DATA MINING CUP 2018

Sales forecast for sporting goods

Dynamic pricing strategies are increasingly common, particularly in online business. For the purposes of simplification, long product lifecycles and an inexhaustible stock are often assumed. Conversely, in the area of fashion it seems expedient to take into account both the product age and the stock. The product lifecycles are mostly very short in this area. In this context it is very important for the retailer to have sold out an item at a particular time if possible, as the subsequent item then appears. It is equally important to know when an item will be sold out, to reorder on cue if the item has not yet reached the end of its life.

Scenario

A sporting goods retailer uses dynamic prices to control when items sell out in their online shop. A good prediction of when items will be sold out is necessary in order to make the most expedient price adjustments. The time a product is sold out depends not just on its price, but also on other product attributes, such as brand, size and product group.
The task for the participating teams is to use the sales data from a period of four months to develop a prediction model, which can be used to predict the products time of sell out in the following month. The aim is to predict as accurately as possible the precise day when items will sell out.

Data

For the task, real anonymized shop data in the form of structured text files are provided. These files contain individual data sets. Below are some restrictions 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 row has the same structure as the data sets, but contains the names of the respective columns (data fields).
3. The top row and each data set contain several fields separated from each other by the „|“ symbol.
4. The floating-point numbers are rounded to two decimal places. The “.” is used as the decimal separator.
5. There is no escape character, quotes are not used.
6. ASCII is the character set used.

The master data attributes of all the products occurring in the learning and prediction time period are listed in the *“items.csv”* file. The *“features.pdf”* file contains a list of the column names that occur in the appropriate order as well as short descriptions and value ranges of the associated fields. A product is uniquely identified by the product number and the size.

The prices for the products over time are included in the file "*prices.csv"*. The format of this file is as follows:

Column name	Description	Value range
pid	Product number	Natural number
size	Size	String
2017-01-10	Product price on the first day of the learning period	Floating point number
…	…	…
2018-02-28	Product price on the last day of the prediction period	Floating point number

The sales data for the learning period is located in the file *"train.csv"*. The file format is as follows:

Column name	Description	Value range
date	Date	Format YYYY-MM-DD
pid	Product number	Natural number
size	Size	String
units	Quantity sold	Natural number

A single data set from the file "*train.csv"* contains information on how much of a product (identified by "*pid*" and "*size*") was sold on a particular day (*"date"*). If a product was not sold on a particular day, this information does not appear in the data. Entries in the *"units"* column are always greater than 0.

Entries

Participants can submit their results up to and including **2018-05-17, 14:00 CEST (2 p.m. UTC+2, or CEST)**. The task description below also explains how to submit entries.

Task

Use historical data to create a mathematical model to reliably predict figures for when items sell out. You are provided with the sales data and the price changes during the four-month learning period. The relevant files *"train.csv"* and "*prices.csv"* are described in the **Data** section. Additional product master data is provided to increase the quality of the prediction. These are included in the file *"items.csv"* and they are also described in the **Data** section.
The task is now to predict the dates of sell out in February on the basis of the prediction models for the products. To this end the stock of all products at the start of the month of February is shown in the "*stock"* column in the product master data in the file *"items.csv".* All products were sold out in the period 2018-02-01 – 2018-02-28, i.e. the quantity in the "*stock"* column of the product master data falls to 0. The product prices over time for the month of February are in the file *"prices.csv"*.
The product number under the attribute "*pid*" and the size of the product under the attribute "*size*" are used as the key to link the pieces of information.

For submitting the solution, a file containing the following information should be used:

Column name	Description	Value range
pid	Product number	Natural number
size	Size	String
soldOutDate	Predicted sold out date	Format YYYY-MM-DD

All combinations of product number and size from the master data in the file "*items.csv*" must appear precisely once in the solution file. The file should continue to comply with the specifications in the **Data** section, as far as they are applicable. Possible values of the *"soldOutDate"* column are time stamps in the format YYYY-MM-DD that represent a day in February 2018. A possible extract from the solution file could look like this:

pid|*size*|*soldOutDate*
15835|*39 1/3*|*2018-02-22*
15835|*40*|*2018-02-03*
15835|*41 1/3*|*2018-02-10*
…

The results file must be uploaded to the DATA MINING CUP website **https://www.data-mining-cup.com/dmc-2018/** in the form of a structured text file (csv) using the form for submission of solutions. Please fill out all obligatory fields on the form and use the password **DMC-3UTK*62&%z8{2018**.

The name of the text file must be made up of the team name and the file type (csv):

"<Team name>.csv" (e.g. TU_Chemnitz_1.csv).

The team name has been sent to the team leader with the entry confirmation.

Evaluation

The solutions received will be graded and compared using the following error function, which should be minimized:

$$E = \sqrt{\sum_i |d_i - \hat{d}_i|}.$$

In this function d_i describes the date on which the last unit of the item i was sold, and $\hat{d}_i$ describes the date of sell out predicted by the team for item i. The amount function calculates the difference between the two time stamps in days. For example, this means that if the actual sold out date is 2018-02-24 and the predicted date is 2018-02-20, the function calculates a value of 2.

The team with the lowest error function value wins. In the event of a draw, the winner will be decided by drawing lots.