Fraud detection

THE TUNISIAN COMPANY OF ELECTRICITY AND GAS (STEG)

Data Science Team

Alida Bouatta
Nur Zulaiha Jomhari
Susanne Ferschl
Alexandros Serafeim

Agenda

- The business problem
- Data overview
- Model selection
- Model performance & interpretation
- Summary & Conclusions

The Business problem

Stakeholder:
Société Tunisienne de l'Électricité et du Gaz (STEG) is responsible for delivering electricity and gas across Tunisia.

Problem:
STEG lost close to 200 millions Tunisian Dinars (59 million Euros) due to fraudulent behaviour of clients.

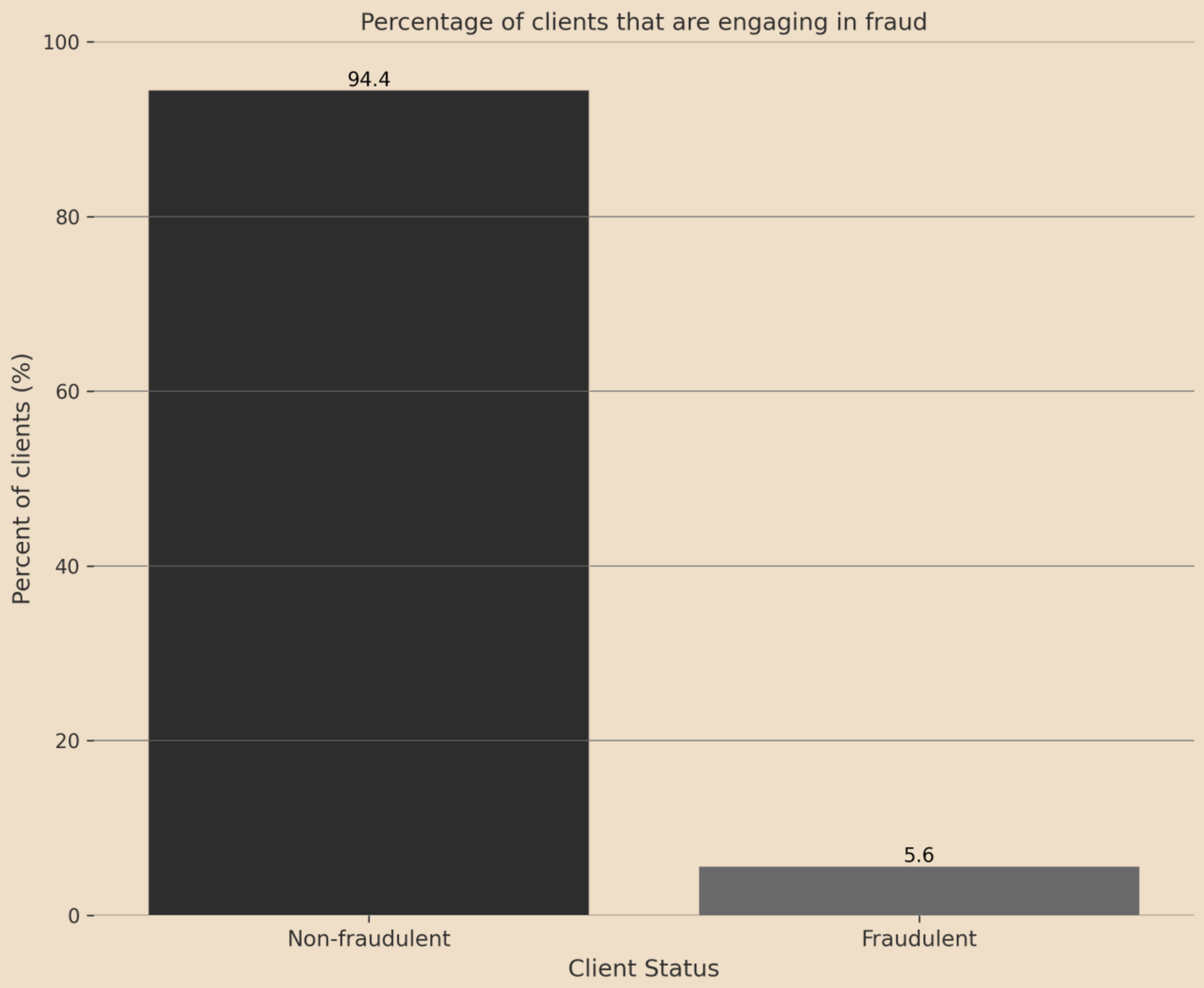
Business question:
How can STEG detect fraudulent activities from their customers while still making their services satisfactory and increasing customer traffic?

Build a model that will help the company classify which customer is likely to commit fraud (classification problem).

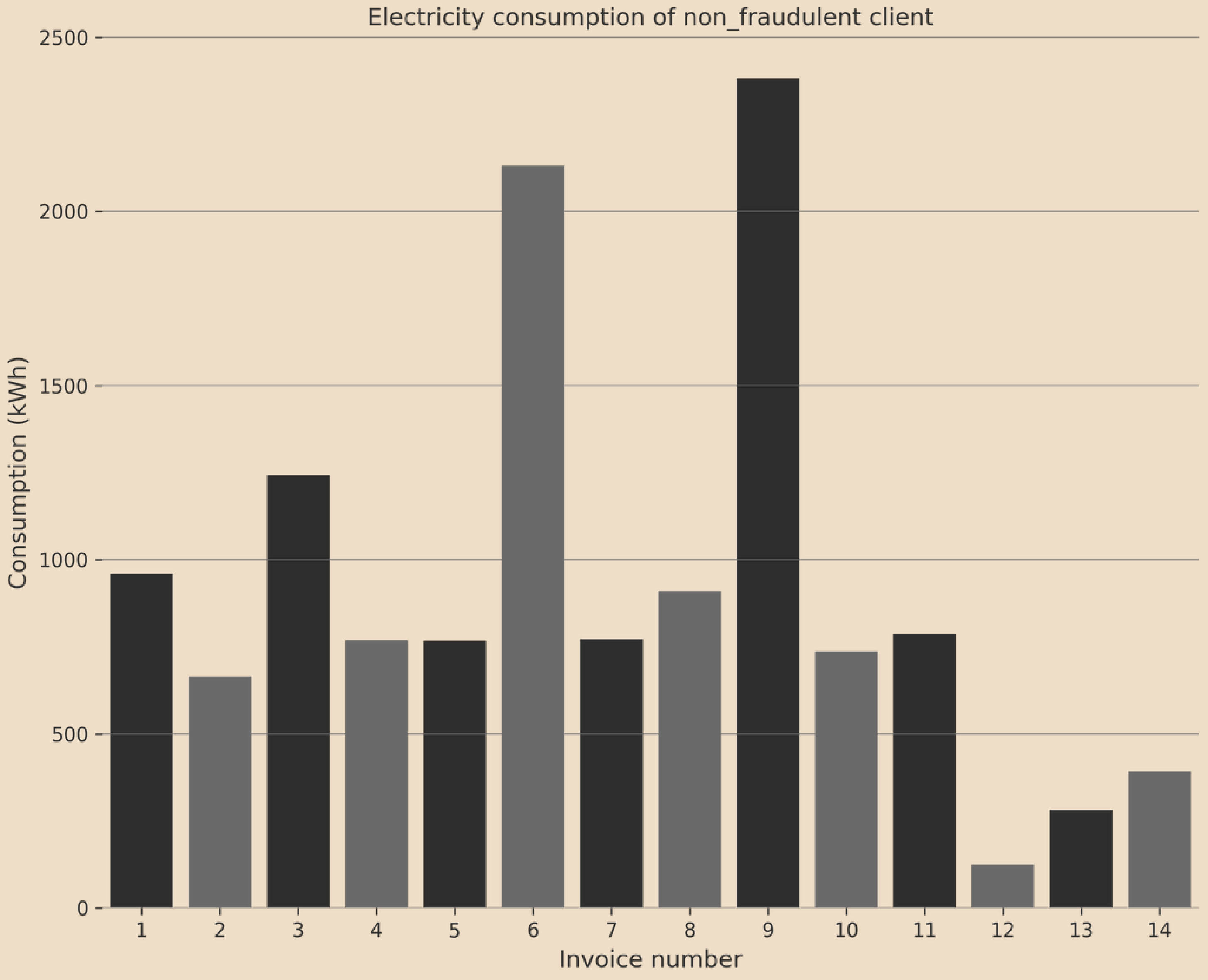
Data overview

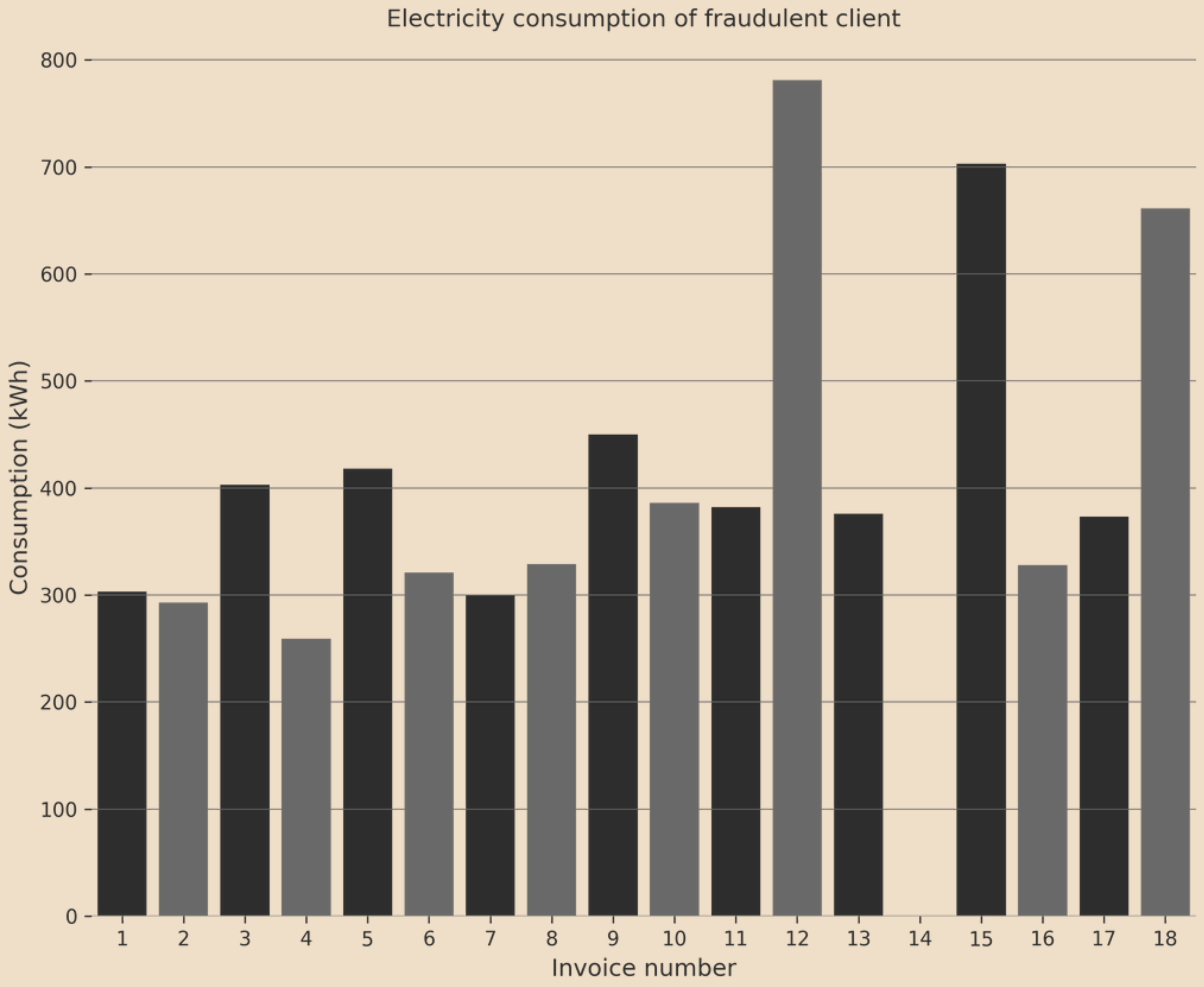
- **Client data:** 135k client data
- **Invoice data:** 4.5 Mil invoices

- **Client ID:** a unique number assigned to one client.
- **Counter Information:** Counter type, ID number, tariff type.
- **Geographical data:** Regions and districts.
- **Consumption informations:** 4 consumption levels in KWh and counter indexes.
- **Datetime information:** number of months between each reading and invoice issue date.
- **Target:** 0 if not fraudulent, 1 if fraudulent.



EDA- Consumption pattern

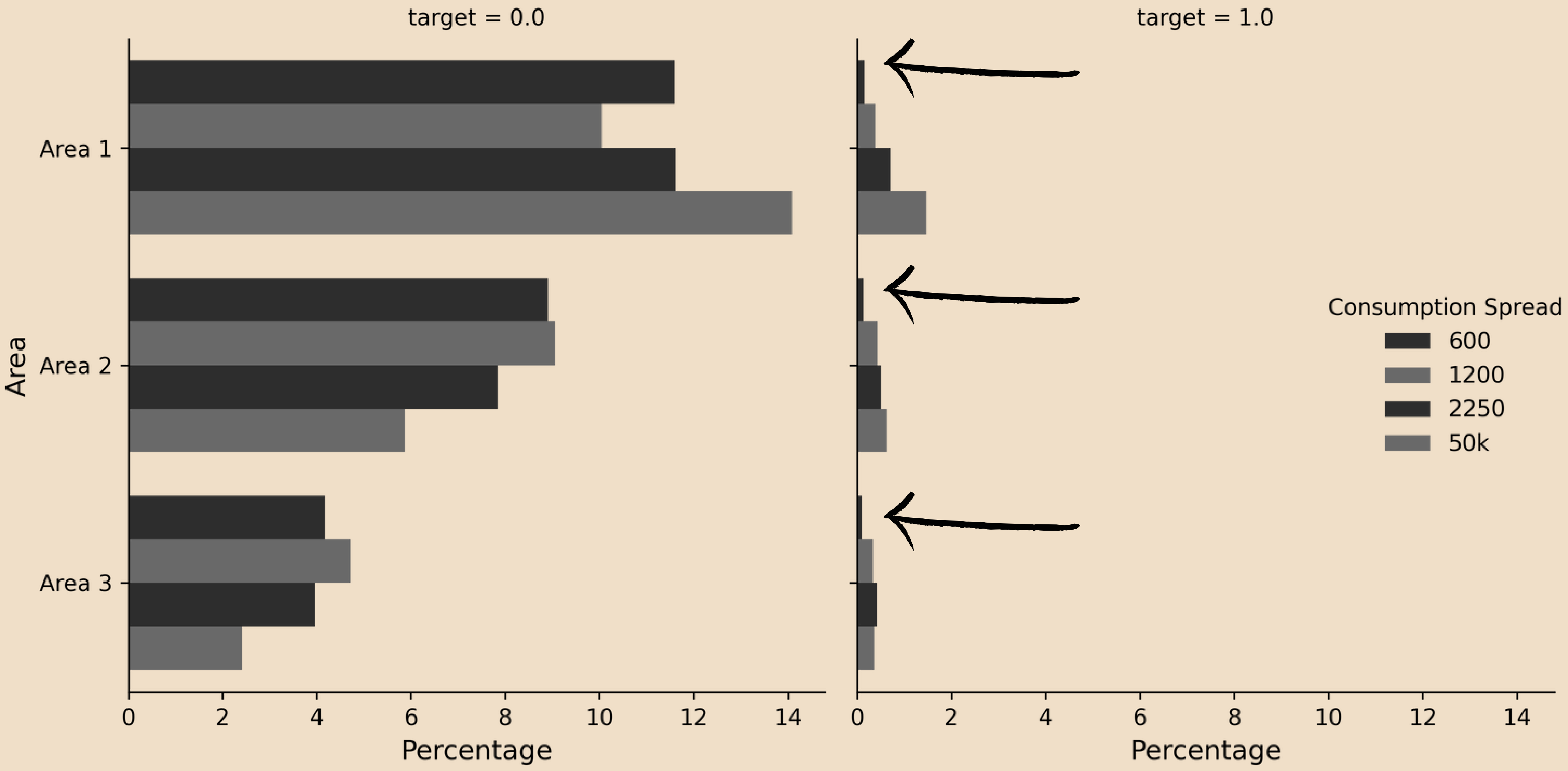




EDA- Consumption spread

- Emergent pattern when analyzing the consumption spread

- Disproportionate amount of fraudulent behaviour on larger and smaller spread



Feature Engineering

Synthetic features

- **Delta time:** the time period between each invoices.
- **Consumption level:** The total consumption for each level
- **Yearly/monthly consumption:** consumption averaged per month/year
- **Invoice issue year/month:** The month/year the invoice was issued
- **Base features**

Aggregating functions:

- **Minimum**
- **Maximum**
- **Median**
- **Mean**
- **Sum**
- **Std**
- **Skew**
- **Max-Min**
- **Std/mean**

- **Label encoding of categorical features**

Model selection

Selected models

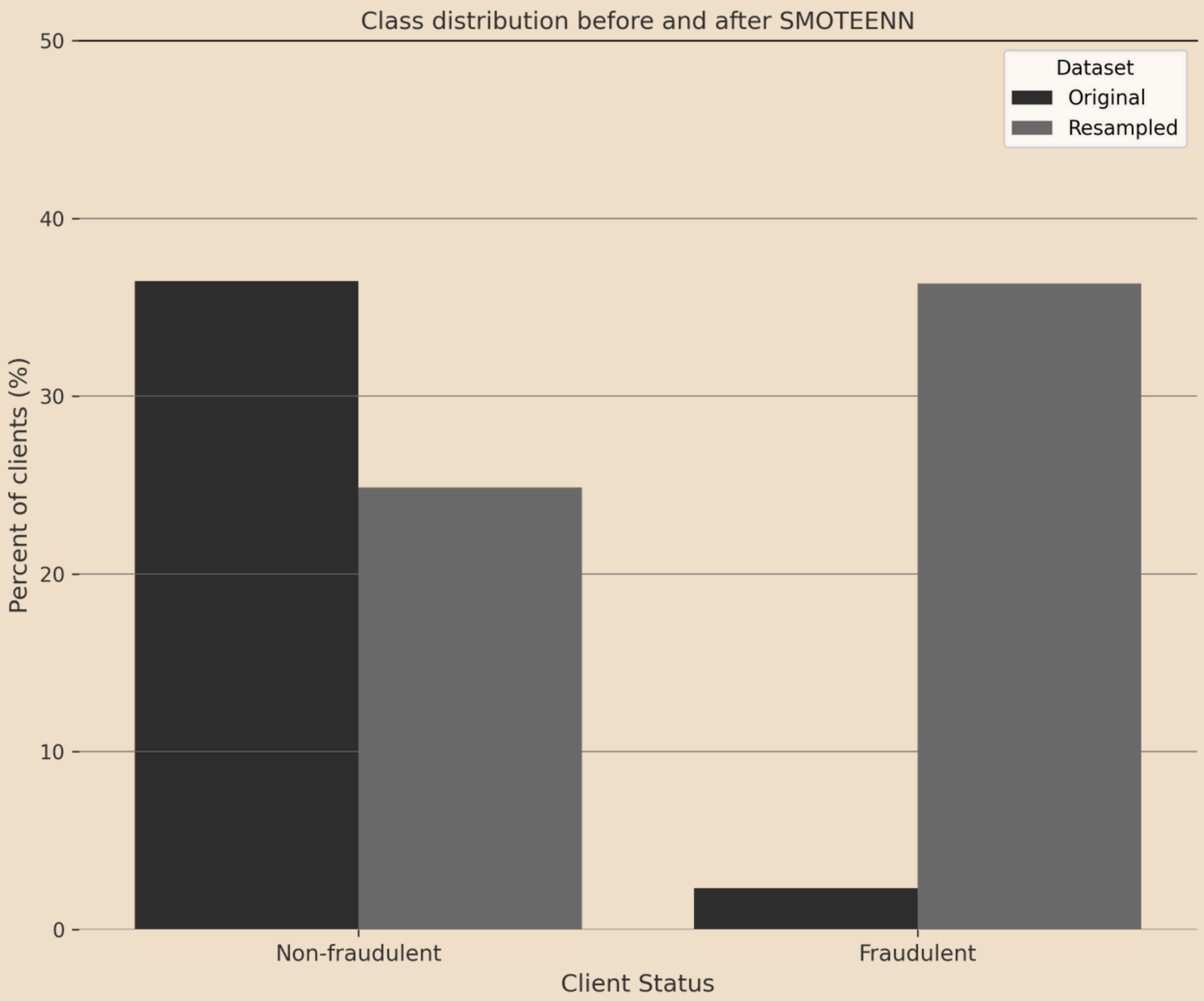
- Decision Tree (DT)
- Random Forest (RF)
- XGBoost (XGB)
- Light gradient boosting machine (LGBM)

Metrics

- Recall : ratio of true positive to true positive and false negatives
- AUC : Area under ROC curve

Misc

- Sample balancing (SMOTEEN)

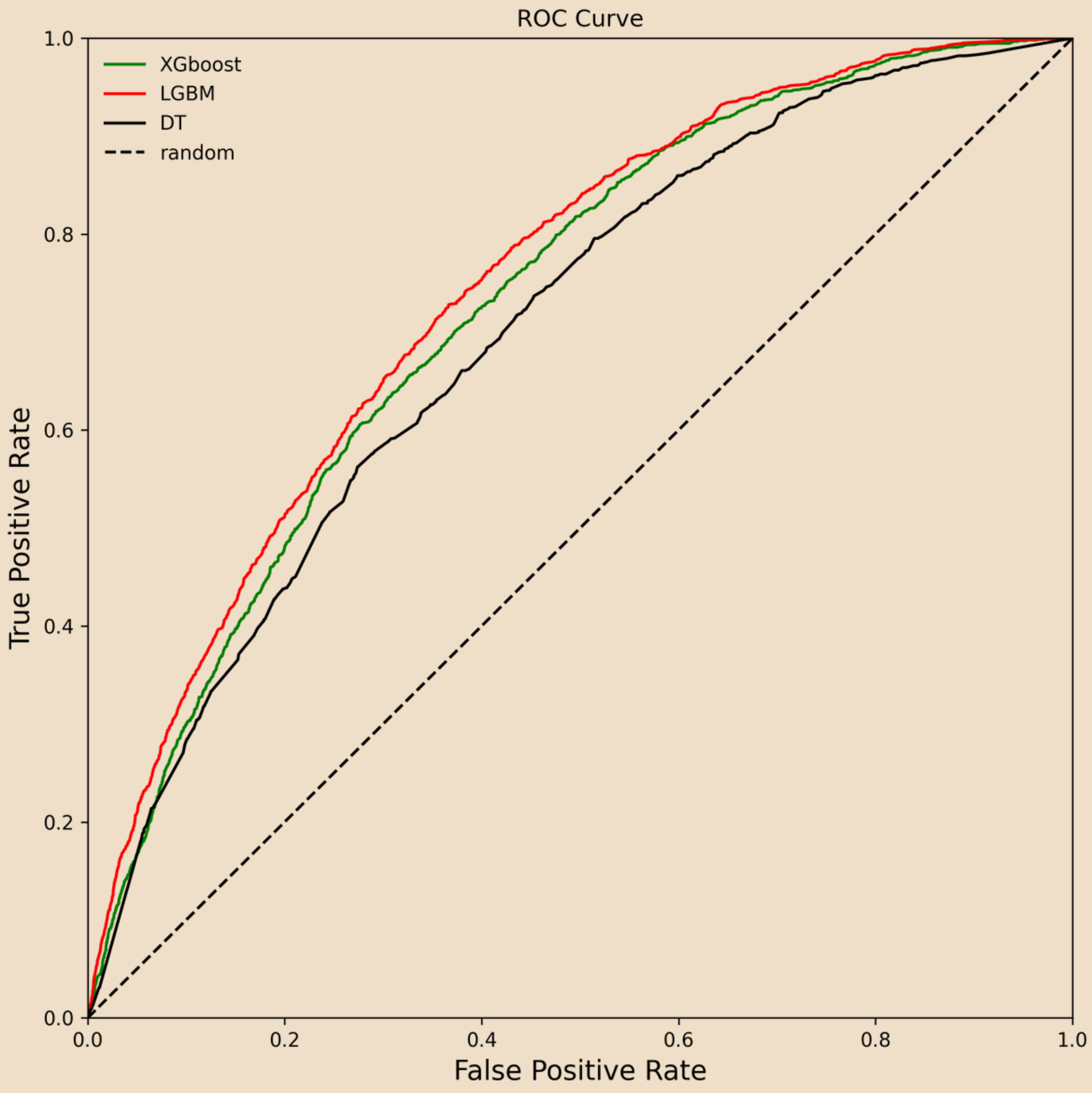


- **Baseline Model**: Consumers with spread more than 600 and belonging to area 1 or 3 are fraudulent.

Model performance

Table: Models and performance metrics

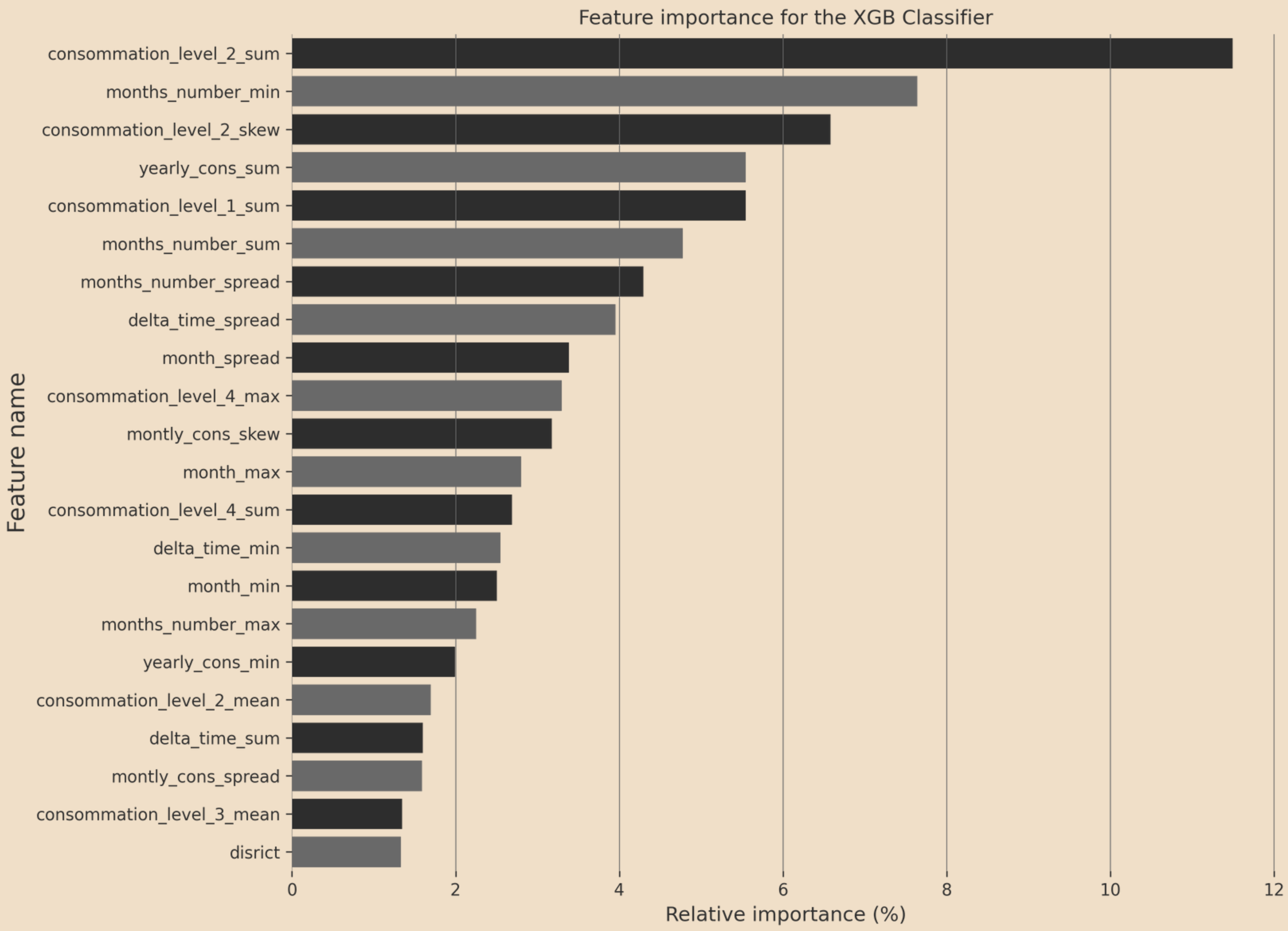
	Baseline model	XGBoost	Decision Tree	Random Forest	LGBM
Test Recall	0.58	**0.62**	0.55	0.44	0.26
Test AUC	0.55	**0.73**	0.70	0.63	**0.75**



Model interpretation

- Based on model training the consumption levels and yearly consumption as well as the interval between invoices are the critical features

- Geographical information was surprisingly less important



Summary & Conclusions

Overview

- Data from over 135k customers and 4.5 Mil invoices.
- No obvious pattern on the data.
- Sector knowledge was not always present

Model Performance

- XGB and LGBM with AUC ~ **0.75**.
- Recall at 0.5 threshold XGB **0.62** and **0.26** for LGBM.
- Threshold can be adjusted depending on company policy.

Business recommendation

- **Meter flagging**: Flag meters as potentially fraudulent and perform check during standard service.
- **Smart meters:** These meters can log more precise electricity consumption which can help detect patterns.
- **Awareness campaigns** that highlight how tampering with meters or illegal connections are detected and punished.

Thank you!