SAMPLE

Associate AI Engineer Technical Test

January 2025

Deadline: [redacted]

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 the following two main tasks:

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

The assignment project background:

Olist is a Brazilian e-commerce marketplace like Lazada, Taobao and Shopee, it is a sales platform that connects small retailers with customers.

The following is a summary of the buying process from a customer perspective. A customer can place a purchase order for multiple items from a vast selection of sellers on Olist. The sales order will then be fulfilled by a logistic partner. The customer will be notified of the estimated delivery date once shipment is confirmed. Upon delivery of the order, the customer will receive a survey to provide feedback.

The datasets provided here are randomly extracted real commercial data from Olist store. The anonymized data consists of order information for about 100k orders from year 2016 to 2018. The datasets include order details such as order status, price, payment, freight information, customer location, product attributes, and customer reviews.

One possible strategic approach to drive profitable sales for an e-commerce platform is to determine who are their high value loyal customers, such that the business can design and deploy more effective and targeted marketing campaigns.

For this assignment, help Olist leverage on data analytics and machine learning, focusing their business objective to build their customer base and increase future sales revenue.

Source:
Brazilian E-Commerce Olist Store

The assignment project objective is:

To identify potential repeat buyers

Please attempt all requirements stated in the following sections and package a submission in zipped file formatting containing the deliverables specified.

1. Data Description

Dataset Download:

Please download the following datasets in csv format:

1. olist_customers_dataset.csv
2. olist_geolocation_dataset.csv
3. olist_order_items_dataset.csv
4. olist_order_payments_dataset.csv
5. olist_order_reviews_dataset.csv
6. olist_orders_dataset.csv
7. olist_products_dataset.csv
8. olist_sellers_dataset.csv
9. product_category_name_translation.csv

Data Summary:

This dataset provides randomly extracted data for orders made between 2016 and 2018.

Dataset Dictionary:

Please refer to the following data dictionary in excel format:

- olist_datadict.xlsx

Data source:

Brazilian E-Commerce Olist Public Dataset

Dataset License:

CC BY-NC-SA 4.0

Instructions:

Please make reasonable assumptions and explain the rationale of the assumptions based on the data provided and problem statement stated.

2. Exploratory Data Analysis (EDA)

Using the dataset provided, perform an EDA and create an interactive notebook that can be used to present and explain the findings of your analysis. (Python or R programming language preferred.)

The report should contain appropriate and sufficient visualisations and explanations to help assessors understand how insights are derived, as well as their implications on the design of your machine learning models.

All analysis related to data preparation, input selection and feature engineering should also be included in the EDA.

(Optional) You can also provide supplementary online dashboard presentations, e.g., online dashboard using Tableau or Power BI. Please include the link to this supplementary dashboard in your submitted **README.md**.

Deliverables:

1. EDA Notebook in Python or R Programming Language: an executable ".ipynb" or ".Rmd" file with the exact naming convention as follows: - **"eda.ipynb" / "eda.Rmd".**

2. Other programming languages are also allowed, but please make sure that all codes are included, with the results and findings write-up shown together with the codes. All scripts included should be running properly.

3. Please provide a copy of the Jupyter Notebook or Rmd (R Markdown) in pdf format as well.

4. Please make sure the EDA notebook is neat and well structured, and insights presented are focused and well-summarized.

5. Do not submit more than 20 pages (A4 size) for your EDA notebook submission. Additional information (beyond the 20 pages) can be included as Appendix in your submission, but Appendix will not be included for assessment.

6. The following is an example of folder structure to be delivered:

```
├── src
│   ├── dataprep
│   └── model
├── saved_model
├── README.md
├── eda.ipynb / eda.Rmd
├── eda.pdf
├── requirements.txt / packrat.lock
└── run.sh
```

Evaluation:

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

~~If you are shortlisted for the interview,~~ you are expected to be able to explain the thought processes and decisions you made throughout the analysis, and to demonstrate that you understand the underlying machine learning concepts. You may be requested to run your EDA notebook during the ~~interview~~ presentation, so please make sure the EDA notebook is able to run on your laptop.

Note: All groups will have to make a presentation

3. End-to-end Machine Learning Pipeline

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

Deliverables: In place of 1-3, you can create a dockerfile, specify instructions to build and run docker image in your README.md

1. A folder named "src" containing Python modules/classes or R scripts, or other programming language scripts. Note that Python or R programming language is preferred, but other languages are also acceptable; all codes must be well structured and documented for **readability** and **can run successfully.**

2. An executable **bash script "run.sh"** at the root folder of your submission.

3. A "**requirements.txt" file or "packrat.lock"** file, or equivalent, at the root folder.

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.

5. The README is expected to contain the following:
   a. Full name (as in NRIC) and email address. Include names and Student ID of all team members

   b. Overview of the submitted folder and the folder structure.

   c. Include information about the **programming language (with version)** used, the run environment prerequisite (including OS platform and version) and the list of libraries or packages (with version) required for both the EDA and pipeline in the submission.

   d. Overview of key findings from the EDA conducted and the choices made in the pipeline based on these findings, particularly any feature engineering. Also include URL link for the supplementary EDA online dashboard, if any (optional)

   e. Instructions for executing the pipeline and modifying any parameters.

   f. Description of logical steps/flow of the pipeline. If you find it useful, please feel free to include suitable visualization aids (e.g., flow charts) within the README.

   g. Explanation of your choice of models for each Machine Learning task.

   h. Evaluation of the models developed. All metrics used in evaluation should be explained.

   i. Other considerations for deploying the models developed.

   j. **Do not submit more than 10 pages (A4 size) for your "README.md" submission**. Additional information (beyond the 10 pages) can be included as Appendix in your submission, but Appendix will not be included for assessment.

Pipeline Requirements:

1. All codes for the pipeline must be submitted. Codes submitted must be structured with welldefined functions, with good documentation.

2. A bash script named "run.sh" to run the above-mentioned modules/classes/scripts. DO NOT submit a Windows batch ("*.bat") script in replacement of the bash script.

3. DO NOT install your dependencies in the "run.sh"; this will be taken care of when we assess the assignment if you have created your "requirements.txt" correctly.

4. Relevant training/evaluation metric(s) outputs to be generated upon completion.

5. Pipeline 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).

6. Python and R Programming Language are preferred for the submission. For Python, use only versions 3.7 and above. For R, use only versions 4.0 and above.

7. Other programming languages are also allowed, but all scripts must be running properly.

8. Please make sure that the pipeline codes can be executable successfully. README.md should include clear and comprehensive setup/running instruction.

9. Include at least one saved model in the folder "saved model" from your pipeline output.

10. DO NOT include the original raw data file in your submission.

11. The following is an example of folder structure to be delivered:

```
├── src
│   ├── dataprep
│   └── model
├── saved_model
├── README.md
├── eda.ipynb / eda.Rmd
├── eda.pdf
├── requirements.txt / packrat.lock
└── run.sh
```

Evaluation:

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

If you are shortlisted for the interview, you are expected to be able to explain the thought processes and decisions you made throughout your code, and to demonstrate that you understand the underlying machine learning concepts. You may be requested to run your pipeline during the interview, so please make sure the scripts are able to run on your laptop.

Submission Format

Submission Specifics:

1. Your work should be uploaded as a **"*.zip" file** to the designated upload link (details below).
2. The zip file size should not exceed **20MB**.
3. The zip file is to be named according to the following naming convention:

EGT309_T<1/2>_<project_team_name>.zip

~~**"<firstname>_<lastname>_<student id>.zip"**~~

~~e.g., "john_lim_2xxxxxT.zip"~~

4. The zip file should have a folder structure similar to the following:

Example:

~~~
```
├── src
│   ├── dataprep
│   └── model
├── saved_model
├── README.md
├── eda.ipynb / eda.Rmd
├── eda.pdf
├── requirements.txt / packrat.lock
└── run.sh
```
~~~

IMPORTANT NOTE:

- Non-conformance to the specified conventions/formats may negatively impact your evaluation.
- AIP will run through plagiarism checks for all submissions.
- **Candidates caught cheating will be disqualified.**

@end