AIAP® Batch 15 Technical Assessment

Deadline: **1900 hrs, 18th September 2023**

Tasks

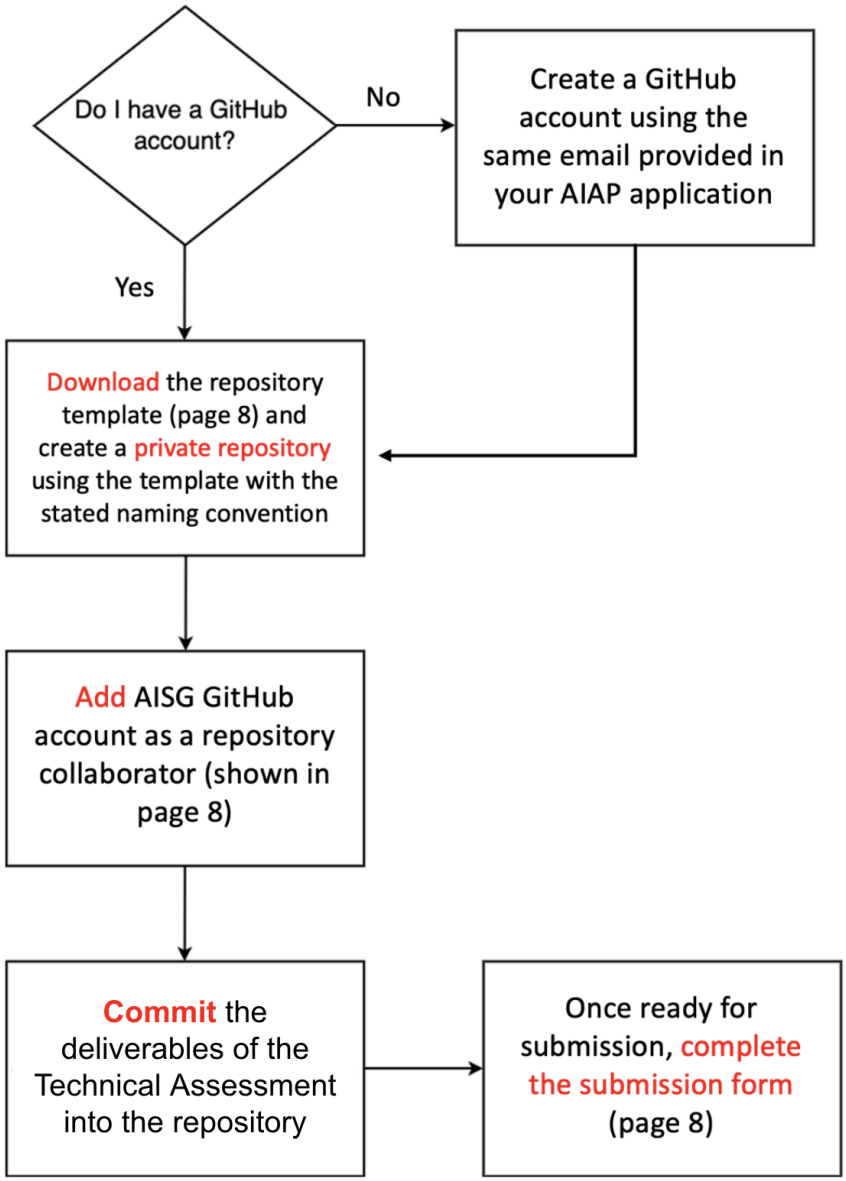
This assessment consists of two parts:

1. **Exploratory Data Analysis in Jupyter Notebook**
2. **End-to-end Machine Learning Pipeline in Python Scripts (`.py`)**

AI Singapore
innovation 4.0, 3 Research Link #02-04 Singapore 117602 | www.aisingapore.org

Technical Assessment Overview

There are two parts to the Technical Assessment: Exploratory Data Analysis and End-to-end Machine Learning Pipeline. You are to attempt both parts and submit the deliverables by uploading them to your own private GitHub repository. The following flowchart outlines the major steps for the Technical Assessment. Details will be provided in the subsequent sections of this document.



Task 1 - Exploratory Data Analysis (EDA)

Using the dataset specified in the **Data** section, conduct an EDA and create an interactive notebook (.ipynb file) in **Python** that can be used as a presentation to explain the findings of your analysis. It should contain appropriate visualizations and explanations to assist readers in understanding how these elaborations are arrived at and their implications.

Deliverable

1. Jupyter Notebook in **Python**: a `.ipynb` file named **`eda.ipynb`**. (do adhere to the naming requirement)

Evaluation

In the submitted notebook, you are required to

1. Outline the steps taken in the EDA process
2. Explain the purpose of each step
3. Explain the conclusions drawn from each step
4. Explain the interpretation of the various statistics generated and how they impact your analysis
5. Generate clear, meaningful, and understandable visualizations that support your findings
6. Organise the notebook so that is it clear and easy to understand

Please note that your submission will be heavily penalised for any of the following conditions:

1. `.ipynb` missing in the submitted repository
2. `.ipynb` cannot be opened on Jupyter Notebook
3. Explanations missing or unclear in the submitted Jupyter Notebook

Task 2: End-to-end Machine Learning Pipeline (MLP)

Design and create a machine learning pipeline in Python scripts (`.py` files) that will ingest and process the entailed dataset, subsequently, feeding it into the machine learning algorithm(s) of your choice.

Do not develop your MLP in an interactive notebook.

The pipeline should be easily configurable to enable easy experimentation of different algorithms and parameters as well as ways of processing data. You can consider the usage of a config file, environment variables, or command line parameters.

Within the pipeline, data (provided in the Data section, Page 6) must be fetched/imported using SQLite, or any similar packages.

Deliverables

1. A folder named `**src**` containing Python modules/classes in `.py` format.
2. An executable bash script `**run.sh**` at the base folder of your submission to run the aforementioned modules/classes/scripts. DO NOT install your dependencies in the `run.sh`; this will be taken care of automatically when we assess the assignment if you have created your `requirements.txt` correctly.
3. A `**requirements.txt**` file in the base folder of your submission.
4. A `**README.md**` file that sufficiently explains the pipeline design and its usage. You are required to explain the thought process behind your submitted pipeline in the README. The README is expected to contain the following:
    a. Full name (as in NRIC) and email address (stated in your application form).
    b. Overview of the submitted folder and the folder structure.
    c. Instructions for executing the pipeline and modifying any parameters.
    d. Description of logical steps/flow of the pipeline. If you find it useful, please feel free to include suitable visualisation aids (eg, flow charts) within the README.
    e. Overview of key findings from the EDA conducted in Task 1 and the choices made in the pipeline based on these findings, particularly any feature engineering. Please keep the details of the EDA in the `.ipynb`. The information in the `README.md` should be a quick summary of the details from `.ipynb`.
    f. Described how the features in the dataset are processed (summarised in a table)
    g. Explanation of your choice of models for each machine learning task.
    h. Evaluation of the models developed. Any metrics used in the evaluation should also be explained.
    i. Other considerations for deploying the models developed.

Evaluation

The submitted MLP, including the `README.md`, will be used to assess your understanding of machine learning models/algorithms as well as your ability to design and develop a machine learning pipeline. Specifically, you will be assessed on

1. Appropriate data preprocessing and feature engineering
2. Appropriate use and optimization of algorithms/models
3. Appropriate explanation for the choice of algorithms/models
4. Appropriate use of evaluation metrics
5. Appropriate explanation for the choice of evaluation metrics
6. Understanding of the different components in the machine learning pipeline

In your submitted Python scripts (`.py` files), you will be assessed on the quality of your code in terms of reusability, readability, and self-explanatory.

Please note that your submission will be penalised for any of the following conditions:

1. Incorrect format for `requirements.txt`
2. `run.sh` fails upon execution
3. Poorly structured `README.md`
4. Disorganised code that fails to make use of functions and/or classes for reusability
5. MLP not submitted in Python scripts (`.py` files), including MLP built using Jupyter Notebooks.

Note for Windows users

DO NOT submit a Windows batch (`*.bat`) script in replacement of the bash script. Use either 'Windows Subsystem for Linux (WSL)' or 'Git Bash'/'cygwin' for the creation of the bash script.

Problem Statement

Objectives

ShipSail is a prestigious cruise company that aims to provide our customers with the best experiences possible, vying to be the top choice for travellers worldwide. Our company is dedicated to constantly improving our service, tailoring offerings to match guests' needs and tastes, hence ensuring an unforgettable cruising experience. As such, we create an interaction-rich ecosystem where they cherish customer involvement at every touchpoint of their journey synergizing the offline and online experiences.

In our quest to elevate the guest experience and meet evolving demands, our company regularly undertakes pre-purchase surveys on our website, incentivising future customers with attractive vouchers and upgrades. The survey requires potential guests to rate their preferences on a range of indicators critical in ensuring a memorable cruise journey - "Onboard Wifi Service", "Embarkation/Disembarkation time convenient", "Ease of Online booking", "Gate location", "Onboard Dining Service", "Online Check-in", "Cabin Comfort", "Onboard Entertainment", "Cabin service", "Baggage handling", "Port Check-in Service", "Onboard Service" as well as "Cleanliness". These preferences provide ShipSail with comprehensive insights into what our potential guests value the most, ensuring we meticulously tailor offerings to guest desires.

Simultaneously, after concluding each journey, ShipSail collects post-trip data such as "Cruise Name", "Cruise Distance", "WiFi", "Dining", "Entertainment" travelled to cross-reference and contextualise guest preferences along with the realities of their chosen itineraries. This information shapes the foundations for our company's data repository further enriching our understanding of guest preferences and patterns. It proves to be invaluable as it empowers our company with insights necessary to formulate efficient and compelling marketing strategies, amplifying our appeal in the cruising market.

As the newly hired AI Engineer, now you are entrusted with an ambitious project. Harnessing the collective power of the pre-purchase and post-trip data, you are to predict the type of tickets potential customers are most likely to purchase. By predicting guests' preferred ticket type, the company aims to customise the experiences and amenities, masterfully aligning them with the guests' comfort and preferences to maximise potential revenue.

In your submission, you are expected to **evaluate at least three suitable models for predicting the ticket type that potential customers will purchase**, enriching ShipSail's marketing strategy and decisively enhancing guests' happiness.

Dataset

The datasets provided contain information that ShipSail has collected for pre-purchase and post-trip. Do note that there could be synthetic features in the dataset. Therefore you would need to state and verify any assumptions that you make.

You can query the datasets using the following URLs.

Pre-purchase data:
https://techassessment.blob.core.windows.net/aiap15-assessment-data/cruise_pre.db

Post-trip data:
https://techassessment.blob.core.windows.net/aiap15-assessment-data/cruise_post.db

Instructions for setting up SQLite and querying the database

The datasets can be accessed through the `cruise_pre.db` and `cruise_post.db` file. You may find either of the following packages, `SQLite` or `SQLAlchemy`, useful for accessing this database.

You should place the `cruise_pre.db` and `cruise_post.db` file in a `data` folder. Your machine learning pipeline should retrieve the dataset using the relative path `data/cruise_pre.db` and `data/cruise_post.db`.

DO NOT upload the `cruise_pre.db` and `cruise_post.db` file onto your GitHub repository.

List of Attributes

Attribute (Pre-Purchase)	Description
Gender	The gender of the cruise passenger.
Date of Birth	The date of birth of the passenger.
Source of Traffic	The channel via which the passenger heard about or booked the cruise.
Onboard Wifi Service	Importance scale of the onboard Wifi service for the passenger. Reference scale: 1: 'Not at all important', 2: 'A little important', 3: 'Somewhat important', 4: 'Very important', 5: 'Extremely important'
Embarkation/Disembarkation time convenient	Importance scale of the embarkation/disembarkation time convenience for the passenger. Reference scale same as above.
Ease of Online booking	Importance scale of the ease of online booking for the passenger. Reference scale same as above.
Gate location	Importance scale of the gate location for the passenger. Reference scale same as above.
Logging	Timestamp of when the passenger's information was logged.
Onboard Dining Service	Importance scale of the onboard dining service for the passenger. Reference scale same as above.
Online Check-in	Importance scale of the online check-in process for the passenger. Reference scale same as above.
Cabin Comfort	Importance scale of the cabin comfort for the passenger. Reference scale same as above.
Onboard Entertainment	Importance scale of the onboard entertainment for the passenger. Reference scale same as above.
Cabin Service	Importance scale of the cabin service for the passenger. Reference scale same as above.
Baggage Handling	Importance scale of the baggage handling for the passenger. Reference scale same as above.
Port Check-in Service	Importance scale of the port check-in service for the passenger. Reference scale same as above.

Onboard Service	Importance scale of the onboard service for the passenger. Reference scale same as above.
Cleanliness	Importance scale of the cleanliness for the passenger. Reference scale same as above.
Ext_Intcode	Internal code for the passenger.

Attribute (Post-Trip)	**Description**
Cruise Name	The name of the cruise that passenger took.
Ticket Type	The type of the ticket purchased by the passenger.
Cruise Distance	The total distance covered by the cruise.
Ext_Intcode	Internal code for the passenger.
WiFi	Satisfaction level of WiFi. Legend: 0: 'dissatisfied', 1: 'satisfied', NA: 'Not Applicable'
Dining	Satisfaction level of dining. Legend: 0: 'dissatisfied', 1: 'satisfied', NA: 'Not Applicable'
Entertainment	Satisfaction level of entertainment. Legend: 0: 'dissatisfied', 1: 'satisfied', NA: 'Not Applicable'

Submission Format

Create a GitHub account using the **same** email provided in your AIAP application form.

Download the repository template from:
https://techassessment.blob.core.windows.net/aiap15-assessment-data/aiap15-NAME-NRIC.zip

The downloaded repository template contains a hidden folder: `.github`. The `.github` folder contains scripts to execute your end-to-end machine learning pipeline using GitHub Actions. Specifically, it will first install the required dependencies using your `requirements.txt` and subsequently, execute your bash script (`run.sh`). You can manually trigger the pipeline under Actions in your repository.

Using the downloaded template, create a **private** repository using the following naming convention:

aiap15-<full name (as in NRIC) separated by dashes>-<last 4 characters of NRIC>

For example, `*aiap15-john-lim-der-hui-321A*`

Add the following account as a collaborator in your private repository:

- Username: **AISG-AIAP**
- Email: **aiap-internal@aisingapore.org**

Your repository is to have the following structure:

```
├─── .github
├─── src
│    └─── (python files constituting the end-to-end ML pipeline in .py format)
├─── README.md
├─── eda.ipynb
├─── requirements.txt
└─── run.sh
```

We encourage you to adhere to Git best practices and commit your work to the repository regularly during the assessment period. Once your repository is ready for submission, complete the following form using the following URL: https://forms.gle/BenbaRZ1BMGTfGcBA

NOTE: During the assessment period, you are still allowed to make changes to your repository after submitting the form.