

Fraud Detection in Electricity and Gas Consumption

Machine Learning project

Jin-Ho Lee
Veranda Osmani
Franziska Schulze Bockeloh

Outline

- **Introduction**

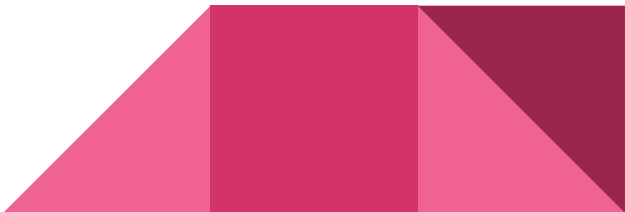
- Tunisian Company of Electricity and Gas
- Our aims
- Introduction to dataset
- Exploratory data analysis

- **ML Model**

- Baseline model
- Final models

- **Conclusion**

- **Future work**



The Tunisian Company of Electricity and Gas (STEG)

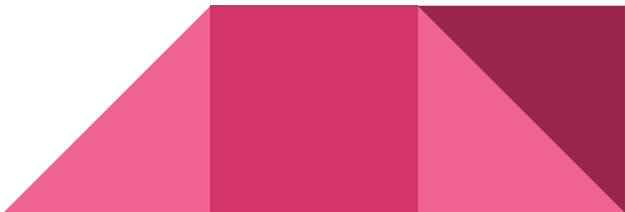
- established 1962
- public, non-administrative
- provider of electricity and gas across whole Tunisia
- second-largest Tunisian company by revenues in 2009



Significant losses of 200 million TND (about **64 million USD**) due to fraudulent manipulations of meters by consumers.



Our aims

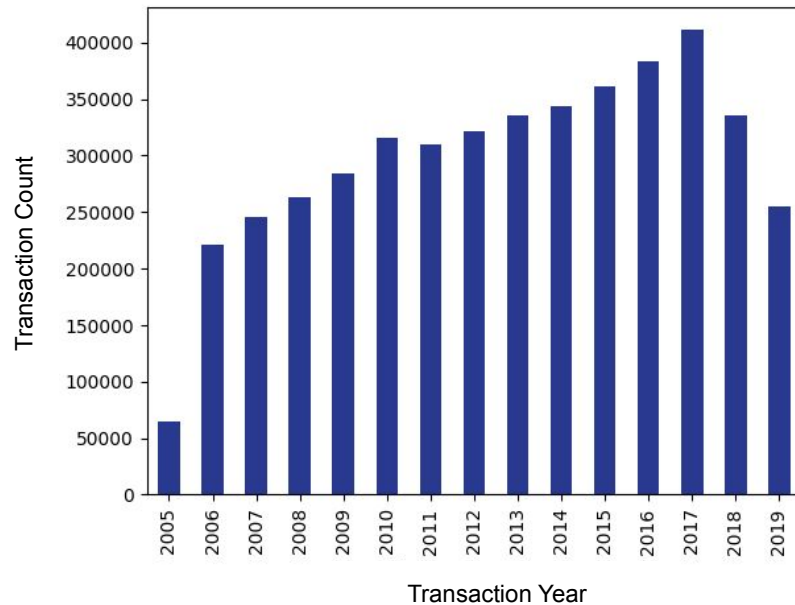
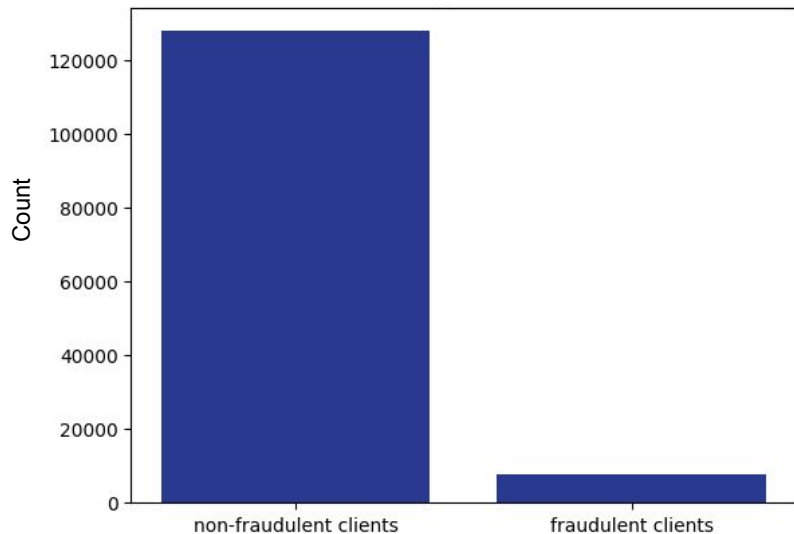
1. **Help improve revenues** of the Tunisian Company of Electricity and Gas (STEG)
 2. **Reduce the losses** caused by clients involved in fraudulent activities
 3. **Detect / predict fraudulent clients** using a data science (machine learning) based model
 4. Maintain the company's **customer satisfaction**
- 

The data set

- **Billing history** of about **4,500,000 transactions** from **136,000 clients** provided in two separate data tables with:
 - a. **client data** with 6 features (**fraud labels**, regional details, etc.)
 - b. **invoice data** with 16 features (billing amount, payment details, **meter readings**, etc.)
- Whole time period of documented transactions: **2005 - 2019**



Exploring the data



The trend of transaction count over the years.

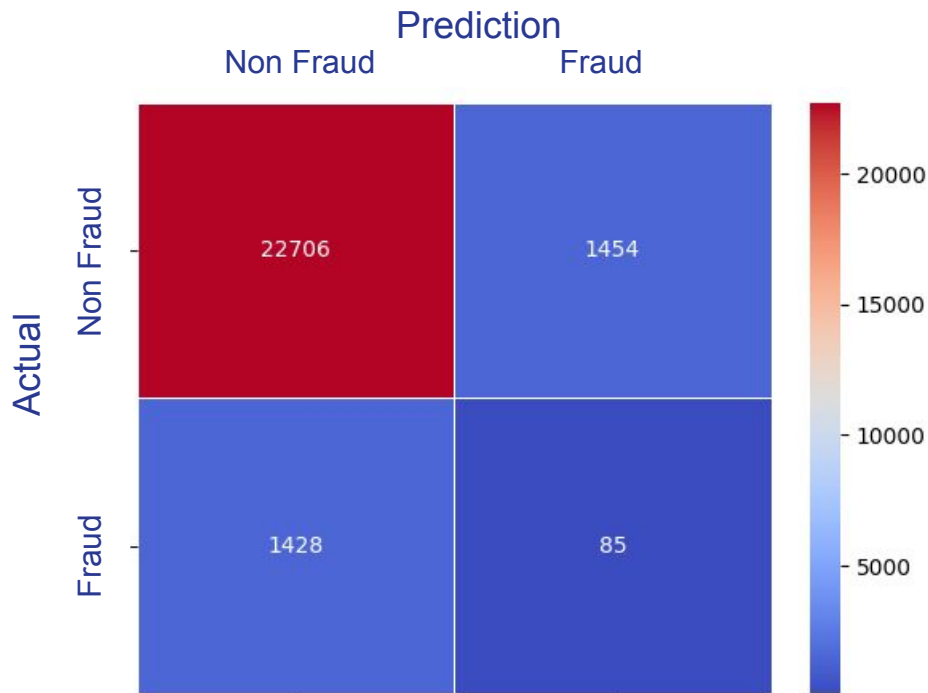
Due to the

1. **high imbalance** in our prediction target and
2. requirement to maintain **customer satisfaction**

we decided to focus on the **harmonic mean between sensitivity and precision** (F1-score).

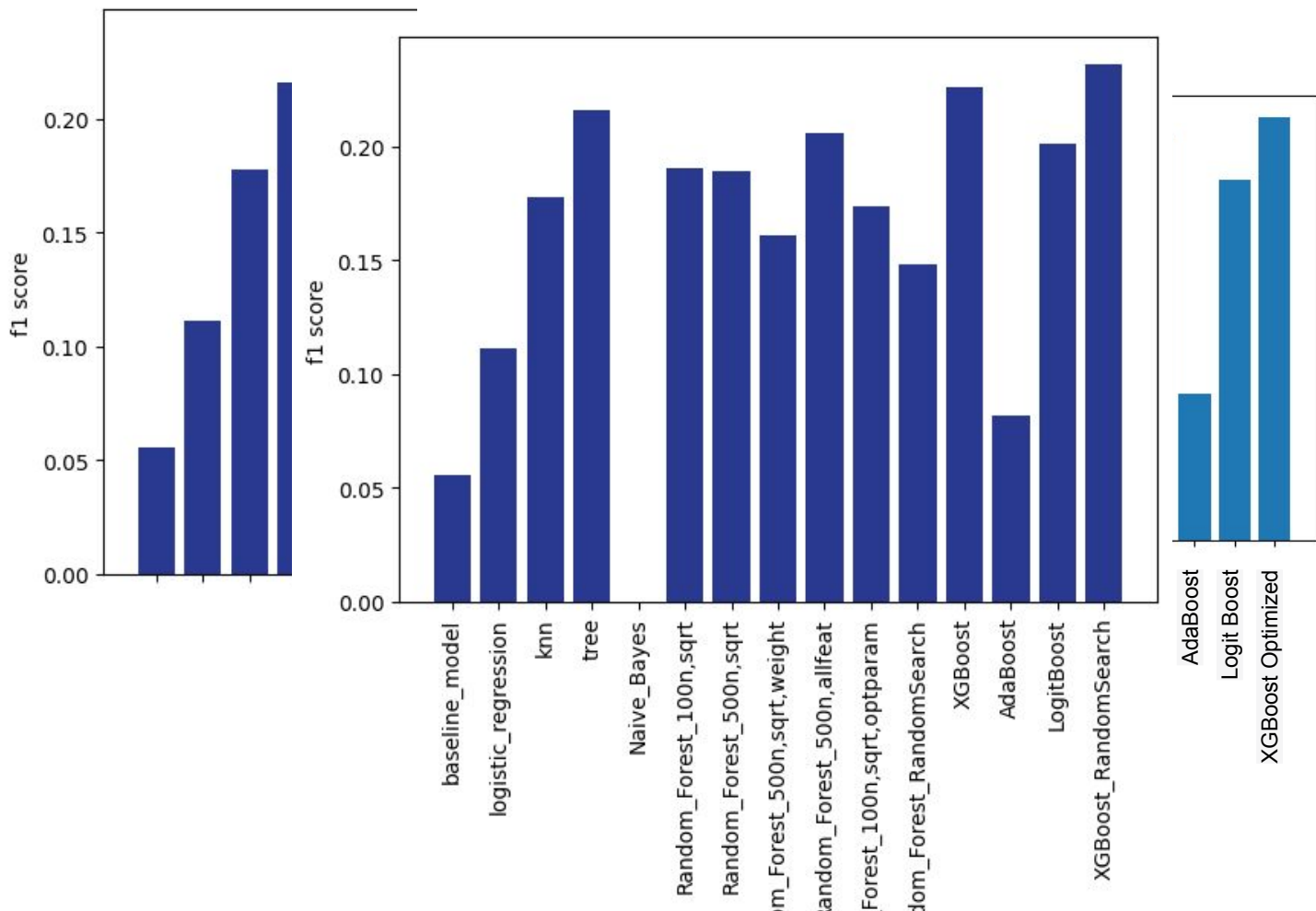
Baseline model

- Simple baseline model as **benchmark** for further machine learning models
- Use **proportion of fraud clients** for baseline
- F1-score of 0.05
- Prediction of baseline
 - **6% of fraud clients** can be predicted



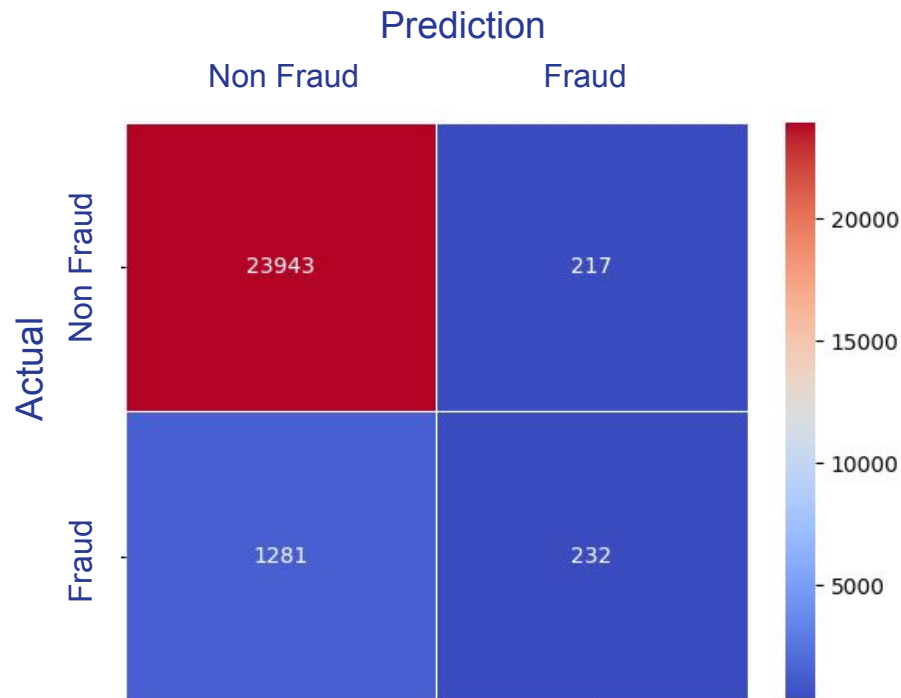
Compari

- All models (except Naïve Bayes)
- Improvement
- Further work



Final model

- **XGBoost model** with optimized parameters performed best
- F1 score = 0.233
- Predictions of model
 - **15% of fraud clients** can be detected



Conclusion

- XGBoost is the best performing model for detecting fraudulent clients
 - Final model of initial analysis predicts 15% of fraud clients
 - Number of wrong accusation (False Positive) is low to maintain customer satisfaction
-

Future work

1. Extend our domain knowledge to create new variables, or transform existing variables that may be more informative
2. Data augmentation; to overcome with imbalanced data
3. Combining multiple models to make predictions, which can help to improve the f-1 and robustness of the model





Thank you for your attention!

References

1. <https://zindi.africa/competitions/fraud-detection-in-electricity-and-gas-consumption-challenge>
2. <https://towardsdatascience.com/baseline-models-your-guide-for-model-building-1ec3aa244b8d>
3. https://scikit-learn.org/stable/supervised_learning.html#supervised-learning





Back-up

Feature Importance

