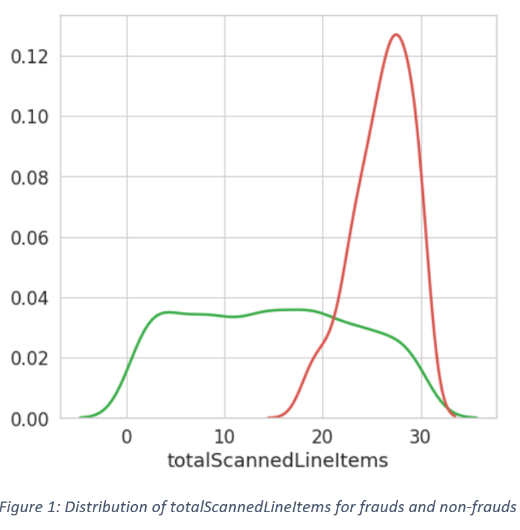
**Approach**

DATA MINING CUP 2019 – HS\_Karlsruhe\_1

**Feature Engineering and Data Preprocessing**

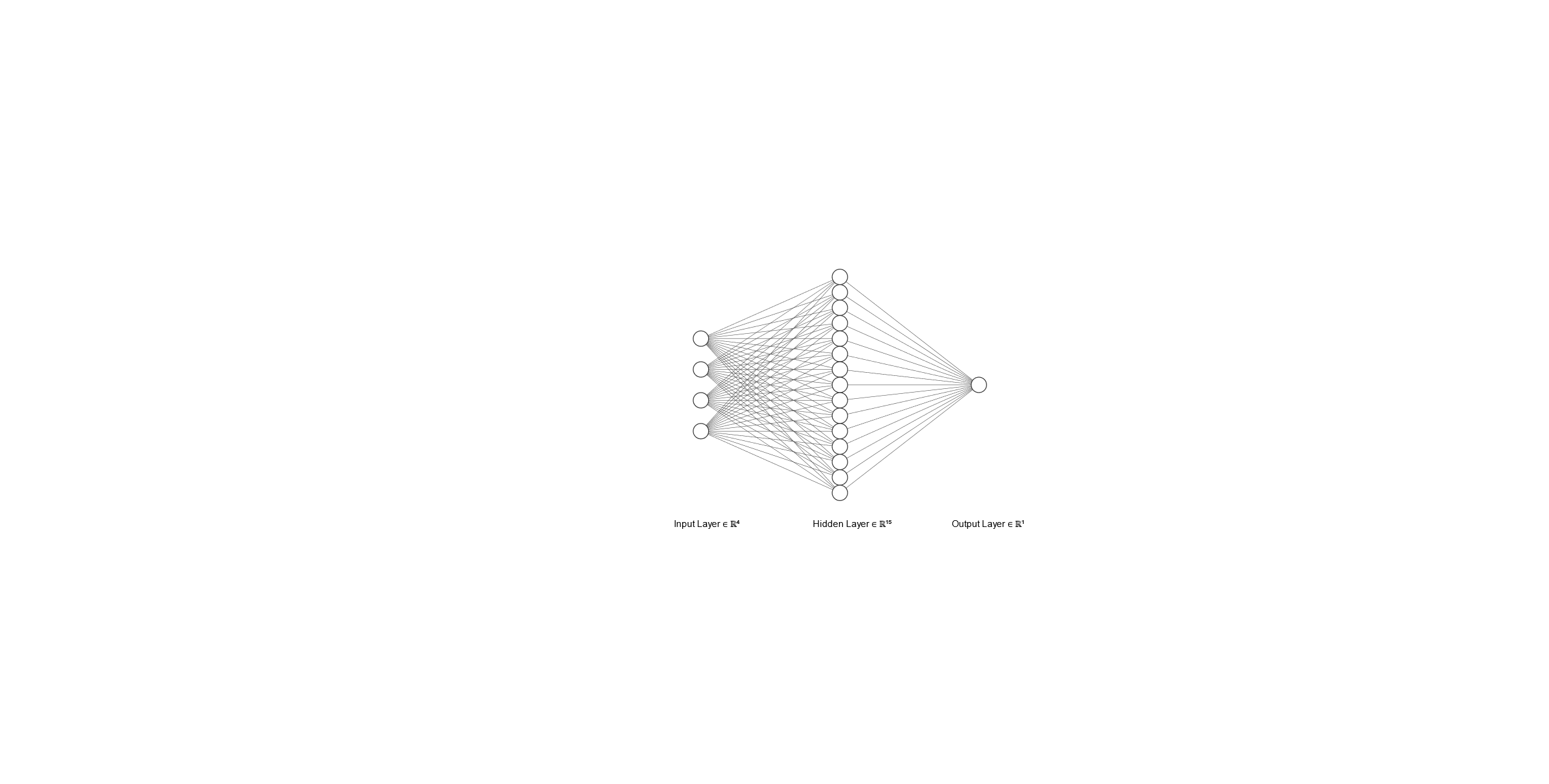
Adding a new feature *totalScannedLineItem*s, which is the result of the product between *scannedLineItemsPerSecond* and *totalScanTimeInSeconds,*leads to better seperationability of frauds and non-frauds.

Because of the susceptibility of a few classifiers to different scales, we need to keep the original dataset and a version 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**

**Semi-Supervised Learning**

One of the main issues was 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 makes the model more robust.

**Linear Support Vector Machine**

Support vector machines 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. A semi-supervised approach prevents this by adding new training data from the test set. In addition, the dataset was scaled, because the linear SVM performs a lot better with scaled data

**Gradient Boosting Algorithm**

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

The two classifiers where chosen, because they complement each other. Only in some rare cases, they make the same mistake by choosing the false class.

**Validation**

Multiple validation routines are used to 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 use the predictions for training. It turned out, that an addition of about 500 test samples further improved the algorithm. If more test data was added, the weighting would shift too far from the original training data.

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