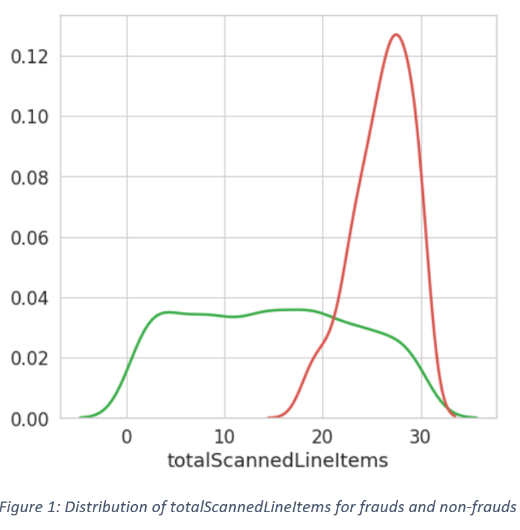
**Approach**

DATA MINING CUP 2019 – HS\_Karlsruhe\_1

**Feature Engineering and Data Preprocessing**

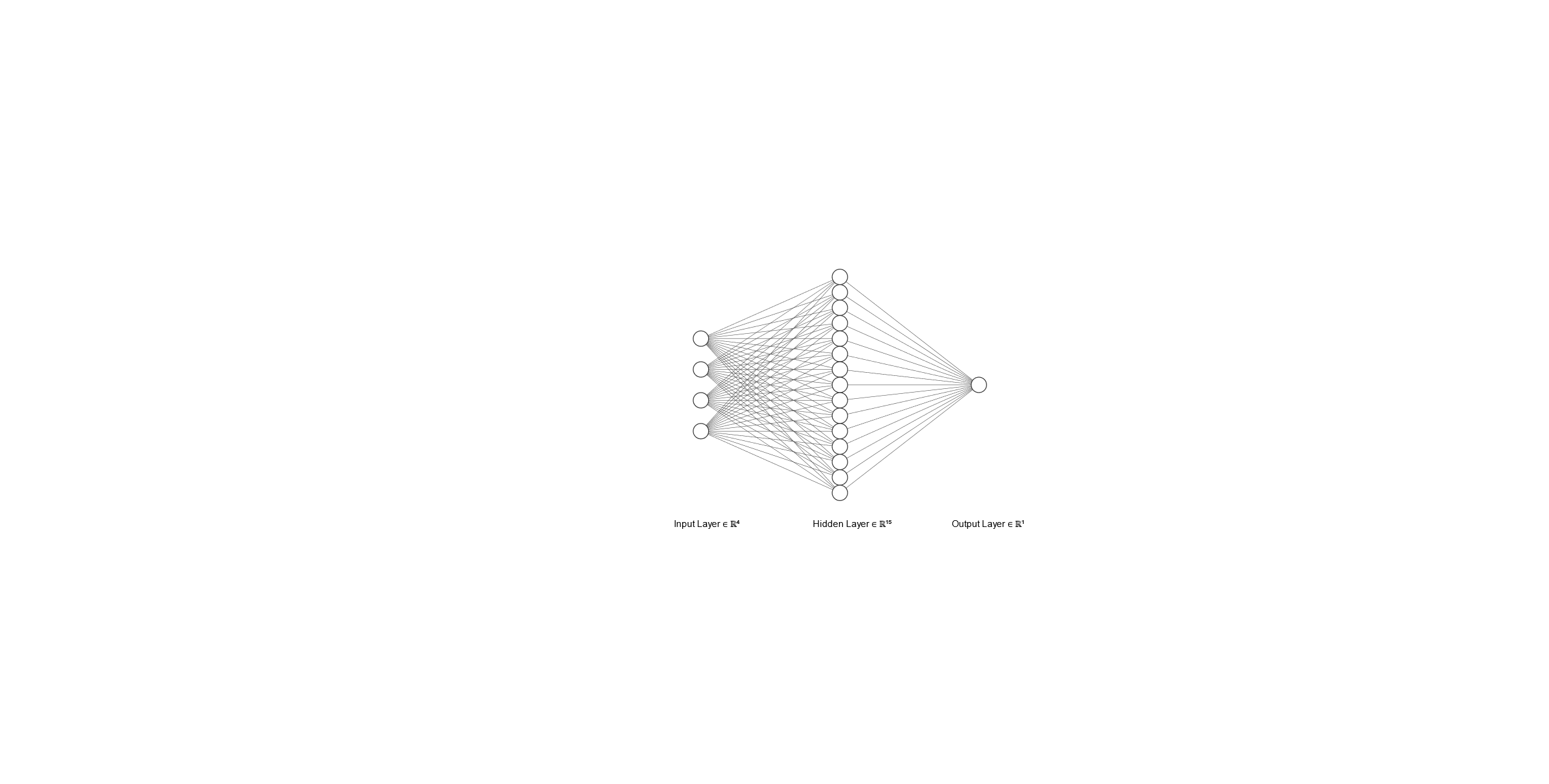
Adding a new feature *totalScannedLineItem*s which is the result of the product between *scannedLineItemsPerSecond* and *totalScanTimeInSeconds*leads to better seperationability of frauds and non-frauds.

Because of the susceptibility of a few classifiers against different scales, we need to keep the dataset a second time, scaled with mean and variance.

A split of the training data is also required, to ensure that the classifiers do not overfit.

**The Used Models**

A shallow neural net is used to process the predictions and prediction probabilities of two base classifiers to predict the final class (fraud/non-fraud).



The base classifiers are a linear support vector machine and a gradient boosting classifier. Both were trained independently with the training set and a subset of the test set.

**Methods**

**Linear Support Vector Machine**

Support vector machines tries to find the biggest margin between two classes to separate them. It could be possible, to misalign the class border into sparsely occupied feature space. The semi-supervised approach should avoid this by adding new training data from the test set. The linear SVM performs a lot better with scaled data, whereas the gradient boosting algorithm performs not as good as with unscaled data.

**Gradient Boosting Algorithm**

For the gradient boosting classifier, a tree boosting algorithm is used as base. The both classifiers where chosen, because the complement each other. Only in some rare cases, they make the same mistake an choose the false class.

**Semi-Supervised Learning**

One of the main issues where the small size of the training dataset. To get more training data, a semi-supervised learning approach is used. This way, a few of the test samples were used together with pseudo labels, predicted by the model. This approach not only increased the accuracy, it also make the model more robust.

**Validation**

Multiple validation pipelines prevent us from choosing a mistakenly good performing model. Besides cross validation we used an additional train/val-split to make sure that the good result is not only based on the training data.

**General**

**Training Data**

The sizes of the training and test dataset (1,9k vs. 500k) were an early indicator that we need to apply additional methods to use this large test set for training. Semi-Supervised Learning allows us to take a subset of the test set, predict the labels and using the predictions for training. It turned out that an addition of about 500 test samples could further improve the algorithm. If more test data is added, the weighting shifts too far away from the original training data.

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