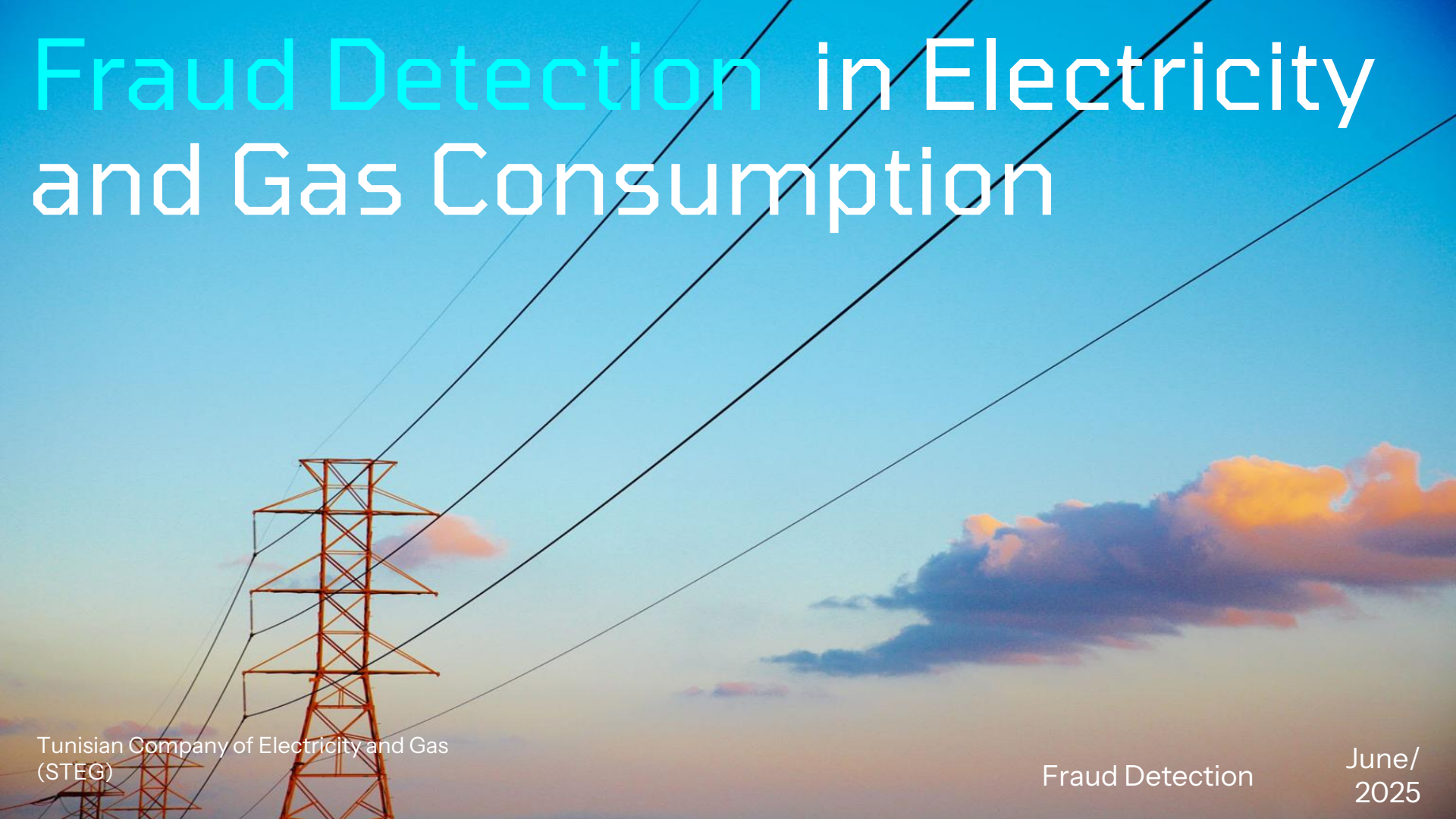
Fraud Detection in Electricity and Gas Consumption



Tunisian Company of Electricity and Gas (STEG)

Key Objective

Why we're here: To save money lost to energy consumption fraud

Tunisian Company of Electricity and Gas (STEG)

Table of Contents

01 Business and Data Understanding

- **Background**: Our Client is the Tunisian Company of Electricity, and Gas (STEG) – a public and non-administrative company responsible for delivering electricity and gas across Tunisia
- **Business Problem**: STEG suffered tremendous losses (approx. 200m Tunisian Dinars) due to fraudulent manipulations of meters by consumers
- **Proposed Solution:** Build a fraud detection model that can detect and recognize client accounts involved in fraudulent activities
- **Value Proposition:** The model will enhance revenues and reduce losses due to fraud

Tunisian Company of Electricity and Gas (STEG)

02 Data Analysis and Modelling Methodology

I. **Data Preparation (pre-processing)**

   I. **Exploratory Data Analysis:** We clean and transform the data into a format appropriate for modelling (changed date formats, renamed columns). We also make some preliminary findings on the dataset (class imbalance)

   II. **Feature Engineering:** We aggregate rows in the invoice datasets to create new columns and merge the client and invoice datasets

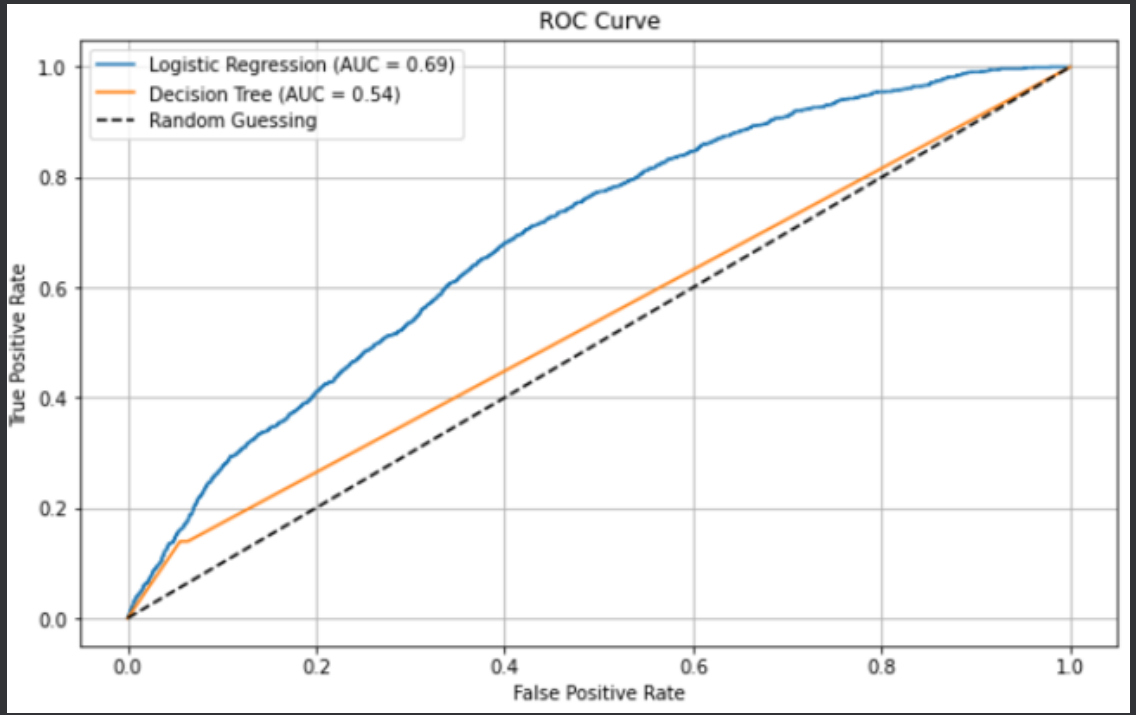
II. **Data Modelling and Evaluation**

   I. **Training:** We train 2 sets of models – **Untuned** and **Tuned**. The first set is composed of 2 models (Logistic Regression, Decision Tree) while the second set is composed of 3 models (first 2 + XGBoost)

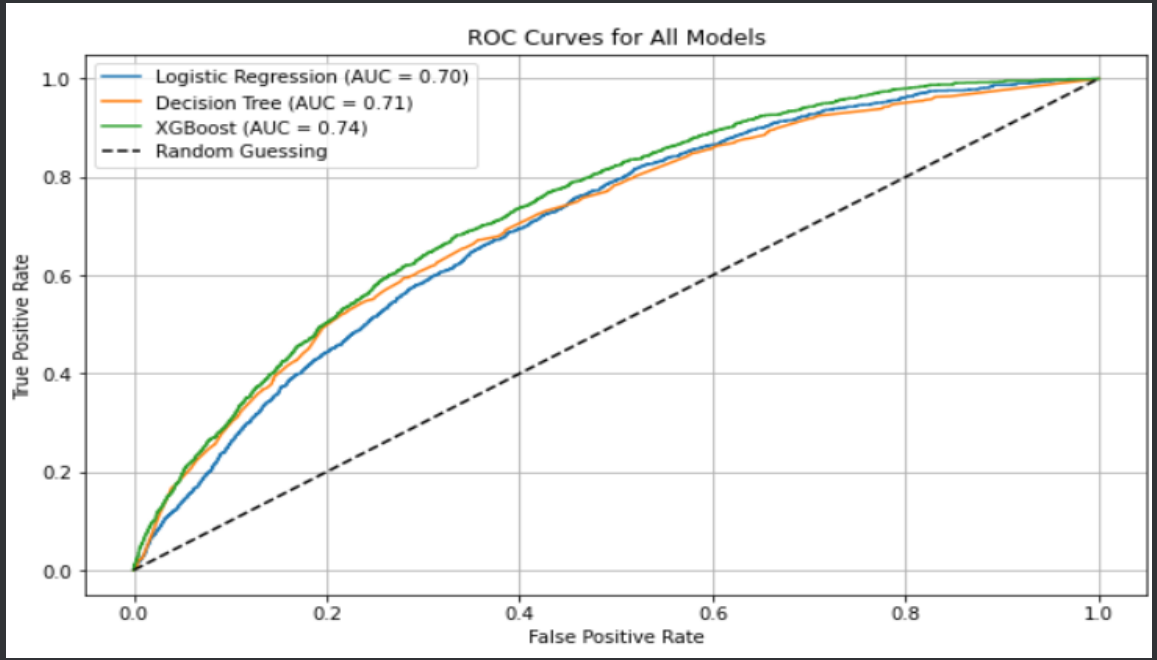
   II. **Evaluation:** Our evaluation metric is **AUC score –** we pick the best out of the 3 tuned models and perform further tuning of hyperparameters via **GridSearchCV** to optimize performance further. This final model is our deliverable to the client

03 Modelling and Evaluation Results (Untuned Models)

- ✔ Model with the best AUC test score: **Logistic Regression (69%)**
- ✔ Logistic Regression also has stable generalization (AUC train score = 68%)
- ✔ Decision tree barely performs better than a random model

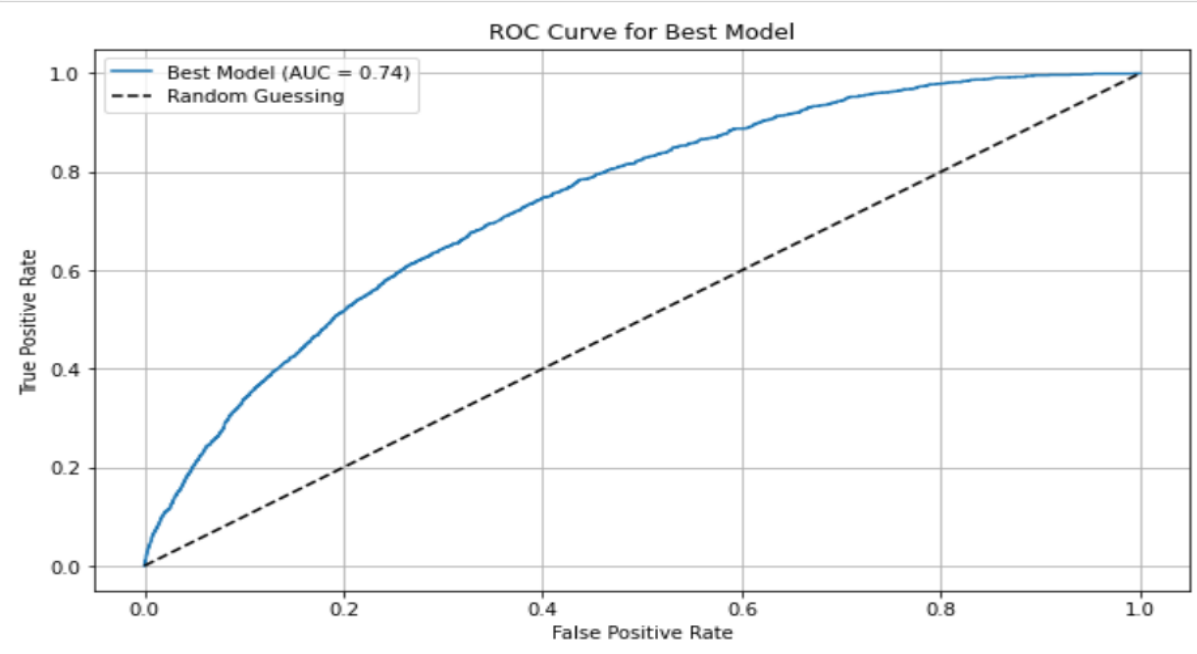


03 Modelling and Evaluation Results (Tuned Models)



- ✔ Model with the best AUC test score: **XGBoost (74%)**
- ✔ **XGBoost** also has the highest F1 Score (**21%**) and solid generalization (AUC train score = 80%)

03 Modelling and Evaluation Results (BEST MODEL: XGBOOST-GRIDSEARCHCV)



- ✔ Best Model: **Tuned XGBoost (test AUC = 74%)**
- ✔ **Tuned XGBoost** has a test F1 Score of **22%** - highest
- ✔ Solid generalization with a narrow train/test gap (**train AUC = 76%**)

04 Conclusions and Next Steps

- The model with the highest predictive power/performance is:
  - Name: Tuned XGBoost
  - Tuning technique: GridSearchCV
  - Best Model (ROC)AUC Score: 74%
- Next Steps:
  - **Feature transparency:** Identify the top features most predictive of fraud: region or district? Consumption threshold?
  - Incorporate Precision-Recall AUC (better than ROC AUC for imbalance)
  - Deploy the model onto the client systems

Thank you

Questions? Please contact
timavedi@gmail..com