Prof. Dr. J. Jacobs
Betriebliche Informationssysteme im Marketing
Wintersemester 2020

Project Description

A food retailer wants to optimize his self-scanning system that allows customers to scan their items using a handheld mobile scanner. Some customers commit fraud by not scanning all of the items in their cart. The food retailer wants to identify those frauds by targeted follow-up checks. The challenge is to keep the number of checks as low as possible to avoid unnecessary added expense as well as to avoid affronting innocent customers due to false accusations. For this purpose the retailer has collected historical data (see description on last page).

The objective is to develop a software system, which uses the historical data to reliably classify scans as fraudulent or not fraudulent.

Project Task:

a) Develop a fraud detection system (see description above) which optimizes the balanced accuracy. The system has to be developed with the R package caret. Please prefer algorithms which have been presented in the lectures. The training data is contained in the file train.txt encoded in Windows-1252 (ANSI). The system shall be able to process unseen records provided in a file unseen.txt containing the same attributes (with the exception of target attribute in the last column) in the same order and encoding as in the file train.txt and containing about the same percentage of frauds. The system output shall be a text file classified.txt containing all records of the file unseen.txt (**all rows and columns in the same order**) completed by the predicted value of the target attribute.
b) Write a project documentation covering the following topics:
    (1) data pre-processing,
    (2) algorithm(s) used,
    (3) training results obtained,
    (4) system user guide.

The project files and documentation shall be delivered via eMail (documentation in hardcopy as well) by Friday, January 15th, 2021.

Description of data:

Header	Description	Value range	Missing Values
trustLevel	A customer's individual trust level. 6: Highest trustworthiness	1,2,3,4,5,6	Yes
totalScanTimeInSeconds	Total time in seconds between the first and last product scanned	Positive whole number	No
grandTotal	Grand total of products scanned	Positive decimal number with maximum two decimal places	No
lineItemVoids	Number of voided scans	Positive whole number	No
scansWithoutRegistration	Number of attempts to activate the scanner without actually scanning anything	Positive whole number or 0	No
quantityModifications	Number of modified quantities for one of the scanned products	Positive whole number or 0	No
scannedLineItemsPerSecond	Average number of scanned products per second	Positive decimal number	No
valuePerSecond	Average total value of scanned products per second	Positive decimal number	No
lineItemVoidsPerPosition	Average number of item voids per total number of all scanned and not cancelled products	Positive decimal number	No
fraud	Classification as fraud (1) or not fraud (0)	0,1	No