Fraud Detection in Electricity and Gas Consumption in Tunisia

Team Members:

Ahmed Saad	201-800-251
Mazen Hassan	201-801-897
Youssef El Desoky	201-800-504

Contents

I - Data Problem Overview

II - Data Overview

III - Exploratory Data Analysis

IV – ML Models

V - Business Insights

Business Problem Overview

STEG is a public and 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

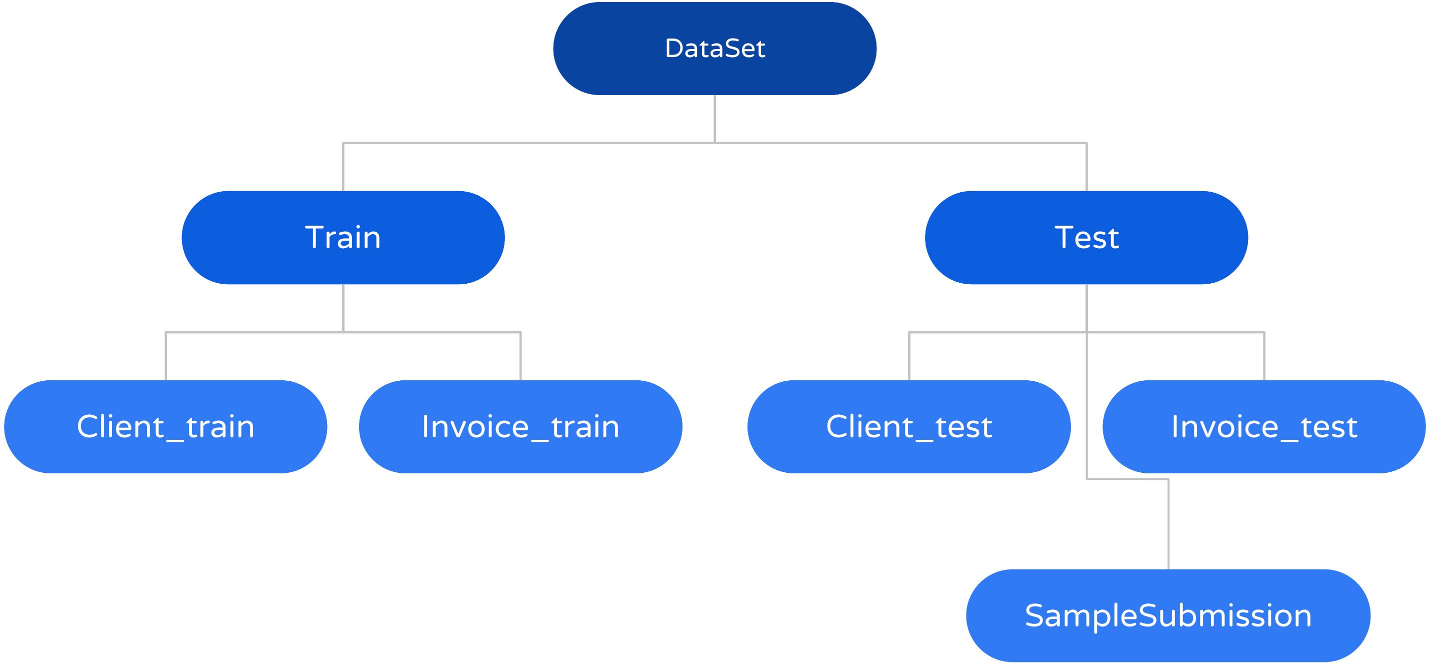
Société Tunisienne
de l'Électricité et du Gaz

الشركة التونسية
للكهرباء والغاز

Using the client's billing history, the aim of the challenge is to detect and recognize clients involved in fraudulent activities

The solution that we propose will enhance the company's revenues and reduce the losses caused by such fraudulent activities.

Data Overview



Data Overview

Client Data

Variable	Description
Client_id	Unique id for client
District	District where the client lives
Client_catg	Category client belongs to
Region	Area where the client lives
Creation_date	Date client joined
Target	Fraud:1 , not Fraud: 0

This Data describes each Client Information only and the result as being a Fraud or Not, however It does not have any information to detect the Fraud process.

Data Overview

Invoice Data

Variable	Description
Client_id	Unique id for client
Invoice_date	Date of the invoice
Tarif_type	Type of tax
Counter_number	Counter_number
Counter_statue:	The Counter health Statue (Working - not working - etc)
Counter_code	Counter_code

This Data describes each Invoice for each Client, therefore each client has several Invoice data which describes their consumption and their usage information.

Data Overview

Invoice Data

Variable	Description
Reading_ remarque	notes that the STEG agent takes during his visit to the client
Counter_ coefficient	An additional coefficient to be added when standard consumption is exceeded
Consommation_ Level_1 : 4:	Consumption_level_1 : 4
Old / New_index	Old / New_index
Months_number:	Months_number
Counter_type	Counter_type

This Data describes each Invoice for each Client, therefore each client has several Invoice data which describes their consumption and their usage information.

Data Manipulation

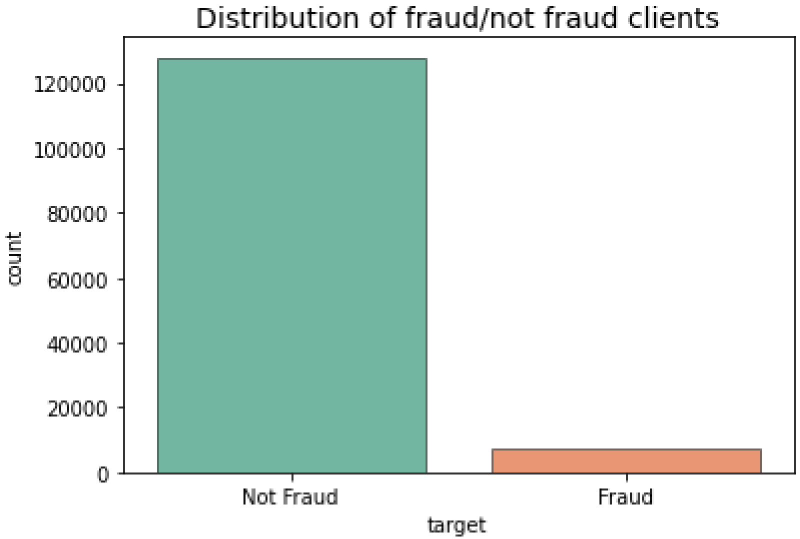
- We merged The Client Data and the Invoice Data Based on the Client ID
- We checked and dropped the duplicated rows
- We checked the null values and imputed them with appropriate techniques.
- We added some features depending on the group of invoices for each customer
    - Consumption level “1:4” Max - mean - min - STD
    - Reading remarque Max - mean - min - nunique

Exploratory Data Analysis

We have performed Different type of Exploratory Data Analysis:

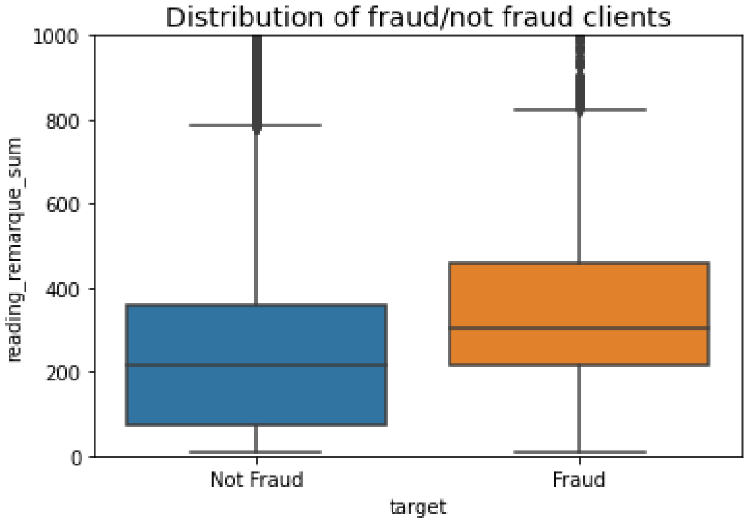
- Univariate Analysis
- Bi-variate Analysis

univariate Analysis



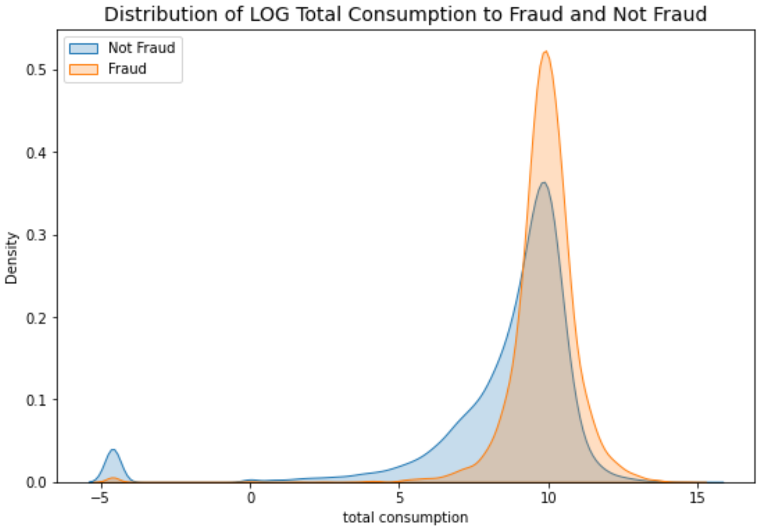
The distribution of target variable is unbalanced.

bivariate Analysis



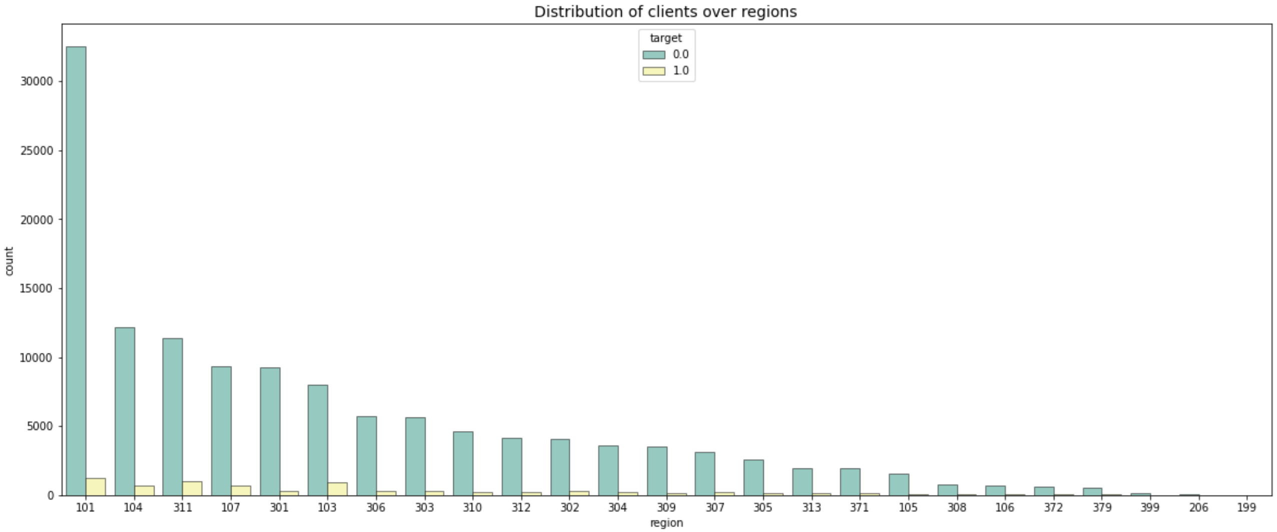
According to the above figure, the STEG agent will be almost able to give a good reading that differentiates between fraud and not fraud.

bivariate Analysis



we notice that total consumption of fraud clients are larger than usual.

bivariate Analysis



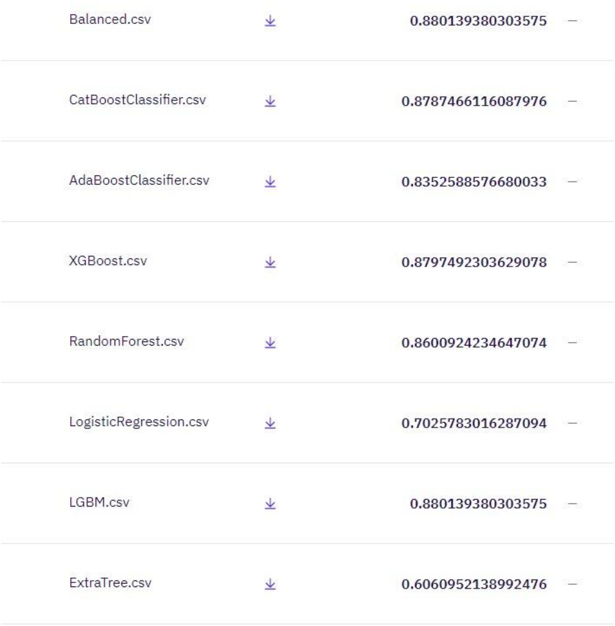
All districts/regions seem to have the same ratio of fraud/not fraud clients.

Models

We used different type of Machine learning Models in order to detect the fraud process:

- Boosting Models
  - XGBoost Model
  - LGB Model
  - Adaboost Model
  - Catboost Model
- Bagging Models
  - Random Forest
  - Decision Tree

Models Results



Balanced.csv	⤓	0.880139380303575	–
CatBoostClassifier.csv	⤓	0.8787466116087976	–
AdaBoostClassifier.csv	⤓	0.8352588576680033	–
XGBoost.csv	⤓	0.8797492303629078	–
RandomForest.csv	⤓	0.8600924234647074	–
LogisticRegression.csv	⤓	0.7025783016287094	–
LGBM.csv	⤓	0.880139380303575	–
ExtraTree.csv	⤓	0.6060952138992476	–

LGBM out performed all the other models with success rate 88% to detect fraud process

Business Insights

- The Fraud clients have usually higher consumption than normal clients
- Districts/regions have no direct relation with the fraud process
- The consumption level 1 has the highest effect on the Fraud detection
- The STEG agent will be almost able to give a good reading that differentiates between fraud and not fraud.