Dataset

We provide two versions of the dataset:

- The Synerise dataset contains all interactions from the online retailer's website.
- The preprocessed dataset is derived from the raw data and was used specifically for the RecSys 2025 Challenge.

Synerise dataset

SYNERISE_DATASET (DOWNLOAD BUTTON HERE)

The Synerise dataset is based on user click behavior that was recorded at a real-world online retailer's website, covering a period of six months. The data consists of five types of events and product attributes:

product_buy

```
client_id (int64): Numeric ID of the client (user).
timestamp (object): Date and time of the event in the format YYYY-MM-DD HH:mm:s
sku (int64): Numeric ID of the item.
```

add_to_cart

```
client_id (int64): Numeric ID of the client (user).
timestamp (object): Date and time of the event in the format YYYY-MM-DD HH:mm:s
sku (int64): Numeric ID of the item.
```

remove_from_cart

```
client_id (int64): Numeric ID of the client (user).
timestamp (object): Date and time of the event in the format YYYY-MM-DD HH:mm:s
sku (int64): Numeric ID of the item.
```

page_visit

```
client_id (int64): Numeric ID of the client.
timestamp (object): Date and time of the event in the format YYYY-MM-DD HH:mm
url (int64): Numeric ID of a visited URL. The explicit information about what (
```

search_query

```
client_id (int64): Numeric ID of the client.
timestamp (object): Date and time of the event in the format YYYY-MM-DD HH:mm:s
query (object): The textual embedding of the search query, compressed using the
```

product_properties

```
sku (int64): Numeric ID of the item.
category (int64): Numeric ID of the item category.
price (int64): Numeric ID of the item's price bucket.
embedding (object): A textual embedding of a product name, compressed using the
```

The Synerise dataset preprocessed for RecSys 2025 challenge

CHALLENGE_DATASET (DOWNLOAD BUTTON HERE)

For the specific purposes of the RecSys 2025 Challenge, the Synerise dataset was heavily preprocessed. Temporally, the events were split into three distinct parts:

- Pre-training time frame – includes all interactions from the first five months of the provided data.
- Training time frame – covers the two-week period immediately following the pre-training time frame.
- Evaluation time frame – covers the two-week period immediately following the training time frame.

The dataset directory contains a restricted version of the additional item attributes file (`product_properties.parquet`), only containing properties for items which appeared in pre-training timeframe. Other files are stored in two further subdirectories: `input` and `target`, whose contents we describe in the following sections.

`input`

This directory contains the same types of events (product_buy, add_to_cart, remove_from_cart, page_visit, search_query) as the raw dataset, but here they are restricted to pre-training time frame.

In addition the directory a NumPy file `relevant_clients.npy` containing a subset of 1M client_ids. In the challenge, participants were required to create Universal Behavioral Profiles for the clients whose `client_id` is listed in `relevant_clients.npy`.

`target`

This directory contains two event-related files.

- `train_target.parquet`: Contains interactions of clients during training time frame.
- `validation_target.parquet`: Contains interactions of clients during evaluation time frame.

Note that these event logs contain all types of interactions.

In addition, the directory contains NumPy files, which are used by the provided evaluation scripts to score submissions. In particular, they are used to store information on items/ categories/prices for which ground truth labels are computed.

- `propensity_category.npy`: Contains a subset of 100 categories for which the model is asked to provide predictions
- `popularity_propensity_category.npy`: Contains popularity scores for categories from the `propensity_category.npy` file. Scores are used to compute the Novelty measure.
- `propensity_sku.npy`: Contains a subset of 100 products for which the model is asked to provide predictions
- `popularity_propensity_sku.npy`: Contains popularity scores for products from the `propensity_sku.npy` file. These scores are used to compute the Novelty measure.
- `propensity_new_sku.npy`: Contains a subset of 20 products not contained in pre-training time frame for which the model is asked to provide predictions
- `popularity_propensity_new_sku.npy`: Contains popularity scores for products

from the `propensity_new_sku.npy` file. Scores are used to compute the Novelty measure.
- `propensity_price.npy`: Contains a set of 100 price buckets for which the model is asked to provide predictions
- `popularity_propensity_price.npy`: Contains popularity scores for price buckets from the `propensity_price.npy` file. These scores are used to compute the Novelty measure.

Finally the directory contains `active_clients.npy` a subset of relevant clients with at least one product_buy event in history. Active clients are used to compute churn target.

GitHub repository

We recommend using the preprocessed dataset together with the challenge code.

CHALLENGE_REPOSITORY (BUTTON WITH LINK HERE)