National AI Student Challenge 2024

Technical Assessment

Deadline: Sunday, 10 Mar 2024, 1130pm

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:

PetFinder.my is one of the leading pet adoption portals in Malaysia and a non-profit organisation for animal welfare. In recent years, it has faced falling donations and slower adoption rates. The organisation is working on a new business model that will allow it to be more self-sustainable and is looking at ways to increase revenue through advertising and sponsorships.

The adoption rate is one of the key metrics for the organisation. With better adoption rate, the organization can have continuous fresh new contents for the portal, which in turn help to boost revenue from sponsorship and partnership. More importantly, better adoption rate means more animals can find new home sooner.

For this project, Petfinders will want predict adoption for the pet listing and to understand what factors affect the adoption rate. For example, whether more images or videos help to improve adoption chances; how to optimise the hosting and streaming of images/videos, without affecting listings appeal.

Please note that the development and use of such a prediction model might involve several ethical and safety concerns that may need to be carefully considered so that the system may be trusted. So while technical efficiency and accuracy are crucial, we are equally concerned about the ethical implications of such a model. For example, the prediction model might unintentionally learn and amplify biases present in the training data. For example, if the data shows a preference for certain breeds, colours, or ages of pets, the model might disproportionately recommend these pets, leading to unequal chances of adoption for others. This can result in certain types of pets (like older animals or specific breeds) being overlooked or discriminated against.

Being a non-profit organization for animal welfare, the primary focus should always be on the welfare of the animals being adopted. The system should not compromise the well-being of pets in favour of adoption rates or user preferences. For instance, it shouldn't encourage the adoption of pets to unfit homes just to increase adoption statistics.

Data Source:
https://www.kaggle.com/c/petfinder-adoption-prediction/data

The assignment project objective is:

To predict adoption rate and better understand the adopter's preferences.

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:

- pets_prepared.csv
- breed_labels.csv
- color_labels.csv
- state_labels.csv

https://learn.aisingapore.org/aip-downloads/

Data Summary:

The data source for this project was obtained from PetFinder.my, which has been Malaysia's leading pet adoption portal since 2008 with a database of more than 150,000 animals. The prepared dataset included tabular data, with text inputs included.

Optional Data:

Additional image datasets can also be downloaded from the original data.
Data Source: https://www.kaggle.com/c/petfinder-adoption-prediction/data

For pets that have photos, they will be named in the format of PetID-ImageNumber.jpg.

The image datasets are **not** required for this assessment and thus not included in the prepared dataset download folder. Image can be an optional data for candidate who want to further enhance their analysis with image information, and it can be included as part of the assessment submission.

However, image dataset is **not** part of requirement, and questions related to images will not be supported.

Dataset Dictionary:

Please refer to the following data dictionary in pdf format:

- Data_Dictionary_Pets.pdf

https://learn.aisingapore.org/aip-downloads/

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, how your model integrates ethical considerations into its design 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 make sure the EDA notebook is neat and well structured, and insights presented are focused and well-summarized.

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

```
├── src
│  ├── dataprep
│  └── model
├── saved_model
├── README.md
├── eda.ipynb / eda.Rmd
├── 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, 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, so please make sure the EDA notebook is able to run on your laptop.

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:

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 acceptable; all codes must be well structured and documented for readability and can run successfully.

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

3. A "requirements.txt" file or "packrat.lock" file, or equivalent, at the base 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.

    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.

Pipeline Requirements:

1. All codes for the pipeline must be submitted. Codes submitted must be structured with well-defined 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
├── 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, 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. (If you are using any libraries that exceed this file size limit, do not include it in the zip file submission.)

3. The zip file is to be named according to the following naming convention:
   "<full_name>.zip"
   e.g., "john_lim_guo_ren.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
├── requirements.txt / packrat.lock
└── run.sh
```

5. Once you have packaged your submission, please access the following link to upload your zip file: https://learn.aisingapore.org/aip-upload/

6. If your file has been successfully uploaded, you should receive an acknowledgement email.

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.