DATA MINING CUP 2019

Fraud detection at self-checkouts in retail

The number of self-checkout stations is on the rise. This includes stationary self-checkouts, where customers take their shopping cart to a scan station and pay for their products. Secondly, there are semi-stationary self-checkouts, where customers scan their products directly and only pay at a counter. The customers either use their own smartphone for scanning or the store provides mobile scanners. You will probably have encountered this already.

This automated process helps avoid long lines and speeds up the paying process for individual customers. But how can retailers prevent the trust they have placed in customers from being abused? How can they decide which purchases to check in an effort to expose fraudsters without annoying innocent customers?

This is the topic of this year’s DATA MINING CUP.

Scenario

An established food retailer has introduced a self-scanning system that allows customers to scan their items using a handheld mobile scanner while shopping.

This type of payment leaves retailers open to the risk that a certain number of customers will take advantage of this freedom to commit fraud by not scanning all of the items in their cart.

Empirical research conducted by suppliers has shown that discrepancies are found in approximately 5 % of all self-scan transactions. The research does not differentiate between actual fraudulent intent of the customer, inadvertent errors or technical problems with scanners.

To minimize losses, the food retailer hopes to identify cases of fraud using targeted follow-up checks. The challenge here is to keep the number of checks as low as possible to avoid unnecessary added expense as well as to avoid putting off innocent customers due to false accusations. At the same time, however, the goal is to identify as many false scans as possible.

The objective of participating teams is to create a model to classify the scans as fraudulent or non-fraudulent. The classification does not take into account whether the fraud was committed intentionally or inadvertently.

To create this model, the teams receive information about the scans and their classification in a learning set.

Data

Real anonymized data in the form of structured text files (csv) 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. Floating point numbers are not rounded. 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 *“DATA-MINING-CUP-2019-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 data sets based on which the models are to be created are listed in the “*train.csv” file*.

The data sets in the “*test.csv*” file should be used for the prediction.

A single data set in the files “*train.csv”* or “*test.csv“* contains information about one scanning process for one customer, including the number of scanned items per second and the total time spent in the store.

For learning purposes, the “*train.csv“* file also contains a column with the classification into fraud and not fraud.

Entries

Participants may submit their results up to and including **May 16, 2019, 15:00 CEST (2 p.m. UTC+2, or CEST)**. The task description below explains how to submit entries.

Task

Use historical data to create a mathematical model to reliably classify scans as fraudulent or not fraudulent.

Complete data from the testing period will be provided.

Data will also be provided for the subsequent testing period.

The files "*train.csv*" and "t*est.csv*" are identical in structure but differ in that the classification column is missing from the latter.

Both files comply with the properties listed in the **Data** section.

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

Column name	Description	Value range
fraud	Fraud classification	{0, 1}

There is no key attribute. The number and order of the predicted classifications must correspond to those in the associated data sets or scans in the "*test.csv*" file.

The possible values for the "*fraud*" column are the integer values 0 or 1.

A possible excerpt from the solution file could look like this:

fraud
1
0
0
…

The solution file must comply with the specifications described in the **Data** section, as far as they are applicable. Incorrect or incomplete submissions cannot be evaluated.

The solution file must be uploaded as a structured text file (csv) to the DATA MINING CUP website at **https://www.data-mining-cup.com/dmc-2019/.**

Please use the password e4=d7QA{3End and make sure that the required fields on the form are filled out correctly and completely for the data upload.

The name of the text file consists of the team name and the file type:

"<Teamname>.csv" (e.g. TU_Chemnitz_1.csv)

The team name was communicated to the team leaders upon confirmation of registration.

Evaluation

The solutions submitted will be assessed and compared based on their monetary value for the food retailer. This can be calculated using the following cost matrix based on empirical observation.

Prediction	**Actual value**		
		0 (no fraud)	**1** (fraud)
	0 (no fraud)	*€ 0.0*	*€ -5.0*
	1 (fraud)	*€ -25.0*	*€ 5.0*

Thus, the food retailer receives a profit of € 5 for every correctly identified fraud attempt. However, for every fraud case that goes unexposed he loses € 5.

A costumer falsely accused of fraud, might not return to this store, which is denoted by a € 25 loss for the retailer.

An honest customer identified correctly means neither profit nor loss for the retailer.

The sum of the costs or profit of all scans is the monetary value of the submitted solution.

The winning team is the one whose solution achieves the highest monetary profit. In the event of a tie, a random draw will determine the winner.