Group Assignment 2

Write the name of all your group members at the top of the report.

Attach all pieces of the code to the respective part of the question.

Whenever you run a logistic regression or a similar analysis, take a screenshot of the R output and report it in the respective part of the question.

In this assignment we are tasked with analyzing customer level data with the logistic regression method. The focal company is in the business of selling beverages to different businesses where customers are mainly restaurants and bars. The company has launched a new beverage product and wants to promote the new beverage to their established customers. To do so, the marketing team has decided to run a campaign of "sending customers a small free sample" with the hope that they might like the new product and hence put an order for the product. To understand whether such a promotional strategy is profitable, the company has decided to run a test campaign where they randomly selected 40,000 customers to send them the free sample. The dataset contains information on whether a customer is a buyer (i.e., ordering the new product within a month after running the free sample campaign), store characteristics, and past purchase behavior.

The cost of sending the free sample to everyone in the dataset consists of a $11 cost of the product (sample) and a $19 cost of shipping the product on average. If the customer decides to put an order to buy the new product (i.e., buyer = 1), then the company can sell a case of the new beverage for $710, but the company has to incur two different costs: a cost of production which is $110 and a shipping cost of $40. The company's total number of customers (excluding the 40,000 used in the test campaign) is 220,000. The company wishes to use the results of the marketing campaign on the sample of 40,000 random customers as a basis to run the free sample campaign in the future for the sample of 220,00 customers.

Given the information provided, please answer the following questions:

1) Run a logistic regression of "buyer" variable on all variables in the dataset. Interpret all results using odds ratio. You should provide a marketing perspective when interpreting results. Based on the coefficients obtained for beverage 1, ..., beverage 6, make an argument of the similarity of the buying behavior of the new product with existing beverages offered by the company.

2) Use the predicted probabilities according to your logit model to create deciles (quintiles of order 10). Then calculate and report the average response rate within each decile. Based on the created 10 quintiles, also plot a Gain graph. In other words, look at each created decile to record how many customers and how many buyers there are and use that information to plot the Gain graph. You are allowed to use Excel to create a Gain graph.
3) Using the cost, as well as the profit margin if the customer buys the product - explained in the second paragraph, calculate break-even rate. Then make targeting decisions based on whether the response probability of the logistic regression quintiles is above or below the break-even rate. Calculate the profit for the case of the entire 220,000 customer base.
4) Now instead of break-even rate to dictate our targeting behavior, we allow for logistic regression decision-making based on threshold. In other words, the customer is classified as a buyer if the predicted probability of logistic regression is above a threshold. Assuming that the cost of making mistakes (cost of making a false positive and a false negative classification) is the same, calculate the optimal threshold for the logistic regression. Using the targeting rule that “a customer that is classified as a buyer must be targeted and a buyer classified as a non-buyer should not be targeted”, calculate the profit for the case of the entire 220,000 customer base. **Hint:** To find the optimal threshold, try all values between 0 and 1 with increment of 0.001.
5) Compare the profit obtained in Question 3 and Question 4 to each other. Which one yields a higher profit for the company? Explain a justification for your observation.
6) Using a 5-fold cross validation, calculate the accuracy score of the model. To calculate the accuracy score, use the optimal threshold you obtained in question 4. Compare the accuracy to two models of "everyone is a buyer" and " no one is a buyer". Do you find any counterintuitive results? Comment on whether you think the accuracy of the model is acceptable.
7) Plot the ROC curve of the logistic regression based on the optimal threshold.

Data Dictionary

Variable	Description	
Customer_type	Whether the customer is a bar or restaurant	
Last_purch	How long ago was the last purchase from the company in terms of weeks	
dollars	The amount spent by the company on all products of the company	
Beverage1-6	How many cases of beverage 1, …, beverage 6 the customer purchased	
buyer	1= the customer put an order, 0= customer did not buy	