AISG AI Internship Programme (AIIP)

AIIP Batch 2 Technical Assessment

Deadline**: <u>1900hrs, 22nd March 2024</u>**

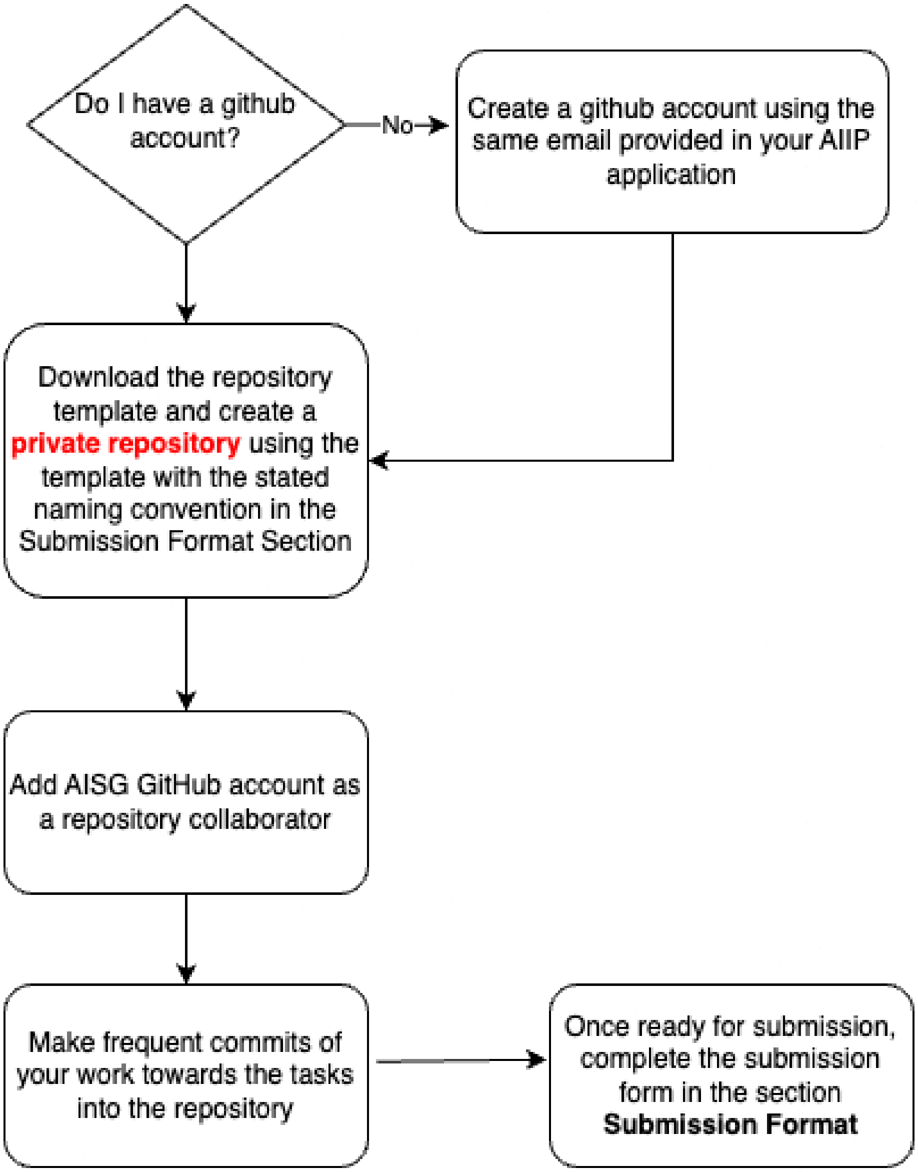
<u>Tasks</u>

This assessment consists of two main parts:

1. Exploratory Data Analysis (EDA)
2. Machine Learning Model Training

Technical Assessment Overview

There are two parts to the Technical Assessment: EDA and MLP. 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 Assessment. Details will be provided in the subsequent sections of this document.


Do I have a github account?
No
Create a github account using the same email provided in your AIIP application
Download the repository template and create a **private repository** using the template with the stated naming convention in the Submission Format Section
Add AISG GitHub account as a repository collaborator
Make frequent commits of your work towards the tasks into the repository
Once ready for submission, complete the submission form in the section **Submission Format**

Task 1 - Exploratory Data Analysis (EDA)

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

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

The goal is to train at least three Machine Learning Models of your choice and do a post-hoc analysis of the performance of the trained models. You should be able to contrast the tradeoff between the complexity of the model and the performances of each of them.

For each of your chosen models, an explanation should be included as to why you chose each model to train. The Python code for training these models should be written in **`models.ipynb`** which takes in the processed output of your EDA and any feature engineering from the first task.

While a notebook is sufficient for this task, applicants should take note of Python coding conventions and clean code practices. A heuristic is that if you were to translate your notebook's code to **modular** Python scripts/functions/classes, it should not take an extraordinary amount of effort. Sectioning your notebook is also recommended to allow for easier reading.

Deliverables

1. A `**requirements.txt**` file at the base folder of your submission.
2. Jupyter Notebook in **<u>Python</u>**: a `.ipynb` file named **`models.ipynb`**. (do adhere to the naming requirement)
3. 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 visualization 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`, this section should be a quick summary.
    f. Describe how the features in the dataset are processed (summarized 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 README will be used to assess your understanding of machine learning models/algorithms and your ability to design and develop a machine learning pipeline. In particular, you will be assessed on

1. Appropriate data preprocessing and feature engineering
2. Appropriate use and optimisation 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 Jupyter Notebooks(.ipynb 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 heavily penalised for any of the following conditions:

1. Incorrect format for `requirements.txt`
2. Poorly structured `README.md`
3. Disorganised code that fails to make use of functions and/or classes for reusability

Problem Statement

Objectives

Your objective is to predict the No-Show of customers (using the dataset provided) to help a hotel chain to formulate policies to reduce expenses incurred due to No-Shows. In your submission, you are to evaluate at least 3 suitable models for estimating the customers' No-Show.

Dataset

The dataset contains the customer records from a hotel chain. Do note that there could be synthetic features in the dataset. Hence, please ensure that you state and verify any assumptions that you make.

You can retrieve the dataset using the following URL:
https://techassessment.blob.core.windows.net/aiip-intake2-assessment-data/noshow.db

Instructions for setting up SQLite and querying the database

The dataset can be accessed through the \`noshow.db\` file. You may find either of the following packages, \`SQLite\` or \`SQLAlchemy\`, useful for accessing this database.

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

DO NOT submit the \`noshow.db\` in your final submission.

List of Attributes

Attribute	Description
booking_id	Unique customer booking ID
no_show	If the customer is a No-Show: 0 = Show, 1 = No-Show
branch	Hotel branch
booking_month	Month the booking was made by the customer
arrival_month	Month the customer plan to arrive at the hotel
arrival_day	Day date the customer plan to arrive at the hotel
checkout_month	Month the customer plan to checkout of the hotel
checkout_day	Day date the customer plan to checkout of the hotel
country	Nationality of the customer
first_time	If it is the first time customer staying in the hotel
room	Room type booked by the customer
price	Price of the room booked by the customer
platform	Platform used to book the room by the customer
num_adults	Number of adults staying
num_children	Number of children staying

Submission Format

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

Create a **private** repository using the following naming convention:

aiip2-<full name (as in NRIC) separated by dashes>-<last 4 characters of NRIC>

For example, `*aiip2-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 (**as an example**):

```
├── src
│   └── (Additional python files)
├── README.md
├── eda.ipynb
├── models.ipynb
├── requirements.txt
```

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/Unygo3iahEqh3YKh8

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