the middle of plot section and that is exactly 50% or 0.5, so we calculate the area above the area and we have total 87.72% from my model which is 37.72% more than straight line.
14. What threshold value produces the maximum Accuracy? 0.3814781
15. What is the Accuracy value when using the threshold you found in the preceding question? (provide a numerical value) 0.845
16. Can we be reasonably sure that a logistic curve specifies an ***incorrect*** model for these data? (Yes/No) NO Why/Why not? After using Hosmer\_Lemeshow from resourceselection package, the p-value was 0.1441 which is bigger than 0.05 and when the test value is bigger than 0.05, we think it is good fit to the data of using logistic curve.

Paste your full R script immediately below this line:

# I used Session menu to set up working directory.

# Import flixIT INC data

flixit <- read.csv("FlixIt\_2022.txt",header = TRUE, sep = " ")

# Checking the data.

head(flixit,5)

tail(flixit,5)

summary(flixit)

flix.logit <- glm(Partic~Age, data = flixit, family= binomial(link="logit"))

summary(flix.logit)

AgeVals <- seq(20,70,1)

NewData <- data.frame(AgeVals)

NewData['PredVal'] <- predict(flix.logit, list(Age = NewData$AgeVals), type = "link")

NewData['PredProb'] <- exp(NewData$PredVal) / (1+ exp(NewData$PredVal))

plot(NewData$Age, NewData$PredProb, pch = 16, xlab = "Age", ylab = "Predicted Probability")

# Question 2

install.packages("pscl")

library(pscl)

pR2(flix.logit)

# Question 3

summary(flix.logit)

# Log = -12.96833 + 0.27695 \* Age

# odds = exp(-12.96833 + 0.27695 \* Age)

# prob = odds / (1 + odds)

# Question 4

install.packages("InformationValue")

library(InformationValue)

flixit['PredVal'] <- predict(flix.logit, list(Age = flixit$Age), type = "link")

flixit['Predprob'] <- predict(flix.logit, list(Age = flixit$Age), type = "response")

head(flixit)

summary(flix.logit)

confusionMatrix(flixit$Partic,flixit$Predprob,0.5)

sensitivity(flixit$Partic,flixit$Predprob,0.5)

specificity(flixit$Partic,flixit$Predprob,0.5)

precision(flixit$Partic,flixit$Predprob,0.5)

npv(flixit$Partic,flixit$Predprob,0.5)

(134 + 29) / (134+29+15+22) \* 100

# Question 5 ~ 8

# 0 1

# 0 134 22

# 1 15 29

# 5.

# Participant & Incorrectly non-participant (FN / FN + TP)

22 / (22+29) \* 100

# 6.

# Non - Participant & Incorrectly participant (FP / FP + TN)

15 / (15 + 134) \* 100

# 7.

# Participant & incorrectly classify

15 / (15 + 29) \* 100

# 8.

# non - participant & Incorrectly classify

22 / (134 + 22) \* 100

#Question 9 ~

logit <- predict(flix.logit, newdata = data.frame(Age = c(30,40)), type = "link")

odds <- exp(logit)

odds

# 40 years old participates compared to 30 years old

odds[2] / odds[1]

# probability

prob <- odds / (1+odds)

prob

0.009468481 /(1+0.009468481 )

# Question 13

install.packages("pROC")

library(pROC)

ROC.curve = roc(Partic~Age , data = flixit)

plot(ROC.curve, col = "red")

auc(ROC.curve)

# Question 14

# install.packages("InformationValue")

library(InformationValue)

optimalCutoff(flixit$Partic,flixit$Predprob)

# Question 15

confusionMatrix(flixit$Partic,flixit$Predprob,0.3814781)

(132+37) / (132+37 + 17 + 14)

# Question 16

install.packages("ResourceSelection")

library(ResourceSelection)

hoslem.test(flixit$Partic, fitted(flix.logit), g = 10)

0.1441 > 0.05