Technical Assessment

Deadline**: 1900hrs, 17 August 2020**

Submissions after the deadline will not be accepted. Please submit the completed assessment early to ensure a smooth submission process.

Tasks

This technical assessment consists of two main parts:

1. Exploratory Data Analysis (EDA)
2. End-to-end Machine Learning Pipeline (MLP)

You are to attempt both parts and package a submission containing deliverables for each of the tasks.

Task 1 - Exploratory Data Analysis (EDA)

Using the dataset specified in the **Data** section below, 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. Notebook in **Python**: An `.ipynb` file named **`eda.ipynb`**.

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

You will be assessed on the usefulness and clarity of visualisations, accuracy and depth of your insights, presentation flow, and structure of your analysis.

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

1. .ipynb missing from submission folder
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

Design and create a simple machine learning pipeline that will ingest/process the entailed dataset and feed it into the machine learning algorithm(s) of your choice.

The pipeline should be easily configurable to enable easy experimentation of different algorithms and parameters as well as different ways of processing data (e.g. usage of a config file, environment variables, or command line parameters). Within the pipeline, data must be fetched/imported using SQLite (provided in the `Data` section).

Deliverables

1. A folder named `src` containing Python modules/classes. Use only Python 3.6.7/3.6.8.
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 at 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.
    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. Explanation of your choice of models for each machine learning task.
    g. Evaluation of the models developed. Any metrics used in the evaluation should also be explained.
    h. Other considerations for deploying the models developed.

Evaluation

The submitted README will be used to assess your understanding of machine learning models / algorithms and ability to design and develop a machine learning pipeline. In particular, you will be assessed on

1. Appropriate use of algorithms/models
2. Appropriate explanation for the choice of algorithms/models
3. Appropriate use of evaluation metrics
4. Appropriate explanation for the choice of evaluation metrics
5. Understanding of the different components in the machine learning pipeline

In your submitted Python scripts, you will be assessed on the quality of your code in terms of clean separation of functionality and ease of use. Code reusability between the two tasks will be viewed favourably.

Please note that your submission will be heavily 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

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 creation of the bash script.

Data

URL

https://techassessment.blob.core.windows.net/aiap7-assessment-data/home_sales.db

Instructions for setting up SQLite and querying the database

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

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

DO NOT submit the `**home_sales.db**` in your final submission.

~~~~
```
├── src
│   ├── module1.py
│   └── module2.py
├── data
│   └── home_sales.db
├── README.md
├── eda.ipynb
├── requirements.txt
└── run.sh
```
~~~~

Dataset description

This dataset contains the sale prices for houses in a county within the USA and information related to the unit sold. The attributes are described in the table below.

Objectives

The objective is to predict housing prices in order to help a real estate company advise their clients on suitable housing prices. In your submission, you are to evaluate **two** different approaches for estimating housing price. For **each approach**, you are required to assess **multiple** different machine learning models and determine the most suitable model.

Approach 1 - Regression

In the first approach, you are to develop a regression model to predict prices. If you find it helpful, you may transform the ***price*** attribute. However, all metrics reported must compare the model's predictions (raw predictions or transformed) against the true house price (***price*** attribute).

Approach 2 - Classification

In this second approach, you are to make use of classification models to estimate the price. Bin the ***price*** attribute into an appropriate number of categories and use them as the targets. You should also consider how to use the models' predictions to advise real estate agents on a suitable housing price.

Overall evaluation

Once you have identified the most suitable model for each approach, evaluate which of these two approaches is more suitable for the task.

List of attributes

Attribute	Description
id	Transaction ID
date	Date of sale
price	House price (USD)
bedrooms	Number of bedrooms
bathrooms	Number of bathrooms
floors	Number of floors / storeys
waterfront	Indication of whether the unit has a waterfront view
view	Internal rating of view from the house
condition	Internal rating of the house condition
review_score	Review scores from an independent organisation
basement_size	Basement area (sqft)
built	Year built
renovation	Year of renovation
zipcode	Postal Code
latitude	Latitude coordinate
longitude	Longitude coordinate
living_room_size	Living room area (sqft)
lot_size	Entire unit area (sqft)

Submission Format

Your work should be uploaded as a `*.zip` archive to AI Singapore's designated blob store (detailed below). The archive file is to be provided with the following naming convention:

`<firstname>_<lastname>_<last 5 characters of NRIC>.zip` e.g. `john_lim_4321A.zip`

The submission folder is to have the following structure (as an example):

~~~
```
├── src
│   ├── module1.py
│   └── module2.py
├── README.md
├── eda.ipynb
├── requirements.txt
└── run.sh
```
~~~

Once you have packaged your submission, you are to upload your submission by following the steps detailed below:

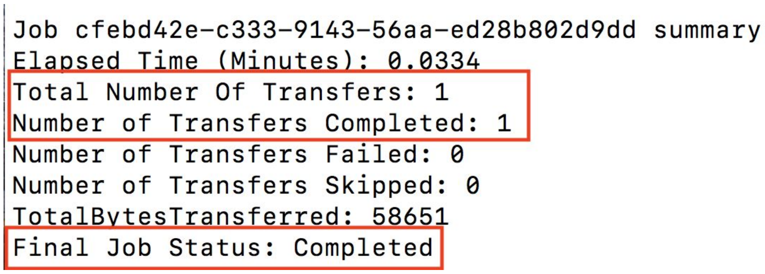
1. Download the Azure **azcopy** tool.

2. Use the following URL and the **azcopy** tool to upload your files through the command line. You are expected to follow the instructions under "**Option 2: Use a SAS token**" to make use of the **azcopy** tool. No other steps are required. You do not need to log into an Azure account or obtain a subscription to use the tool.

   The URL includes the required SAS token. Please ensure that you copy the link correctly and remove any white spaces.

   https://techassessment.blob.core.windows.net/aiap7-assessment-submission?sv=2019-12-12&ss=bfqt&srt=co&sp=rwacx&se=2020-08-17T11:00:00Z&st=2020-08-11T13:33:32Z&spr=https&sig=OAAwaLzqFAE9NaT3Q0BttlIzYNvU4OrRMsHUegMncSA%3D

3. If your file has been successfully uploaded, you should observe an output that is similar to what is shown below:



```
Job cfebd42e-c333-9143-56aa-ed28b802d9dd summary
Elapsed Time (Minutes): 0.0334
Total Number Of Transfers: 1
Number of Transfers Completed: 1
Number of Transfers Failed: 0
Number of Transfers Skipped: 0
TotalBytesTransferred: 58651
Final Job Status: Completed
```

Note: The ability to use this tool is considered as part of the technical assessment and evaluation. There will be automated checks that will assess the conformance of your uploaded submission to aforementioned specified instructions. Non-conformance to specified conventions/formats will negatively impact your overall score.