DATA MINING CUP 2022

Scenario

This year’s scenario is all about Pia and Philip, a married couple. They started their new e-commerce business during the pandemic in 2020 by offering convenience goods online. They began by selling an assortment of masks and disinfectants, but quickly expanded to a wider range of various everyday commodities.

Having both a background in traditional and online retail, they are aware of how distant and impersonal online shopping can feel and, at the same time, how important customer guidance and recommendations are for long-term customer loyalty.

To differentiate themselves from the many other commodity shops, they decided to put an even more significant emphasis on personalized recommendations and offers.

One key element of this strategy is a customized weekly newsletter that personally addresses each of their clients. The newsletter includes user favorites, products similar customers liked, new additions, and special offers.

However, they quickly noticed a problem: repeated recommendations of recently purchased products. One quick workaround for this issue was implementing a filter that would exclude products from the recommendation for a fixed number of days. This, however, did not meet the high standards of Pia and Philip.

They are instead looking for a model that can reliably predict the week that a returning customer might repurchase one of their frequently purchased items.

By knowing the estimated week of replenishment, products can be added to the newsletter as a reminder, thus increasing basket sizes and profits.

Since the owners are only interested in the best possible solution, they organized a contest to benchmark competing prediction approaches.

Data

To create a prediction model, the participants will receive historical transaction and descriptive item data in the form of structured text files (.csv).

The data is provided in four individual files. One file containing the orders (“orders.csv”), one containing descriptive item data (“items.csv”), one containing item and category hierarchy information (category_hierarchy.csv), and a template for the result submission (“submission.csv”).

Here are some points to note about the files:

1. The first line (top line) has the same structure as the data sets but contains the names of the respective columns (data fields).
2. A list of all the column names, which occur in the appropriate order, can be found in the “features.pdf” file along with brief descriptions and value ranges for the associated fields.
3. The top row and each data set contain several fields separated from each other by the “|” (pipe) symbol.
4. There is no escape character: quotes are not used.

5. The encoding is “utf-8”.

The “items.csv” file is a master data set that contains descriptive features. All features are categorical but numerically encoded. The list of features is explained in the “features.pdf” file. Each data line contains the description for one single item.

The “orders.csv” file contains information about user-specific orders across eight months. Each line displays one transaction for one single item. All attributes are described in the “features.pdf” file.

The “category_hierarchy.csv” file contains two columns of encoded categories, that maps each category to its parent category.

The “submission.csv” file is the reference for submission and contains a predefined subset of userID and itemID combinations as well as an empty “prediction” column for the participants to fill in.

Entries

Participants must submit their results by **2 p.m. on 28 June 2022 (UTC+2 or CEST)**. The task description below explains how to submit entries.

Task

The participating teams’ goal is to predict the user-based replenishment of a product based on historical orders and item features. Individual items and user specific orders are given for the period between 01.06.2020 and 31.01.2021. The prediction period is between 01.02.2021 and 28.02.2021, which is exactly four weeks long.

For a predefined subset of user and product combinations, the participants shall predict if and when a product will be purchased during the prediction period.

The prediction column in the “submission.csv” file must be filled accordingly.

- 0 - no replenishment during that period
- 1 - replenishment in the first week
- 2 - replenishment in the second week
- 3 - replenishment in the third week
- 4 - replenishment in the fourth week

The different columns are separated by the “|” symbol. A possible example of the solution file might look like this:

userID|itemID|prediction
12|6723|0
20|8272|1
28|9873|4
…

The solution file must match the specifications described in the Data section. Incorrect or incomplete submissions cannot be assessed.

Participants must upload the solution file as a structured text file (csv) to the DATA MINING CUP website: **https://www.data-mining-cup.com/dmc-2022/.**

Please make sure that the mandatory boxes on the form are correctly and fully completed before uploading the data.

The name of the text file consists of the team's name and the file type:

“<Teamname>.csv” (e.g. TU_Example_1.csv)

The team's name was communicated to the team leaders when their registration was confirmed.

Evaluation

The solutions submitted will be assessed and compared based on the number of accurate predictions.

The teams will receive one point for every item that is correctly identified for replenishment (1-4) or no replenishment (0). Three points will be awarded if the correct week is predicted.

The team with the highest number of points wins.