MSBA7002 Business Statistics - HW 2

Name: _____ Student ID: _____

1 November 2022

Overview / Instructions

This homework will be *due on 21 November 2023 by 11:55 PM* via Moodle.

You are required to submit 1) original R Markdown file and 2) a knitted HTML or PDF file. Please provide comments for R code wherever you see appropriate. Nice formatting of the assignment will have extra points.

In general, be as concise as possible while giving a fully complete answer. All necessary data are available in Moodle.

Remember that the Class Policy strictly applies to homework. You are encouraged to work in groups and discuss with fellow students. However, each student has to know how to answer the questions on her/his own. Note that the final exam is individual based.

Question 0

Review the lectures.

Question 1: Default Data from ISLR

Q1.1

Fit a logistic regression with **student** as the X variable and **default** as the response variable. Interpret the coefficients and discuss whether the X variable is significant.

Q1.2

For the above logistic regression, one can actually obtain explicit expression of the maximum likelihood estimates of the coefficients. Please do the following

i. Write down the logistic regression model by coding **student** using one dummy variable: **0 for students and 1 for non-students**.

ii. Write down the corresponding likelihood function.

iii. Obtain expressions for the coefficient estimates, and compare them with the answers in Q1.1.

Q1.3

i. Consider all the variables and obtain the final logistic regression model.

ii. Compare the ROC curves for the model in Q1.1 and Q1.3, along with the corresponding AUC.

iii. Consider a threshold of 0.5 on the probability of default. Calculate the corresponding specificity, sensitivity, false positive rate, true positive rate.

Question 2: Lost Sales

In many industries throughout the world, suppliers compete for business by submitting quotes for work, services or products. A key criterion used to determine the winning quote is the dollar amount of the quote, but other factors include expected quality, estimated delivery time of the product, or quoted completion time of the work.

The focus of this case is a supplier of equipment to the automotive industry. The products of interest in this case are various precision metal components used in a range of automotive applications, such as braking systems, drive trains, and engines. Some of the products will be used in the manufacture or assembly of new automobiles (i.e. original equipment), while others will be used as replacement parts in automobiles already on the road (i.e. aftermarket).

The supplier wants to increase sales and expand its market position. Many of the quotes provided to prospective customers in the past haven't resulted in orders. Do the data provide any indication why? Are there certain situations that make it more or less likely that a customer will place an order?

Please fit a model using the available data to explore these questions. Based on the fitted model, please provide your answers to the above questions. Drop insignificant variables for variable selection if you like.

The data set contains 550 records for quotes provided over a six month period. The variables in the data set are:

Quote: The quoted price, in dollars, for the order

Time to Delivery: The quoted number of calendar days within which the order is to be delivered

Part Type: OE = original equipment; AM = aftermarket

Status: Whether the quote resulted in a subsequent order within 30 days of receiving the quote: Lost = the order was not placed; Won = the order was placed.

Question 3: Wine Quality

The wine quality data contain information on quality ratings for 6,497 different wines, along with measures of wine properties. The response variable is **Quality**. Please build a model to predict wine quality.

Conduct exploratory data analysis and see if there are redundant or irrelevant information in the data. If so, remove them. Then consider two possible modelling choices:

i. Multinomial logistic regression

ii. Ordinal logistic regression

Write up a summary about characteristics of **Good** quality wine for each model. Do you roughly get similar conclusions using those two models?