Mango - Datathon UPC

Challenge Description

The goal of the challenge is to provide a production quantity that we will send to our supplier, who will manufacture the number of units determined by the algorithm for each product in the new season. The output dataset must contain two columns:

- ID

- demand (full estimation of demand and, indeed, what you should buy)

To generate these quantities, participants will receive a dataset with historical data from the last four seasons. Each year includes two seasons (Spring–Summer and Fall–Winter), which means we provide two full years of history plus one validation season.

This dataset is anonymized, as it contains sales and production information. The sales, demand, and production variables are masked with a scaling transformation, so the values visible to participants range from 0 to 1 while preserving their proportions and correlations. The dataset includes:

- Image embeddings: We provide precomputed garment image embeddings to make the modeling process faster. These may be useful for linking products from past seasons.

- Product family: Indicates the category of the garment (e.g., dresses, jeans, coats). Some families may only appear in specific seasons (e.g., swimsuits only in summer).

- Product attributes: Categorical features describing garment characteristics such as sleeves, neckline type, silhouette, etc.

- Phase_in, phase_out, life cycle: These variables define the selling period of a product. Phase_in indicates the Monday of the first week of sales, phase_out indicates the Monday of the last week of sales, and life cycle is simply the number of weeks between these two dates.

- Number of stores: Indicates how widely the product will be distributed. More stores generally means higher expected sales.

- Number of sizes: Similarly, more sizes imply higher potential production and sales.

- Sales: Actual past sales. Keep in mind sales are naturally capped by the available stock — you can never sell more than you produced.

- Production: The production quantity ordered by Mango teams in past seasons. This can help identify historical success levels.

- Demand: This variable, obtained with an internal Mango algorithm, estimates the maximum potential sales of a product — essentially, what it would sell with infinite stock. This is the target variable of the challenge and the one that must be predicted.

Evaluation

Do you think you can anticipate demand and help us decide how much to produce?

To evaluate your predictions, we will compare the real demand of the validation season with your estimated values. We created a scoring metric that measures how close your predictions are to the real target. But be careful: just like our buyers in real life, the metric penalizes lost sales more heavily than excess stock, since unsold items can still be sold later.

The score ranges from 0 to 100 and summarizes in a single value how accurate all your predictions were during the validation season.

To help you track your progress throughout the datathon, we will reserve a subset of validation products so you can test your predictions.

Submissions are evaluated on the following score function:

Submission File

For each id_model in the test set, you must predict a quantity for the TARGET variable. The file should contain a header and have the following format:

ID,demand

2,0.2

5,0.45

6,0.12

etc.

Overall evaluation of the Mango challenge

The final evaluation of the Mango challenge will be:

80% * Kaggle challenge + 20% * Presentation score

Dataset Description

Files

- **train.csv** - the training set
- **test.csv** - the test set
- **sample_submission.csv** - a sample submission file in the correct format

Columns

- id - Model id
- id_season - Season identifier
- aggregated_family - Highest product hierarchy
- family - Family of the product (Dresses, Coats,...)
- category - Kind of garment (Tops, Bottoms, ...)
- fabric - Kind of fabric (Synthetic Leather, ...)
- color_name - Color
- color_rgb - Color in RGB format
- image_embedding - Pre-calculated image embedding
- length_type - Length attribute (Midi, Crop, ...)
- silhouette_type - Silhouette atribute (Slim, Oversize, ...)
- waist_type - Waist atribute (High waist, Adjustable, ...)
- sleeve_length_type - Sleeve length atribute (Long, Short, ...)
- heel_shape_type - Heel shape atribute
- toecap_type - Toecap atribute
- woven_structure - Woven atribute
- knit_structure - Knit atribute (Slim, Oversize, ...)

- print_type - Print atribute

- archetype - Who the garment is designed for

- moment - By the time the garment is designed (Formal Work, casual event,...)

- ocassion - Aggregation of moment (Urban, Formal)

- phase_in - When you plan to start selling

- phase_out - When you plan to finish selling

- life_cycle_length - Number of weeks you plan to sell

- num_stores - Number of stores you plan to sell

- num_sizes - Number of sizes for the garment

- has_plus_size - Indicates if the garment has plus sizes in the size range

- price - Selling price

- year - Year of sale

- num_week_iso - ISO Week

- weekly_sales - Number of sales in the week

- production - Number of units asked for production

- weekly_demand - Variable to forecast. Weekly full potential of sales