# Carousell Take-home Assignment

In Carousell, we develop Machine Learning (ML) models to improve buying and selling experience on the platform. One of the objectives of our work is to minimise the effort required by sellers in creating a listing. In this assignment, we will focus on the problem of price suggestion within the sell form.
→ *Your task is to predict the prices for listings in the `Video Games > Gaming Consoles` category.*

Implement and evaluate 2 approaches for predicting the prices of Gaming Console listings.

a. Task (a): Propose and implement a method or model for predicting the price of a new listing. How does your model / approach perform on the `test.csv`?
   i. Submit your predictions on `test.csv` as `**submission_a.csv**`.
   ii. Note: You are <u>not</u> permitted to use LLM outputs (embeddings or predictions) for task (a).

b. Task (b): Let's try a different approach with LLM:
   i. Part (i): With the help of LLM, construct and propose a Product Catalogue for Gaming Consoles using only information from listings provided in `train.csv`. A Product Catalogue should consist of a combination of attributes for which each row uniquely identifies a Product SKU.
   ii. <u>Example of a Product Catalogue in Watch category</u>

| **SKU** | **attribute 1** | **attribute 2** | **attribute 3** | **attribute 4** | **…** |
|---|---|---|---|---|---|
| 121 | Rolex | Datejust | Stainless Steel | Yellow Gold | … |
| 122 | Rolex | Daytona | Stainless Steel | Mother of Pearl | … |
| 123 | Rolex | Submariner | Stainless Steel | Champagne | … |
| 132 | Patek Philippe | Aquanaut | Leather | White Gold | … |
| … | … | … | … | … | … |

   1. Write a script that generates a Product Catalogue taking only `train.csv` as input. There should not be any hardcoded attributes. (i.e. your script should work for any set of listings in any given category.)
   2. Submit your generated Product Catalogue as `**product_catalogue.csv**`.

   iii. Part (ii): Using additional information from the Product Catalogue you have created, propose and implement a method or model for predicting the price of a new listing.
      1. For example, given a new listing, you may try to predict the corresponding Product, then use the information to aid in a prediction of its price.
      2. Submit your predictions on `test.csv` as `**submission_b.csv**`.

Evaluate and compare methods (a) and (b) above. Which of the above methods do you prefer and why?

**Notes**:

- You are allowed to use any libraries or pre-trained models in any part of the assignment. However, do showcase how you will train or finetune a model to optimally deal with the task at hand.
- Task (a): You are not permitted to use LLM for this task.
- Task (b): You may choose to use LLM at any point of this task. However, do train your own custom statistical / custom model for any specific task as required.

**Dataset**

You will be provided with 4 files:

- "train.csv": Train Dataset of listings with listing ID, features and price labels.
- "validation.csv": Validation Dataset of listings with listing ID, features and price labels.
- "test.csv": Test Dataset of listings with listing ID, features only, used to create "submission.csv"
- "sample_submission.csv": Sample of how "submission.csv" should look like.

You should construct the taxonomy and train any models using only "train.csv".
You should use "validation.csv" for testing the accuracy of your model / pipeline.
Generate predictions for "test.csv" and submit the "submission.csv" as an output of this task.

All datasets consists of listings within "Video Gamers > Gaming Consoles" category, with the following features:

- Title: A title written by a seller.
- Description: A description written by a seller.
- Condition: One of `Brand New`, `Like New`, `Lightly Used`, `Well Used`, `Heavily Used`
- Image URL
- Date sold
- Price: Transacted price of the listing. (only for Train & Validation datasets)

You may use any combination of the above features in dealing with this task.

**Expected Deliverables**

Please submit a **zip file** containing the following deliverables.

- PDF report or Jupyter notebook, including:
  - Exploratory data analysis
  - Modelling approach and Explanation of design choices
  - Performance evaluation
  - Ideas for Future work
- Code Repository / Folder
  - Clean code with instructions on how to replicate the results
  - We expect proper project structure with code to train, eval and infer.
- `**product_catalogue.csv**`
  - Your generated product catalogue for Video Gaming Consoles.
- `**submission_a.csv**` and `**submission_b.csv**`
  - Submission should follow the format as shown in `sample_submission.csv`, containing the following columns:
    - id : Listing ID
    - predicted_price : Predicted price of the listing

Your submission will be evaluated based on the following:

1. Model(s) performance based on predictions from `submission.csv`

2. ML approach and problem solving skills
3. Code quality and structure

**Guidelines for the assignment**

- Please provide a zip file with all the necessary code to replicate your results:
    - Complete the assignment using Python and tooling of your choice
    - Keep code clean and organised
    - Include `README.txt` to describe how to use the code in your repository
    - Please ensure all packages are included in `*requirements.txt*` and installed with pip
- Consider:
    - Evaluation metrics that are relevant to the problem and model.
    - How you would benchmark your own model performance.
- Please include all links and references to the source/articles/publications you use when working on the assignment

Good luck!