Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft

Analytics Project 2025

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.

This automated process 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 to expose fraudsters without annoying innocent customers?

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 expenses 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 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.

Variables

Variable name	Description	Value range
trustLevel	A customer's individual trust level. 6: Highest trustworthiness	1-6
totalScanTimeInSeconds	Total time in seconds between the first and last product scanned	Positive whole number
grandTotal	Grand total of products scanned	Positive decimal number with maximum two decimal places
lineItemVoids	Number of voided scans	Positive whole number
scansWithoutRegistration	Number of attempts to activate the scanner without actually scanning anything	Positive whole number or 0
quantityModification	Number of modified quantities for one of the scanned products	Positive whole number or 0
scannedLineItemsPerSecond	Average number of scanned products per second	Positive decimal number
valuePerSecond	Average total value of scanned products per second	Positive decimal number
lineItemVoidsPerPosition	Average number of item voids per total number of all scanned and not cancelled products	Positive decimal number
fraud	Target variable: Classification as fraud (1) or not fraud (0)	0 or 1

There are missing values which must be treated adequately. Perhaps not all attributes contribute to the classification.

Try out at least three different classification algorithms and compare them.

What is the business aspect of the problem?

Present your findings with the help of data story telling in a paper and an on-site presentation.

You can find the dataset on moodle. It consists of over 400'000 records.

Source: Data Mining Cup 2019 (adapted)