Fraud detection

by Morad, Daniel and Tom

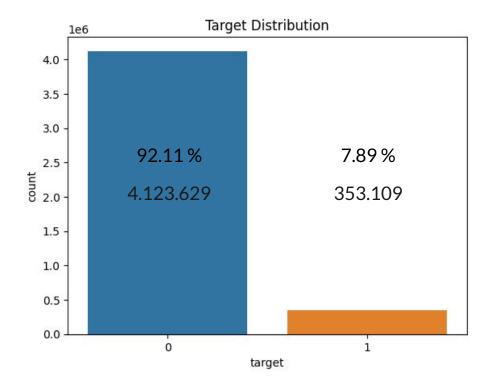
Predicting fraudulent clients

- 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?





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

district -		0.0032	0.024	-0.32			0.003	0.0054	-0.015	0.039	-0.0067	0.017	-0.0083	0.015	0.0098	-0.018	-0.018	-0.0042
client_id -	0.0032	1	-0.009	0.0018	0.0002	0.0014	0.0011	-0.02	-0.0006	0.0011	-0.0022	0.0012	0.00014	-0.00091	-0.0004	-0.0021	-0.0022	-0.0031
client_catg -	0.024	-0.009	1	-0.045		0.00026	0.0011	0.00089				0.045				0.44	0.45	
region -	-0.32	0.0018	-0.045	1	0.02	-0.042	0.017	-0.0051	-0.028	-0.0083	0.001	-0.024	-0.039	-0.044	-0.035	-0.073	-0.074	0.00099
target -		0.0002		0.02	1	0.013	0.028	0.0052	-0.0021	0.0048	-0.0029	0.0055	0.0084	0.024	0.013	0.028	0.028	0.0067
tarif_type -		0.0014	0.00026	-0.042	0.013	1	-0.051	0.0053	-0.83	-0.053	-0.0066	-0.19	-0.06	-0.088	-0.041	-0.19	-0.2	0.014
counter_number -	0.003	0.0011	0.0011	0.017	0.028	-0.051	1	-0.0028	0.03		-0.00073	-0.00053	-0.003	0.016	0.0091	-0.028	-0.027	-0.00097
counter_statue -	0.0054	-0.02	0.00089	-0.0051	0.0052	0.0053	-0.0028	1	-0.0044	-0.012	-0.00015	-0.01	0.00073	-0.005	-0.0016	-0.0059	-0.0061	-0.0011
counter_code -	-0.015	-0.0006		-0.028	-0.0021	-0.83	0.03	-0.0044	í		0.025							-0.00066
reading_remarque -	0.039	0.0011		-0.0083	0.0048	-0.053		-0.012	0.07	1	0.024		0.0025					0.012
counter_coefficient -	-0.0067	-0.0022		0.001	-0.0029	-0.0066	-0.00073	-0.00015	0.025	0.024	1	0.0044	0.031	0.0055	0.015	0.0034	0.0034	0.0018
consumption_level_1 -	0.017	0.0012	0.045	-0.024	0.0055	-0.19	-0.00053	-0.01			0.0044	1			0.025			-0.0069
consumption_level_2 -	-0.0083	0.00014		-0.039	0.0084	-0.06	-0.003	0.00073		0.0025	0.031	0.068	1	0.042	0.012			0.03
consumption_level_3 -	0.015	-0.00091		-0.044	0.024	-0.088	0.016	-0.005			0.0055		0.042	1				
consumption_level_4 -	0.0098	-0.0004		-0.035	0.013	-0.041	0.0091	-0.0016			0.015	0.025	0.012	0.27	1			-0.00057
old_index -	-0.018	-0.0021	0.44	-0.073	0.028	-0.19	-0.028	-0.0059			0.0034				0.28	1	0.99	-0.0057
new_index -	-0.018	-0.0022		-0.074	0.028	-0.2	-0.027	-0.0061			0.0034					0.99		
months number -	-0.0043	-0.0031	0.073	0.00099	0.0067	0.014	-0.00097	-0.0011	-0.00066	0.012	0.0018	-0.0069	0.03		-0.00057	-0.0057	0.07	1
monuis_number -	-0.0042	-5.0031		0.00039	0.0067	0.014	-0.00097	1	-0:00000	0.012	0.0018	0.0009	0.03	0.09		0.0037	0.07	•
	trict	Ę.	catg	lou	rget	.ype	. per	atne	ode	dne	ient		<u>=</u> 2	<u></u>	e 4	dex	xəp	per

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 $\overline{i}d$	0.000201
• counter_code	-0.002109
• counter_coefficient	-0.002853

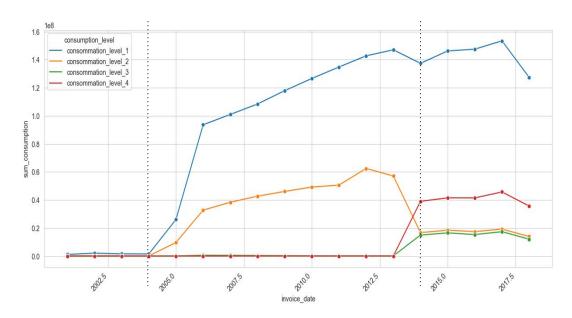
client_catg	11	12	51	
target				
0	92.391495	94.413284	79.046272	%
1	7.608505	5.586716	20.953728	%

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



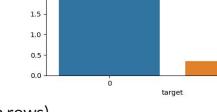
Feature Engineering

Creation of new features:

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

Sampling

Trying different sampling methods:



92.1

3.5

3.0

2.0

Target Distribution

7.89

%

- 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

