MSIN0094 Third Assignment

Due 10 am, 17 Dec

1 Background

Due to its successful operation and marketing management, Amazon has achieved steady growth in its customer base. Yet while sales have grown steadily, profits began falling when the database got larger and when the company increased the number of offers sent to customers. The falling profits have led Tom, Amazon's senior marketing manager, to experiment with different targeted marketing approaches in order to improve Amazon' mailing return on marketing investment (MROI).

Tom began a live market test, involving a random small-sized sample of customers from the existing customer base. In the live test, an marketing promotion offer for the Amazon Prime membership is sent to the sampled customers via Royal Mail, and then the sampled customers' responses, either subscribe or not subscribe, are recorded and used to calibrate a predictive model for the current offering. The response model's results are then used to "score" the remaining customers in the database and select customers from the full customer database for the 'rollout' phase.

The current dataset includes the following variables for 10,000 sampled customers who received the mail offer.

- `user_id`: customer ID
- `gender`: gender
- `first`: number of days since the customer's first purchase
- `last`: number of days since the customer's last purchase
- `electronics`: customer's spending in the electronics product category
- `nonelectronics`: customer's spending in the non-electronics category
- `home`, `sports`, `clothes`, `health`, `books`, `digital`, `toys`: the number of purchases in each of these categories
- `subscribe`: whether the customer subscribed to Amazon Prime
- `city`: the city the customer resides in.

And cost information to assess the return on investment for different targeted marketing strategies:

- Cost of mailing the offer: £2
- Average revenue of goods purchased by new subscribers: £40.00
- Average COGS: 50%

- Average shipping costs: £4.00[1]

2 RFM Analysis (38 pts)

Tom's first attempt is to use an RFM analysis. Marketing managers have used this approach to predict customer behavior for more than 50 years. The approach is intuitive, easy to implement, and produced significant improvements in response rates and profits compared with blanket mailing to Amazon' full database.

1. Within `data_full`, use `dplyr` package: (**6pts** in total)
   a. From existing variables in the data, generate three variables named `recency`, `frequency`, and `monetaryvalue`, that represent the recency, frequency, and monetary value in an RFM analysis. (1pts for correct code for each; 3 pts in total)
   b. report the means of `recency`, `frequency`, and `monetaryvalue`(1pts for each correct mean; 3pts).
2. Based on the generated `recency`, `frequency`, and `monetaryvalue`, categorize customers into RFM groups as follows: (**4pts** in total)
   a. Label customers into 4 Recency groups, with 1 being the most recent customer and 4 being the least recent customer (1pts for correct code)
   b. Within each Recency group, further label customers into 4 Frequency groups, with 1 being the most frequent customer and 4 being the least frequent customer (1pts for correct code)
   c. Within each RF group, further label customers into 4 Monetary Value groups, with 1 being the most valuable customer and 4 being the least valuable customer (1pts for correct code)
   d. Categorize customers into RFM groups (hints: there should be 4 * 4 * 4 = 64 RFM groups); use `RFM_group` to denote the RFM group ID, with a smaller ID representing better customers (1pts for correct code).
3. Based on the `RFM_group` in question 2: (**6pts** in total)
   a. Generate a variable named `avg_response_rate` in `data_full` that computes the average response rate for customers in each RFM group. (2pts for correct code).
      - *Tips:* note that `subscribe` is a string variable in `data_full` and has to be transformed into a numeric variable before it can be used.
   b. **Use R** to answer which RFM group has the highest average response rate? Which RFM group has the lowest average response rate? (1pts for correct codes; 1pts for correct answers)
   c. Discuss if a smaller `RFM_group` ID always leads to a higher average response rate? (2pts for correct discussion)

[1]Note that Amazon Prime members enjoy free shipping, so this shipping cost is for Amazon; COGS does not include shipping costs.

4. Compute in R the break-even response rate based on the cost structure given; assign the value of the break-even response rate into a variable named `breakeven_response_rate`. (**2pts** in total)

5. To do targeted marketing, Tom should only make marketing offers to the groups whose `avg_response_rate` is larger than the break-even response rate (**4pts** in total)
   a. Generate a variable named `is_target_RFM`, which equals 1 if Tom should target this customer and 0 otherwise based on RFM analysis. (2pts)
   b. How many customers are targeted? (2pts)

6. Use R to compute and report the return on investment (ROI) (**6pts** in total)
   a. for the blanket marketing campaign, i.e., offers are sent to every customers in the `data_full`. (1pts for correct codes; 1pts for correct final ROI)
   b. for the targeted customers only based on RFM analysis. (1pts for correct codes; 1pts for correct final ROI)
   c. Discuss whether Tom should go with the RFM targeted marketing? (2pts)

7. Rerun the RFM analysis now with **10 groups** in **each** R, F, and M group (so 1000 RFM groups in the end), and compute the new return on investment for the new RFM analysis. (**10pts** in total)
   a. Report the new ROI (2pts for correct codes and ROI)
   b. Compare the two ROIs from the two RFM analyses, discuss which RFM analysis has a better ROI and why there is a difference? (4 pts)
   c. Comment on the statement: "In an RFM analysis, we should have many quantile groups in each R, F, M group as possible so as to increase the effectiveness of our targeting." (4pts)

3 LPM and Logistic Regression (45 pts)

"Fortunately, I took enough time to learn the RFM analysis instead of watching Squid Game with my flatmates during my MSc BA studies." Tom was staring at the RFM analysis results, immersed in a huge sense of achievement.

Suddenly, his phone rang; it was a call from his previous module leader of Marketing Analytics, Wayne. They have become close buddies since the module was finished. "Tom, let's play some Fortnite at my place tonight."

"Sorry Wayne, but I have an Amazon dataset to analyze tonight, and I can resist any temptation."

"Tom, listen, I've got a brand new PS5 from Black Friday, and you know how hard it is to get this big boy these days."

"Hmmm, Wayne, but I still have lots of . . . "

"I have ordered your favorite QQ Style Milk Tea from T4 Bubble Tea; I will need to drink it for you if you don't . . . "

"Hang on, on my way!"

After a few rounds of Fortnite, Tom told Wayne that, despite the initial success with RFM analysis, he was eager to evaluate the effectiveness of alternate approaches. Amazon carries products in different categories including home, sports, clothes, health, books, digital, and toys; the number of previous purchases in each category is recorded in each customer's record in the database. However, he remembered what Wayne had taught in class: RFM analysis normally does not use this or other customer information such as gender, therefore, Tom was a bit clueless whether a more sophisticated modelling approach could yield superior results to the RFM approach.

"Well, Tom, hope you didn't fall asleep in my logistic regression lecture, did you? Logistic Regression offers a powerful method for modelling response. Logistic regression is similar to linear regression - the key difference is that the dependent variable is binary; for example, purchase or no purchase, rather than continuous. For each customer in the dataset, logistic regression predicts a probability, between 0 and 1, of purchase or response, which can be used for targeting and prediction decisions. Like linear regression, it can accommodate both continuous and categorical predictors, including interaction terms. Its use in database marketing has grown as customer database becomes more readily available and as familiarity with the approach grows."

When Tom arrived home that night, he kept thinking over Wayne's words; he wanted to know whether logistic regression, an analyst-driven model, would beat RFM analysis, a simple heuristic model. Unfortunately, though Tom didn't admit it at Wayne's place, he did fall asleep in that class when Wayne started to talk about the mathematics behind the logistic regression. Despite the regret, Tom is eager to assess the potential value of logistic regression as a method for predicting customer response, so he is texting you, a rising future star of business analytics from the current UCL BA program, to complete the following analyses, with a promise of a considerable consultation fee.

8. Discuss and compare the advantage and disadvantages of RFM analysis, linear probability model, and logistic regression in terms of conducting targeted marketing. (4pts for each model; **12pts** in total)

9. Complete the following code block to split the `data_full` into a training set that accounts for 70% of total data, and a test set that accounts for the remaining 30% of data. (Please do not modify the seed, or you will get different results) (**4pts** in total)

10. Train a linear probability model with the predictors including `gender`, `last`, `electronics`, `nonelectronics`, `home`, `sports`, `clothes`, `health`, `books`, `digital`, `toys` (**5pts** in total)

    → factor!

    11. Train a LPM model (for gender, please treat **male customers** as the baseline group) on the **training set**; generate a variable `predicted_prob_LPM` in the **test set** that predicts the probabilities of subscribing to Amazon Prime for customers in the **test set** (2pts for correct codes)

    12. Which customer in the **test set** has the highest predicted probability of subscribing from LPM? (1pts for correct answer)

- **Tips:** you can use `which.max()` function to quickly find out the row index of the max value.

i. Interpret the coefficient of gender for the LPM model (2pts)

11. If Tom only makes marketing offers to the groups whose `predicted_prob_LPM` is larger than the break-even response rate. (**6pts** in total)

   i. Generate a variable named `is_target_LPM` in the test set, which equals 1 if Tom should target this customer and 0 otherwise based on LPM analysis. Report the number of targeted customers (1pts for correct codes; 1pts for correct count of customers targeted)

   ii. Compute and report the return on investment for targeted marketing using LPM on the **test set** (2pts for correct codes; 2pts for correct results)

12. Train a logistic regression model with the predictors including `gender`, `last`, `electronics`, `nonelectronics`, `home`, `sports`, `clothes`, `health`, `books`, `digital`, `toys` (**7pts** in total)

   i. Train a logistic regression model (for gender, please treat **male customers** as the baseline group) on the **training set**; generate a variable `predicted_prob_logistic` in the **test set** that predicts the probabilities of subscribing to Amazon Prime from logistic regression for customers in the test set (2pts for correct codes)

   ii. Which customer in the test set has the highest predicted probability of subscribing based on logistic regression model? (1pts for correct answer)

   iii. Interpret the coefficient of `gender` for the logistic regression model (2pts)

   iv. Interpret the coefficient of `last` for the logistic regression model (2pts)

13. If Tom only makes marketing offers to the groups whose `predicted_prob_logistic` is larger than the break-even response rate. (**6pts** in total)

   i. Generate a variable named `is_target_logistic` in the test set, which equals 1 if Tom should target this customer and 0 otherwise based on the logistic regression. Report the number of targeted customers (1pts for correct codes; 1pts for correct count of customers targeted)

   ii. Compute and report the return on investment for targeted marketing using logistic regression (2pts for correct codes; 2pts for correct results)

14. Compare and comment on the the ROI between LPM and Logistic regression. Explain the difference. (**4pts** in total)

4 Customer Relationship Management

Thanks to your kind help, Tom succeeded in the data analytics job. As a result, Tom was more than pleased to have offered you a final-round interview opportunity for Amazon's Data Analyst Position based in London, in the hope that you can forget about the consultation fee he owed you.

Today is the day. Your two opponents sit right beside you, both looking ambitious for the position. Peeking at their CVs, you notice that they both just graduated from the MSc Business Analytics program at Imperial College London. No wonder Tom whispered "make UCL proud" in your ears and smirked at you before you entered the interview room.

"My name is Tom, the senior marketing manager and, of course, the future CMO of Amazon UK, which is just a matter of time. Now, it is only a few weeks before Christmas holidays, our data analytics team would like to use predictive analytics to better manage the customer relationship management during this gold period, especially for customer development and customer churn management. Could you offer your insights into this please?"

"I just read from the latest news that, John Lewis has offered free-shipping, price discount voucher, and interest-free installment plan to all customers during Black Friday, and they saw a considerable boost in sales by 23%. If we adopt the same strategy, we will definitely boost our sales by 23% as well, without the need to do any targeted marketing." Says Harry, the first job candidate.

Ron, the second job candidate, interrupts Tom, "I believe we still need to conduct targeted marketing instead of sending offers to everyone. In the `data_full` you just showed us, we can train predictive models based on all customers in the dataset, check the performance of our models, and pick the best one."

Tom, Harry, and Ron now look back at you, waiting for your input.

15. Analyze and point out at least two vulnerabilities in Harry's and Ron's statements. (2 pts for each point; **4 points** in total)
16. Discuss the two fundamental tradeoffs in predictive analytics and how the tradeoffs may affect how we should build the CRM models. (**8 points** in total)
17. Discuss the steps to build a next product to buy model. (**5 points** in total)