DATA MINING CUP Competition 2015

Influence of coupons on order patterns

This year's DATA MINING CUP competition (DMC for short) deals with the topic of coupon generation. What is interesting in this case is the impact of coupons on the shopping basket value as well as the redemption rate of the individual coupons. Coupons have been used as purchase incentives for some time now, not only by mail order companies. This presents interesting challenges, including the question: "Who responds to coupons?", and then the question: "Who would have made the purchase even without the coupon?". Only by giving equal consideration to both of these questions can we arrive at profitable couponing. Data mining processes, in particular scoring processes, can be used to answer these questions. Such processes make it possible to create forecasting models which provide answers to both questions.

Scenario

Using the historical order data from an online shop with accompanying coupon generation, a model should be created that comes up with a prediction for the redeemed coupons and for the shopping basket value for new orders within the shop. The historical data contains both a time stamp of the coupon generation and the orders as well as different product and customer attributes. The information "coupon redeemed yes/no" and the basket value of the order are also known for the historical data.

Data

Real anonymous shop data in the form of structured text files consisting of individual data sets are provided for the task. Below are some points to note about the files:

1. Each data set is in a row of its own, ending with "CR" ("carriage return", 0xD) 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 pipe symbol ("|").
4. There is no escape character, quotes are not used.
5. ASCII is the character set used.
6. Missing values may occur.

In practice 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 attributes "coupon1Used", "coupon2Used", "coupon3Used" and "basketValue".

Entries

Participants can submit their results up to and including May 19, 2015 14:00 CEST (2 o'clock p.m. UTC+2, or CEST). The task below explains how to submit entries.

Task

Historical data used in generating coupons over a period of several weeks is known for the task. One coupon applies to a single product. In addition, the orders are still given in response to the generation including the information as to which coupons have been redeemed and the total basket value of each order. Using this data, a model for predicting coupon redemption and the total basket value should be learned. The target attributes "coupon1Used", "coupon2Used" and "coupon3Used" are described with the value "0" for coupons not redeemed and the value "1" for coupons redeemed. The remaining target attribute "basketValue" contains the total basket value of the order in the form of a real number. An analysis as to whether the coupons will be redeemed as well as the total basket value should be made for a portion of the coupon generation. Predictions are given for each order. In the case of "coupon1Used", "coupon2Used" and "coupon3Used", they should lie in the interval [0,1] and in the case of the total basket value they should be able to accept any real number. The errors compared to actual values should be as low as possible.

A file in the following format should be used to convey the solution data:

Column name	Description	Value range
orderID	Order number	Natural number
coupon1Prediction	Prediction as to whether the first coupon will be redeemed	Real number from [0,1]
coupon2Prediction	Prediction as to whether the second coupon will be redeemed	Real number from [0,1]
coupon3Prediction	Prediction as to whether the third coupon will be redeemed	Real number from [0,1]
basketValuePrediction	Prediction of total basket value	Real number

Each consecutive number of order items from the classification data must occur exactly once. The file should continue to 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|coupon1Prediction|coupon2Prediction|coupon3Prediction|basketValuePrediction
6054|0.4|0.1|0.9|167.85
6055|0.95|0.2|0.0|325
…

The results file must ultimately be sent as an attachment to an email to dmc_task@prudsys.de as a zipped text file. The email must reach prudsys AG by the abovementioned entry deadline. Please take into account potential delays in sending the email and send your solution in a timely manner!

The name of the zip file and the included text file must be made up of the team name and the type:

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

The team name will have 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:

$$E = \sum_{i=1}^{n} \left(\left(\frac{|coupon1Used_i - coupon1Prediction_i|}{\frac{1}{n}\sum_{j=1}^{n} coupon1Used_j} \right)^2 + \left(\frac{|coupon2Used_i - coupon2Prediction_i|}{\frac{1}{n}\sum_{j=1}^{n} coupon2Used_j} \right)^2 + \left(\frac{|coupon3Used_i - coupon3Prediction_i|}{\frac{1}{n}\sum_{j=1}^{n} coupon3Used_j} \right)^2 + \left(\frac{|basketValue_i - basketValuePrediction_i|}{\frac{1}{n}\sum_{j=1}^{n} basketValue_j} \right)^2 \right)$$

Here, $coupon1Used_i$ is the information as to whether or not the first coupon was redeemed in order i (0 means the coupon was not redeemed, 1 means the coupon was redeemed) and $coupon1Prediction_i$ is the probability of redemption predicted by the team for the first coupon in the order i. The terms for the second and third coupons are defined in the same way. The term $basketValue_i$ indicates the actual total basket value of the order i, whereas $basketValuePrediction_i$ contains the predicted total basket value for the order i. The variable n indicates the number of data sets included. The error function is calculated based on the data sets in the classification file. The teams whose error functions have the smallest values will win. If two teams are equal a coin flip will decide the winner.