# Data Science and Machine Learning in Python

## Assignment 6 – Fraud Detection in Retail

Stephan Weyers



Fachhochschule Dortmund
University of Applied Sciences and Arts

# Assignment 6

Fachhochschule Dortmund
University of Applied Sciences and Arts

**Provided files**

- W06_training.txt (labelled training dataset)
- W06_scoring.txt (unlabelled dataset used for evaluation)
- W06_Results.xlsx
- W06_Task.pdf (this document)

**Instructions**

- Work on this assignment with the team mates who have been assigned to you
- Follow the instructions on the next pages
- Solve the tasks using Python and Jupyter Notebook
- Fill in the results into the provided Excel-file
- Evaluate the contributions of all team members also in the Excel file
- Submit both ipynb and xlsx files in ILIAS

**Due date**

- June 28th (23:59 German time)

# Assignment 6 – Exercise 1



- Import the data set "W06_training.txt“
- Use only the first 7 columns (i.e. ignore the „Fraud“ column for now)
- Divide each column by its maximum value, so that each each column is scaled between 0 and 1
- Perform a principal component analysis. Report the explained variance ratio and the components of the first two principal components
- Perform a k-means cluster analysis to derive a solution with 3 clusters. For comparison use KMeans(n_clusters=3, random_state=100) to always get the same results. Report the components of the cluster centers and the size of the clusters (i.e. the number of data points assigned to each cluster)

Please note: Exercise 1 is independent of exercise 2. You might or might not use some of the methods of exercise 1 to solve exercise 2, this is completely up to you. Exercise 1 is a pure technical guided task, while exercise 2 is an open problem, for which you are allowed and need to be creative to get to a very good solution

# Assignment 6 – Exercise 2: Context

Fachhochschule Dortmund
University of Applied Sciences and Arts

Your client is a grocery retailer facing increasing pressure from expanding competition. To reduce personnel costs while at the same time improving service quality, it has been relying on self-service checkout stations for some time now. At these stations, customers can scan their products themselves and pay directly. The self-checkout stations help avoid long queues and speed up the payment process for individuals. However, it also offers the opportunity to cheat when scanning the products.

In order to identify fraudulent actions without upsetting innocent customers through controls, the company has commissioned you and your team as an external consultancy to develop a model that determines the probability of fraud in self-checkout purchases. Your assignment is now to use the collected data and develop a high-performance model for predicting or detecting fraud and to identify the key indicators for fraudulent behavior.

When creating the classification model the following costs and revenues should be considered:

| | | **Reality** | |
|---|---|---|---|
| | | No fraud | Fraud |
| **Prediction** | No fraud | 0 EUR | -5 EUR |
| | Fraud | -25 EUR | 5 EUR |

The table shows that each correctly identified fraud attempt brings in an average of EUR 5 additional revenue. Each fraud attempt that is not detected, however, causes 5 EUR costs. Customers who are falsely accused of fraud may not return, which means an average loss of 25 EUR for the supermarket. Correctly identified honest customers mean neither profit nor loss.

# Assignment 6 – Exercise 2: Task

Fachhochschule Dortmund
University of Applied Sciences and Arts

The aim of the analysis is to use the data set of 300,000 cases (W06_training.txt) to train a model that is suitable for detecting fraud attempts. The prediction of your model is finally checked with the help of another data set with 100,000 purchases for which you do not know the target variable. This data set (W06_scoring.txt) is used to evaluate how well your model's prediction works, using the total cost or total revenue. This means that you must ensure that there is no overfitting when training your model, otherwise the prediction on the new data set will give poor results. To do this, you should split your data set into training and test data or use a suitable cross-validation method to avoid overfitting.

Please proceed as follows:

1. Get an overview and understanding of the available data
2. Clean the data if necessary and derive new "smart" variables (e.g. scale the data or compute ratios of given variables) that can be used for the predictions
3. Determine which characteristics are key characteristics as they are driving fraudulent behavior
4. Use the appropriate classification algorithms you are familiar with and make predictions. If necessary, tune the parameters of the model
5. Measure the quality of the model or compare the quality of the models and choose a final model
6. Use statistical key figures, but especially the total costs or the total revenue
7. Hand-in your fraud predictions (0/1) for each of the 100,000 cases of the scoring dataset

# Assignment 6 – Data

Fachhochschule Dortmund
University of Applied Sciences and Arts

| Column | Description |
|---|---|
| LevelOfTrust | The trust level of a customer.<br>6: Highest trust |
| ScanTimeInSeconds | Seconds between the first and last product scanned |
| TotalBasketValue | Total EUR value of products scanned |
| ScannedProducts | Number of scanned products |
| CountOfVoidedScans | Number of voided scans |
| ActivateWithoutScan | Number of scanner activations without actually scanning anything |
| ModifiedQuantities | Number of modified quantities for one of the scanned products |
| Fraud | Classification as fraud (1) or not fraud (0) |