Vinpearl Recruit Challenge Report

Candidate: Truong Si Thi Vu

I.   Problem understanding

From the description, we can see that the manager of the Rainbow Store wants to understand their buyers better and get a list of customers who are likely to make at least 1 purchase in a given month. The data provided for this problem contains detailed transaction information from five consecutive months in 2018 (February to June). My goal is to use the data to build a model that can make said predictions with high precision, specifically, given a certain month, use the transaction data from the previous two months to indicate customers that will make at least 1 purchase in that month. Additionally, any insight that can be extracted from the dataset is valuable to the manager as well.

For this type of problem, we would usually want to find buying patterns of customers and divide them into groups. Generally, customers who regularly make purchases or spend large amounts of money on items will likely be the people who come back to the store.

Input: A month and transaction data from the previous two months

Output: A list of customers who will make at least 1 purchase in that month

II.   Methodology

One way to approach this problem would be to use a market research technique called RFM to analyze customers. RFM stands for Recency, Frequency, and Monetary.

- Recency: Indicates how recently did a customer buy from the store
- Frequency: Indicates how regularly a customer buys from the store
- Monetary: Indicates how much a customer spends on items from the store

With these 3 features, we can find the group of customers that purchase recently, frequently, and spend lavishly, and from that, figure out which customers will come back in the targeted month.

In general, there are different variants of the RFM method where the segmentation part differs slightly. For this project, the RFM scores range from 1 to 5, with 5 being the highest score for a dimension.

After extracting Recency, Frequency, and Monetary from the customers, we will use the data to label them with "Yes" and "No" for the question "Did they make a purchase in the following month?". From there, we have a binary classification problem with 3 features and we can use machine learning models to make predictions.

III. Workflow

Problem understanding: In this step, I carefully read the description to understand the tasks assigned and to think about the techniques that can be used to solve the problem.

Explanatory Data Analysis (EDA): Since a dataset was already provided, no data collection was required, instead, I analyzed the data by summarizing statistics and making visualizations. This step is essential for understanding the dataset and information gathered from EDA will be valuable for later steps such as data processing.

Data processing: In data processing, I made necessary changes to the data structure as well as extracted the required features from the dataset to prepare it for modeling.

RFM analysis: This step provides information on customer segmentation, which I think is useful if the owner wants to understand their customers more thoroughly.

Predictive modeling: In this part, a few machine learning models were selected to train the data, and after that, results were recorded and compared. One model was chosen to be the main subject.

Inference: Finally, I tested the model on an existing month and a new one to see the results.

IV. Problem discussion

You're asked which customers to send promotional e-mails to next month, based on your model. What is your recommendation?

Based on the trained model, I would recommend that the owner send promotional e-mails to the customers who were predicted by the model to come back the following month. Additionally, we can expand this list by including customers with high RFM ranks, because, in my opinion, they are still likely to make purchases and it is not cost-ineffective to send extra emails with the chance of acquiring more buyers.

What problems do you see in applying the model in this way?

Sending e-mails to predicted customers may have certain limitations. For instance, the owner may be overconfident in thinking that customers who are expected to buy again will want these promotional emails. What if these so-called loyal customers, in fact, do not need e-mails for them to purchase any item from the store in the first place? Conversely, customers who are predicted to be "inactive" may need more attention from the store and hence, they should be the people receiving promotional e-mails. Furthermore, even though the results from past data are promising, there is no guarantee that in the future, the model will continue to be accurate. Our world changes every day and it is unwise to make critical business decisions based on only one model. Hence, in order to make the most out of the model, the owner may need to carefully weigh different options on how to use the insight and consider the pros and cons of every scenario.