**Berlin 2019**

**Application for Fraud Analyst – Case Study**

The attached CSV-file (Latin-1 encoded) contains a dataset of online purchases:

As part of the recruitment process for the Fraud Analyst position we ask you to complete a

case study from home. The case does not contain real data, but the assignments are similar

to actual analysis assignments. We are looking for your ability to analyse data, explain the

results and how they relate to fraud.

* **2 ids:** “Transaction\_id” and “Customer\_id”
* **12 variables**
* **1 target:** “is\_fraud”

Imagine the dataset is created by first collecting 12 variables for each purchase. For each

purchase a new transaction id is provided. The customer info is then matched and if no

match is found a new customer id is provided. Afterwards an assessment of fraud is

performed (is\_fraud).

The assignment is divided in 4 parts:

1. **Basic analysis:** 3 simple questions
2. **Trend analysis:** Single variable analysis
3. **Fraud rule creation:** Threshold setting
4. **Variable creation:** Multiple variable and advanced analysis

We expect you to just provide the numbers in assignment 1, but for assignments 2-4 see the

assignments as first a data question and then an explanation / communication of the result.

Feel free to represent your findings graphically and / or use tables. The dataset contains

special characters - if they cause trouble just ignore them and focus on the data and explanations.

You can review and process the dataset using any software of your choice e.g. Excel, SQL,

SAS, R, Matlab, python etc. We expect the final hand in to be between 2-4 pages.

1. **Basic Analysis**

What is the proportion (%) of fraud to total transactions that meet the following criteria?

1. Purchase in the Clothing Segment
2. Purchases Greater than 1000 SEK
3. Purchases coming from UK
4. **Trend Analysis**

We are interested in finding categories indicative of fraud

1. Which three data subcategories in the 12 variables (e.g. type\_of\_goods\_ = books or ip\_country=SE) have the highest value (SEK) of fraudulent transactions? Example - type\_of\_goods=”Tickets”, the sum of fraudulent transactions = 1.47M SEK.
2. Briefly explain why these subcategories would be indicative of fraud.
3. **Fraud rule creation**

Imagine you can only review a fraction of the transactions for fraud. You want to create

rules that pick out the transaction that are most likely to be fraud for review. A good

rules balances two parameters “hit rate” and “catch rate”

Hit rate = ratio of fraud to total transactions hit by rule. Catch rate = ratio of fraud of total fraudulent.

Example: You have 100 transactions, 20 are fraud. You make a rule that suspect 10 cases

where 5 are fraud. Hit rate = 5/10 = 50%, Catch rate = 5/20 = 25%.

1. What is the hit and catch rate for the 3 data subcategories from assignment 2.
2. Try to combine variables and use cut-offs / sub-categories to improve hit rate and catch rate (provide at least 2 examples of fraud rules).
3. Explain why both hit rate and catch rate are relevant parameters? What others

factors might be important when creating and assessing new fraud rules?

1. **Variable creation**

Some information is very indicative of fraud but hard to extract from the raw variables.

1. Look into how variables (customer\_id included) can be combined or processed to form new variables that are predictive of fraud?
2. Describe a process/method for processing and combining variables in general and how it relates to different types of fraud (stolen identity, providing fake information etc.)