DATA MINING CUP 2017

Revenue forecast as foundation for dynamic pricing

Dynamic pricing strategies have become an essential part of online business. Large online shops with a wide range of products in particular rely on the automatic adjustment of product prices. The focus here is not on price personalization but rather price optimization at the product level. The goal is thus not to provide customer-specific and personalized prices. On the contrary, regular price adjustments lead to a price per product in line with the market that will maximize parameters such as revenue and gross margin.
At this year's DATA MINING CUP we will take a detailed look at these relationships. In the process we will take into account that it is not only the price but also the interaction between different product attributes that influences revenue.

Scenario

A mail-order pharmacy uses a dynamic pricing strategy in the form of daily automatic price adjustments for its online shop. In order to evaluate its success, prices, revenue figures and different product attributes are recorded. User behavior expressed in actions such as clicks on products, assigning shopping baskets and purchases is also recorded.
The challenge for the participating teams is to develop a model based on the data of a period of three months and to use this model to forecast revenue figures for the following month, obtaining the best possible forecast results.

Data

Real anonymized shop data in the form of structured text files are provided for the task. These files contain individual data sets. Below 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. The top line 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 "." symbol is used as the decimal separator.
5. There is no escape character, quotes are not used.
6. ASCII is the character set used.

The *"features.pdf"* file contains a list of all the column names that occur in the appropriate order as well as short descriptions and value ranges of the associated fields.

The attributes of all the products occurring in the learning or classification time period that do not change with time are listed in the *"items.csv"* file.

The information that changes with time for the learning time period is located in the *"train.csv"* file and the information for the classification time period is located in the *"class.csv"* file.

The key to linking information that changes with time and information that does not is the product number under the *"pid"* attribute.

A single data set from the "*train.csv*' file contains information about the action of a user regarding a particular product and other product information that changes with time (e.g. the competitor's price). The "*click*," "*basket,"* and "*order"* columns provide information about the type of action. Only one value per line in these columns can be "1", the others are "0". If a product has only been clicked, the value of the *"click"* column is "1". If the product was added to the shopping basket, the value in the *"basket"* column is "1". If the product was purchased, the value in the *"order"* column is "1". This does not provide information as to the number of units of the product purchased or added to the shopping basket.

Submission

Participants can submit their results up to and including **17 May 2017, 14:00 CEST (2 o'clock p.m. UTC+2, or CEST)**. The task description below explains how to submit entries.

Task

The task is to use historical data to create a mathematical model to reliably predict the revenue. To achieve this there is a full data set provided for the three-month learning period. The relevant files, *"items.csv"* and "*train.csv"*, are described in the **Data** section.
Product attributes and prices for the subsequent classification time period of one month are specified in the *"class.csv"* and *"items.csv"* files which are also described in the **Data** section. Here, there is no information as to whether the product was purchased, added to the shopping basket or just clicked. The main goal is then to predict the revenue per user action in the classification time period.

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

Column name	Description
lineID	Unique key to identify user action
revenue	Predicted revenue

The values in the *"lineID"* column are increasing natural numbers and are taken from the column of the same name in the *"class.csv"* file. All data records from the classification data must occur exactly once only and in the original order, represented by the *"lineID"*.

The file should continue to comply with the specifications in the **Data** section, as far as they are applicable. Possible values in the *"revenue"* column are positive floating-point numbers or "0". A possible extract from a solution file could look like this:

lineID|revenue
1|0
2|12.34
3|5.6
...

The result file must be sent as a zipped structured text file attached to an e-mail to **dmc_task@prudsys.de**. The name of the zip file and the compressed text file must be made up of the team name and the file type (zip or csv):

"<Teamname>.zip" (e.g. *TU_Chemnitz_1.zip*) or
"<Teamname>.csv" (e.g. *TU_Chemnitz_1.csv*).

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

Evaluation

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

$$E = \sqrt{\sum_i (r_i - \hat{r}_i)^2}.$$

The figure r_i corresponds to the actual revenue achieved by the user action i, whereas the figure $\hat{r}_i$ corresponds to the predicted revenue for the user action i. The team with the lowest error function value wins. In case of a draw, the winner will be decided by drawing lots.