Objective:

In online advertising, click-through rate (CTR) is a very important metric for evaluating ad performance. As a result, click prediction systems are essential and widely used for sponsored searches. For this assignment, we have provided 90 days' worth of Zepto data to build and test prediction models. Your task is to build a CTR prediction model on top of the given feature set.

Overview:

This assignment involves extracting insights from the data, performing feature selection and pre-processing, selecting appropriate evaluation metrics, and training a CTR ranker.

Data:

- **data.zip**

Data fields

- **search_term:** Search query.
- **city_id:** Unique IDs for different cities (city identifier).
- **product_variant_id:** Unique IDs for unique products (product identifier).
- **is_clicked:** 0/1 indicating non-click/click (target variable).
- **query_type:** Head/tail (head means popular query, tail means unpopular queries).
- **total_clicks:** Total clicks on a product for a given query at the city level (QxPxC).
- **session_views:** Total views of a product for a given query at the city level (QxPxC).
- **query_products_clicks_last_30_days:** Total clicks on a product for a given query at the city level in the last 30 days (QxPxC).
- **CTR_last_30_days:** Click-through rate of a product for a given query at the city level in the last 30 days (QxPxC).
- **CTR_last_7_days:** Click-through rate of a product for a given query at the city level in the last 7 days (QxPxC).
- **CTR_product_30_days:** Click-through rate of a product at the city level in the last 30 days (PxC).
- **query_product_plt_clicks_60_days:** Total clicks on a product for a given query in the last 60 days across the platform (QxPxPlt).
- **query_product_plt_ctr_60_days:** Total click-through rate of a product for a given query in the last 60 days across the platform (QxPxPlt).
- **CTR_plt_30_days:** Click-through rate of a product in the last 30 days across the platform (PxPlt).

- **predicted_category_name:** Predicted category name for the query.
- **predicted_subcategory_name:** Predicted subcategory name for the query.
- **query_product_plt_clicks_30_days:** Total clicks on a product for a given query in the last 30 days across the platform (QxPxPlt).
- **Product_name:** Name of the product.
- **Brand_name:** Brand of the product.
- **category_name:** Category of the product.
- **subcategory_name:** Subcategory of the product.
- **latest_margin:** Profit margin of the product.
- **savings:** Savings on the product.
- **savings_with_pass:** Savings on the product with a Zepto pass.
- **ad_revenue:** Advertising revenue.
- **total_unique_orders:** Total unique orders (PxC).
- **product_clicks_30_days:** Total clicks on the product in the last 30 days at the city level (PxC).
- **product_clicks_plt_30_days:** Total clicks on the product in the last 30 days on the platform (PxPlt).
- **total_unique_orders_plt_30_days:** Total unique orders in the last 30 days on the platform (PxPlt).
- **product_ctr_city_30_days:** Product click-through rate in the last 30 days for the city (PxC).
- **query_product_similarity:** Cosine similarity of query and product embeddings.

Q -> Query
P -> Product
C -> City
Plt -> Platform

Deliverables:

1. **Exploratory Data Analysis (EDA)**: Perform an EDA of the provided dataset and report relevant findings.
2. **Feature Engineering**: Develop and engineer relevant features to build a prediction model.
3. **Model Building**:
    a. Pre-process and clean the raw features.
    b. Build a model to rank products based on search terms and city level.
    c. Include any metrics used to track the improvement of your solutions.
4. **Documentatio**n: Create a document detailing your approach.

Hints:

- Read about ranking techniques, feature pre-processing, and machine learning architectures.
- Use classical machine learning algorithms instead of deep learning for this task.