DATA MINING CUP Competition 2014

Prediction of returns

In the course of the last years, the topic of returns management, and in particular the avoidance of returns, has become an omnipresent topic in the mail order business. According to the motto "the customer is always right" many mail order companies offer the possibility of easy and in most cases free returns to their customers, and they do not intend to change this strategy in the near future. It is evident that in this way a high proportion of returns is generated. The resulting costs have to be borne by the trader. Especially in the clothing trade returns proportions of partially more than 50 % are not exceptional. The goal for the sender is to lower these proportions without causing deterioration in customer service. When this scenario and possible solutions are discussed it soon becomes evident that preventive measures carried out on the basis of probabilities of returns (restriction with respect to payment options, adjustment of shipping costs, sizing guides, …) could become a target-oriented strategy. The topic of this year's DMC task is the calculation of probabilites of returns.

Scenario

On the basis of historical purchase data of an online shop a model is to be learned generating a prediction of the probability that a certain purchase is converted into a return on the basis of new purchase data of the shop. For this purpose the historical data contain as well purchase and shipping data as different product and customer attributes. The information "return yes/no" is known, too, for the historical data.

Data

For the task anonymized real shop data are provided in the form of structured text files consisting of individual data sets. For the data, in particular the following applies:

1. Each data set is in an individual line that is closed by "CR" ("carriage return", 0xD), or "CR" and "LF" ("carriage return" and "line feed", 0xD and 0xA).
2. The first line is structured analog to the data sets but contains the names of the respective columns (data arrays).
3. The header and each data set contain several arrays separated by a semicolon.
4. There is no escape character, and no quota system is used.
5. ASCII is used as character set.
6. There may be missing values. They are coded by the symbol "?".

In concrete terms, only the array names of the attached document "*features.pdf*" in their respective sequence will be used as column headings. The corresponding value ranges are listed there, too.

The training file for the DMC ("*orders_train.txt*") contains all data arrays of the document, whereas the corresponding classification file ("*orders_class.txt*") does not contain the target attribute "returnShipment".

Submission

The participants can submit their results up until 14 May 2014. The following task description explains how to submit the results.

Task

For the task, historical data of one year are known (approximately 481,000 order items) by means of which a model for the prediction of the returns can be learned. The target attribute "returnShipment" of the order item is known here, and it is described by the parameter value "0" in the case "item kept" and the parameter value "1" in the case "item returned". For the purchases of one month (about 50,000 order items) it is to be assessed in each case whether the item will be returned or not. For this purpose for each order item a prediction is to be made the value of which is within the interval [0,1]. The higher the value, the more probable is the return. The error with respect to the real outcome concerning the return of the order item should be as small as possible.

In order to submit the data of the solution a file with the following format is to be used:

Column name	Description	Range of values
orderItemID	Running number of the order item	Natural number
prediction	Prediction whether an item will be returned	Real number from [0,1]

Each sequential number of the order items from the classification data must appear here exactly one time. Furthermore, the file should comply with the specifications defined in the paragraph "Data" insofar as they are applicable. An exemplary extract of this file could look as follows:

orderItemID;prediction
1;0.4
2;0.95
…

The result file must finally be submitted as e-mail attachment in form of a zipped text file to dmc_task@prudsys.de. The name of the zip file and of the included text file must be composed of the team name and the file type:

"<Teamname>.zip", (e.g. *TU_Chemnitz_1.zip*), and *"<Teamname>.txt"*, (e.g. *TU_Chemnitz_1.txt*).

The team name was submitted to the team leader together with the registration confirmation.

Evaluation

The submitted solutions will be evaluated and compared by means of the following error functional that is to be minimized:

$$E = \sum_i \left| returnShipment_i - prediction_i \right| .$$

Here, $returnShipment_i$ is the information whether order item i represents a return (0 means "item kept", 1 means "item returned"), and $prediction_i$ is the predicted return probability for the order item i . The team whose error functional reaches the smallest value will win. In case of a tie the decision will be made by drawing lots.