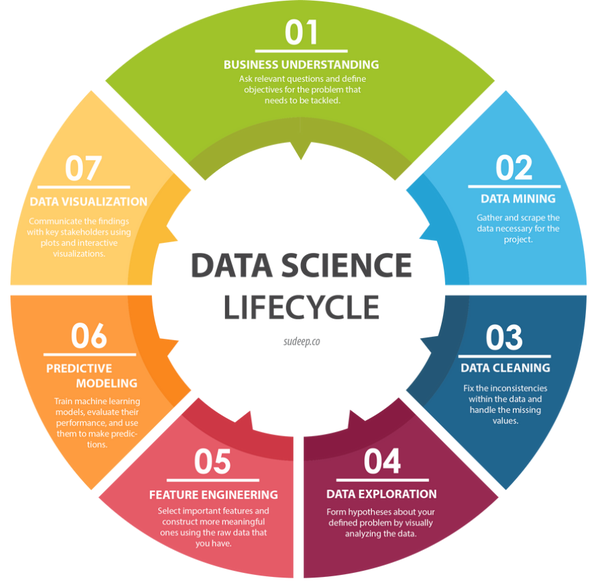
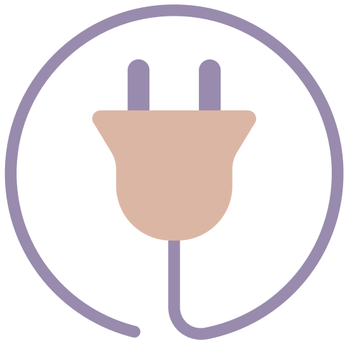
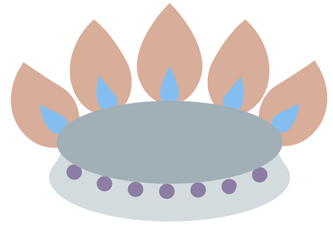


STEG - Société Tunisienne de L'Électricité et du Gaz

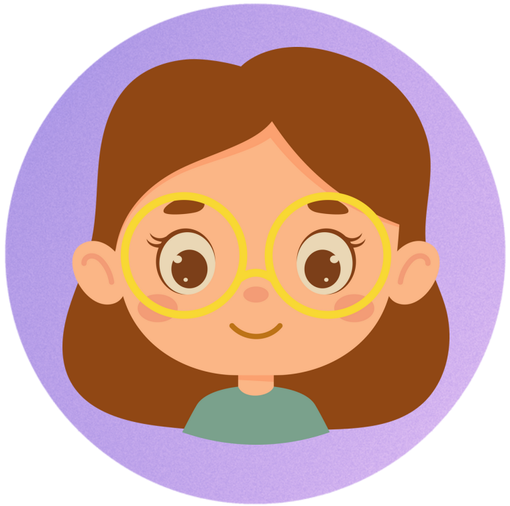
Fraud Detection with Applied Data Science



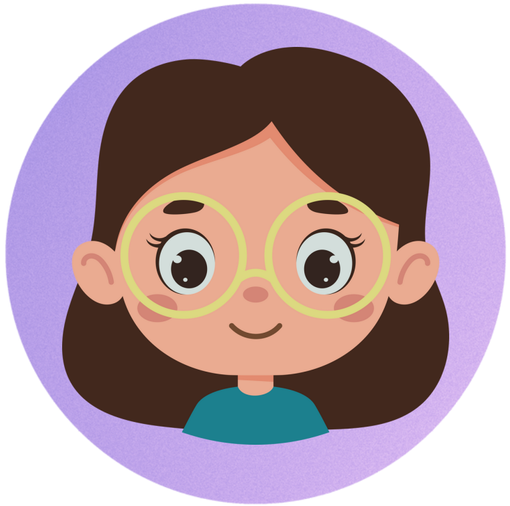


The Team

ANAS



LANA



ANNA

GRACE

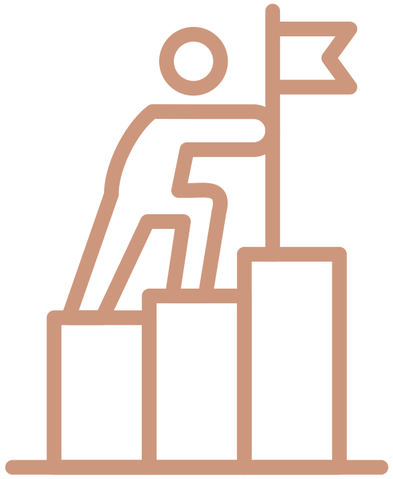
Problem at Hand

- The company suffered tremendous losses in the order of 200 million Tunisian Dinars due to fraudulent manipulations of meters by consumers.

Objective

- Build a model to predict clients that are likely committing fraud by manipulation of their gas or electricity meters.

- Our goal is to apply machine learning to correctly predict fraud (prevent financial damage for the company) while limiting the number of falsely accused clients (prevent reputation damage).



Your Data

- Clients Data:
  - 135 500 Clients
  - Distributed over 25 regions in 4 districts across Tunisia
  - Company entry between 1977 and 2019

- Billing History:
  - 4 500 000 invoices for electricity and gas consumption
  - Invoices from 1977 to 2019
  - Maximum duration as client: 42 years

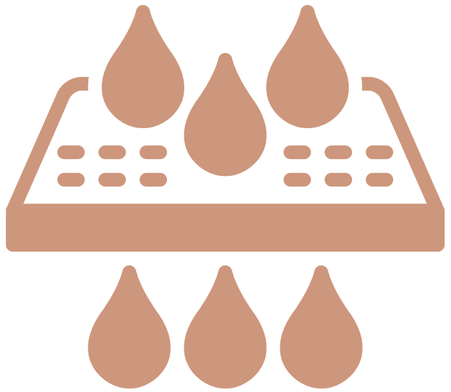
02

DATA MINING

Gather and scrape the data necessary for the project.

Cleaning and Preprocessing

- Combine client and invoice data to gain insights
- Drop duplicates and faulty data (less than 0.0002 %)
- Clean data types of entries
- Drop features not of relevance after feature engineering

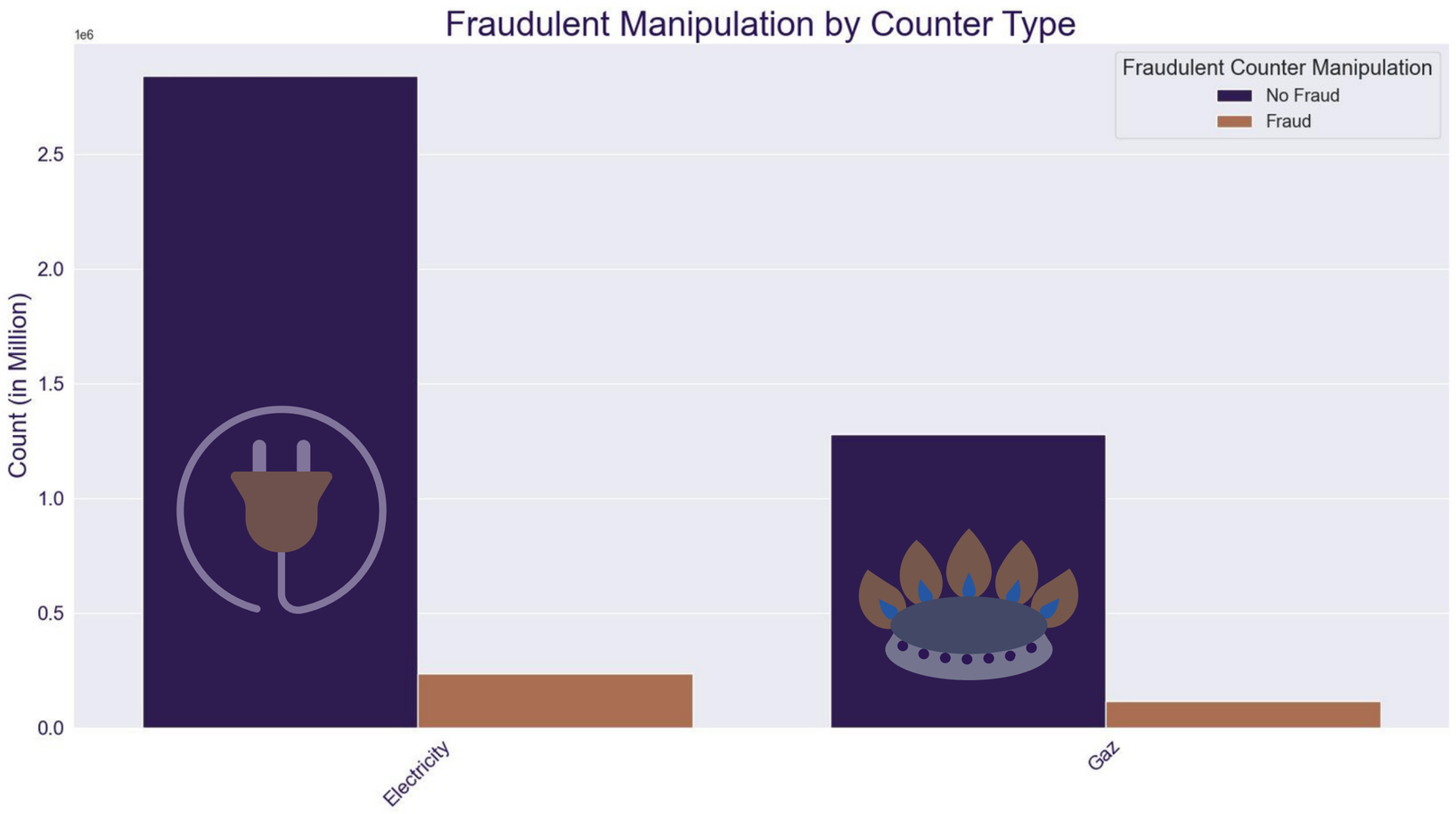


03

DATA CLEANING

Fix the inconsistencies within the data and handle the missing values.

Frequency of Fraud



- The fraud rate is 8%
- Fraud rate gas: 9%
- Fraud rate electricity: 7%

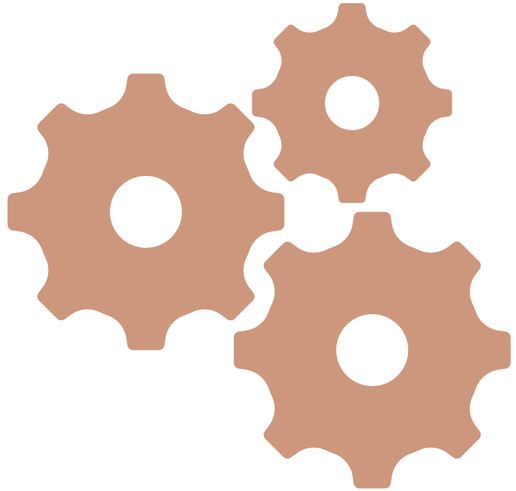
- 68% of invoices issued for electricity consumption
- 32% of invoices for gas consumption

04

DATA EXPLORATION

Form hypotheses about your defined problem by visually analyzing the data.

Feature Engineering



- Goal:
  - Gain insights from newly created features
  - More meaningful features
  - Better predictions

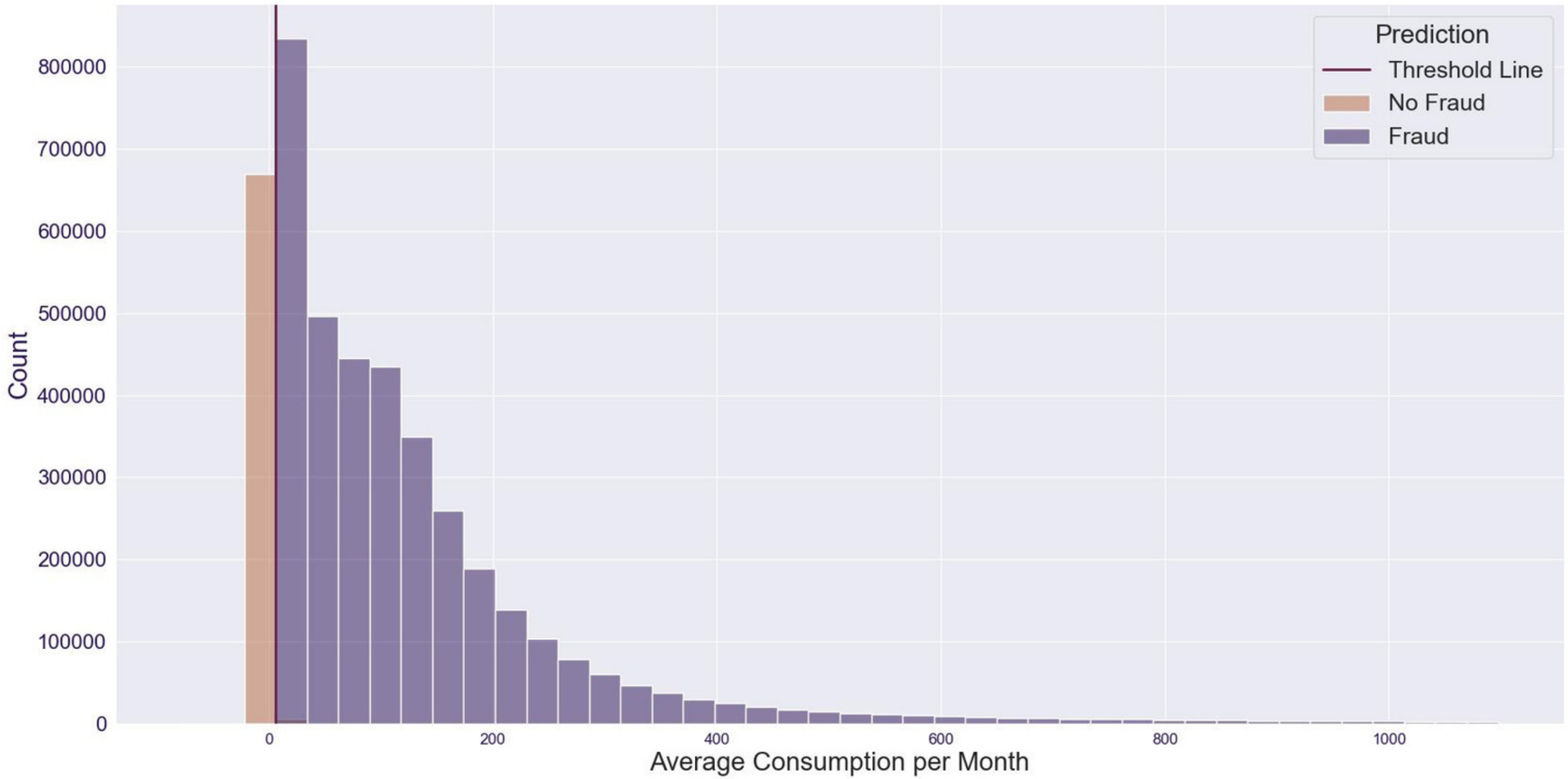
- New features:
  - Years as client
  - Group very rare tariffs together
  - Average Consummation per month
  - Number of counters per client

05

FEATURE ENGINEERING

Select important features and construct more meaningful ones using the raw data that you have.

Baseline Model



- A very low amount of used electricity or gas per month is deemed suspicious.
- The baseline model predicts the lowest 15% of monthly consummation to be fraudulent.

06

PREDICTIVE MODELING

Train machine learning models, evaluate their performance, and use them to make predictions.

Machine Learning Model: Easy Ensemble

- We trained several machine learning models for you
- Picked the one with the best performance

- The EasyEnsembleClassifier
  - Made for Imbalanced Data
  - Based on Combination of Decision Trees
  - Focuses on minority class
  - Iteratively focuses on false predictions in every estimation step

06

PREDICTIVE MODELING

Train machine learning models, evaluate their performance, and use them to make predictions.

Model Comparison

How many clients were correctly detected as fraudulent?

Base Model — ***14%*** *of fraudulent activities correctly detected*

ML Model — ***63%*** *of fraudulent activities correctly detected*

How many clients were correctly detected as non-fraudulent?

Base Model — ***85%*** *of non-fraudulent activities correctly detected.*

ML Model — ***61%*** *of non-fraudulent activities correctly detected.*

Model Comparison

How many frauds were missed?

Base Model — ***86%*** *of the fraudulent activities missed*

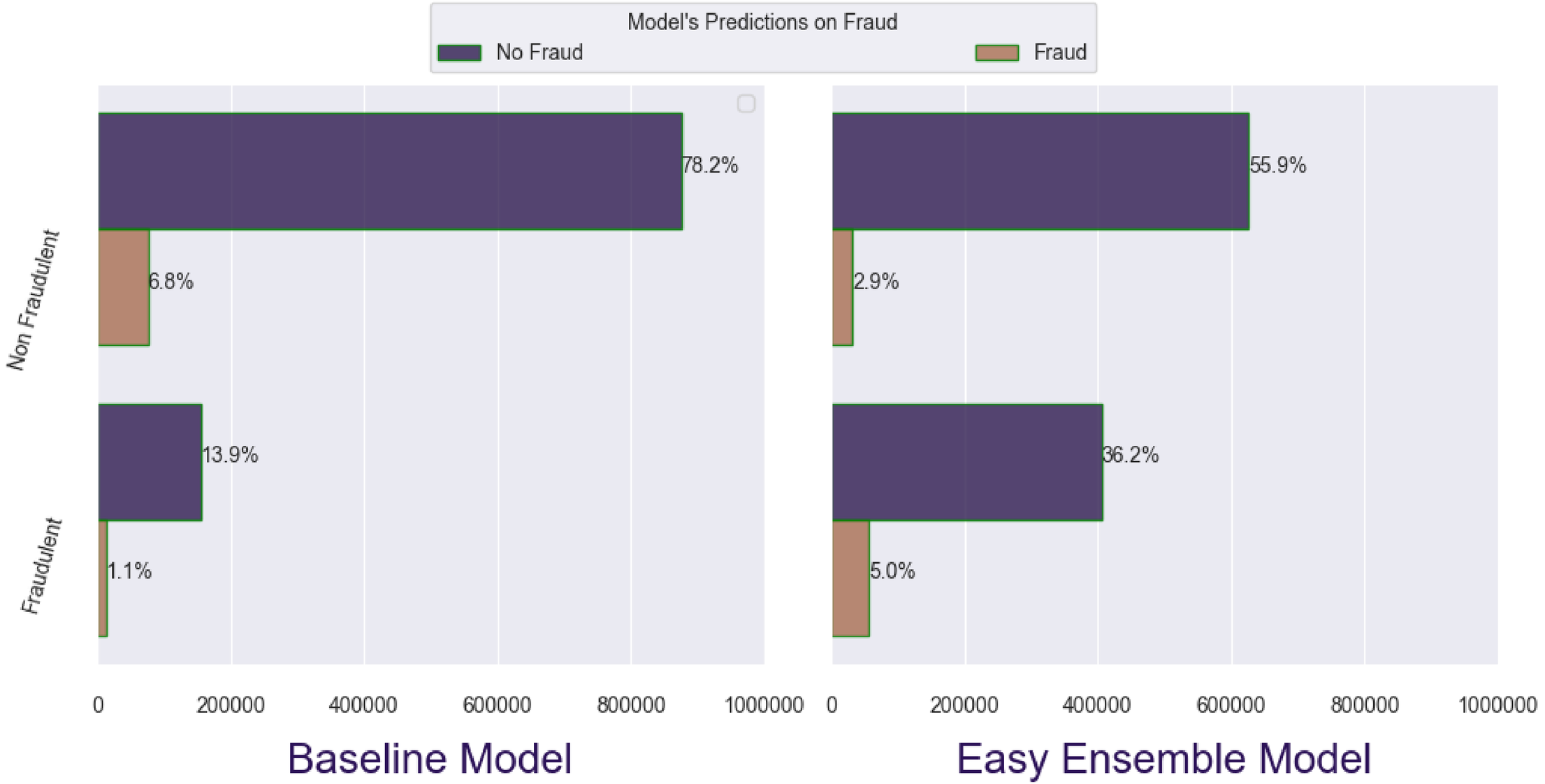
ML Model — ***37 %*** *of the fraudulent activities missed*

How many non-fraudulent clients were thought fraudulent?

Base Model — ***15 %*** *of non-fraudulent activities misclassified as fraud.*

ML Model — ***39 %*** *of non-fraudulent activities misclassified as fraud.*

Model performance based on billing transactions



06

PREDICTIVE MODELING

Train machine learning models, evaluate their performance, and use them to make predictions.

Model Error Analysis

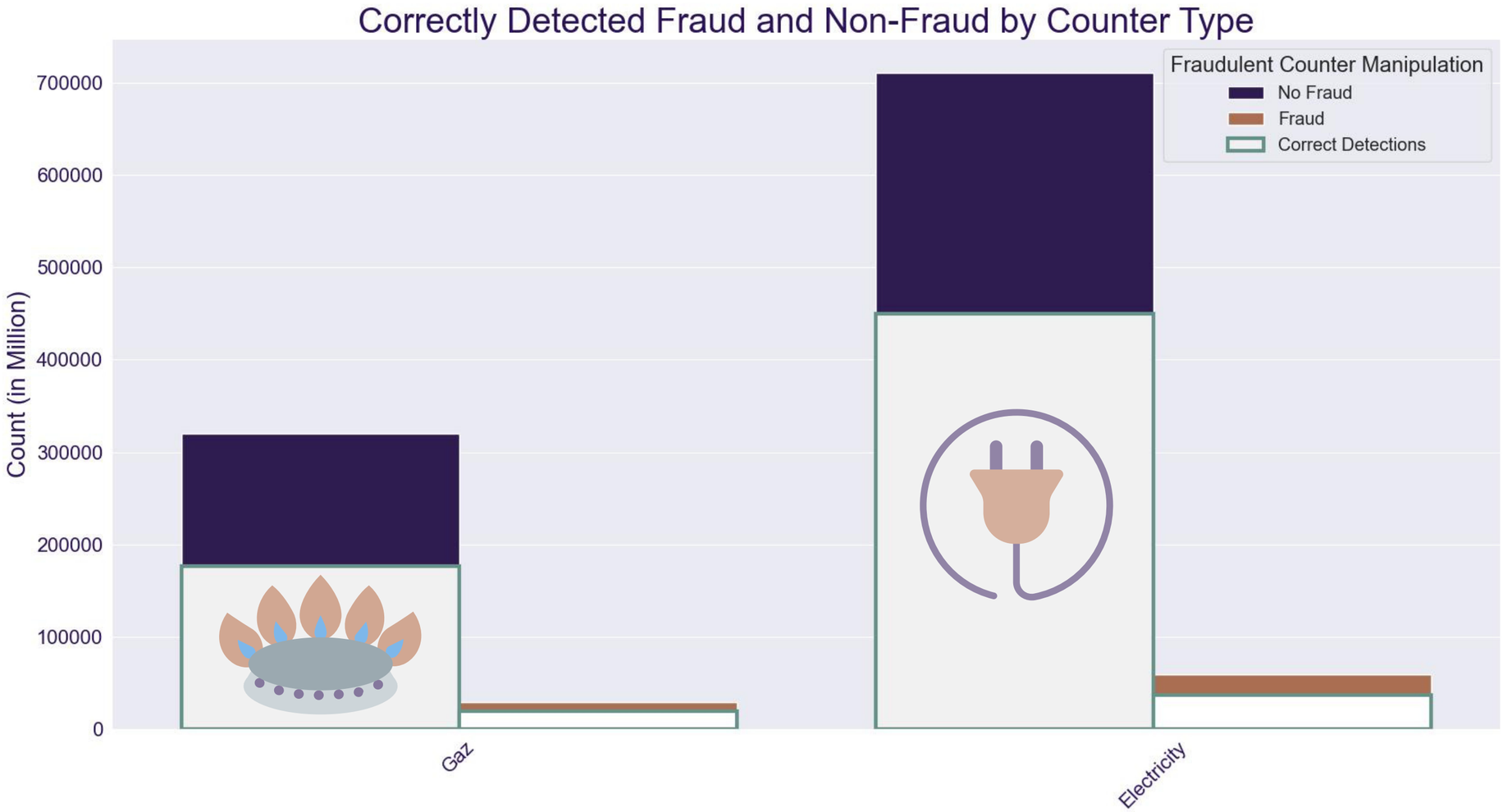
- The model predicts 63% of the frauds correctly, but falsely accuses 39% of honest clients.
- Raising or lowering the threshold (how sure the model has to be to predict fraud) to over, resp. under 50% did not improve the model.
- We compared the descriptive statistics of classifications:
  - missed fraud and wrong accusations: no substantial differences
  - correct predictions and errors: no substantial differences
- We found that the model performs better on electricity then on gas predictions.

06

PREDICTIVE MODELING

Train machine learning models, evaluate their performance, and use them to make predictions.

Model Error Analysis



- The model performs better on Electricity

- Gas: 56% correct
- Electricity: 63% correct

06

PREDICTIVE MODELING

Train machine learning models, evaluate their performance, and use them to make predictions.

Business Value

Deploying the model will decrease the financial damage by 63%

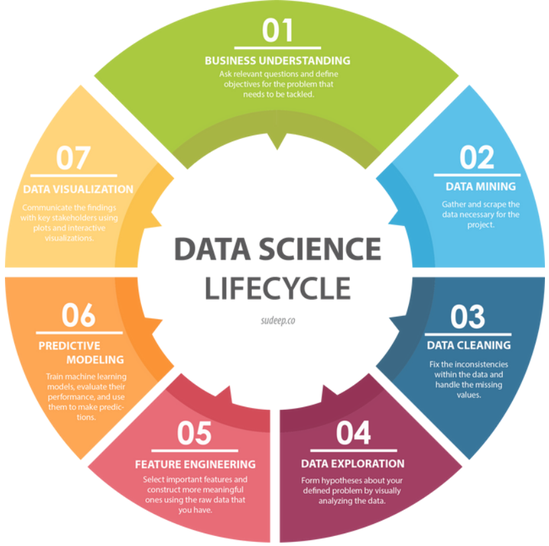
- Fraud lost the company 200 million Tunisian Dinars
- Out of 135 493 clients, 7 566 committed fraud (5.6 %)

- Every detected fraudulent client saves the company 26 434 Tunisian Dinars
- Estimation: The Model will help save 126 million Tunisian Dinars (detected fraud) in the next billing period

Model Usage

- Proceed with caution: We do not want to lose non-fraudulent clients
- For clients predicted to be fraudulent:
  - avoid direct accusations
  - possible strategy:
    - shorten timespan to next meter reading
    - advise employees to carefully look for signs of fraud

Future Work



- Further iterations of the Data Science Lifecycle
- Separate models for electricity and gas fraud detection
- Mine relevant data, e.g. crime rate of different regions
- Gain more business insight for better data cleaning (differentiate between faulty and fraudulent data)
- Get more computing power
- Train further machine learning models

Thank you!