Candidate Name:

Please note the following:

1) Both questions use the provided datasets
2) All source code must be provided with comments along with instructions on how to execute the code.
3) R is preferred, but you are free to consider other tools. Regardless of the tool you use, source code / workings must be provided and clearly documented
4) Two datasets are provided **spenddata.csv** and **testdata.csv.** You are free to decide how best to use them. The labels are masked for confidentiality reasons.
5) Refer to **mock survey data 3.xls** for Question 2. You can refer to the sheet "dictionary" for a description of the fields to aid in your analysis.
6) You do not need to augment this dataset with any external data. However, if you choose to do so, you must document the reasons.

Question 1

Dataset: spenddata.csv and testdata.csv

You are given a set of survey data which captures spend amount among other data points. Some of the respondents have been tagged as belonging to group 1 – 6. However, due to a data calculation issue, some of the respondents have had their groups (**pov6**) missing.

Build a model that will classify these respondents back into one of the 6 groups.

a. Please explain the choice of metric / evaluation criterion used

As pov6 has 6 categories and imbalance samples between classes, I have chosen f1 score as the evaluation metric, because f1 score can capture the miss classification for minority class much better than accuracy.

b. What are the assumptions you made when building this model?

Assumption 1: we assume it's customer index as this variable looks like customer index + year, and will drop it from the data frame.

Assumption 2: To apply modelling, we assume the training data and test data are i.i.d, and there are strong correlation between predictors and target variable.

c. What were the approaches you considered? Please explain the reason for the technique / approach used as well as the pros and cons.

To classify the pov6, I have trained a XGBoost model for classification. The pros for this model is that it can handle missing values very well, which fits to this problem. And it has very good performance on wide tabular data.

The cons of this model is the explainability compared to Multiple-linear regression. Therefore I have applied SHAP framework to interpret its predictions.

d. Please explain under what conditions will the model you choose **be not appropriate**

A few cases where XGBoost will not perform well:

1. Small sample size dataset
2. Unstructured data like Image or text
3. High cardinality data

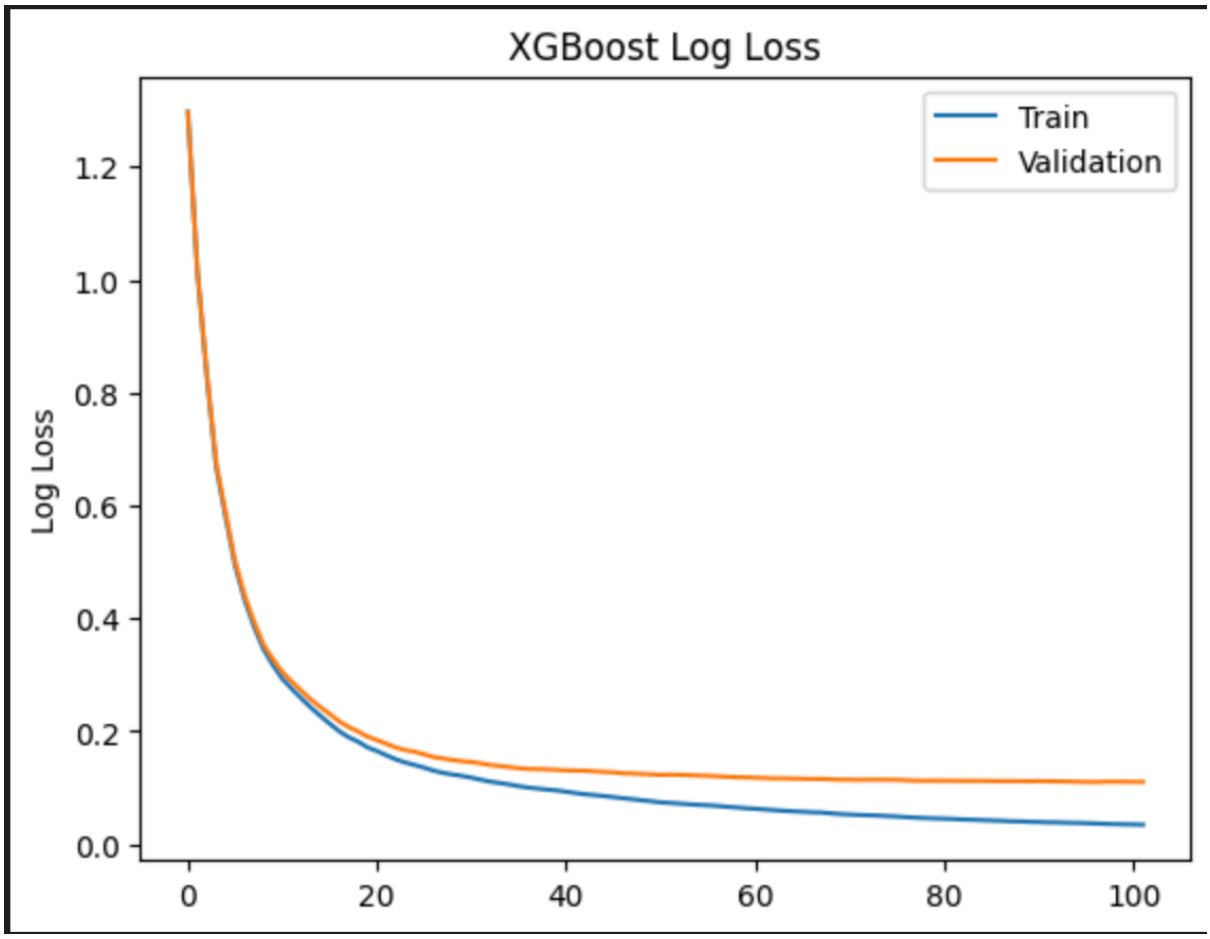
e. How confident are you of the model's robustness and how would you explain the model's performance?

In order to ensure the robustness and accuracy, I have split the dataset into three sets:

1. Train: used to train model
2. Validation: used for cross validation to tune hyper-parameters, and evaluate f1 score
3. Test: used for evaluating model performance

The model's f1 score is above 0.96 on test set and eval set, which says the model is fitted well and can predict with high precision and recall.

In addition, I have plotted the training curve, which shows good fit to the training and evaluation dataset, and has minimum overfit and under fit problem(see blow).



f. Why is your model performing well / not well?

The model performs quite good, the reasons are:

1. The first reason is fine tuning and correct data cleaning works.

2. The second reason is there exists strong correlation between predictors and target variable.

3. The given data is quite clean with no much noises.

Was any feature engineering required? If yes, what were they. If no, why?

In this project, I have only done simple feature engineering like imputing missing values, encode categorical values. As we do not know the meaning of each column, we are not able to perform finer engineering works for each column, e.g. derive the weekdays based on date.

Question 2

Using the mock survey data (**mock survey data 3.xlsx**), answer the following questions.

1) Based on the data, answer the following business questions:

    a. What can we learn about our visitors from the survey data that will help us better understand them?

Based on the visitors information, and the total spend, I have built a XGBoost model and correlation matrix to check their correlation and feature importance. By doing this, I can evaluate the important features determining the total spend, as well as drawing some actionable insights to attract visitors to spend more during their visit to SG.

    b. Is there a correlation between travel companions and choice of hotel?

Based on Phik Correlation Matrix, we can see there is a 0.4 score between MainHotel and Travel companion, so we conclude they are weakly correlated, which can be interpreted as visitors with companion could consider to choose certain type of hotel, but companion is not the dominating factor for hotel choice.

    c. With your findings from (a - b), what other insights can be derived from analysing business and leisure visitors and how will your insights help STB attract visitors to spend in Singapore. **Prepare a short PowerPoint presentation to share this**. *Hints:*

        i. *The type of visitors can be found through purpose of visit)?*

        ii. *You are encouraged to explore the data **beyond** the questions asked in (a) – (b)*

        iii. *Use appropriate charts / visuals to communicate your story*

        iv. *Note the chart-ink ratio*