DATA MINING CUP 2019 – HS_Karlsruhe_1

Approach

Feature Engineering

A new feature called *totalScannedLineItems* (the product of *scannedLineItemsPerSecond* and *totalScanTimeInSeconds*) is added. Training with this feature leads to a better separability of frauds and non-frauds.

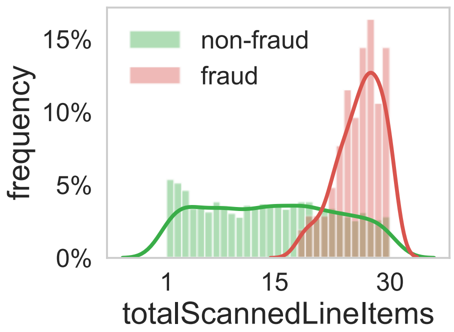


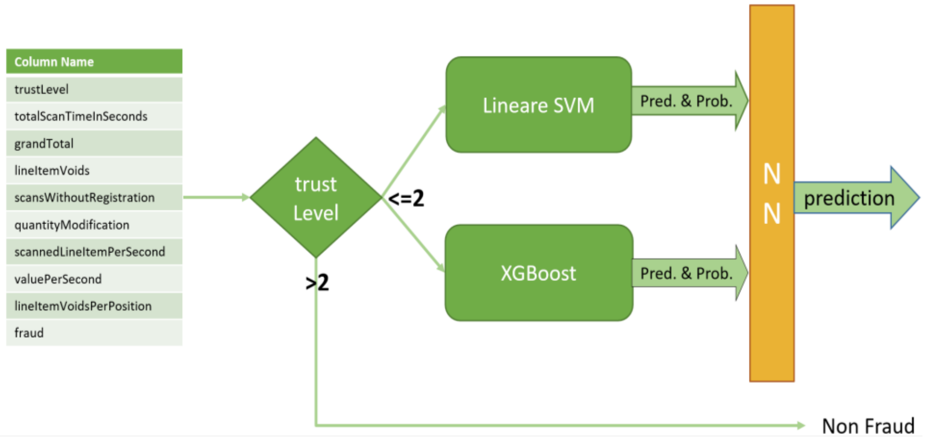
Figure 1: Class distribution of new feature totalScannedLineItems

Data Pre-Processing

Some of the used classifiers are scale variant. Therefore, both the original dataset and a scaled version are kept. To prevent overfitting, the datasets are also being split into multiple subsets (as explained later).

The Models

Two base classifiers (a linear support vector machine (SVM) and a gradient boosting classifier) and an additional shallow neural network are used to process the resulting predictions and their probabilities to predict the final classes (fraud/non-fraud). Both base classifiers are being trained independently with the training set and a subset of the test set (→ *Methods*).



Column Name
trustLevel
totalScanTimeInSeconds
grandTotal
lineItemVoids
scansWithoutRegistration
quantityModification
scannedLineItemPerSecond
valuePerSecond
lineItemVoidsPerPosition
fraud

Figure 2: Classification process

Methods

Semi-Supervised Learning (SSL)

One of the main obstacles encountered during the competition is the small size of the training dataset. To get more training data, a semi-supervised learning approach called pseudo labeling is being used. Predicted test samples are combined with existing training data to create a new training dataset. This approach does not only increase the accuracy, but it also makes the model more robust.

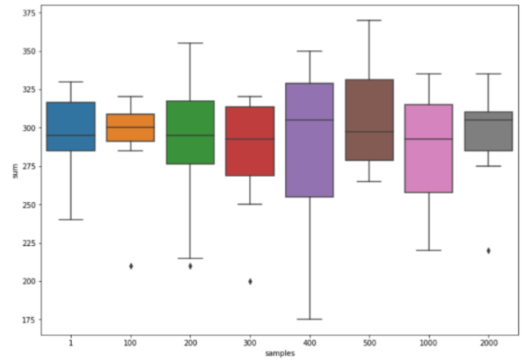


Figure 3: DMC Score for different test sample sizes

Linear Support Vector Machine

Support Vector Machines (SVMs) try to find the biggest margin between two classes to separate them. It may be possible that the SVM misaligns the class border into a sparsely occupied feature space. SSL occupies those by adding new training data from the test set. In general, SVMs perform better on scaled datasets, so the scaled version of the data is used.

Gradient Boosting (GB) Algorithm

For the gradient boosting classifier, a tree boosting algorithm is used as a base. In the case of this competition, the GB algorithm performs better on unscaled data.

Validation

Multiple validation routines are used to prevent choosing a mistakenly good performing model. Besides cross validation an additional train/validation-split is used to validate the final model on completely unseen data.

General

Training Data

Comparing the sizes of the training (1,900) and test (500,000) dataset was one of the first things that was done. The huge size difference has led to the idea of additionally using the test data for the classification. It was decided to use Semi-Supervised Learning which allows to take a subset of the test set, predict its rows labels and use those predictions for another training. Different test sample sizes were tested, and it turned out that an addition of about 500 test samples leads to the best improvements. With higher sample sizes, the weighting would shift too far from the original training data.

Classifier Selection

The two classifiers are chosen because they complement each other very well. There are only a few rare cases in which they both choose the wrong class.

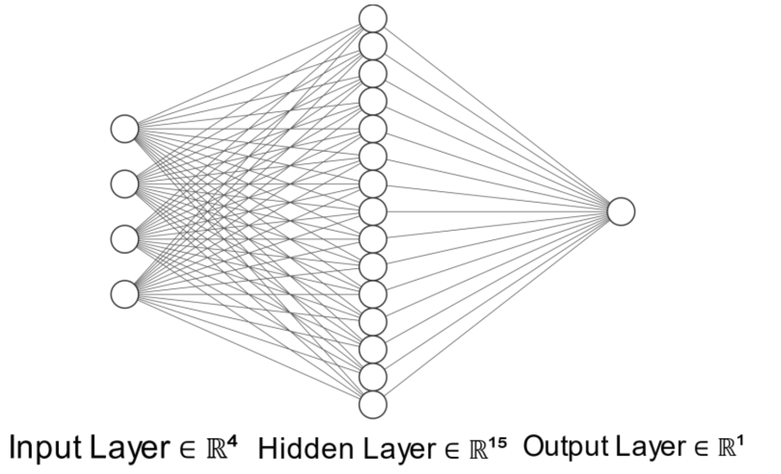


Figure 4: Architecture of the shallow neural net for final classification

DATA MINING CUP 2019
International Student Competition

University	Hochschule Karlsruhe
Country	Germany (DE)
Team Leader	Alexander Melde
Team Member	Lukas Theurer, Christian Wernet