AI SINGAPORE

AICE (Associate) Technical Test

April 2020

Deadline: 2359hrs, 26 April 2020

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

This technical assessment consists of three parts:

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

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

Main Task

The main task is to predict the *total number of active e-scooter users* given the available attributes. You are required to propose and develop a machine learning pipeline that addresses this problem statement using the given dataset.

1. Data Extraction

Summary:
This dataset provides hourly values for the number of active users for an e-scooter and e-bike rental service in a city. The features include the date, time and various weather parameters.

Attributes:

Attribute	Description
date	Date in YYYY-MM-DD.
hr	Hour (0 to 23).
weather	Description of the weather conditions for that hour.
temperature	Average temperature for that hour (Fahrenheit).
feels-like-temperature	Average feeling temperature for that hour (Fahrenheit).
relative-humidity	Average relative humidity for that hour. Measure of the amount of water in the air (%).
windspeed	Average speed of wind for that hour (arbitrary units).
psi	Pollutant standard index. Measure of pollutants present in the air (0 to 400).
guest-scooter	Number of guest users using the rental e-scooters in that hour.
registered-scooter	Number of registered users using the rental e-scooters in that hour.
guest-bike	Number of guest users using the rental e-bikes in that hour.
registered-bike	Number of registered users using the rental e-bikes in that hour.

Deliverable:
The data is hosted on AI Singapore's databases. In a python script (as part of a data ingestion and preprocessing pipeline), write an SQL query to extract a dataset with the following criteria.

- The dataset should only consist of data recorded between the years 2011 and 2012.
- Extract all columns except *guest-bike* and *registered-bike*.

In your submission, the dataset must be extracted from the database. However, you may save a copy of the dataset on your local machine to work on the subsequent tasks.

Connection details:

server	aice.database.windows.net
database	aice
username	aice_candidate
password	@ic3_a3s0c1at3
table name	rental_data

2. Exploratory Data Analysis (EDA)

Using the dataset specified on page 2, conduct an EDA and create an interactive notebook in Python that can be used as part of a presentation of your findings. The notebook should contain appropriate visualisations and explanations to assist readers in understanding your findings as well as their implications.

Deliverable:
Notebook in Python: an `.ipynb` file named **`eda.ipynb**``

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

3. End-to-end Machine Learning Pipeline (MLP)

Design and create a simple machine learning pipeline that will ingest/process the dataset and feed it into appropriate machine learning algorithm(s), returning suitable metrics as outputs.

Deliverables:

1. A folder named `mlp` containing Python modules/classes.
2. An executable bash script `run.sh` at the base folder of your submission.
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. An explanation of your choice of model(s) and an evaluation of the model(s) developed should also be included in the README.

Pipeline Requirements/Specifics:

- Structured as Python modules/classes with well-defined functions.
- A bash script named `run.sh` 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.
- Relevant training/evaluation metric(s) outputs to be generated upon completion.
- Made easily configurable to enable easy experimentation of different algorithms and parameters. as well as different ways of processing data (e.g. use of a config file, environment variables, or command line parameters).
- For Python, use only versions 3.6.7/3.6.8.
- Within the pipeline, data must be fetched/imported from the database (information provided on pages 2-3).
- DO NOT include the data file in your submission.

Evaluation:

You will be assessed on the quality of your code in terms of clean separation of functionality, creativity, and ease of use. Code reusability between the tasks will be viewed favourably.

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.

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):

```
├── mlp
│   ├── module_script
│   └── module_script
├── README.md
├── eda.ipynb
├── requirements.txt
└── run.sh
```

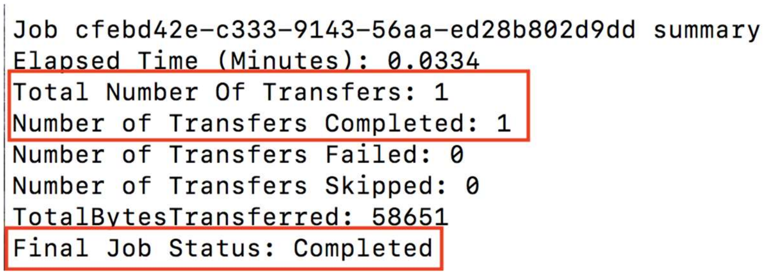
Once you have packaged your submission, you are to upload your submission by following the steps detailed below:

1. Download and configure/install the Azure `azcopy` tool

2. Use the following URL and the `azcopy` tool to upload your files through the command line. The URL includes the required SAS token. Please ensure that you copy the link correctly and remove any white spaces.

   https://aisgaice.blob.core.windows.net/aice-associate?sv=2019-02-02&ss=bfqt&srt=co&sp=rwac&se=2020-04-26T15:59:59Z&st=2020-04-18T09:45:14Z&spr=https&sig=zRGXnR9czev%2FTIhhVrEdkQ9kwoAnZsDIx2%2FJycTxCg4%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 part of the technical assessment and evaluation. There will be automated checks that will assess the conformance of your uploaded submission to the aforementioned instructions. Non-conformance to specified conventions/formats will negatively impact your overall score.