DATA MINING CUP Competition 2016

Predicting returns for the sales of discounted articles and voucher redemptions

The central topic of this year's task in the DATA MINING CUP competition is the prediction of the return rates for a fashion distributor. Returns have been a major cost driver for online shops for many years. This is particularly the case for assortments in the fashion industry. Many approaches to solve this problem are based on forecasting models. Predicting returns was the central topic in the DATA MINING CUP of 2014. This year we look at the topic in more detail focusing in particular on the influence of discounted items and vouchers on return rates.

Scenario

A fashion distributor sells articles of particular sizes and colors to its customers. In some cases items are returned to the distributor for various reasons. The order data and the related return data were recorded over a two-year period. The aim is to use this data and methods of data mining to build a model which enables a good prediction of return rates.

Data

For this task real anonymized shop data are provided in the form of structured text files consisting of 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 has the same structure as the data sets, but contains the names of the respective columns (data fields).
3. The header line and each data set contain multiple fields separated from each other by a semi-colon (;).
4. There is no escape character, quotes are not used.
5. ASCII is the character set used.
6. Missing values may occur. These are coded using the character string "NA".

Actually only the field names from the attached document "*features.pdf*" can appear as column headings in the order used in that document. The associated value ranges are also listed.

The DMC training file ("*orders_train.txt*") contains all data fields from the document whereas the associated classification file ("*orders_class.txt*") does not contain the target attribute "returnQuantity".

Task

The task is to use known historical data from January 2014 to September 2015 (approx. 2.33 million order positions) to build a model that makes predictions about return rates for order positions. The attribute "returnQuantity" in the given data indicates the number of returned articles for each order position (the value "0" means that the article will be kept while a value larger than "0" means that the article will be returned). For sales in the period from October 2015 to December 2015 (approx. 340,000 order positions) the model should then provide predictions for the number

of articles which will be returned per order position. The prediction has to be a value of the set of natural numbers including "0". The difference between the prediction and the actual return rate for an order position (i.e. the error) must be as low as possible.

Submission

Participants can enter their results until **18 May 2016 14:00 CEST** (2 o'clock p.m. UTC+2, or CEST). The solution data must be reported in the following format (key columns and the prediction column) and sent in a file using:

Column name	Description	Value range
orderID	Order number	Natural number preceded by the letter "a"
articleID	Article number	Natural number preceded by the letter "i"
colorCode	Color code	Natural number
sizeCode	Size code	Natural number or string of letters
prediction	Prediction of the quantity of returns	Natural number and "0"

All data sets from the classification data must exactly occur only once. Furthermore the file should comply with the specifications in the "Data" section, as far as they are applicable. An example of an excerpt from the file could look like this

orderID;articleID;colorCode;sizeCode;prediction
a1744178;i1002632;3097;I;0
a1744178;i1003278;1097;40;1
...

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

"<teamname>.zip", (e.g. TU_Chemnitz_1.zip) or *"<teamname>.txt"*, (e.g. TU_Chemnitz_1.txt).

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

Evaluation

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

$$E = \sum_i |returnQuantity_i - prediction_i|.$$

The function E indicates the error that results by comparing the quantity of articles actually returned for the order position i ($returnQuantity_i$) with the quantity of returned articles for the order position i as predicted by the team ($prediction_i$). The team with the lowest error will win the competition. In the event of a draw, the winner will be decided by drawing lots.