Fraud detection

by Morad, Daniel and Tom

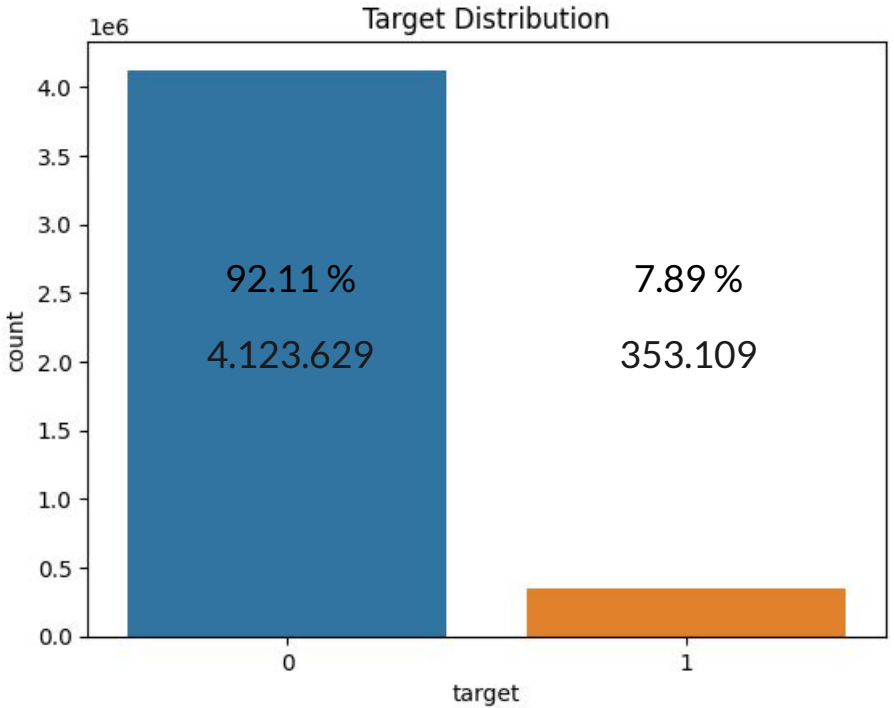
Predicting fraudulent clients

- The Tunisian Company of Electricity and Gas (STEG) is a public and a non-administrative company, it is responsible for delivering electricity and gas across Tunisia.
- The company suffered tremendous losses in the order of 200 million Tunisian Dinars due to fraudulent manipulations of meters by consumers.

EDA

Some insights into the data:

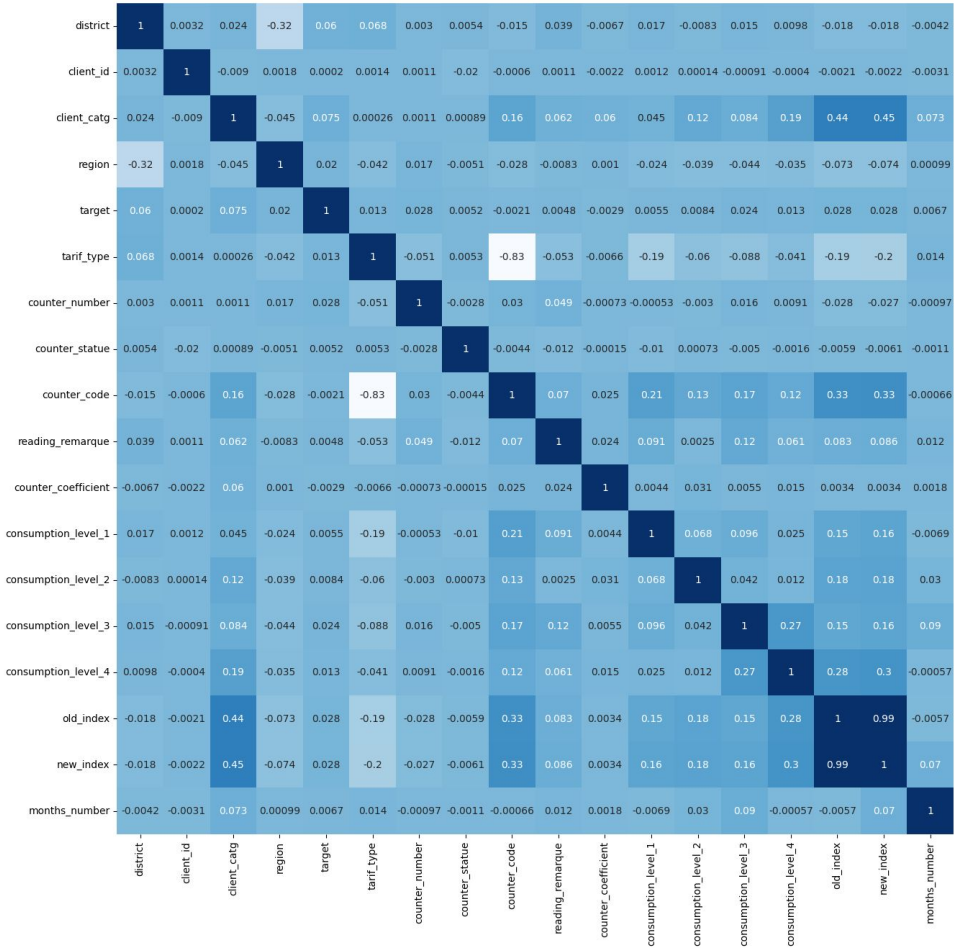
- The data is unbalanced
- upsampling or downsampling the data?



EDA

Correlations with target:

target	1.000000
client_catg	0.074530
district	0.059542
new_index	0.028188
counter_number	0.028005
old_index	0.027520
consumption_level_3	0.023713
region	0.019523
tarif_type	0.013384
consumption_level_4	0.013094
consumption_level_2	0.008421
months_number	0.006669
consumption_level_1	0.005515
counter_statue	0.005183
reading_remarque	0.004769
client_id	0.000201
counter_code	-0.002109
counter_coefficient	-0.002853



EDA

Correlations with target:

- target 1.000000
- client_catg 0.074530
- district 0.059542
- new_index 0.028188
- counter_number 0.028005
- old_index 0.027520
- consumption_level_3 0.023713
- region 0.019523
- tarif_type 0.013384
- consumption_level_4 0.013094
- consumption_level_2 0.008421
- months_number 0.006669
- consumption_level_1 0.005515
- counter_statue 0.005183
- reading_remarque 0.004769
- client_id 0.000201
- counter_code -0.002109
- counter_coefficient -0.002853

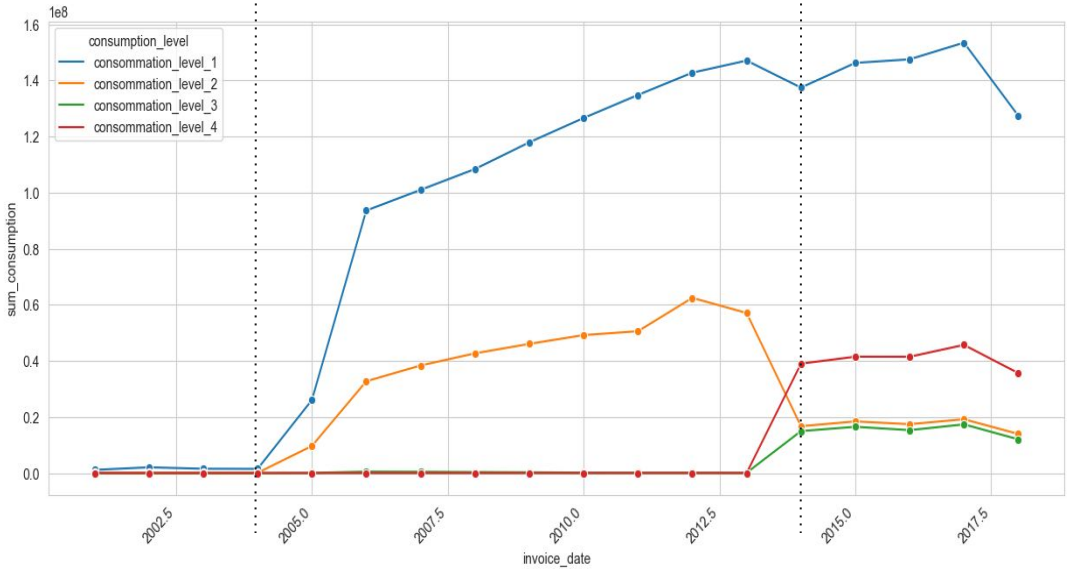
client_catg	11	12	51	
target				
0	92.391495	94.413284	79.046272	%
1	7.608505	5.586716	20.953728	%

district	60	62	63	69	
target					
0	95.1773	92.737725	90.62507	90.01014	%
1	4.8227	7.262275	9.37493	9.98986	%

EDA

Correlations with target:

target	1.000000
client_catg	0.074530
district	0.059542
new_index	0.028188
counter_number	0.028005
old_index	0.027520
consumption_level_3	0.023713
region	0.019523
tarif_type	0.013384
consumption_level_4	0.013094
consumption_level_2	0.008421
months_number	0.006669
consumption_level_1	0.005515
counter_statue	0.005183
reading_remarque	0.004769
client_id	0.000201
counter_code	-0.002109
counter_coefficient	-0.002853

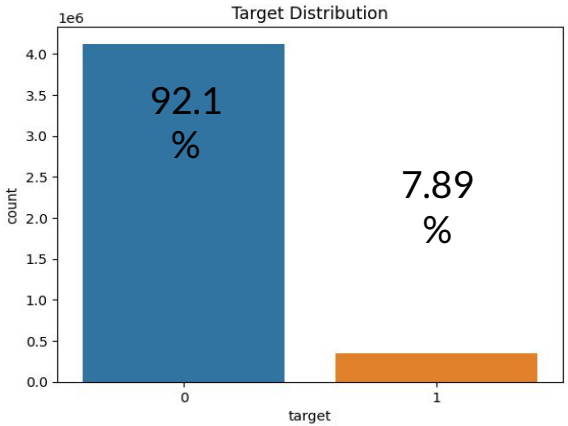


Feature Engineering

Creation of new features:

- total consumption
- mean consumption per year
- historical mean consumption

Sampling



Trying different sampling methods:

- Upsampling: too much data -> takes too long to compute (~8 Million rows)
- Downsampling: loss of data (invoice history, from >4 Million rows to ~200.000) for non-fraudulent clients (Target 0)
- Customized sampling -> Downsampling by Client_id

Milestone 3:
Model 1 Error Analysis

- XGBOOST

Accuracy: 0.71

Precision: 0.71

Recall: 0.72

F1 Score: 0.72

ROC AUC Score: 0.71

Milestone 4:
Model 2 Error Analysis with ‘customized’ sampling

- XGBoost

Accuracy: 0.62
Precision: 0.66
Recall: 0.74
F1 Score: 0.70
ROC AUC Score: 0.60

Data product

- Predict fraudulent clients and make them pay to reduce losses for the Tunisian Company of Electricity and Gas

prediction per customer vs. prediction per invoice