Fraud Detection in Electricity and Gas Consumption Challenge

Marius Bosch, Felix Geyer, Daniel Marques Rodrigues, Nikita Wilms

Overview

1. Intro to the topic
2. Baseline model and hypothesis validation
3. Initial machine learning approach
4. Necessary adjustments and dealing with imbalanced data
5. Refined machine learning approach
6. Summary

The Tunisian Company of Electricity and Gas (STEG)

- STEG lost 200 million Tunisian Dinars(~60 million Euro) due to **fraudulent manipulations of meters** between 2005 and 2019
- Our model **detects fraudulent customers** by using client billing history
- The solution aims to **reduce STEG's losses** and enhance data transparency

https://www.wikidata.org/wiki/Q3488699

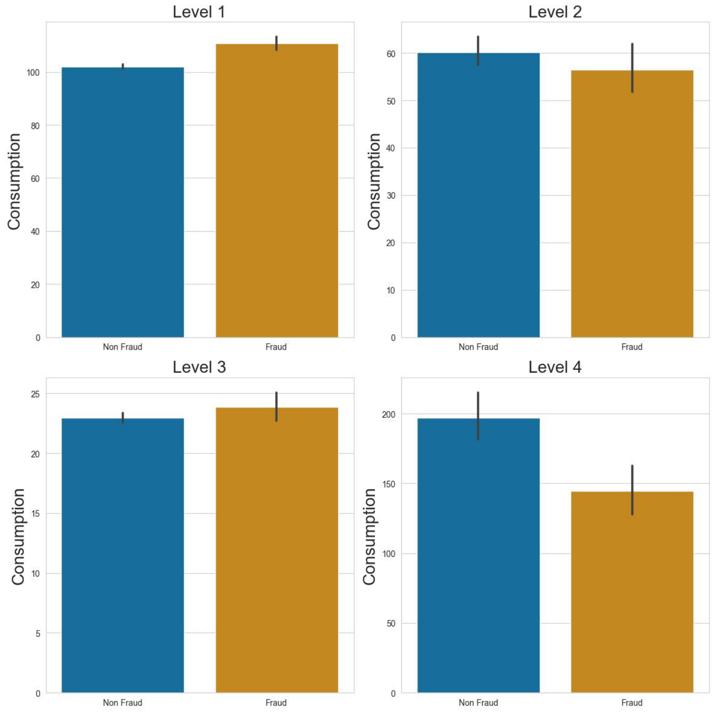
Initial data exploration identifies approx. 6% of customers with fraudulent behavior

Data consists of:

- **Customer meta data** (e.g. geographical, invoice frequency, tariff types, etc.)
- **Technical equipment data** (Gas, electricity, meter types)
- **Meter readings** Measurement information (remarks by technician, consumption levels)

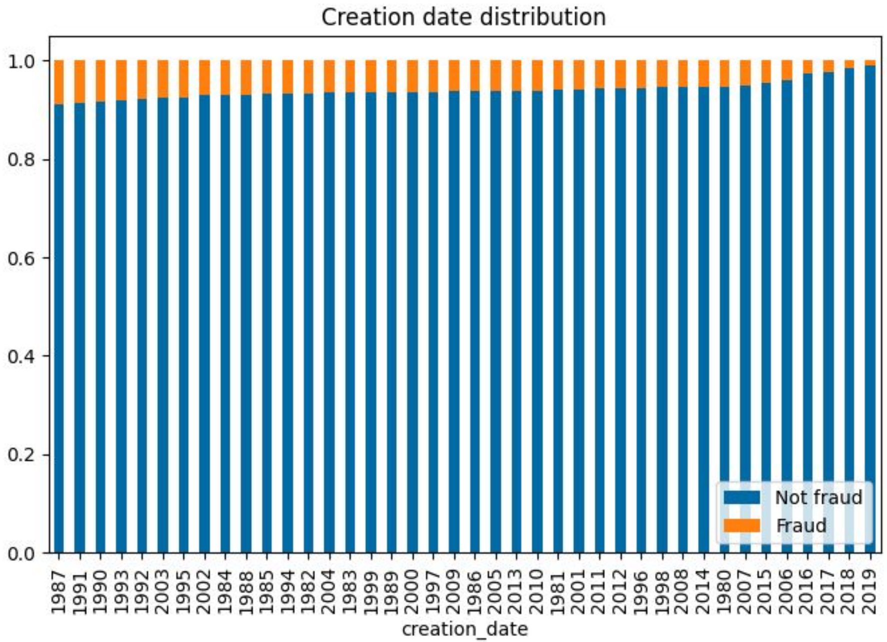
Which customer types tend to have fraudulent activity?

- Are customers with **higher consumption** more likely to be fraudulent?
  - **Not verified**

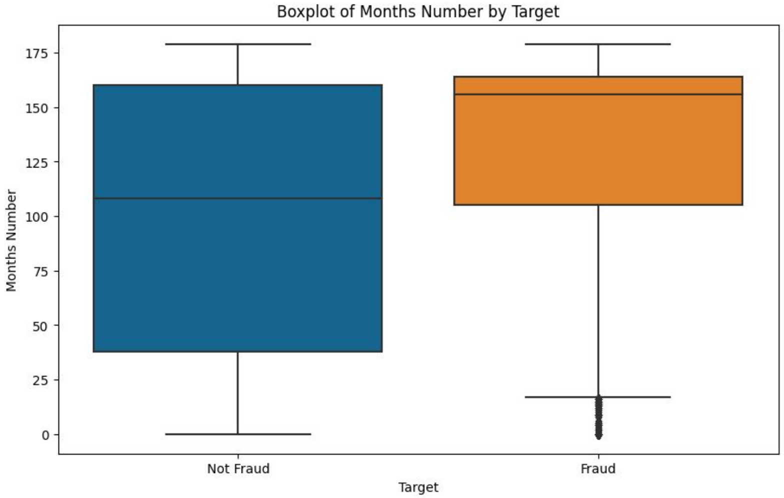


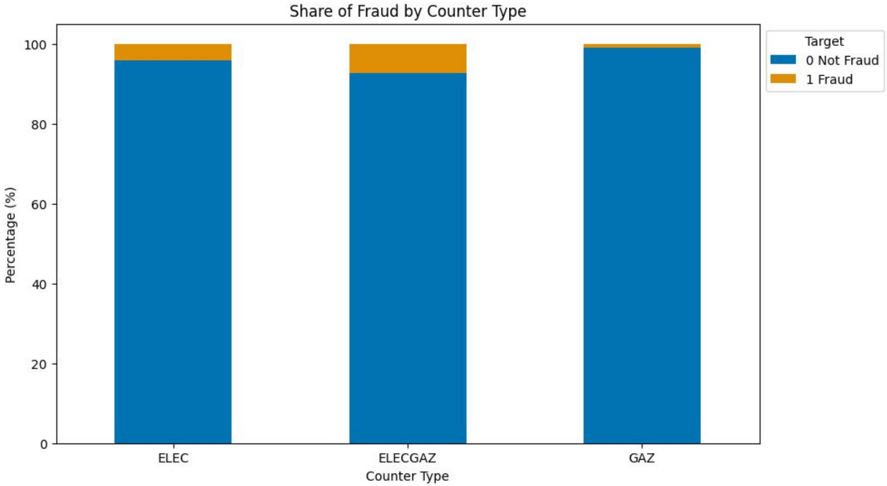
Which customer types tend to have fraudulent activity?

- Are customers with **newer contracts** more likely to be fraudulent?
    - **Not verified**

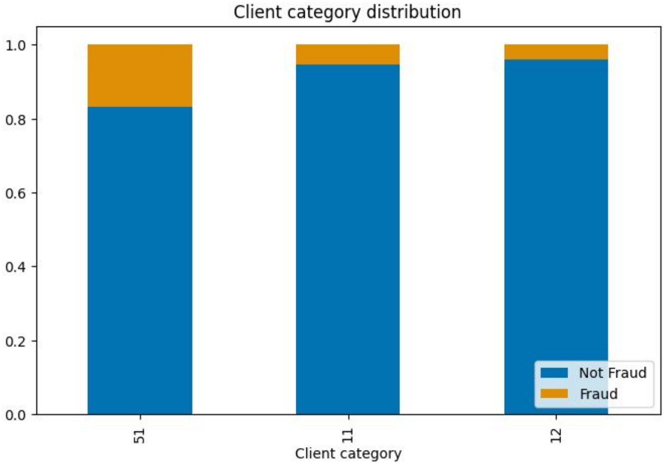


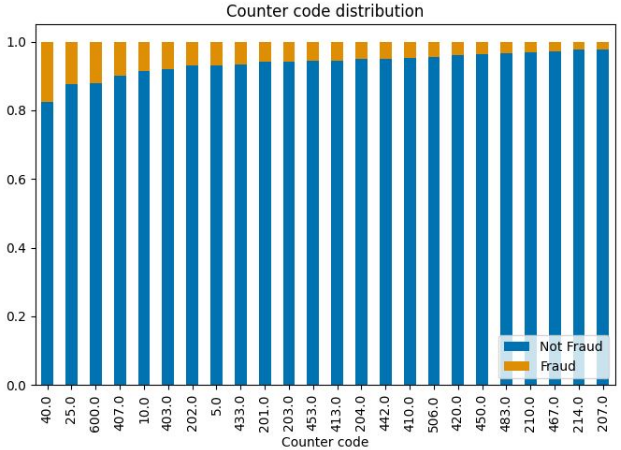
Long-term customers with both electricity and gas contracts tend to be more fraudulent





Incorporating these features resulted in the first baseline model





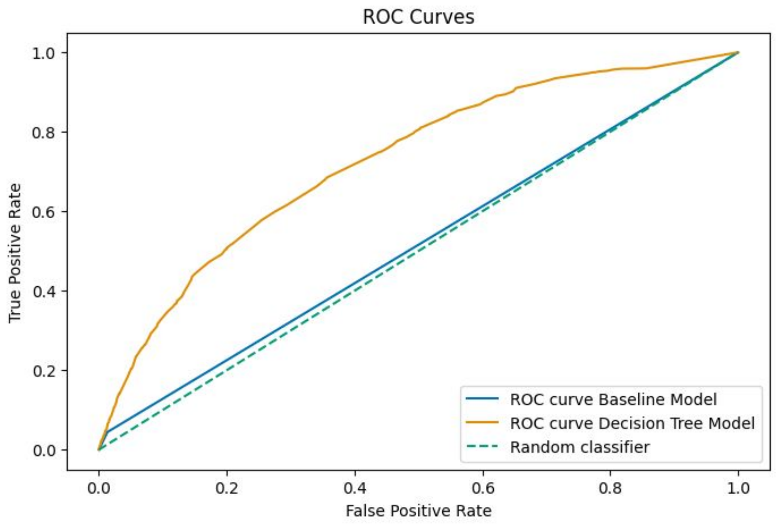
- **Client category 51** experiences higher activity
- **Counter codes 40 and 25** experiences higher activity

ROC-AUC score serves as our metric to determine **how well our model identifies fraudulent cases** (0.5 relates to random)

Heuristic baseline model with **ROC-AUC-score: ~0.52**

Extensive feature engineering shows promising results with initial Machine Learning Model

- **Decision tree** model with no in-depth hyperparameter tuning results in **improved prediction quality**
- Model yields higher quality predictions, yet still a **large number of false negative cases** (fraudulent activity classified as honest customers)
- **Only 9% of frauds classified correctly**



Decision tree model with **ROC-AUC-score: ~0.73**

Various adjustments to the training data set are necessary to boost prediction quality

- As most customers are truthful (95.5%) - **data set is highly imbalanced**

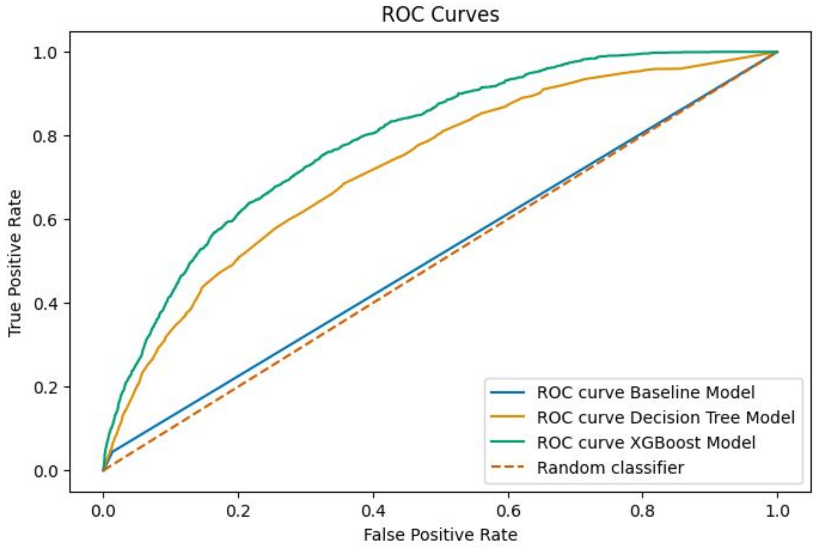
  ➔ **Comprehensive resampling using** Synthetic Minority Over-sampling Technique (**SMOTE**) to adjust imbalance

- Variable transformation & aggregation necessary - **bias towards long-term customers**

  ➔ Invoice data is aggregated using weighted monthly averages, separation of customers by contract type (carrier technologies)

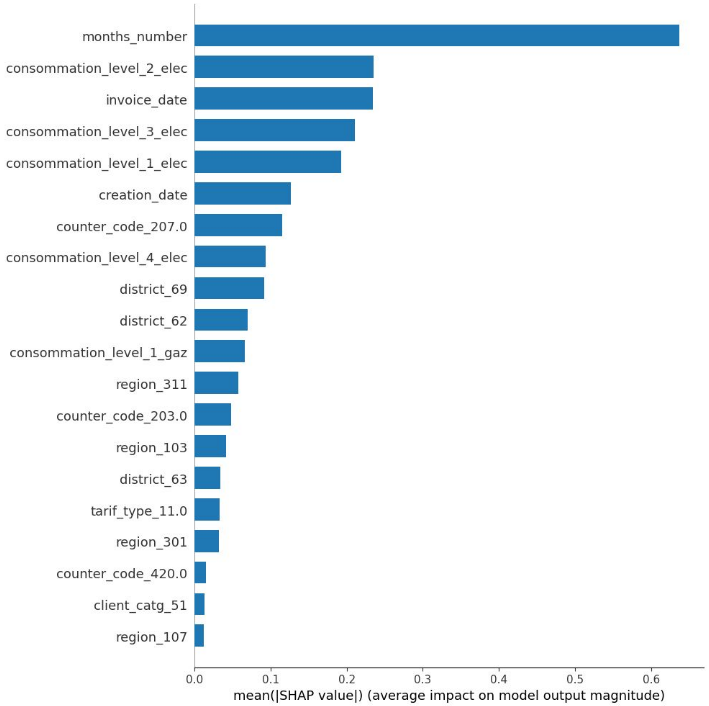
The refined data set allows for a more detailed ML-model with expressive results

- A **Gradient boosting approach** (XGBoost) is used and further tuned using (Gridsearch)
- The results show **no overfitting** of the model and guarantee a **high quality of predictions**
- **Significant improvement** of prediction results to previous model iterations



XGBoost model with **ROC-AUC-score: ~0.79**

We identified features that can be particularly interesting to analyse going forward



- Highest impact features:
  - Consumption of electric contracts
  - Contract duration in months
  - Creation Date of the Invoice

Summary

Where did we start?

- **STEG with significant losses** due to fraudulent activities from customers (6%) manipulating meter readings

Where did we end up?

- The model can confidently predict **24% of fraudulent cases** potentially recouping approx. **45 million Tunisian Dinars**

What are the future prospects?

- Initiate tool development (early classification & detection system for new and existing customers)

Thank you for your attention!

Acknowledgements:

ZINDI

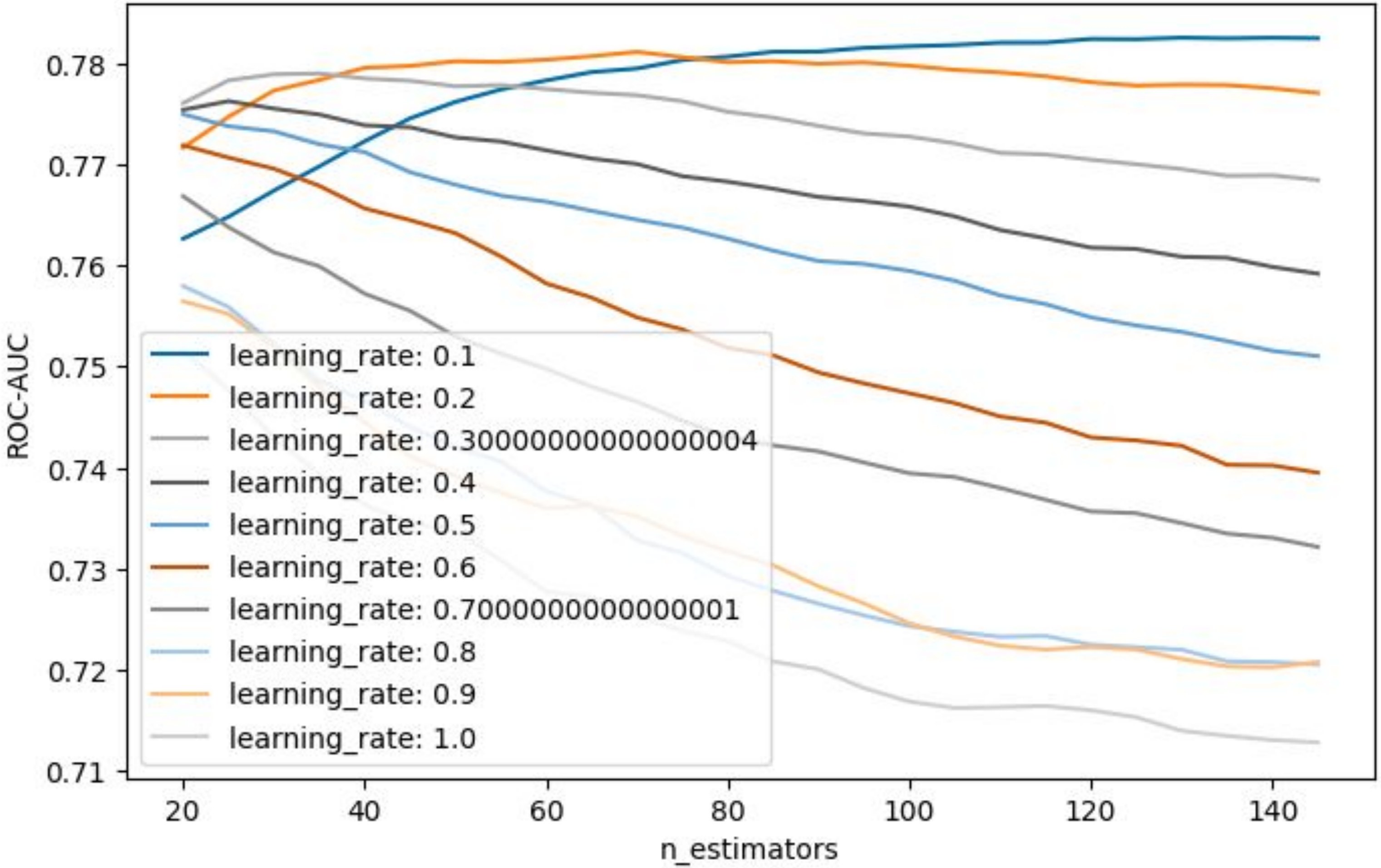
TEAM CRIME

The audience

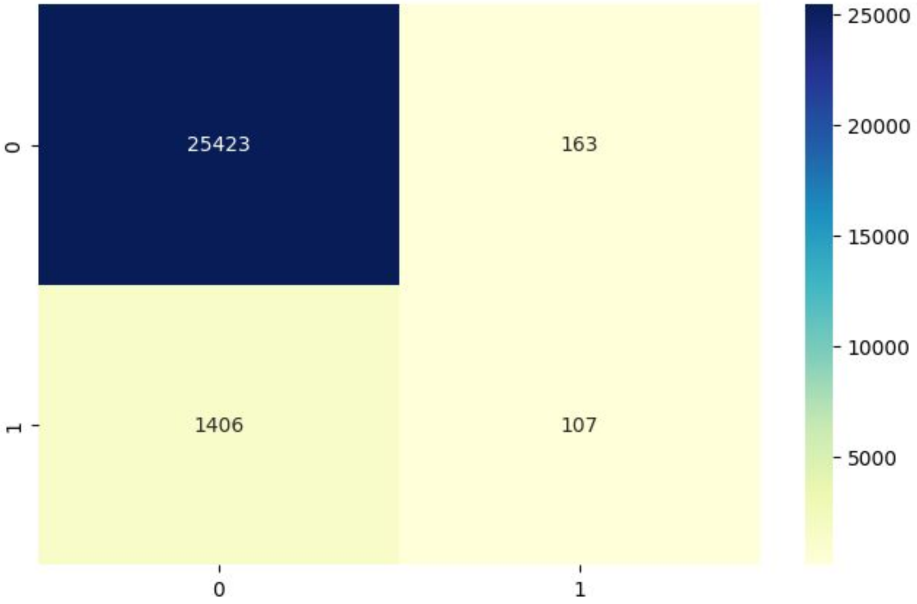
https://zindi.africa/competitions/fraud-detection-in-electricity-and-gas-consumption-challenge/data

https://stock.adobe.com/nl/search/images?k=neon+question+mark

Backup


ROC-AUC
0.78
0.77
0.76
0.75
0.74
0.73
0.72
0.71
20
40
60
80
100
120
140
n_estimators
learning_rate: 0.1
learning_rate: 0.2
learning_rate: 0.30000000000000004
learning_rate: 0.4
learning_rate: 0.5
learning_rate: 0.6
learning_rate: 0.7000000000000001
learning_rate: 0.8
learning_rate: 0.9
learning_rate: 1.0

Example Confusion Matrix XGBoost-Model



Feature Importance by value