ഭാരതീയ സാങ്കേതികവിദ്യാ സ്ഥാപനം പാലക്കാട്
भारतीय प्रौद्योगिकी संस्थान पालक्काड
Indian Institute of Technology Palakkad
Nurturing Minds For a Better World

IIT PALAKKAD

DEPARTMENT OF DATA SCIENCE

DS5610 BUSINESS ANALYTICS

TEST 1 (OPEN BOOK/NOTES)

28 March 2025 **Marks: 30 (Weight: 15 Marks)** **Duration: 50 Mins**

PLEASE READ THE FOLLOWING CAREFULLY

- ***Answer all the questions***. There are no choices.
- ***Answer all parts of the question together***. For example, if Q. No. 2 contains sub-questions such as 2a, 2b and 2c, all these questions should be answered together in sequence in one place. **Please do not write 2a on page 1 and 2c on the last page. SUCH ANSWERS WILL NOT BE EVALUATED.**
- Draw diagrams neatly and legibly.
- ***THIS IS AN OPEN-BOOK EXAMINATION. Textbook, hand-written notes in bounded form, and printed slide handouts are allowed. NO OTHER MATERIAL, LOOSE SHEETS are allowed.***
- **The use of a scientific calculator is permitted.** However, the **use of mobile phone, laptop and other electronic device are NOT PERMITTED during the examination.**
- **No exchange of scientific calculators, books, printed slides, loose sheets, etc. are permitted.**
- **If you find any of the data missing for any of the questions, MAKE YOUR OWN ASSUMPTIONS AND HIGHLIGHT THE SAME BY DRAWING A BOX AROUND IT.**
- **WRITE THE ANSWER IN A BLUE/BLACK PEN. Writing by Pencil or Red/Green Pen is not allowed.**

Q. No. 1. Consider the "Default" dataset available in the "ISLR" library. It provides information on 10,000 customers. Furthermore, the dataset contains four variables as follows:

Table 1: Variable Description

Variables	Description
Default	A factor with levels "No" and "Yes" indicating whether the customer defaulted on their debt
Student	A factor with levels "No" and "Yes" indicating whether the customer is a student
Balance	The average balance that the customer has remaining on their credit card after making their monthly payment
Income	Income of customer

The objective is to predict whether an individual will default on his/her credit card payment. The data set is divided into two parts: a training set consisting of 5000 observations and a test set consisting of the remaining 5000 observations. The following results are available:

Table 2: Logistic Regression Model (Model 1) for predicting "default" using "student" based on the Training Data Set

	Estimate	Standard Error	z value	p-value
Intercept	3.482	0.099	35.312	0.000
studentYes	0.307	0.166	1.845	0.065

Note that "studentYes" is a dummy variable which takes the value 1 if the individual is a "student" and 0 otherwise.

Furthermore, another logistic regression model using the variables "student" and "balance" is fitted based on the training data set. The details are provided below:

Table 3: Logistic Regression Model (Model 2) for predicting "default" using "balance" and "student" based on the Training Data Set

	Estimate	Standard Error	z value	p-value
Intercept	−11.020	0.7130	−15.446	0.000
balance	0.006	0.0003	17.383	0.000
studentYes	−0.822	0.3401	−2.416	0.016

Based on the fitted logistic regression model (Model 2), a confusion matrix is obtained for the test dataset using a threshold of 0.5. The confusion matrix is shown below.

Table 4: Confusion Matrix Based on Test Data Set for Logistic Regression

		True "Default" Status	
		No	Yes
Predicted "Default" Status	No	4808	120
	Yes	23	49

Based on the above output, answer the following questions (**no need to fit any of the models**):

a. Write the equation of the fitted Model 1 (summary provided in Table 2). Calculate the predicted default probabilities for an individual who is a student and for an individual who is not a student. Who is riskier for the credit card company? [5]

b. Consider the fitted Model 2 (summary provided in Table 3). Interpret the coefficients. Discuss the difference between Models 1 and 2. [6]

c. Suppose the credit card company wants to provide a credit card only to those customers who have the predicted default probability below 0.05. Recently, a student has approached the credit card company. Based on the fitted Model 2 (summary provided in Table 3), calculate the maximum allowed "balance" for such an individual. [5]

d. Based on the confusion matrix shown above (in Table 4), compute the *sensitivity*, *specificity* and *total error rate* for the logistic regression model. Interpret each of these performance metrics. Which one will you use in this context? Why? [6]

e. Discuss in detail how the performance of the logistic regression model can be improved. [4]

f. How will you use such a logistic regression model for decision-making in this context? [4]

*****GOOD LUCK*****