DATA MINING CUP 2020

Forecasting Demand for Optimized Inventory Planning

It is no secret that the ability to optimize stocks provides many benefits for retail companies. Different advantages may accrue, depending on the type of company, its strategy and its situation. It allows store-based retailers to downsize their storage space and increase the sales area to provide a more open and inviting shopping experience, for example.

Online retailers, on the other hand, may be able to upscale their business without relocating their entire operations to a larger facility.

Overall, optimized inventory planning helps to reduce the number of slow-moving goods, because retailers only stock products that people actually buy. This, in turn, means that it is not necessary to send customers away because products are temporarily unavailable; this increases both revenues and customer satisfaction. Moreover, fewer slow-moving goods means less reorganization, accounting and clearance and this also reduces the work time required and the outlay for logistical services.

For these reasons, forecasting demand is the focus of this year’s DATA MINING CUP.

Scenario

An established retailer wants to optimize its inventory planning to not only significantly reduce storage space, but also its costs and need for logistical operations. It plans to restock its inventory every other week and only keep in stock the items that it has actually sold during that period.

The goal of the participating teams is to create a machine learning model to predict the demand for every product over the two-week period. It is important to point out that some products will be promoted for limited periods of time. Products that are promoted during the simulation period will be earmarked. However, the transaction data needs to indicate whether a product is being promoted during the training period. Finally, the model does not need to be able to respond to price changes during the simulation period. To simplify matters, prices will not be changed during the period.

In order to create this model, the teams obtain information about the exact time of every transaction during a period of six months and about other features that describe the products.

Data

Real anonymized data in the form of structured text files (csv) is provided for this task. There are three individual files containing master data (*"items.csv"*), transaction data ("*orders.csv")* and an info file (*"infos.csv"*) for the simulation period.
Here are some points to note about the files:

1. Each data set is on a single line ending with "CR" ("carriage return", 0xD), "LF" ("line feed", 0xA) or "CR" and "LF" ("carriage return" and "line feed", 0xD and 0xA).
2. The first line (top line) has the same structure as the data sets, but contains the names of the respective columns (data fields).
3. A list of all the column names, which occur in the appropriate order, can be found in the *"features.pdf"* file as well as brief descriptions and value ranges of the associated fields.
4. The top row and each data set contain several fields, which are separated from each other by the "|" symbol.
5. Floating point numbers are not rounded. "." is used as the decimal separator.
6. There is no escape character: quotes are not used.
7. The character set used is ASCII.

The *"items.csv"* file is a master data set that contains descriptive features. The features may be categorical or numerical. The list of features is explained in the *"features.pdf"* file. Each data line contains the description for one single item.

In addition to other information, the "*orders.csv"* file contains every order with its dedicated timestamp for the 6-month period. Each line displays one transaction for one single item. All the attributes are described in the *"features.pdf"* file.

The *"infos.csv"* file contains a list of all the items, their current sales price as well as the dates of scheduled promotions during the simulation period. Every line contains the price and promotion dates for one single item.

Entries

Participants may submit their results by **2 p.m.** on **30 June 2020 (UTC+2 or CEST)**. The task description below explains how to submit entries.

Task

Historical data must be used to create a machine learning model to reliably forecast the demand for each item in the "*items.csv*" file for a period of 14 days. Use the period starting on 30 June 2018 00:00:00, the day after the last date from the transaction files.
The historical demand for an item (e.g. daily) can be derived from the "*orders.csv*" file by aggregating the orders for each item (daily). The "*orders.csv*" file is not already aggregated (e.g. on a daily basis); as a result, the participant can choose the scope of its time steps more freely.
In addition to time-dependent features, participants are allowed to use any attribute provided by the *"items.csv"*, *"orders.csv"* and *"infos.csv"* files.

One file containing the following information should be used to send the solution data:

Column name	Description	Data type
itemID	Unique identifier for each product	String
demandPrediction	Demand prediction for 14 days	Integer

The key attribute for every prediction is the "*itemID*", which can be found in the "*items.csv*" file. It is therefore not necessary to match the order provided in the "*items.csv*" file.
Possible values for the "*demandPrediction*" column are positive integer values, which predict the total demand for a single item over a period of 14 days. The two different columns are separated by the "|" symbol. A possible extract from the solution file might look like this:

```
itemID|demandPrediction
995539|34
1000002|42
995554|10
…
```

The solution file must match the specifications described in the **Data** section, if they are relevant. Incorrect or incomplete submissions cannot be assessed.

The solution file must be uploaded as a structured text file (csv) to the Data Mining Cup website: **https://www.data-mining-cup.com/dmc-2020/.**
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 name and the file type:

"<Teamname>.csv" (e.g. TU_Chemnitz_1.csv)

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

Evaluation

The solutions submitted will be assessed and compared on the basis of their monetary value for the retailer. The monetary value is determined by the predicted revenue and an overstocking fee for overestimating the demand for any products. The demand prediction for every product is therefore compared with the actual number of orders within the same time frame.

If the demand is predicted correctly, the monetary value for that product is simply the revenue (i.e. price x demandPrediction). If the prediction is lower than the actual number of units sold, we assume that only the number of predicted stocks for this product will be available and the company will fail to generate the potential revenue. As a result, the monetary value is once again the revenue (i.e. price x demandPrediction). If the demand prediction is higher than the actual number of units sold, we assume that the remaining stock will generate an overstocking fee of 0.6 x price x (orders - demandPrediction).

The total revenue minus the additional costs for stock clearance is the monetary value of the solution that is submitted.

The winning team is the one whose solution achieves the highest monetary value. In the event of a dead heat, a random draw will decide which team wins.