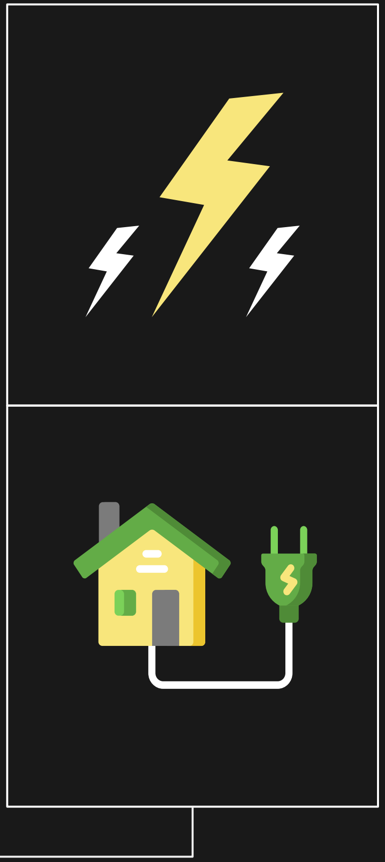
Fraud Detection in Electricity and Gas Consumption



Vanessa Lampe and Vadym Khvoinytskyi

Société Tunisienne
de l'Electricité et du Gaz

Introduction

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.

Goal

Our aim is to find fraudulent transactions, save money, avoid reputation damage and prevent money laundering.

Our data

- Contains billing history from 1977 till 2019
- Includes 135,493 clients and 4,476,749 invoices
- 4 features for client data and 17 for invoice data
- Target: fraud or not fraud

Steps to the best model

01.

EDA

Overview of data

02.

Feature engineering

Derive new insights from features

03.

Baseline model

Develop a simple model

04.

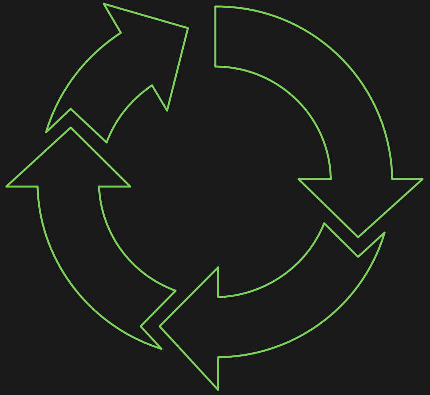
Modeling

Search for better solution for business problem

05.

Evaluation

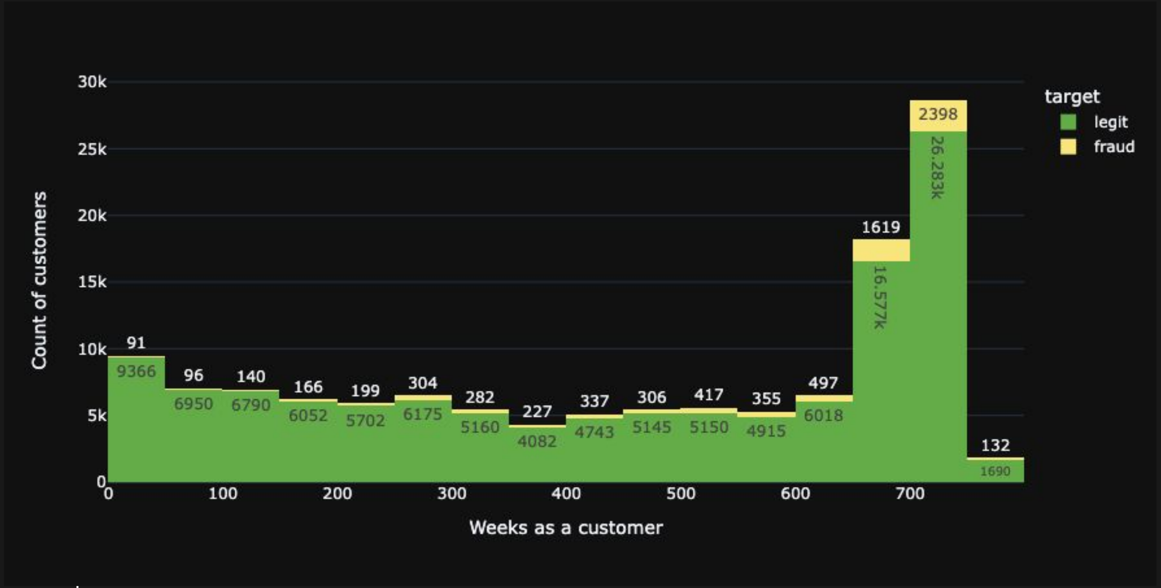
Assess model predictive ability



01.

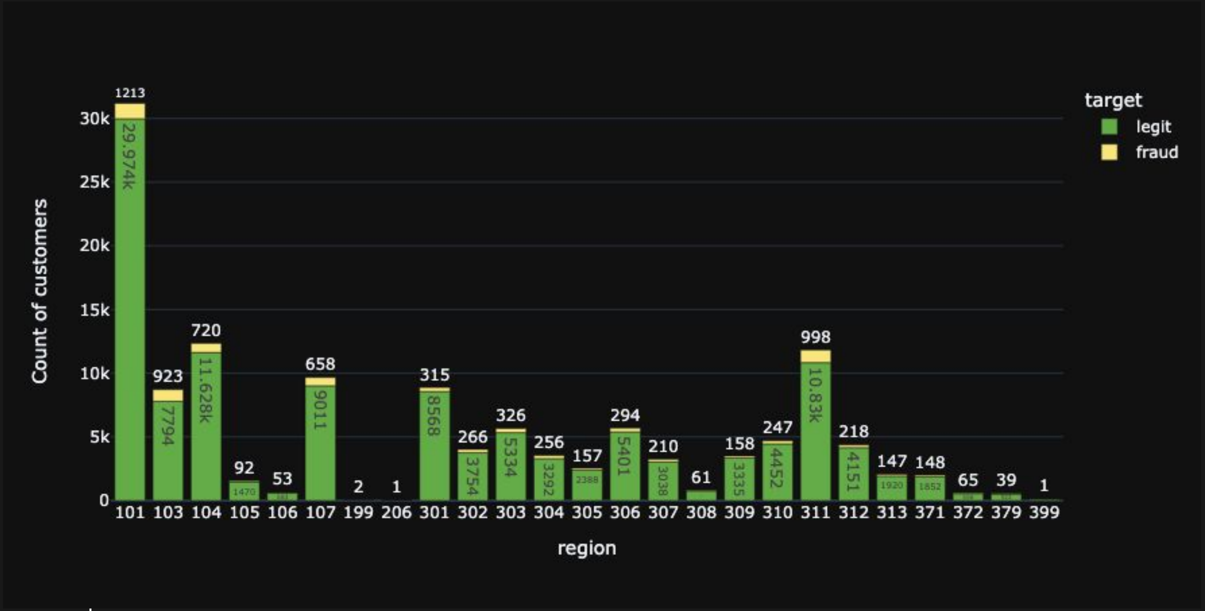
Exploratory Data Analysis

Loyal customers are less honest?



- 5.6% of customers have commited a fraud
- The percentage of frauders is highest for long-term customers

Is fraud a regional problem?



- Certain regions exposed to higher percentage of fraudsters
- Customers who use both gas and electricity commit more fraud

02.

Feature Engineering

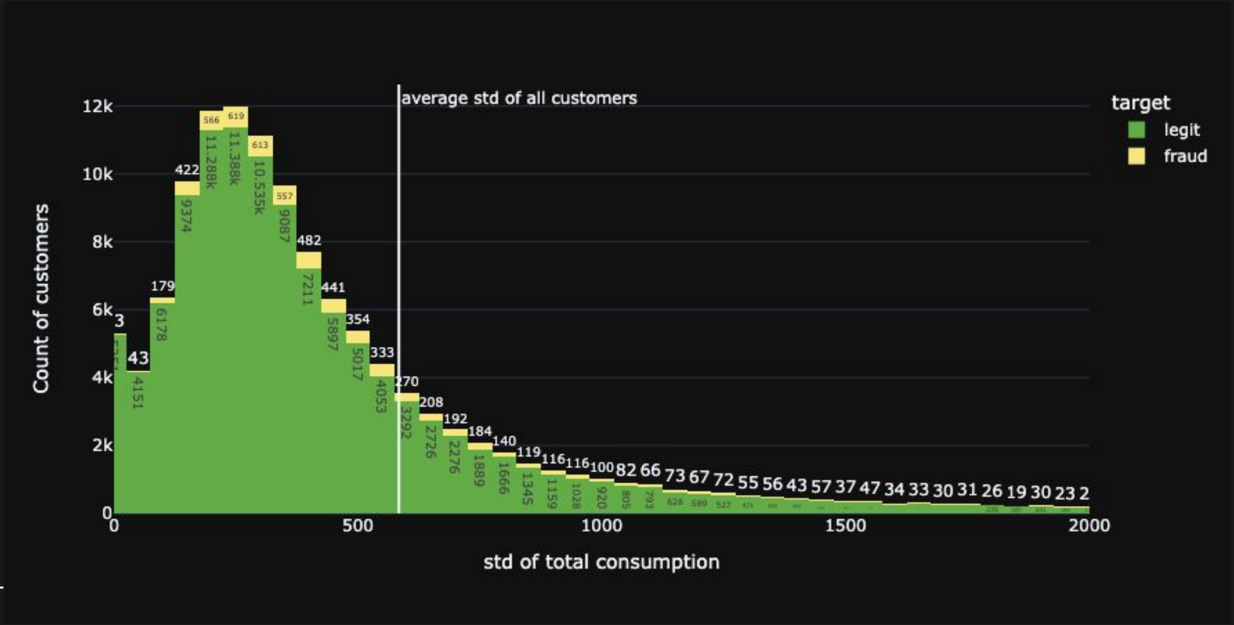
New features

- Mean total consumption
- Range of total consumption
- Standard deviation of total consumptions
- Customer´s number of counters
- Mode counter statue
- Mode reading remarque
- Weeks as a customer
- Energy type (electric only, electric and gas)
- Number of invoices with counter mismatch
- Total mismatch of consumption and counter indexes

03.

Baseline model

Our baseline model



A customer with exceptional fluctuations in monthly consumption is likely to be a fraud.

If the clients STD is higher than the mean STD of all clients, they will be flagged as fraud.

Score: AUC = 0.57

04.

Modeling

Modeling

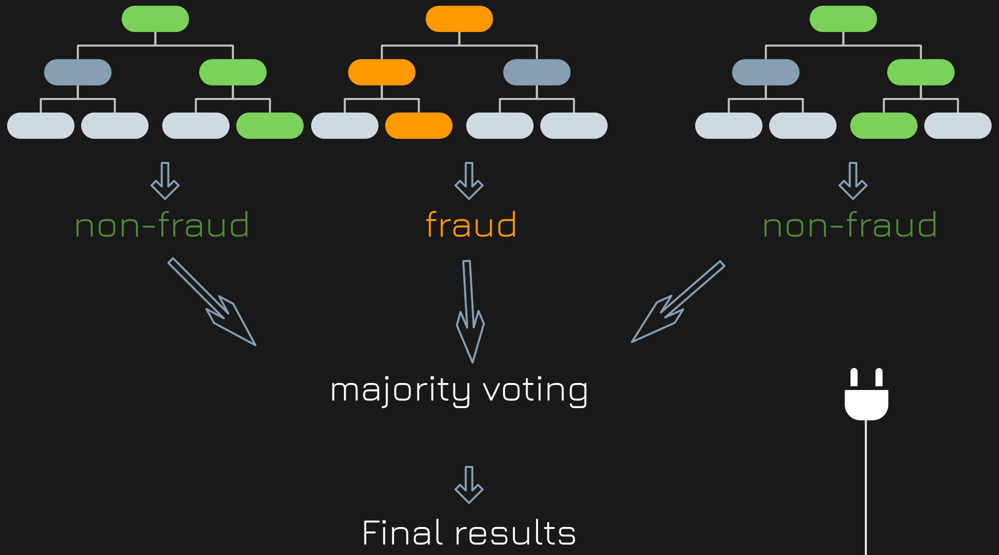
Sampling	Normalization	Model	Result, roc_auc
❌	✅	Logistic Regression	👎 (0.74)
✅	✅	KNN	👎 (0.78)
✅	❌	Decision Tree	👍 (0.80)
✅	❌	Balanced Random Forest	👍 (0.81)

Balanced Random Forest

It builds multiple decision trees and use majority voting to determine outcome.

The model use sampling to overcome challenges of imbalance dataset.

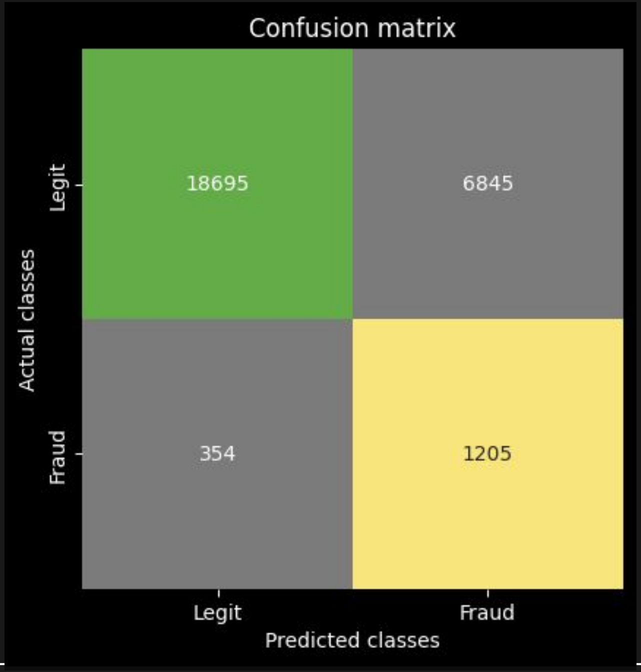
It makes the model robust to overfitting what provides reliable results on new data.



05.

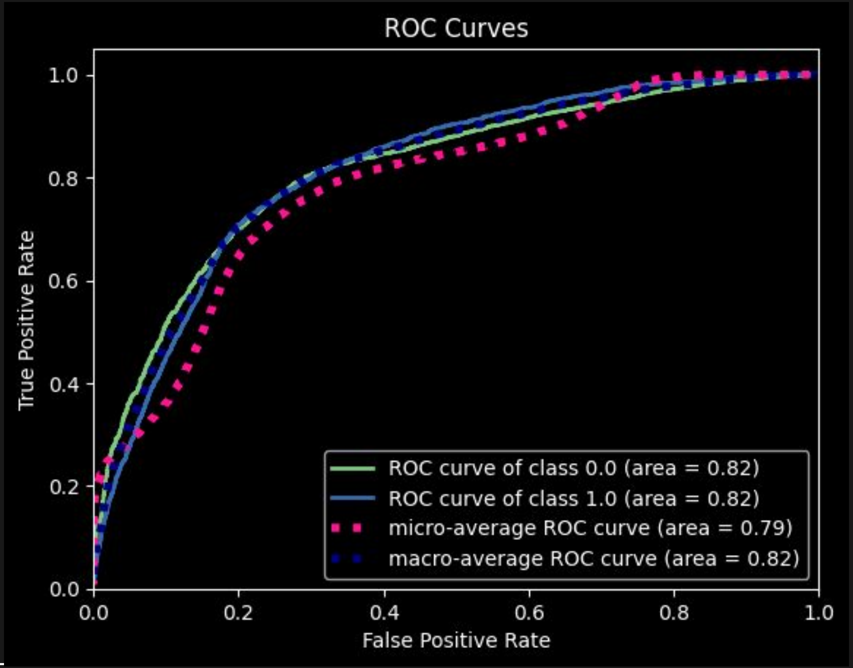
Evaluation

Model performance

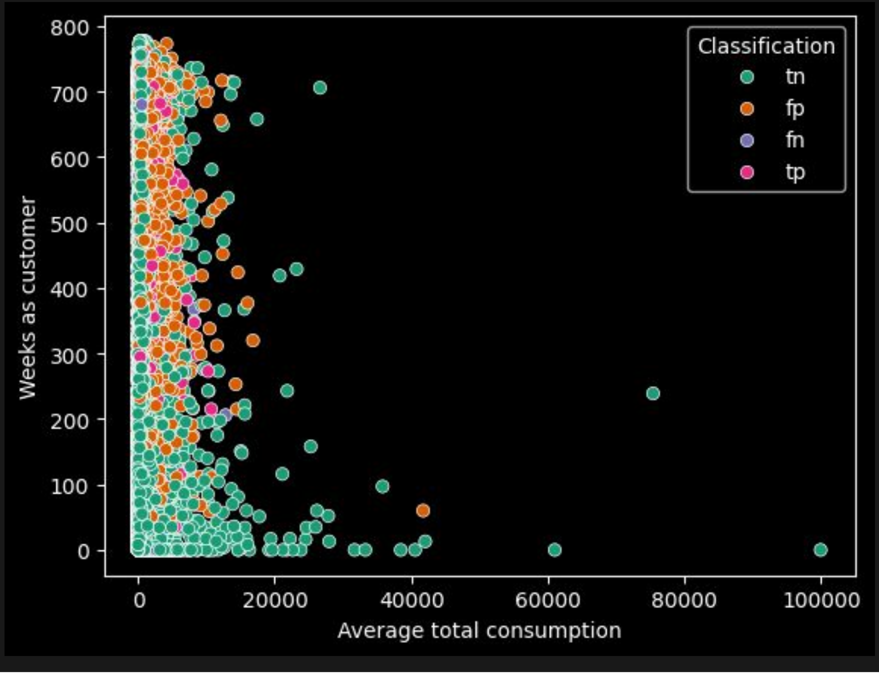


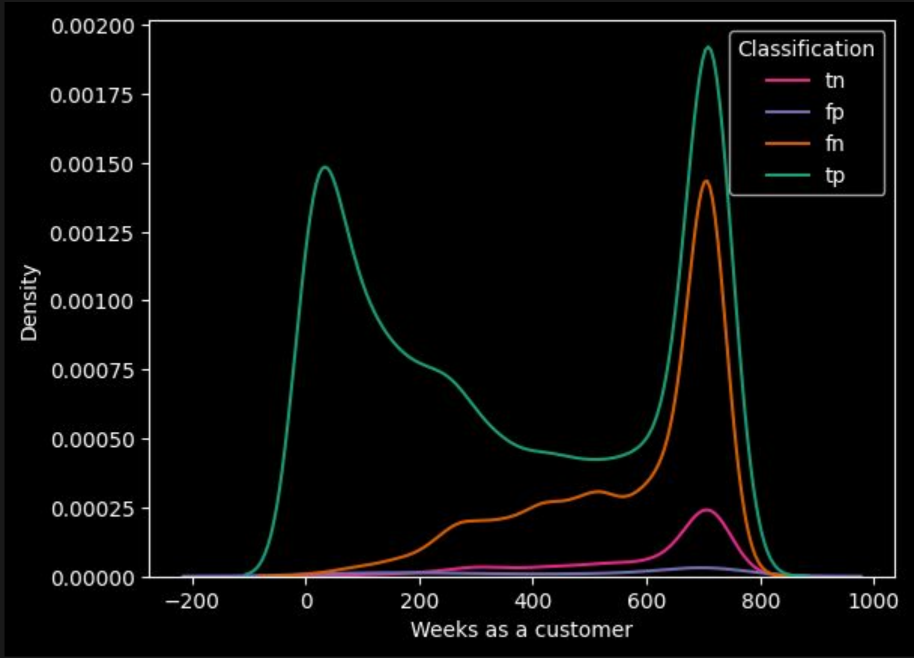
Confusion matrix

Actual classes \ Predicted classes	Legit	Fraud
Legit	18695	6845
Fraud	354	1205



Error analysis





Conclusion

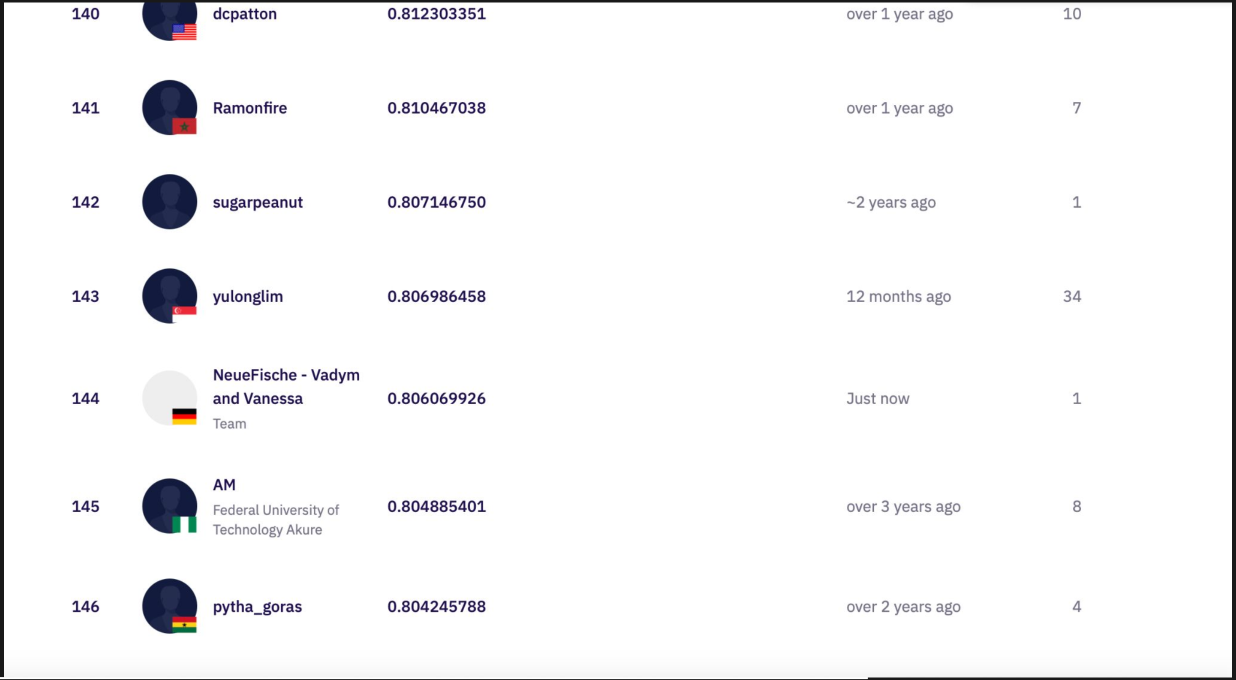
Before

5.7% of customers were frauders who were never catched

Using our model

Only **1.2%** of customers can possibly remain as uncatched frauders

Zindi Leaderboard



140	dcpatton	0.812303351	over 1 year ago	10
141	Ramonfire	0.810467038	over 1 year ago	7
142	sugarpeanut	0.807146750	~2 years ago	1
143	yulonglim	0.806986458	12 months ago	34
144	NeueFische - Vadym and Vanessa Team	0.806069926	Just now	1
145	AM Federal University of Technology Akure	0.804885401	over 3 years ago	8
146	pytha_goras	0.804245788	over 2 years ago	4

Thanks!

Do you have any questions?

linkedin.com/in/vadym-khvoinytskyi/
linkedin.com/in/vanessa-lampe

github.com/vlampe/fraud-detection-ML-project

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**