Fraud in Electricity and Gas Consumption Dataset

|  |  |
| --- | --- |
| **Size** | 40 MB |
| **Dataset Characteristics:** | Multivariate |
| **Attribute Characteristics:** | Numerical, Categorical, Ordinal |
| **Associated Tasks:** | Classification |
| **Number of Instances:** | Around 21000 clients in total |
| **Number of Attributes:** | 5 features in “client.csv”, 16 features in “invoice.csv” |
| **Number of Classes** | 2 (“target” column of “client.csv”) |
| **Area:** | Business |
| **Additional Details:** | highly imbalanced dataset (very low number of fraud cases) |

# Dataset Information:

This dataset contains the client information for over 21000 people, each of them can have 3-50 invoices (see “invoice.csv”). The common column between the two csv files is “id”. Your job is to analyze the pattern of their consumption and find out which clients are fraud. Remember that you need to classify the client and not the transaction (a client performing 3-50 transactions). You are expected to apply various innovative techniques in – (1) handling the data imbalance issue and (2) summarizing the variable number of transaction info for each client.

Attribute Information:

## client .csv

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Type** |
| id | Unique id for client | str |
| dis | The district where the client is | int |
| catg | Category client belongs to | int |
| region | Area where the client is | int |
| date | Date client joined | str |
| target | fraud:1, not fraud: 0 | int |

**invoice.csv**

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Type** |
| id | Unique id for the client | str |
| date | Date of the invoice | str |
| Tarif\_type | Type of tax | int |
| counter\_number | Number | float |
| counter\_statue | takes up to 5 values such as working fine, not working, on hold statue, ect | int |
| counter\_code |  | int |
| reading\_remarque | notes that the STEG agent takes during his visit to the client (e.g: If the counter shows something wrong, the agent gives a bad score) | int |
| counter\_coefficient | An additional coefficient to be added when standard consumption is exceeded | int |
| consommation\_level\_1 | Consumption\_level\_1 | int |
| consommation\_level\_2 | Consumption\_level\_2 | int |
| consommation\_level\_3 | Consumption\_level\_3 | int |

|  |  |  |
| --- | --- | --- |
| consommation\_level\_4: | Consumption\_level\_4 | int |
| old\_index: | Old index | int |
| new\_index: | New index | int |
| months\_number: | Month number | int |
| counter\_type: | Type of counter | str |