BUS 336: Data Analytics & Visualization

Gohram Baloch

Assignment 3

Logistic Regression

Total Marks: 60

Due date: 11:59 PM July 12, 2024

Instructions:

Please review the instructions carefully and make sure to follow the submission requirements.

1. This assignment is based on the data file `Student_Dropout.csv`, which has been uploaded to Canvas. The assignment focuses on Logistic Regression and should be completed individually.
2. You will need to submit:
    a. A single PDF file with R codes on Crowdmark. While answering each question, you report must include lines of R code used and the resulting output (e.g., graphs, model summary etc). For report format, please refer to the previous assignment solutions.
    b. The complete .R/RMD file to be uploaded on Canvas.
3. Late submissions will not be accepted.

You are doing great in the course. KEEP UP THE GOOD WORK!

Enhancing Student Retention: Predictive Modelling for Dropout Prevention at Simon Fraser University

SFU is experiencing an annual student dropout rate of approximately 12%, which impacts graduation rates, increases recruitment costs, and reduces overall university ranking. To address these issues, the university has hired you to develop a predictive model to identify students with high-risk of dropping out ahead of time. The insights gained from this analysis will guide the management in implementing targeted changes to reduce student dropout. Furthermore, SFU seeks to identify the most critical variable that requires immediate attention.

Data:

We'll use the student dropout data `Student_Dropout.csv` uploaded on Canvas. It provides personal student information and whether they dropped out in the past year. The variables in this dataset are:

Table 1: Data Codebook

Feature	Description
StudentID	Unique identifier for each student
Sex	Male (M) or Female (F)
BirthYear	Birth Year for the student
Major	Major subject of the student
GPA	Student's cumulative graph point average
CreditsCompleted	Number of credit hours completed
PartTime	Whether the student is a part-time student
Scholarship	Whether the student has a scholarship
Dropout	Whether the student dropped out in the past year

Part 1: Model Building [20 points]

a) How many students do we have data for? What percentage dropped out? **[2 points]**

b) Split the data into a training set and a testing set, putting first 75% of the rows in the training set. Then, build a best-fit logistic regression model to predict the variable `Dropout`. You should use the training dataset to build the model. **[5 points]**

c) Describe your resulting model. Which variables are significant in your model? Which variables are insignificant? **[5 points]**

d) What can we say based on the coefficient of the `GPA` variable? **[3 points]**

e) Consider a 23-year male student studying in Beedie School as a full-time student with a GPA of 3.1 after completing 100 credit hours having GPA of 3.1. The student also receives scholarship from Canadian Government. According to your model, what is the probability that this individual will dropout from the university? **[5 points]**

Part 2: Model Performance on Training Dataset [15 points]

a) Compute the model's predicted probabilities for student in the training set. Then create a confusion matrix for the training set using a threshold of 0.5. Calculate model's specificity, sensitivity, and accuracy on the training set? **[5 points]**

b) Plot an AUC curve, and use it to choose the best threshold. **[5 points]**

c) Repeat part (i) using the threshold selected in part (ii). **[5 points]**

Part 3: Performance Evaluation on Testing Dataset [25 points]

a) Now, compute the model's predicted probabilities for students in the testing set. Create a confusion matrix for the testing set using the best threshold in previous part. Calculate model's specificity, sensitivity, and accuracy on the test set? **[5 points]**

b) Compare your accuracy on the test set to a baseline model that predicts every student in the test set will be retained, regardless of the values of the independent variables. Does your model improve over this simple model? **[5 points]**

c) Compute the AUC of the model on the test set, and interpret what the number means in this context. **[5 points]**

d) Considering the AUC, accuracy, sensitivity, and specificity compared to the baseline model, and what happens when the threshold is adjusted, do you think this model is of value to SFU? Why or why not? **[10 points]**