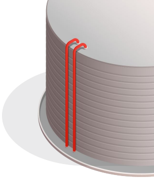
Fraud Detection in Electricity and Gas Consumption.

STEG Electricity and gas distribution Company.

Oasis
ANALYTICS



Meet the Team

Milkah

Gloria

Ibrahim

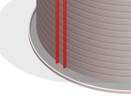
Dorcas

Stella

Sodiq

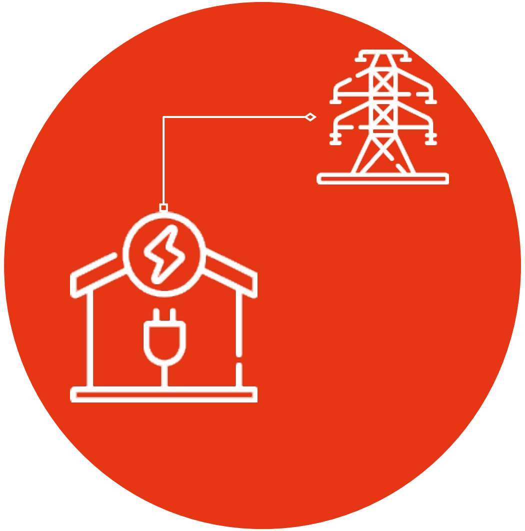
TABLE OF CONTENTS

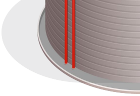
1. Introduction
2. Problem Statement
3. Data Sets
4. Exploratory Data Analysis
5. Features Engineering
6. Modelling
7. Deployment
8. Conclusion



INTRODUCTION

- What is Fraud?
- Fraud Detection.
- Impact of fraud.

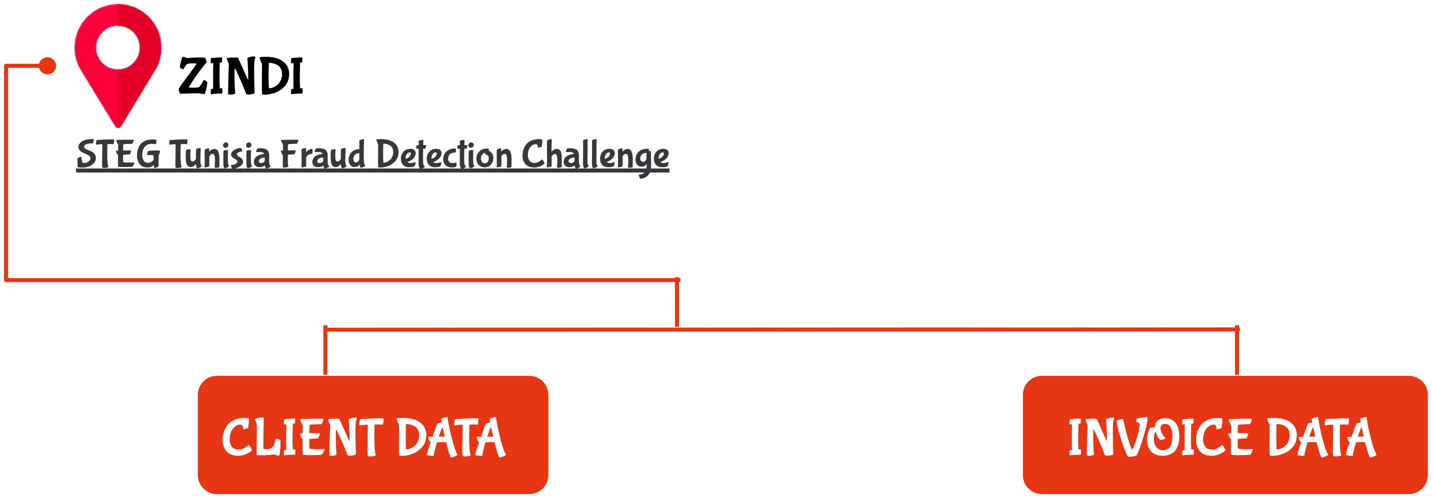




PROBLEM STATEMENT

- Tremendous Loss in Revenue due to fraudulent meter manipulations.
- Detect fraudulent customers based on billing histories.

DATA SETS



ZINDI

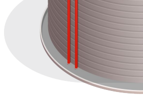
STEG Tunisia Fraud Detection Challenge

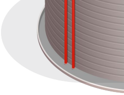
CLIENT DATA

INVOICE DATA

- Contains personal information of each client. E.g. client_id, target… e.t.c.
- Samples – 135,493.

- Contains transactions information performed by each client. E.g. client_id, invoice_date, meter_type, consumption_level e.t.c.
- Samples – 4,476,749.





EXPLORATORY DATA ANALYSIS

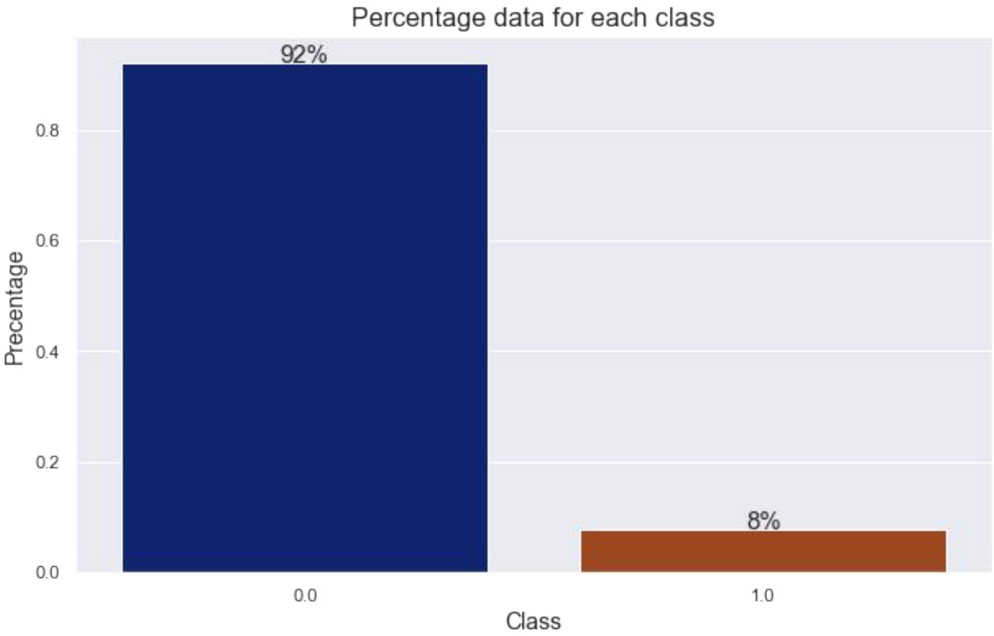
Basic Analysis

1. Merge the client data and invoice data.
2. Data Consist of 4,476,749 samples and 21 features.
3. Data Type – “Numerical”, “Categorical”.
4. Missing Value Count: “Zero”
5. Target Variable: “0.0 – Non Fraud”, “1.0 – Fraud”.

EXPLORATORY DATA ANALYSIS

Advanced Analysis

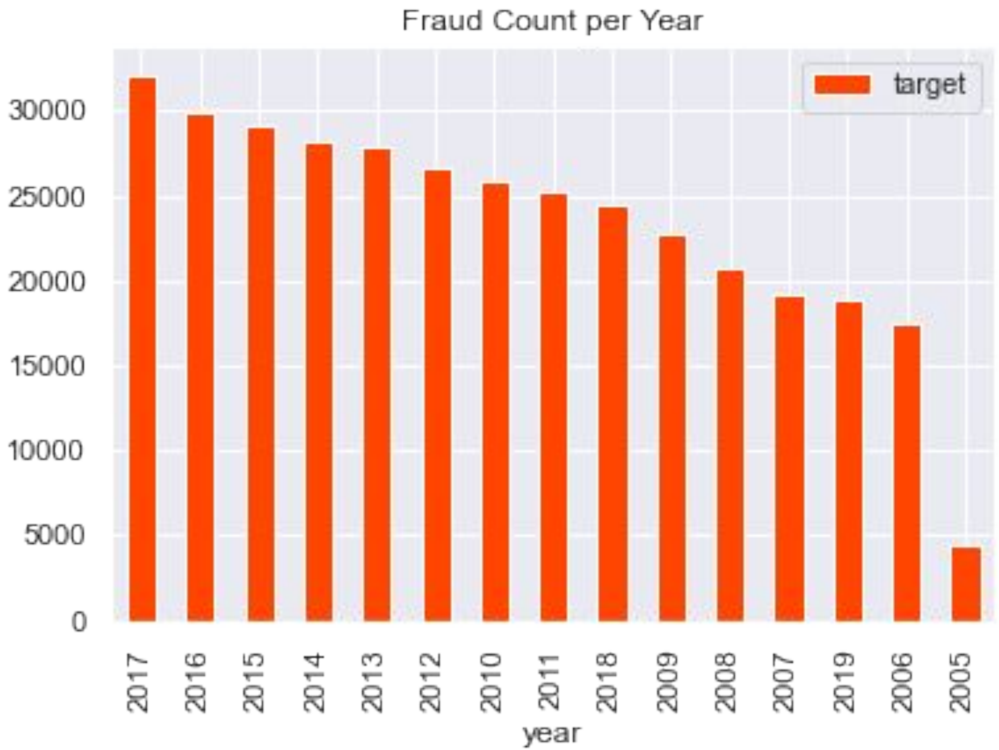
1 Examine the *Target* variable.



EXPLORATORY DATA ANALYSIS

Advance Analysis "fraud cases"

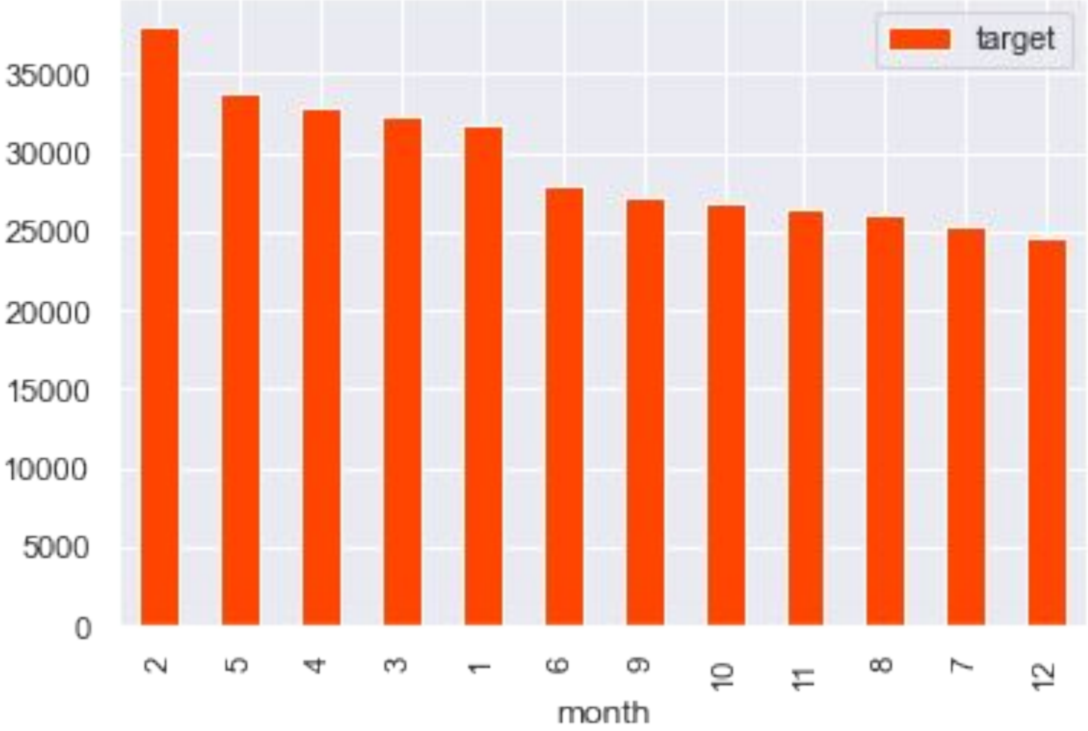
2 Fraud cases per year.



EXPLORATORY DATA ANALYSIS

Advance Analysis "fraud cases"

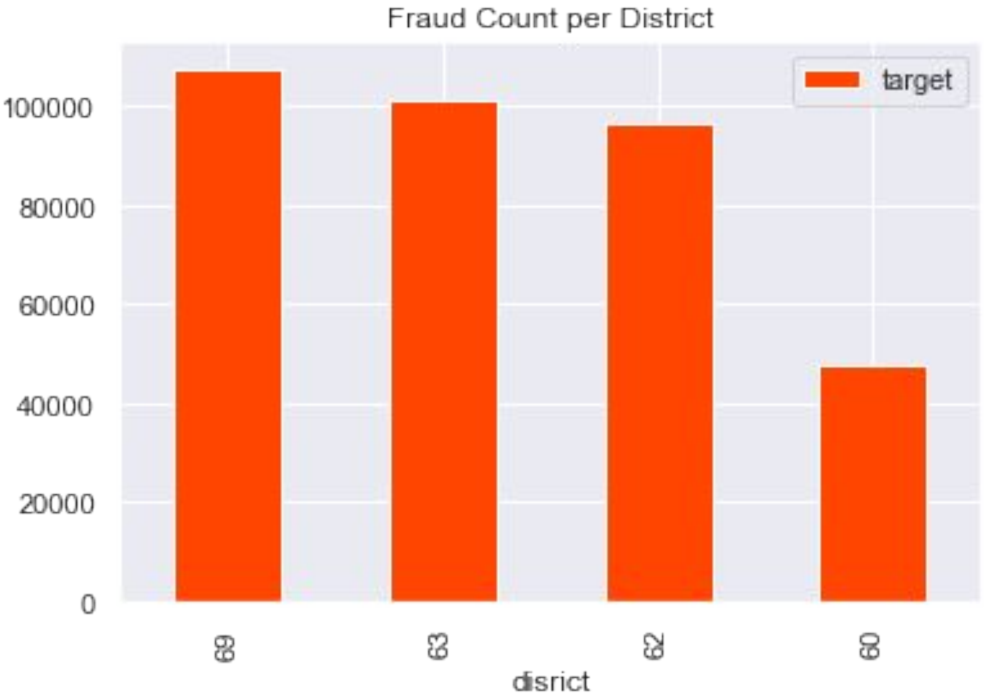
3 Fraud cases per month.



EXPLORATORY DATA ANALYSIS

Advance Analysis “fraud cases”

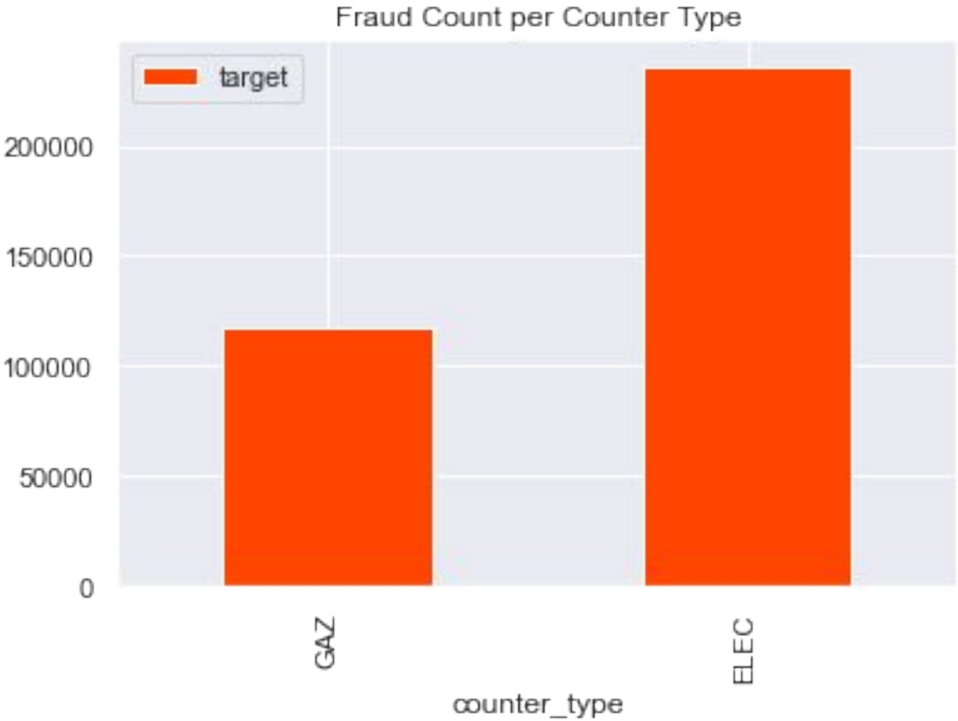
4 Fraud cases per district.

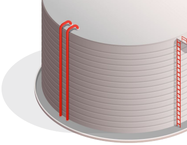


EXPLORATORY DATA ANALYSIS

Advance Analysis "fraud cases"

5 Fraud cases per counter type..





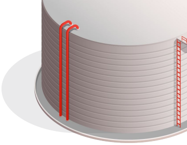
FEATURES ENGINEERING

1 Data Type Change.

Datetime, Categorical, Numerical.

- ❖ To datetime ➪ Creation_date, Invoice_date
- ❖ To Categorical (object) ➪ District, Region, Counter_type
- ❖ To Integer ➪ Counter_statue, Target

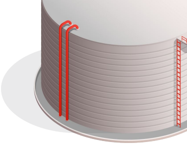
FEATURES ENGINEERING



2 Split data into training and validation set.

80/20 split.

- Shuffled and stratified based on the target
- Split into 80% training and 20% validation set

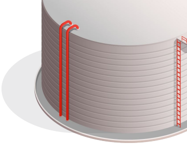


FEATURES ENGINEERING

3 Features Scaling.

Standard Scaler.

- ❖ Scaled numerical features
- ❖ Tool used ➪ Sklearn 's StandardScaler



FEATURES ENGINEERING

4 Features Encoding.

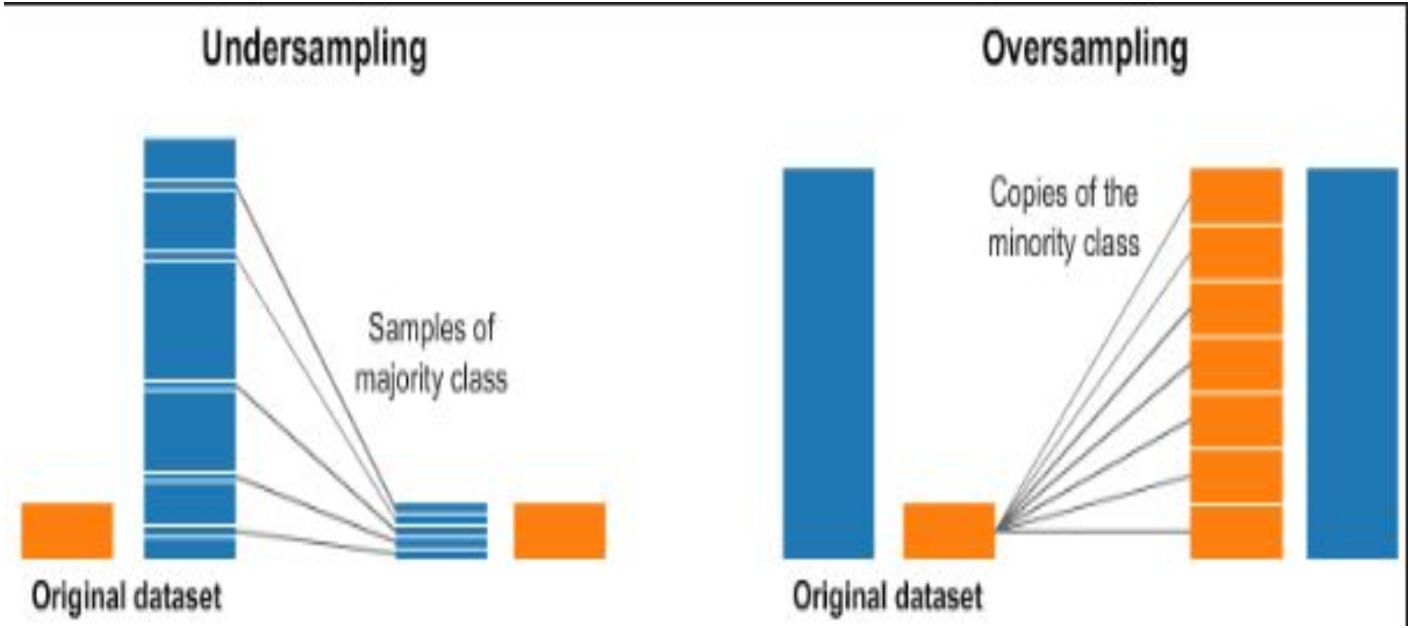
Onehotencoder.

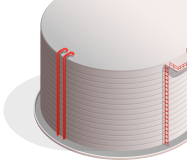
- ❖ Encoded categorical features
- ❖ Tool used ⇨ Sklearn 's OneHotEncoder

FEATURES ENGINEERING

5 Data Resampling.

Random Sampler.





MODELING

1. Logistic Regression

"lbfgs"

2. Decision Tree

"Entropy"

3. Support Vector Machine

"SGDClassifier"

4. Naive Bayes

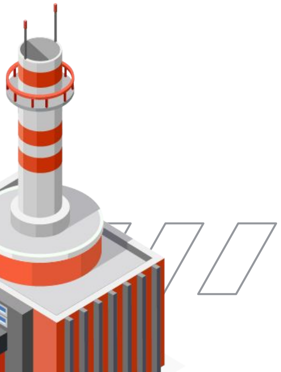
"Gaussian"

5. Random Forest

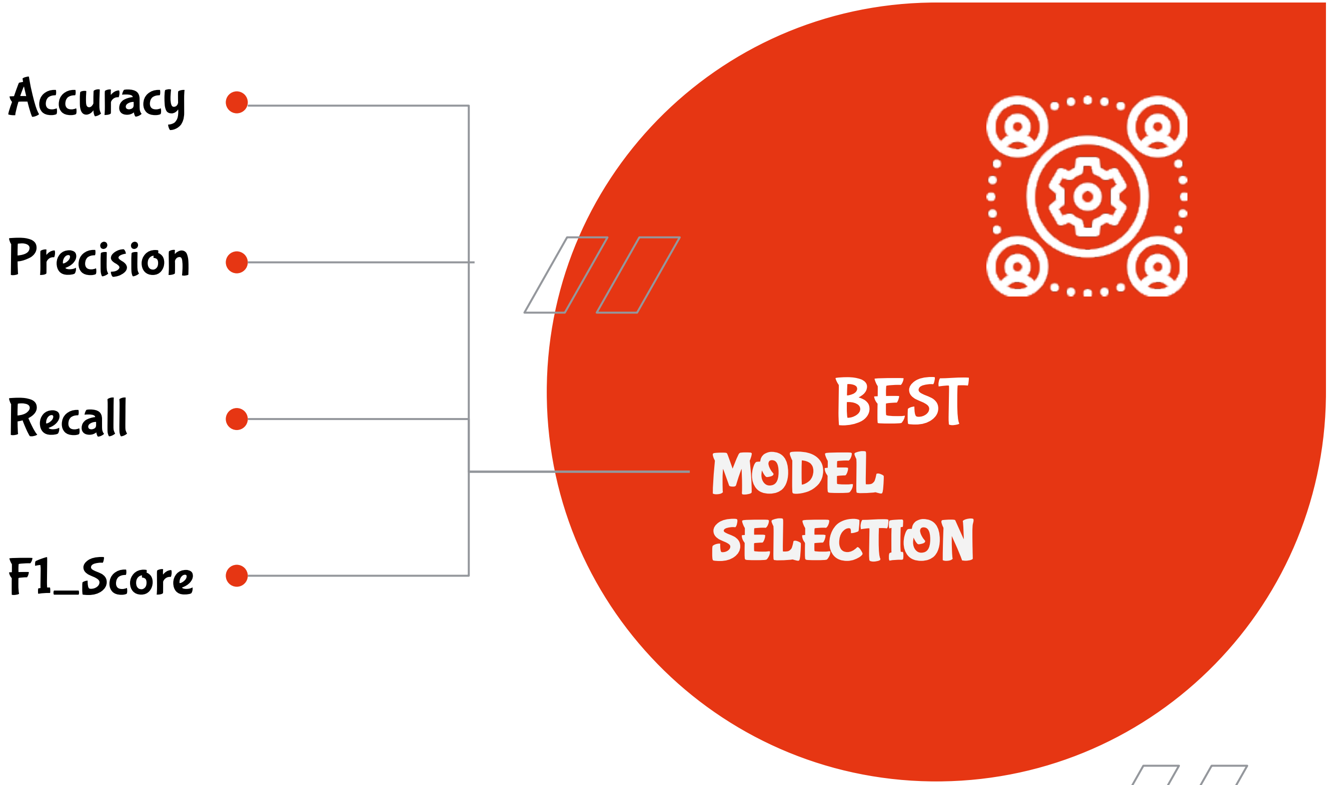
"Gini"

6. Bagging

"Decision Tree"



PERFORMANCE METRICS



Accuracy

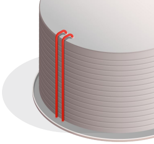
Precision

Recall

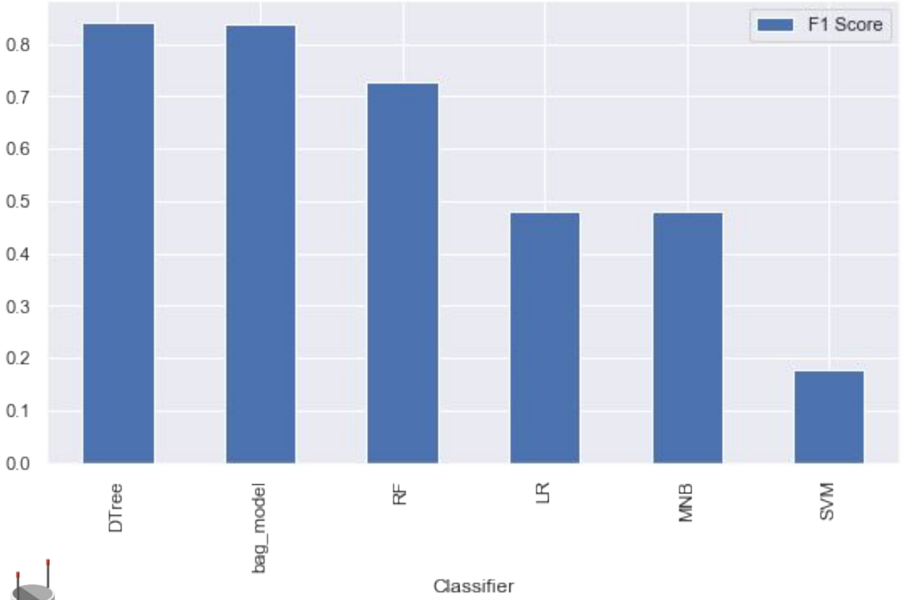
F1_Score

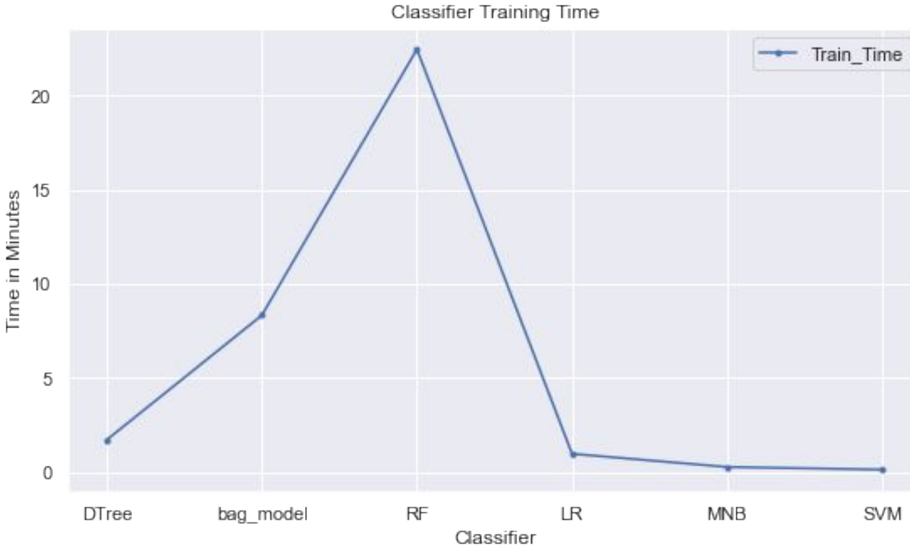
BEST MODEL SELECTION

MODEL PERFORMANCE



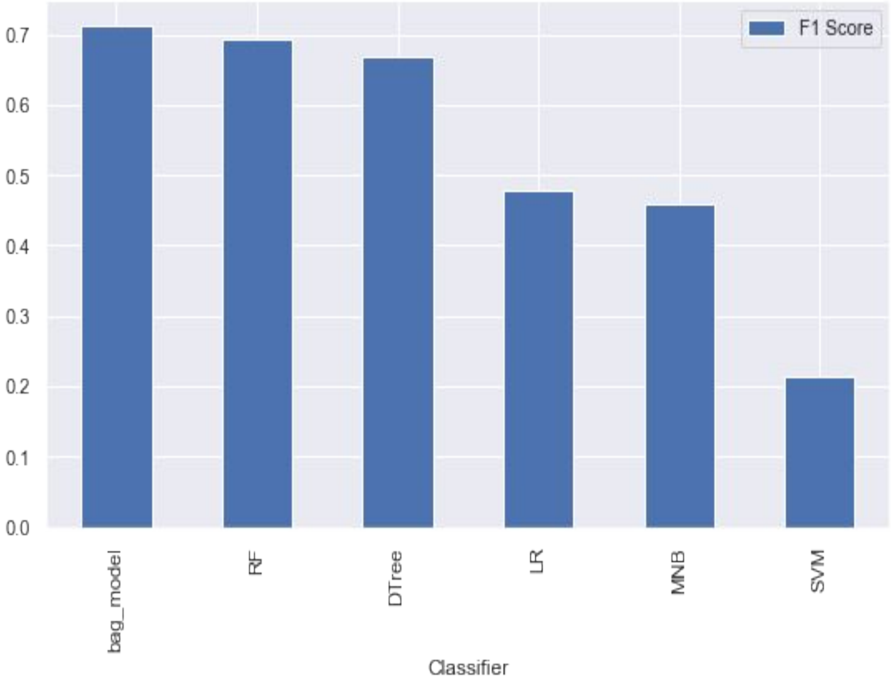
1 No Sampling Done.

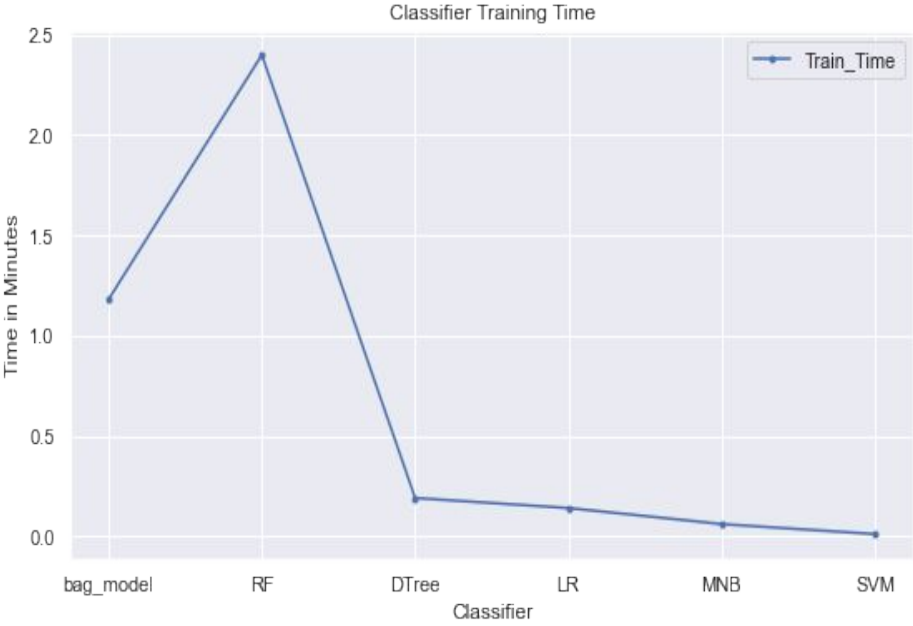




MODEL PERFORMANCE

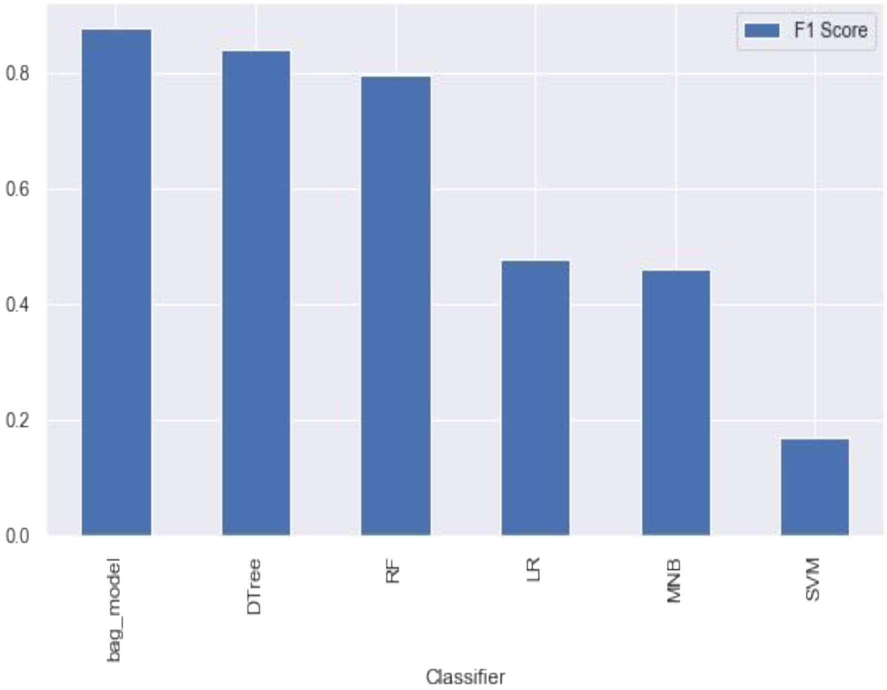
2 Undersampling.

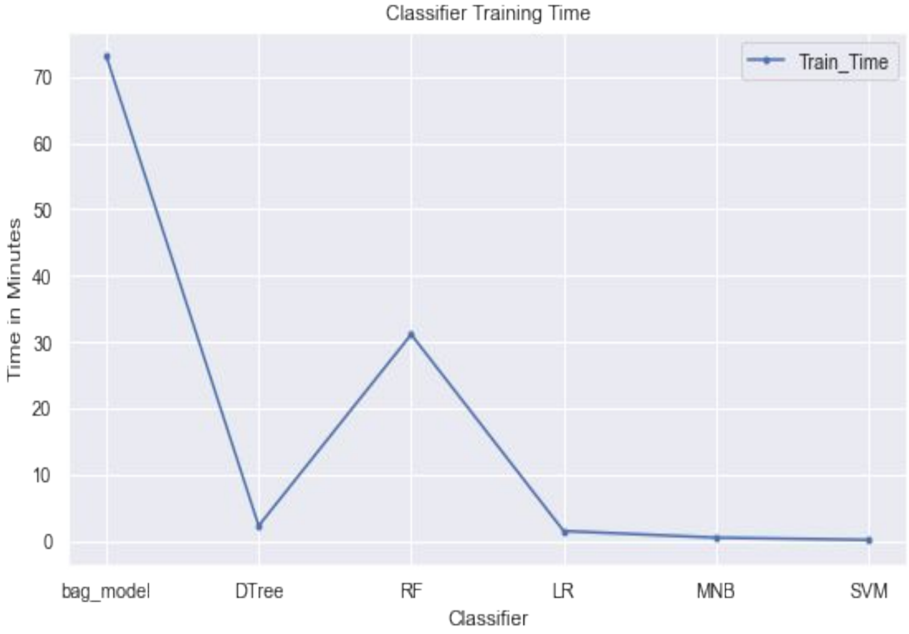




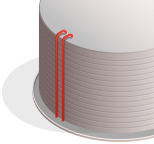
MODEL PERFORMANCE

3 Oversampling.





MODEL PERFORMANCE



Bagging Classifier

F1-Score: 0.87

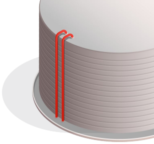
Train Time: 73.13mins

Best Model:

Decision Tree Classifier

F1-Score: 0.84

Train Time: 2.30mins



DEPLOYMENT

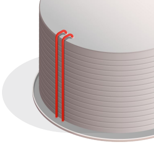
1 Model Deployment.

(a) AWS ’s EC2 Instance Using Python ’s Streamlit App.

2 Application Prediction:

(a) Client Selection from test dataset.
(b) Entering client ’s metadata for new clients.

CONCLUSION



- Model Predictive Accuracy → 84.2%.

- A solution that withstands the test of time.

- Revenue Increase.

THANK YOU.

EXPLORE | DATA SCIENCE ACADEMY