Electricity Theft and Energy Fraud Detection

Ali Jassim Rajab
ajr5102@rit.edu

Follow this and additional works at: https://scholarworks.rit.edu/theses

Recommended Citation

This Master’s Project is brought to you for free and open access by RIT Scholar Works. It has been accepted for inclusion in Theses by an authorized administrator of RIT Scholar Works. For more information, please contact ritsscholarworks@rit.edu.
Electricity Theft and Energy Fraud Detection

by

Ali Jassim Rajab

A Capstone Submitted in Partial Fulfilment of the Requirements for
the Degree of Master of Science in Professional Studies:
Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology
RIT Dubai
May 04, 2023
Master of Science in Professional Studies:
Data Analytics

Graduate Capstone Approval

Student Name: Ali Jassim Rajab
Graduate Capstone Title: Electricity Theft and Energy Fraud Detection

Graduate Capstone Committee:

Name: Dr. Sanjay Modak Date:
Chair of committee

Name: Dr. Ioannis Karamitsos Date:
Member of committee
Acknowledgments

I would like to extend my heartfelt gratitude to the capstone project advisor Dr. Ioannis Karamitsos for the unwavering support and guidance that have been invaluable to me throughout this project and to Dr. Sanjay Modak for his help, recommendation, and advice, which contributed to the successful completion of the project.

My thanks are extended to all the instructors and colleagues in cohort 8 for their efforts, knowledge sharing, and collaboration that helped me grow both personally and professionally, and also to the Rochester Institute of Technology and the Department of Graduate Programs & Research for providing this opportunity.

Last but not least, I would like to thank my workplace for the invaluable support, encouragement, and approval that made it possible to successfully reach this stage and complete all the requirements of the program.
Abstract

Electricity theft is a widespread problem with significant negative economic and financial impacts. This problem is still a challenge for energy utilities all around the world. Estimates place the cost of power theft and fraud in the energy sector at $96 billion yearly, with costs of up to $6 billion in only the United States, making it the third largest form of theft in the country. Therefore, tackling electrical fraud and theft is more necessary now than ever.

Different methods are being used or developed to detect this practice and lessen its effects. The project aims to detect electricity theft and fraud through machine learning by analyzing customers’ consumption patterns, among other features like bill history, reading remarks, and regions. The dataset is provided in Kaggle by the Tunisian Company of Electricity and Gas (STEG), containing 43 years of records of more than 135,000 customers with 21 different attributes. The project will adopt CRISP-DM, as the methodology used for the project completion, which provides a structured approach to data mining project planning.

The adapted supervised machine learning models for this project were decision tree, random forest, and support vector machine since they are considered the most common models used in fraud detection based on the conducted literature review. The final model selection was based on different metrics, the accuracy, precision, recall, and F1 score of the model. The random forest model surpassed the other two models, achieving an accuracy of 82.73%.

Fraud detection will reduce the significant impact of financial losses and enhance the service quality provided by electric utilities. Furthermore, the obtained results from the selected model can be used by electric utilities, especially in developing countries, to prioritize their Advanced Metering Infrastructure (AMI) installation plans in areas where fraud is detected, which will smoothen the transformation journey financially.

Keywords: Electricity theft, Fraud Detection, Machine Learning, Conventional meter, Non-technical Losses
# Table of Contents

ACKNOWLEDGMENTS.............................................................................................................. II
ABSTRACT ............................................................................................................................... III
LIST OF FIGURES..................................................................................................................... V
LIST OF TABLES....................................................................................................................... VI
LIST OF EQUATIONS................................................................................................................ VI

CHAPTER 1 .................................................................................................................................. 7
  1.1 BACKGROUND .................................................................................................................. 7
  1.2 STATEMENT OF PROBLEM ......................................................................................... 9
  1.3 PROJECT GOALS .......................................................................................................... 10
  1.4 METHODOLOGY ......................................................................................................... 10
  1.5 LIMITATIONS OF THE STUDY .................................................................................. 13

CHAPTER 2 – LITERATURE REVIEW .................................................................................... 14

CHAPTER 3 - PROJECT DESCRIPTION .................................................................................. 25
  3.1 OVERVIEW .................................................................................................................. 25
  3.2 DATA ANALYTICS TOOLS AND SOFTWARE USED .............................................. 26
  3.3 DATASET DESCRIPTION ......................................................................................... 26

CHAPTER 4 - PROJECT ANALYSIS ..................................................................................... 28
  4.1 EXPLORATORY DATA ANALYSIS AND VISUALIZATION ...................................... 28
  4.2 DATA PREPARATION ............................................................................................... 39
    4.2.1 DATA PREPROCESSING ....................................................................................... 39
      4.2.1.1 Data Cleansing ............................................................................................ 39
      4.2.1.2 Data Transformation .................................................................................. 42
      4.2.1.3 Feature Engineering .................................................................................... 42
      4.2.1.4 Feature Selection ....................................................................................... 44
  4.3 MODELING .................................................................................................................. 45
    4.3.1 DECISION TREE ............................................................................................... 47
    4.3.2 RANDOM FOREST ............................................................................................. 49
    4.3.3 SUPPORT VECTOR MACHINE .......................................................................... 50
  4.4 FINAL MODEL EVALUATION AND SELECTION ..................................................... 51
  4.5 MODEL DEPLOYMENT ............................................................................................... 52

CHAPTER 5 CONCLUSION .................................................................................................... 53
  5.1 CONCLUSION ............................................................................................................. 53
  5.2 RECOMMENDATIONS ............................................................................................ 54
  5.3 FUTURE WORK ....................................................................................................... 54

BIBLIOGRAPHY .................................................................................................................... 55
List of Figures

Figure 1. CRSIP-DM .................................................................................................................................. 10
Figure 2. Client Dataset Structure ........................................................................................................... 28
Figure 3. Client Dataset Summary ............................................................................................................. 28
Figure 4. Invoice Dataset Structure ........................................................................................................ 29
Figure 5. Invoice Dataset Summary ........................................................................................................ 29
Figure 6. Number of Records for Target, Region, & Client Category ....................................................... 30
Figure 7. Number of Records for Tarif Type, District & Reading Remarks ............................................. 30
Figure 8. Customer Count Per Year 1977 - 2019 .................................................................................... 31
Figure 9. Counter Type Distribution ......................................................................................................... 32
Figure 10. Yearly Consumption Pattern ................................................................................................... 32
Figure 11. Target Attribute Distribution .................................................................................................. 33
Figure 12. Target Distribution Counter Type Wise ................................................................................... 33
Figure 13. Target Distribution District Wise ............................................................................................ 34
Figure 14. Target Distribution Client Category Wise ............................................................................... 35
Figure 15. Target Distribution Region Wise ............................................................................................ 36
Figure 16. Target Distribution Tarif Type Wise ......................................................................................... 37
Figure 17. Correlation Matrix .................................................................................................................. 38
Figure 18. Missing Values Check ............................................................................................................. 39
Figure 19. Counter Status Count .............................................................................................................. 40
Figure 20. Duplicated Rows Check ......................................................................................................... 40
Figure 21. Consumption Levels Boxplot .................................................................................................. 40
Figure 22. Yearly Average Consumption Level 2 ..................................................................................... 41
Figure 23. Dataset Down-Sampling .......................................................................................................... 41
Figure 24. Dataset Split ............................................................................................................................ 45
Figure 25. Training & Testing Dataset Balance Check ............................................................................... 46
Figure 26. Decision Tree Model .............................................................................................................. 47
Figure 27. Decision Tree Accuracy .......................................................................................................... 48
Figure 28. Decision Tree Model Precision, Recall & F1 Score ................................................................. 48
Figure 29. Random Forest Accuracy ....................................................................................................... 49
Figure 30. Random Forest Model Precision, Recall & F1 Score ............................................................. 49
Figure 31. Support Vector Machine Models' Accuracy Per Cost .............................................................. 50
List of Tables

Table 1. Attribute Description ..................................................................................................................26
Table 2. Dataset Overview Stage 1 ...........................................................................................................43
Table 3. Dataset Overview Stage 2 ...........................................................................................................43
Table 4. Final Dataset Overview ..............................................................................................................43
Table 5. Support Vector Machine Models' Accuracy, Precision, Recall & F1 Score .................................51
Table 6. Machine Learning Models' Performance Comparison ...............................................................51

List of Equations

Accuracy Equation ..................................................................................................................................46
Recall Equation ......................................................................................................................................46
Precision Equation .................................................................................................................................46
F1 Score Equation ..................................................................................................................................46
Electricity is an essential part of modern life. Energy utilities' primary objective is to ensure its delivery through three main stages: generation, transmission, and electricity distribution to customers. The electricity journey will encounter technical and non-technical losses; the technical losses are related to the energy dissipated in the network, including the transmission lines and transformers, which depend on the network characteristics and the mode of operation, which are taken into consideration.

On the other hand, the non-technical losses (NTL) can be installation errors, faulty meters, billing errors, or energy fraud and electricity theft, significantly impacting energy utilities’ revenue. Electricity theft is a growing problem across the globe with significant negative economic and financial impacts. It contributes to raising the electricity cost paid by the customer and creates the risk of fire and electrocution.

Electricity theft can take many forms, including meter tampering through physical manipulation or software-based attacks to show lower power readings. Illegal connections are another form where a line or cable will be rigged from the power source bypassing the meter. In addition to that, unpaid bills and billing irregularities are also considered another form of electricity theft, leading to financial losses.

Estimates place the cost of power theft and fraud in the energy sector at $96 billion yearly, according to (Networked Energy Services, 2020). It is considered the third-largest form of theft in the United States, according to (Pepco, n.d.). Whereas in the United Kingdom, more than £400 million worth of energy is stolen yearly, and estimates suggest an addition between £20 and £30 to “every honest customer’s bill” (LeicestershireLive, 2022).
Electricity theft can cause the generation units to be overloaded. Additionally, since the power utilities lack an estimate of the amount of electricity to be given to both legitimate and illegitimate consumers, the quality of the electricity supply is negatively impacted. This overload may lead to or cause an overvoltage, hinder performance, and possibly damage customers' equipment. This enormous volume of Non-Technical Losses could lead to a generator trip, interrupting electricity to all consumers.

The Installation of Advanced Metering Infrastructure (AMI) was one of the approaches of many countries to counteract electricity theft and fraud, among other advantages such as network problem identification and quick power restoration. The AMI allows for frequent monitoring and collection of data to be analyzed and used for fraud detection using machine learning. However, it’s still a challenge, especially for developing countries, to adopt this approach due to the installation cost (Nadeem & Arshad, 2021). Alternatively, data collected from conventional meters - despite limitations - in such countries can be used for fraud detection, like the data provided in Kaggle by STEG, to reduce the financial impacts and prioritize the installation plan of AMI in the area where fraud is detected, resulting in a more feasible transition.

Electricity theft and fraud must be tackled due to the above-mentioned problems, and many tools are now being used or developed to detect this activity and reduce its impact. Through machine learning algorithms, the collected data are used for analysis, hidden patterns discovery, and insights obtainment to detect fraudulent customers. Realistically, Power utilities will never be able to eliminate electricity theft and fraud completely, but they can take measures and action to detect and reduce it.
1.2 Statement of problem

Electricity theft is a growing problem across the globe, classified as non-technical losses that are defined as unmetered use of electricity, and causes a substantial economic and financial impact. Energy utilities around the world continue to struggle with such losses. Estimates indicate the cost of electricity theft and fraud in the energy sector worldwide at $96 billion annually, with expenses of up to $6 billion in the United States alone (Networked Energy Services, 2020). Additionally, in the United Kingdom, for example, more than £400 million worth of energy is stolen yearly, and estimates suggest an addition between £20 and £30 to “every honest customer’s bill” (LeicestershireLive, 2022).

Electricity theft can take many forms, including meter tampering through physical manipulation or software-based attacks to show lower power readings. Illegal connections are another form where a line or cable will be rigged from the power source bypassing the meter. In addition to that, unpaid bills and billing irregularities are also considered another form of electricity theft, leading to financial losses. Such financial losses caused by electricity theft are critical for many utilities. Lost earnings can result in a lack of profits, a fund shortage for power systems investments, and it may lead to bankruptcy from the impact of year-over-year losses in the worst-affected countries (Networked Energy Services, 2020).

Therefore, electricity theft and fraud must be tackled, and many tools are now being used or developed to detect this activity and reduce its impact. Through machine learning algorithms, the collected data can be used for analysis and discover hidden patterns to detect fraudulent customers and reduce the significant effects they cause. The results and information obtained from the models selected can also be used by the utilities, especially in developing countries, to prioritize the installation of smart meters in areas where fraud is detected, which will smoothen the transition journey financially.
1.3 Project goals

The project aims to detect electricity theft and fraud by analyzing the electricity consumption patterns of customers along with other important features and factors, such as bill history, reading remarks, and areas or regions of the load in question, and by detecting anomalies through machine learning. Detecting fraudulent customers will reduce the impact of the non-technical significant financial losses caused by consuming electricity illegally and will enhance the quality of the service provided by the electric utilities. Furthermore, the obtained results can be used by electric utilities, especially in developing countries, to prioritize their smart meter installation plans in areas where fraud is detected, which will smoothen the transformation journey financially.

1.4 Methodology

The methodology adopted for the capstone project will be based on the CRoss Industry Standard Process for Data Mining (CRISP-DM), which provides a structured approach to data mining project planning. CRISP-DM has six different phases or stages which describe the process of the project:

1- Business understanding
2- Data understanding
3- Data preparation
4- Modeling
5- Evaluation
6- Deployment

Figure 1. CRISP-DM
**Business understanding**

In this phase, the focus is on understanding the project objectives and requirements from a business perspective in terms of what the customer wants to accomplish, the availability of the resources, the selection of the tools and technologies, and the plan for each phase of the project.

Through the machine learning algorithm selected during the project, the electric utilities should be able to detect fraudulent customers who are stealing and using electricity illegally and are considered electricity thieves and can develop a plan for smart meters installation and prioritize the execution on areas where fraud is detected especially for developing countries. Electricity theft and fraud detection will reduce the significant financial impact they cause and enhance the quality of service the electric utility is adopting the approach.

**Data understanding**

This phase focuses on identifying, collecting, and analyzing the dataset, which will be used for accomplishing the project objective. It includes the collection of the initial data, the data description, the data exploration, and the understanding of the relationship between the different variables.

The dataset selected is provided in Kaggle by the Tunisian Company of Electricity and Gas (STEG) through the link [here](#), which contains 21 different features, including the district, client ID, client category, region area, creation date, invoice date, tariff type, counter number, counter status, counter code, reading remark, counter coefficient, consumption levels 1-4, old index, new index, months number, and counter type.

**Data preparation**

In this phase, the data is prepared for modeling, it includes five different tasks including the dataset final selection, data cleaning, where the detection and correction of data occurs, data construction, where new helpful attributes are derived, data integration, where different data sources are used and combined, and finally, the reformatting of data which include the conversion of values. However, the tasks to be completed depend on the project and the selected dataset. Therefore, not all tasks are required to be followed or completed.
Modeling

In this phase, different supervised machine learning models such as the support vector machine, decision tree, and random forest will be built and assessed. It includes four different tasks: the selection of the model technique, and the determination of which model to use, the selection of the training, and testing data and how they are going to be split, the model building through code such as R language, and finally, the assessment of the models’ results being tested based on the domain knowledge, success criteria, and the evaluation metrics.

Evaluation

In this phase, the models will be assessed broadly based on the business meeting criteria. It includes three different tasks, the evaluation results’ interpretation based on the business meeting success criteria, the process reviewal and findings summary, and finally, determining the next step to be followed, either to proceed with the deployment or iterate further.

Deployment

In this phase, the selected model will be accessible to the customer to use through the deployment. The phase’s complexity varies wildly. It can include four different tasks, to develop a plan for the deployment, to monitor and maintain, to document a summary of the project, and finally, to identify the lessons learned and what could have been better. The plan for the selected model for this capstone is to deploy it as an API to be accessed and used by the stakeholders.
1.5 Limitations of the Study

Energy fraud detection and electricity theft are difficult, complex problems, and the data currently accessible on these subjects is either insufficient or of low quality. The validity of the findings may also be impacted by the fact that the data is not always correct or full. This presented difficulty in searching for an appropriate dataset for the project.

However, even after the dataset's final selection for the project, it did lack a clear and full description of the different attributes used, this includes, for example, the attributes reading remarks, where no descriptions or explanation were provided on the different values in the attribute, therefore it presented a business wise challenge, where no understanding could be achieved in terms of the meaning of the different values.

The technical complexity of the subject matter is considered one of the major challenges to conducting a study and research on electricity theft and energy fraud detection, where it requires advanced technical knowledge and expertise of how the electrical system operates and the various type of theft and fraud.

Conducting a comprehensive study on electricity theft and energy fraud detection requires a significant amount of time. Timing constraints restrict the study's scope, the amount of data that can be gathered, and the depth of the analysis. The balance between work and research posed an additional challenge; appropriate time management was critical in overcoming such a challenge and being able to complete the research and project.
Chapter 2 – Literature Review

In recent years, the need for electrical power has increased globally. This rise may be attributed to new technology made available to society in high-income countries, an increase in industrial production, and the economic revival of low-income nations. Since some of the electricity provided by concessionaires reaches the final customer despite rising demand, this expansion presents economic and regulatory issues that are directly tied to the sustainability of distribution systems. The distribution system loses a significant quantity of electrical energy, and these losses can be classified as either technical or non-technical. Technical losses are unavoidable in transmission and are mostly caused by the loss of power during transit, transformation, distribution, and energy measurement. The term "non-technical losses" (NTL) refers to any electrical energy used but not billed (Savian et al., 2021). These may result from improper connections, problems with energy meters such as installation delays or reading errors, contaminated, malfunctioning, or unsuitable measurement equipment, extremely low valid usage estimates, flawed connections, and unappreciative clients, and electricity theft.

According to the comparative analysis carried out by (Smith, 2004), the evidence indicates that power theft is on the rise in numerous nations, and the financial losses from some systems are so significant that the utility is in financial difficulty. Theft losses can cause millions of dollars in lost revenue annually, including effective systems. For many electric power organizations, financial losses are crucial. Loss of revenue can lead to a lack of profits, a lack of funding for investments in the enhancement of the power system, and the need to increase generation capacity to make up for the power losses. Some electrical networks in the worst-hit nations are on the verge of bankruptcy. Only by embracing a robust and proactive stance, power industry organizations can minimize and keep electricity theft under control. An in-depth comprehension of the unique characteristics of the theft problem should serve as the foundation for the strategy and the action to be taken. In power systems with a good governance culture, power theft can be reduced and kept below reasonable limits more successfully.

According to (Depuru et al., 2011), several things motivate people to steal electricity. Some of the socioeconomic elements that greatly affect people's decision to steal electricity include the prevalent belief among illegal customers that it is dishonest to steal from a neighbor but not from
the government or a publicly owned utility provider, the rising unemployment, and a challenging consumer economy, and the lesser consumer illiteracy rate, as they could not be aware of the problems, regulations, or offenses associated to stealing. Electricity theft is openly evident at several distribution feeders in many developing nations. Many technical and non-technical methods are used to estimate and regulate theft in order to control power theft. Despite being crucial, locating the location of theft is not a way to stop theft. Other non-technical approaches include giving low-income clients incentives in the form of cheaper rates, which may give them the impression that power is affordable and tempt them to use genuine electricity. Also, effective business procedures and accurate auditing of electricity consumption at the distribution level discourage illicit consumers from tampering with the energy meters. Furthermore, illegal use of any form of energy owned by the utility company must be made punishable by law, and electrical theft must be treated as a major crime.

Since electricity is a major component in many production processes, ensuring a steady supply is essential. Power theft is a significant problem because it interferes with the method power companies distribute their electricity by overloading or short-circuiting their systems. This frequently causes a disruption (i.e., a partial or whole loss) in the energy supply to legitimate consumers (Lewis, 2015). Due to technical and non-technical electrical inefficiencies, not all the electricity produced gets delivered or supplied to the end user. In other words, some level of electricity theft is unavoidable. For both utility companies and their legitimate consumers, electricity theft has at least four costly effects. it unnecessarily drives up the cost of electricity for genuine consumers because utility companies are typically compelled to pass along the costs of energy lost (due to theft) and the higher maintenance costs needed for their distribution networks. Second, power theft degrades the quality of the electricity supply by overloading the system, which frequently results in sporadic power outages for both legal and illegal users, loss of output, and damage to electrical appliances. Thirdly, power theft reduces the amount of reinvestment and employment that can occur in the electrical sector by limiting the amount of potential money that utility companies may earn. These restrictions reduce the amount of money that can be used to finance the growth or expansion of generating capacity, which contributes to load-shedding. Finally, power theft, particularly when done through unlawful connections, poses a risk of fire and sporadically leads to the death of the power thieves or even innocent bystanders who unintentionally become electrocuted after becoming entangled with illegally strung throw-ups.
Utility companies in developing nations are frequently faced with a significant additional problem due to non-technical loss of energy on top of their existing struggles with overburdened infrastructure and poor investment levels. A utility is heavily burdened financially as a result of the lost revenue from non-technical losses, which typically results in higher tariffs for paying customers and generally poorer service. According to (Carr & Thomson, 2022), The assumption that the poorest households are the primary culprits of non-technical loss is simplistic; research has revealed that the majority of non-technical loss is frequently caused by wealthier clients, including state-owned companies and big businesses. Although customer poverty and a lack of income are important variables, it seems that customers who are legitimately unable to pay for electricity tend to cut their consumption or transfer to alternate energy sources. Although average residential customers make up the majority of the customer base in terms of numbers, they might only be responsible for a tiny percentage of non-technical electricity loss. In order to reduce non-technical losses, utility firms should concentrate their efforts on major consumers and those who can afford to pay cost-reflective pricing. Enforcement must be done consistently.

According to (Savian et al., 2021) researchers are experimenting with several approaches for effectively identifying clients who commit fraud. The currently used NTL detection techniques can be roughly divided into hardware-based and non-hardware-based methods. Hardware-based solutions mostly concentrate on installing meters with certain technology to allow Power Distribution Companies to spot any fraudulent behavior by customers. New non-hardware-based NTL detection techniques have been created as a result of recent developments in communication and data processing of energy consumers' behavior. These alternatives to hardware are divided into three main groups: data-based methods, network-based methods, and hybrid methods. The data-based methods merely rely on data analytics and machine learning methodologies. The use of the network in the power grid is the primary distinction between data-based and network-based solutions. The data-based methods are further split into supervised and unsupervised methods. Network-based methods rely on data obtained from smart meters and the calculation of various physical electrical network metrics for effectively recognizing NTL activity. Whereas, hybrid-based methods are used to accurately incorporate the techniques and algorithms of both data-oriented and network-based methodologies for NTL identification.
A methodology has been developed and tested by (Buzau et al., 2018) on real smart meter data of all Endesa's (Spain’s largest electricity utility) industrial and commercial clients for detecting non-technical losses by the usage of supervised learning. It generates an in-depth study of customers' consumption patterns using all the data that the smart meters capture, including energy consumption, alerts, and electrical magnitudes. Additionally, it makes use of auxiliary databases to offer more details on the precise location and technical specifications of each smart meter. The model was trained, validated, and tested using data from over 57,000 on-site inspections, with a clear data imbalance that was reduced using under-sampling techniques. The methodology starts with smart meter data cleaning, which is then used for feature extraction based on energy consumption, smart meters alarm, and electrical magnitude measurements with the additional usage of auxiliary databases. The data is pre-processed through the standardization of features into zero mean and unit variance, the conversion to numerical variables from categorical, and the replacement of discrete and continuous missing values with the most frequent and the mean value, respectively. The features are inserted into different machine-learning algorithms, including support vector machine, logistic regression, k-Nearest Neighbors, and XGBoost for model selection and evaluation, performed with 5-fold nested cross-validation. The selection of the model depends on the required standard by the utility. The successful model will output a ranked list of customers, whereas a repeated cycle will be performed if the standard is not met. XGBoost was found to outperform the rest of the classifiers in terms of AUC, precision, and recall, with KNN scoring the lowest.

A communication layer provided by advanced metering infrastructure (AMI) creates a more effective administration and monitoring system for the utilities, which improves customer service. AMI also enables utilities to provide an additional layer of protection in cases of widespread system inefficiencies. After a predetermined amount of time, the utilities receive recorded electricity consumption data. Electricity theft is further decreased by the utility provider being alerted right away if meter manipulation is discovered. Nevertheless, even with the system's advantages from AMI, developing nations cannot afford to build such infrastructure on a large scale. A data-driven method of detecting outliers using monthly electricity consumption is developed by (Nadeem & Arshad, 2021) to tackle these challenges. Utility providers will be able to identify clients with unusual load profiles using this strategy and give these customers priority for smart meter installation. This method enables targeted AMI deployment to a portion of the
vulnerable population, lowering non-technical losses and producing a more reliable distribution system that is simple to control.

Additionally, it will enable utility firms to upgrade their infrastructure gradually and painlessly without experiencing too much financial strain. The algorithm examines historical monthly electricity consumption data from residential users to identify outliers with irregular fluctuations in electricity use over the observation horizon. Each region or so-called subdivision shares several characteristics, including similar lifestyles, weather patterns, and neighborhoods which considerably affect electricity consumption. A threshold has been marked to identify the outlier exceeding it, which is directly proportional to the number of clusters pre-defined for k-means clustering. This approach to finding outliers makes the assumption that a region has more regular clients than anomalies. Because the trend is set by the majority, if there are more colluders, the pattern will be disturbed, and the suggested method will not be able to find outliers. This is a reasonable assumption, though, given that colluders are usually outnumbered by loyal customers. The regular customers' patterns in this analysis were identified, and outliers based on deviations and suspicious acts were identified and marked as potential electricity thieves.

Machine learning algorithms have been a widely used strategy for identifying energy fraud. Recognizing and using patterns in energy use can enable us to recognize odd behavior. This study (Cody et al., 2015) describes a novel way to use decision tree learning to profile typical energy use patterns and identify potentially fraudulent behavior. The paper discusses two types of fraud, the less consumption reported by the smart meter than used (type 1) and the higher reported consumption due to rogue connections (type 2). To anticipate future energy use and identify fraudulent activities, the decision tree machine learning method is used to model consumers' typical energy consumption behavior. The dataset is a collection of 5,000 residential and 650 business smart meters, with consumption recorded every 30 minutes for approximately two years (2009 – 2011). In the pre-processing data phase, two main techniques were utilized: feature selection and data aggregation. The features selected shall only be relevant to learn an accurate model and to reduce the model's complexity. The month feature among the day, week, time, and energy consumption features was found to be invaluable in prediction results; therefore, it was removed. Furthermore, in order to smooth out the variations that could mistakenly be detected as fraud, the data were aggregated into a larger interval. For anomaly detection, the Root Mean
Squared Error was used, which is an indication of the difference between the predicted and the actual value, that will be compared to a threshold indicating possible fraud (found to be 0.4kwh). To simulate the two types of fraud mentioned earlier, a random value deviating from 0 to 0.5 kWh was subtracted and added, imitating type 1 and type 2, respectively. The original dataset was used to learn the model (noise-free) and simulated fraudulent data (with noise) were used in validating the model. Three different experiments have been conducted using the decision tree (i.e., the same month of a year, subsequent weeks, and the same weather season). The experiments' findings show that it is possible to anticipate energy use with accuracy.

In the paper published by (Depuru et al., 2011), the support vector machines (SVMs) are trained with the smart meters’ collected data, which reflects all potential forms of theft and are tested on a variety of customers. It presents a methodology for training and testing the dataset to identify and classify illegal users. Different types of electricity theft are discussed in this paper, including directly obtaining electricity from the distribution feeder and tampering with the meter by interchanging its connections, resetting it, or inserting unwanted harmonics, etc. Specific criteria were set for training the model on the dataset based on the electrical consumption, such as geographic locations, year's season, and different customer classifications (commercial, residential, etc.) to create noise-free data to train and test the support vector machines. The consumption data of different customers in terms of load size and appliances used, the season of the year, and the time was plotted to understand the patterns. The customers were classified into three different classes, D, S, and I, based on specific rules related to consumption patterns. A customer is rated as genuine if their consumption is continuous and matches the patterns. If not, then the profile is suspicious and will be checked against the pre-selected rules to classify it as a specific class—the proposed methodology classified data with an accuracy of 98.4% for 220 customers.

(Gunturi & Sarkar, 2021) Suggests ensemble machine learning models for energy theft detection using the consumption patterns of the customers. Ensemble machine learning models are meta-algorithms that combine several machine learning approaches smartly into one predictive model reducing the reducible error: bias and variance. Six different algorithms, including random forest, extra trees, as well as adaptive, categorical, extreme, and light boosting, were tested to find their detection rates and false positive. The data were obtained from the smart meter program of the
Commission for Energy Regulation, an imbalance data which was pre-processed using data standardization and SMOTE to get balanced data. The ensemble machine learning classifier was then trained using this data and verified by the testing data. In the data pre-processing stage, three different phases are defined: Data cleaning where null values and errors are found, Data standardization where the outliers are detected and transformed into acceptable limits; and finally, Feature engineering where essential features are extracted. The power theft attempts on the Advanced Metering Infrastructure can be categorized into direct attacks where meters are tempered by different methods, such as using a strong magnet or flipping the meter, and second, network attacks where the smart meters’ firmware and storage are accessed to alter their reading. In this paper, two different balancing techniques have been used: Near-miss (under-sampling) and SMOTE (oversampling), Standard evaluation metrics, including recall, precision, and f1-score were calculated for the validation and found using SMOTE, and the ensemble machine learning models outperform the usage of near-miss. Six different theft cases were used to produce malicious samples, where each case has a different number of users, thus a different number of samples. The gradient-boosting techniques outperformed the bagging methods for the first single-user case with an achieved 0.80 AUC. However, with the increased number of users, the AUC decreased in boosting techniques and increased in the bagging type, where it achieved 0.90 AUC for both the random forest and extra trees for the sixth case. Thus, bagging methods like the random forest and extra tree are preferred for real-time use.

To detect NTL operations in electric utilities, the framework presented by (Nagi et al., 2010) looks for customers who have abnormal usage patterns. To find clients that have irregularities and conduct fraud, a support vector machine (SVM) and load profiles combined with an automatic feature extraction method are utilized. This study makes use of historical consumer consumption information gathered from various places in Malaysia's peninsula. Data mining and statistical techniques are used to extract customer usage patterns, which define load profiles. SVM divides client load profiles into two categories: normal and fraud, presuming that load profiles contain anomalies when a fraud occurrence takes place. Although there are numerous sorts of fraud that might happen, this research solely focuses on situations where sudden changes appear in. The dataset used in this study contained 265,870 customers’ data such as monthly consumption, meter reading type, and date, etc. between July 2006 – 2008. Additional data which contains information about 105,525 detected fraud cases (by on-site inspection) was obtained to enhance the detection
accuracy, some cases, however, were duplicated up to 35 times. Thus, SQL was used to group all individual cases as single records decreasing the number to 32,972 cases. Additionally, as part of customers’ filtering and selection, repeated customers, missing data, and customers with no consumption records were removed, leading to 186,968 customers. The data were normalized by using the min-max approach to be used for the SVM training. Features such as the creditworthiness rating (CWR) were selected due to their significance in detecting fraud. The proposed FDM can be utilized to accurately detect anomalies and fraud activities within electrical supply utilities, according to experimental results. The technique of utilizing SVM for fraud customer identification has shown to be quite promising.

An electricity theft detection method based on the random forest is proposed (Hu et al., 2020). To assess the elements influencing electricity theft, an evaluation index system that includes the power consumption slope index, the warning information index, and the line-loss index was developed. In the occurrence of electricity theft, such parameters of the distribution network will vary. However, the amount of power used by consumers cannot accurately reflect the phenomena of electricity theft. The cause is that a variety of events, including the weather, holidays, and unexpected failure, have an impact on users' power usage. Additionally, the warning system may contain outdated or incorrect information. Building the evaluation index system before the model training process is therefore crucial for the detection of electricity theft. Prior to the training procedure, data cleaning and missing value processing are carried out taking the features of the data into consideration. The removal of useless data is the purpose of data cleaning. Banks, schools, and industrial users are examples of non-residents who often do not steal electricity and whose data must be removed from the data collection. Additionally, there is a significant difference in the amount of energy used on weekends and on workdays. Therefore, holidays are removed to improve the detection performance. Additionally, Lagrange's interpolation is used to handle the missing value to improve the modeling impact. The model is then trained with the continuous adjustment of the maximum depth of decision trees until the highest detection accuracy of 2.85% was achieved with a maximum depth of 4. The ROC curve was used as an evaluation criterion comparing the random forest with the decision tree and GDBT. The random forest scored 0.95, the highest in terms of the area, with a higher and shorter training time compared with the decision tree, and GDBT, respectively. Thus, the random forest was found to outperform the others in detecting electricity theft.
Different techniques have been developed to find electricity theft automatically. Most of these techniques solely evaluate records of electricity consumption. However, because there are many different types of theft and the unpredictability of electricity theft behavior, it is difficult to identify fraudulent customers by merely looking at consumption records. Due to unbalanced data, several approaches also have poor classification accuracy. To address the aforementioned problems, (Arif et al., 2022) suggest two solutions: Tomek Link Borderline Synthetic Minority Oversampling Technique with Support Vector Machine (TBSSVM) used for resampling the data by balancing the two different classes' (minority and majority) observations, and Temporal Convolutional Network with Enhanced Multi-Layer Perceptron (TCN-EMLP) used for the classification. Two different datasets were used for this approach the State Grid Cooperation of China (SGCC), and Pakistan Residential Electricity Consumption (PRECON). In this paper, the focus was to discover and detect the irregular behavior of electricity thieves. As part of data pre-processing, the missing values were filled by linear interpolation using the scikit-learn class. The data were normalized using the min-max approach due to the sensitivity of neural networks to diverse data.

Furthermore, to balance the dataset, a combination of oversampling and under-sampling techniques were used through Borderline-SMOTE-SVM and Tomek Links (TL) to obtain an equal distribution of both classes. Deep learning models are known to increase classification accuracy. Nevertheless, they sometimes show high variance. To tackle this challenge, ensemble learning is used to reduce it in neural network models by training multiple models and integrating the results. AUC was used as a performance metric, and the results show that the proposed method surpassed the other used methods like logistic regression, support vector machine, and random forest, with an additional 2% better results compared to the baseline models with 80% training samples after auxiliary data integration.

In general, on-site inspection is crucial to ascertain a customer's sincerity in consuming. Utility companies search for innovative strategies to narrow the scope of inspections to situations with a higher likelihood of fraud because examining every customer is costly. Making use of machine learning (ML) techniques to analyze consumption patterns is one way to narrow the scope of inspection. However, the insufficiency of data plays an important role. (Emadaleslami & Haghifam, 2021) has introduced a novel two-stage ML-based approach to identify distribution network fraud. An Artificial Neural Network (ANN) is trained to model fraudulent clients in the
first step, and then it is utilized to forecast theft scenarios for regular consumers to handle data insufficiency. A Support Vector Machine (SVM) classifier is trained in the second step to discriminate between normal and suspicious customers. The evaluation and comparison of the suggested algorithm to conventional models demonstrates its great performance on an actual data set with more than 5000 consumers. The proposed model consisted of four steps, data preprocessing, which includes data cleaning, Feature Engineering, and normalization, the second step included clustering customers using silhouette and K-Shape clustering, next the dataset was divided into train/test with a ratio of 80/20 and was used to train and test the ANN models with different configurations, and to evaluate its performance, the generated dataset was then used to train the SVM classifier. The model demonstrates its superiority in terms of high detection rate, low false positive rate, and satisfactory accuracy. Given that LSTM and CNN networks perform better in time-series regression than the proposed model. it was recommended that a combination of these deep networks be employed in future studies to address the minor limitations of the model.
Key Takeaways

- Electricity theft is on the rise in numerous nations around the globe, and the financial losses from some systems are so significant that the utility is in financial difficulty. Such losses can cause millions of dollars in lost revenue annually and require an up-to-date solution to overcome those challenges.

- Identifying and detecting fraudulent behavior or customer by solely looking into the electrical consumption records is challenging. Other relevant features and data are required to be integrated to support the detection of fraud and electricity theft, and this can include bill history, the addition of statistical features, weather information, and the type of residents.

- The datasets used in many research papers are found expectedly to be imbalanced, where the fraud class is a minority, and the honest class is the majority. Thus, resampling techniques such as under-sampling and oversampling are mandatory to be applied to prevent a biased machine learning model.

- The selection of the machine learning model for the detection of electricity theft and fraud was conducted based on standard evaluation metrics, including accuracy, recall, precision, F1-score, and AUC, where the performance of each model is calculated and compared before the final selection. Common models used include decision tree, random forest, and support vector machine, which will be used in the project.
Chapter 3 - Project Description

3.1 Overview

In the modern world, electricity is a necessity since almost all facets of daily life are powered by it, including homes, businesses, and hospitals. It gives us access to information, keeps us in contact with our loved ones, and powers the technology we use for work and leisure. Our contemporary way of life would come to a complete stop without power in many areas.

Electricity theft is a serious issue that jeopardizes the future sustainability of the energy sector. Either by people or businesses, it strains the power grid and lowers the amount of money utilities gain. Because of this, consumers pay more for energy, and utility providers may deliver less quality service as a result.

The sustainability of the energy sector depends on the detection and prevention of electricity theft. The use of machine learning is one of the best methods for identifying electricity theft. In this project, different machine learning algorithms, including decision tree, random forest, and support vector machine, are built and trained on an open-source dataset provided in Kaggle, which will be described in the below section.

The dataset will first be analyzed through data visualization. Different attributes will be plotted to understand and gain insight into how those attributes correlate with each other and the target class, which classifies fraud and non-fraud consumers. Next, pre-processing and cleaning of data will be performed prior to the model building to ensure that models are trained on a well-cleaned dataset with no missing values to obtain the highest accuracies. The built models will then be evaluated, and the best-performed model will be chosen and deployed as an API for the utilities.
3.2 Data Analytics Tools and Software Used

The following are the tools that were used to execute and complete the project:

- R Studio: It is a free and open-source integrated development environment for R, a tool that is used for data mining, visualization, model building, and deployment. The R Studio is the primary tool that was used to execute and complete the project.
- Tableau: It is a visualization tool that allows the creation of different charts, graphs, and dashboards, to analyze the data, identify patterns, and obtain valuable insights. In this project, some visualization charts were created using Tableau.

3.3 Dataset Description

The selected dataset for this project, titled Fraud Detection in Electricity and Gas Consumption, was provided through Kaggle by the Tunisian Company of Electricity and Gas (STEG), a non-administrative public company. Established in 1962, it oversees gas and energy distribution throughout Tunisia. The company experienced significant losses of almost 200 million Tunisian Dinars due to consumers' manipulation of meters.

The dataset is originally divided into two training datasets: client train and invoice train, and two testing datasets: client test and invoice test. A total of 4.5 million records from 1977 to 2019 representing more than 135,000 users and their consumption details were imported from STEG's dataset.

There are 21 different attributes that make up this dataset (Client and Invoice), and they are mentioned in the below table:

Table 1. Attribute Description

<table>
<thead>
<tr>
<th>No.</th>
<th>Attribute Name</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>district</td>
<td>Client’s District</td>
<td>Discrete</td>
</tr>
<tr>
<td>2</td>
<td>client_id</td>
<td>Client’s Unique ID</td>
<td>Discrete</td>
</tr>
<tr>
<td>3</td>
<td>client_catg</td>
<td>Client’s Category</td>
<td>Discrete</td>
</tr>
<tr>
<td>4</td>
<td>region</td>
<td>Client’s Area / Region</td>
<td>Discrete</td>
</tr>
<tr>
<td></td>
<td>Field</td>
<td>Description</td>
<td>Type</td>
</tr>
<tr>
<td>---</td>
<td>------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td>5</td>
<td>creation_date</td>
<td>Client’s Joining Date</td>
<td>Continuous</td>
</tr>
<tr>
<td>6</td>
<td>target</td>
<td>Fraud or Not Fraud</td>
<td>Discrete (Binary)</td>
</tr>
<tr>
<td>7</td>
<td>invoice_date</td>
<td>Date of the Invoice</td>
<td>Continuous</td>
</tr>
<tr>
<td>8</td>
<td>tariff_type</td>
<td>Type of Tax</td>
<td>Discrete</td>
</tr>
<tr>
<td>9</td>
<td>counter_number</td>
<td>Counter’s Number</td>
<td>Discrete</td>
</tr>
<tr>
<td>10</td>
<td>counter_code</td>
<td>Counter’s Code</td>
<td>Discrete</td>
</tr>
<tr>
<td>11</td>
<td>reading_remarque</td>
<td>Notes that the STEG agent takes during inspection (e.g., If the counter shows something wrong, etc.)</td>
<td>Discrete</td>
</tr>
<tr>
<td>12</td>
<td>counter_statue</td>
<td>Status of the counter which has 5 values such as working fine, not working, on hold status, etc.</td>
<td>Discrete</td>
</tr>
<tr>
<td>13</td>
<td>counter_coefficient</td>
<td>An additional coefficient to be added when standard consumption is exceeded</td>
<td>Discrete</td>
</tr>
<tr>
<td>14</td>
<td>consommation_level_1</td>
<td>Consumption level 1</td>
<td>Continuous</td>
</tr>
<tr>
<td>15</td>
<td>consommation_level_2</td>
<td>Consumption level 2</td>
<td>Continuous</td>
</tr>
<tr>
<td>16</td>
<td>consommation_level_3</td>
<td>Consumption level 3</td>
<td>Continuous</td>
</tr>
<tr>
<td>17</td>
<td>consommation_level_4</td>
<td>Consumption level 4</td>
<td>Continuous</td>
</tr>
<tr>
<td>18</td>
<td>old_index</td>
<td>Old Index</td>
<td>Continuous</td>
</tr>
<tr>
<td>19</td>
<td>new_index</td>
<td>New Index</td>
<td>Continuous</td>
</tr>
<tr>
<td>20</td>
<td>months_number</td>
<td>Number of Month</td>
<td>Discrete</td>
</tr>
<tr>
<td>21</td>
<td>counter_type</td>
<td>Type of Counter (ELEC or GAS)</td>
<td>Discrete (Binary)</td>
</tr>
</tbody>
</table>
Chapter 4 - Project Analysis

This chapter analyzes the data using various visualization methods and exploratory data analysis techniques. Additionally, different machine learning models are fitted on the Fraud Detection in Electricity and Gas Consumption dataset, and their classification performance is evaluated.

4.1 Exploratory Data Analysis and Visualization

In this section, different attributes will be explored and visualized to gain some interesting insight which will help us discover the hidden patterns and the correlations of the attributes with the target (class label).

First, R programming language was used through multiple functions to discover both datasets’ (Client train and Invoice train) main information and features. Such as data structure (number of rows, number of columns, and data structure) and a summary that contains (Min, Median, Mean, Max, 1st Qu, and 3rd Qu):

**Client Train Dataset**

<table>
<thead>
<tr>
<th>district</th>
<th>client_id</th>
<th>client_catg</th>
<th>region</th>
<th>creation_date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. :60.00</td>
<td>Length:135493</td>
<td>Min. :11.00</td>
<td>Min. :101.0</td>
<td>Length:135493</td>
</tr>
<tr>
<td>1st Qu.:62.00</td>
<td>Class :character</td>
<td>1st Qu.:11.00</td>
<td>1st Qu.:103.0</td>
<td>Class :character</td>
</tr>
<tr>
<td>Median :62.00</td>
<td>Mode :character</td>
<td>Median :11.00</td>
<td>Median :107.0</td>
<td>Mode :character</td>
</tr>
<tr>
<td>Mean :63.51</td>
<td></td>
<td>Mean :11.51</td>
<td>Mean :206.2</td>
<td></td>
</tr>
<tr>
<td>3rd Qu.:69.00</td>
<td></td>
<td>3rd Qu.:11.00</td>
<td>3rd Qu.:307.0</td>
<td></td>
</tr>
<tr>
<td>Max. :69.00</td>
<td></td>
<td>Max. :51.00</td>
<td>Max. :399.0</td>
<td></td>
</tr>
<tr>
<td>target</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min. :0.00000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st Qu.:0.00000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median :0.00000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean :0.05584</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3rd Qu.:0.00000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max. :1.00000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Client Dataset Structure

Figure 3. Client Dataset Summary
The obtained statistics for both datasets provide insight into how the different attributes are distributed. In the next step, a few attributes, including the target class, have been chosen to get more statistical details in terms of their quantity and ranking.
The number of records was calculated to the following six attributes:

<table>
<thead>
<tr>
<th>target</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>127927</td>
</tr>
<tr>
<td>1</td>
<td>7566</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>region</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>33770</td>
</tr>
<tr>
<td>104</td>
<td>12865</td>
</tr>
<tr>
<td>311</td>
<td>12406</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>client_catg</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>131494</td>
</tr>
<tr>
<td>12</td>
<td>2321</td>
</tr>
<tr>
<td>51</td>
<td>1678</td>
</tr>
</tbody>
</table>

**Figure 6. Number of Records for Target, Region, & Client Category**

The target class, as expected, is unbalanced, where only 7,566 out of the 135,493 customers are classified as fraudulent, making it only 5.6% of total customers. Furthermore, in terms of region, the top three regions were 101, 104, and 311. In the first region, 101 made up 24.9% of total records regarding customer count, 104 made up 9.49%, and 311 scored 9.16%.

Furthermore, in the client category attribute, 11, 12, and 51 were the different three categories, and as per the statistics above, the dominance is to category 11, with a percentage of 97% in terms of total client records of 135,493. Categories 12 and 51, however, make up an approximate percentage of 1.71% and 1.25%, respectively.

<table>
<thead>
<tr>
<th>tariff_type</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>122799</td>
</tr>
<tr>
<td>40</td>
<td>61678</td>
</tr>
<tr>
<td>10</td>
<td>12798</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>district</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>40353</td>
</tr>
<tr>
<td>69</td>
<td>34231</td>
</tr>
<tr>
<td>60</td>
<td>31922</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>reading_remark</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>119781</td>
</tr>
<tr>
<td>9</td>
<td>116596</td>
</tr>
<tr>
<td>8</td>
<td>95591</td>
</tr>
</tbody>
</table>

**Figure 7. Number of Records for Tariff Type, District & Reading Remarks**

Additionally, regarding tariff type, the highest three types were considered 11, 40, and 10. The first type has a record of 122,799 which is 59.9% of the total record of 204,902 taken for distinctive client IDs and tariff types. Similarly, tariff types 40 and 10 comprise approximately 30.1% and 6.3%, respectively. Additionally, the top three districts in terms of client count are district 62, with a percentage of 29.8%, and districts 69 and 60, with 25.3% and 23.6%, respectively.

In terms of reading remarks, however, reading remark numbered six is the highest in terms of count, with a percentage of 36% considering a total record of 332,451 taken for distinctive client IDs and reading remarks. Similarly, remarks nine and eight comprise 35% and 28.8% respectively. However, there is no description of the meaning of each remark number.
Different visualization plots will be created to match those findings, which will help understand the data and derive conclusions from them. In Figure 8, for example, the number of customers between 1977 – 2019 is shown in terms of the top obtained client category.

The highest number of customers for more than 40 years is clearly for the 11th category as shown by Figure 8, which continuously increased, especially from 2005, however, after four years it started to decline for some reason. On the other hand, no clear relationship between categories 12 and 51 from the first figure could be obtained. Thus, log10 has been applied to the plot, which illustrated that the 12th category surpassed the 51st till 2009, from which it started to decrease.
Figure 9 below is a pie chart that illustrates the customers' distribution in terms of counter type, where 68% of counters were found to be ELEC with a record of 3,079,406 and the remaining 32% were GAZ with a record of 1,397,343 out of a total of 4,476,749 records.

![Counter Type Distribution](image)

**Figure 9.** Counter Type Distribution

To understand how electrical and gas consumption patterns or behavior differ between an honest and a fraudulent customer, the consumption of two randomly selected customers was plotted for a duration of one year, as shown in Figure 10, along with the monthly average consumption.

![Yearly Consumption Pattern](image)

**Figure 10.** Yearly Consumption Pattern
An interesting pattern has been noticed in Figure 10, where the honest customer consumption pattern somehow followed approximately the average consumption for that period, whereas the fraudulent customer had a strange zigzag pattern. In most cases, honest customers will always follow the average pattern, in contrast with fraudulent customers, whose consumption patterns will be different and noticeable, which can be used to detect fraud in suspected customers.

Following the previously obtained statistics about the datasets, which illustrated how different attributes were distributed, in Figure 11 a bar chart is obtained, which displays how imbalanced the target class in the dataset is, which is already expected in a fraud detection domain. Only 5.6% of all customers were fraudulent with a total of 7,566, and the remaining 94.4% were honest.

The target class was then divided in terms of counter type and plotted as the below bar chart in Figure 12, which as expectedly, shows that 64.6% of honest clients were electricity consumers, and the remaining 29.3% were gas consumers since the majority of the records in the dataset is of ELEC counter type as illustrated previously. Additionally, in terms of fraudulent customers, 3.8% of the total 5.6% were of ELEC counter type, and the remaining were of GAZ counter type.
Further bar charts are plotted to get deeper into how the different attributes are related to the target class, starting with the district attribute in the figure below.

Of the different districts (60, 62, 63, and 69) that were plotted, the 62nd district has the highest percentage of total customers with a total of 40,353, and the least was in the 63rd district with a total of 28,987. The 69th district was noticed to have the highest percentage in terms of fraudulent customers, with a total of 2,447, which is a 1.8% of the 5.6%, and the least district was the 60th, with a percentage of 0.8%. On the contrary, the highest percentage of honest customers was in the 62nd district, with a total of 38,270 which is 28.2% of the 94.4%, and the least was in the 63rd with a percentage of 20%.

Figure 13. Target Distribution District Wise
The next plotted attribute with relevance to the target class is the client category, as shown in Figure 14 below, and as discovered earlier, the three client categories are categories 11, 12, and 51.

![Figure 14. Target Distribution Client Category Wise](image)

Figure 14 clearly shows the dominance of category 11, with a total of 124,303 and a percentage of 91.7% out of the 94.4% in terms of honest customers and compared with a total of 2.6% for categories 12 and 51. Similarly, in terms of fraudulent customers, category 11 comes in first, with a total of 7,191 and a percentage of 5.3% out of the total 5.6%, with the remaining 0.1% and 0.2% for categories 12 and 51, respectively. It is interesting to discover that 97% of the Tunisian Company of Electricity and Gas (STEG) clients are from category 11.
Furthermore, the attribute region was plotted, with different bar charts indicating how clients are distributed in different regions. The summary of data obtained earlier indicates that regions 101, 104, and 311 were at the top of the list in terms of the number of customers, which also matches Figure 15 below.

![Figure 15. Target Distribution Region Wise](image)

As indicated by Figure 15, the highest region in terms of both honest and fraudulent customers is region 101, with a percentage of 24% and 0.9%, respectively. Furthermore, region 104 got 9.5% of the total customers, where only 0.5% of them were fraudulent. However, in region 311, this percentage is higher at 0.7%, whereas honest customers were allocating 8.4%.

Moreover, it is interesting to see that regions 199 and 206 have no records of registered clients at all, and in regions 106, 308, 379, and 399, there are no records of any fraudulent customers. In general, it is clear from the data distribution above that it is unbalanced.
Moreover, the attribute tariff type has been plotted with all 17 different types ranging between type 8 to type 45, with each type indicating the target distribution in terms of fraudulent customers and honest customers as shown below in Figure 16:

![Figure 16. Target Distribution Tarif Type Wise](image)

From the above plots, an expected connection has been noticed for most of the tariff types, including types 8, 9, 10, 11, 12, 13, 14, 15, 29, 40, and 45, where the majority of the customers were honest customers with a percentage up to 95.65% within those types. However, few interesting relationships have been noticed in types 24, 30, and 42 where all customers under those types were honest. On the other hand, in tariff type 18, only one customer was registered as per the records, and this customer happened to be fraudulent, thus a 100% percent is allocated for this specific bar chart.
The correlation matrix in Figure 17 was created to understand the relationship between different attributes within a dataset. First, the average cumulative sums of consumption levels 1 to 4 were noticed to be all positively inter-correlated, and the approximate range of correlation is between 0.1 to 0.7. Additionally, the same attributes are also positively correlated with the average cumulative sums of new and old indexes, where the approximate range of correlation is between 0.2 to 0.6. Furthermore, an approximate range of correlation between 0.1 and 0.6 was also noticed between the same attributes and the average consumption levels 1 to 4.

Furthermore, a negative correlation has been noticed between some attributes including the average difference of consumption level 1 and the old index for a gas counter type, which is the highest negative correlation detected with a value of -0.3, which means that the increase of one attribute will result in decreasing the other attributes and vice versa.
4.2 Data Preparation

Data preparation is a crucial stage and the foundation of any data analytics project, to ensure accurate and trustworthy data analysis. The objective is to assure that data is correct, reliable, and free from errors, missing values, or duplicates. Thus, biases are reduced, and the accuracy and dependability of the insights drawn from the data are ensured. The two common problems that can arise in data analytics projects due to inadequate data preparation and preprocessing are the overfitting and underfitting of data, which indicates the significance of this stage.

4.2.1 Data Preprocessing

Data Preprocessing involves many stages that must be performed before the model-building stage since the outcome of Preprocessing stage will determine the trustworthiness, dependability, and accuracy of the results obtained from the built models.

4.2.1.1 Data Cleansing

In this step, the data is scanned for any missing values, duplicates, or outliers. In the dataset collected from STEG, a total of 13 missing values were detected only among 4.5 million records, and all were identified to be in one attribute, the counter status, whereas all other attributes were free from any missing values as shown below in Figure 18:

```
r missing values
sum(is.na(rit_client_train))
sum(is.na(rit_invoice_train))
sapply(rit_client_train, function(x) sum(is.na(x)))
sapply(rit_invoice_train, function(x) sum(is.na(x)))
```

```
[1] 0 13
```

<table>
<thead>
<tr>
<th>district</th>
<th>client_id</th>
<th>client_catg</th>
<th>region</th>
<th>creation_date</th>
<th>target</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>client_id</td>
<td>invoice_date</td>
<td>tarif_type</td>
<td>counter_number</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>counter_status</td>
<td>counter_code</td>
<td>reading_remark</td>
<td>counter_coefficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>consommation_level_1</td>
<td>consommation_level_2</td>
<td>consommation_level_3</td>
<td>consommation_level_4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>old_index</td>
<td>new_index</td>
<td>months_number</td>
<td>counter_type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 18. Missing Values Check*
Due to the insignificance of the missing values detected compared to the total number of records, the missing values could be neglected, or replaced by the most frequent observation in the counter status attribute, which is 0 according to the obtained count below in Figure 19, and this was the approach selected.

<table>
<thead>
<tr>
<th>counter_status</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;dbl&gt;</td>
<td>&lt;int&gt;</td>
</tr>
<tr>
<td>0</td>
<td>4379008</td>
</tr>
<tr>
<td>1</td>
<td>74036</td>
</tr>
<tr>
<td>5</td>
<td>20639</td>
</tr>
</tbody>
</table>

**Figure 19.** Counter Status Count

Furthermore, with the usage of the duplicated function, it revealed that there are no duplicated rows present in the datasets as shown below in Figure 20:

```r
duplicated values
rit_client_train %>% duplicated() %>% sum()
rit_invoice_train %>% duplicated() %>% sum()
```

```
[1] 0
[1] 0
```

**Figure 20.** Duplicated Rows Check

The outliers in the dataset were checked for user consumption levels 1, 2, 3, and 4 through the usage of the boxplot in Figure 21. The outliers were from different clients, in different years and were available in all consumption levels except for level 3.

**Figure 21.** Consumption Levels Boxplot
The outliers' values were identified, and to determine the replacement value of each outlier, the yearly average consumption was plotted for each client that had a record of an outlier, and in Figure 22 below is a sample of three different clients. The replacement value of the outlier was calculated as the average consumption for the year before and after the outlier since the values were unrealistically high, which is suspected to be a typo taken during the on-site inspection planned for the consumption level data collection.

For example, for client 113523, the average consumption in 2012 had a value of 275,019, which is where the outlier value was detected, the value will be replaced by the average of consumptions in 2011 and 2013, which happened to be 1,993.5. This has been applied for all the detected outliers, the new value will match the trend in consumption of the client and therefore will contribute to obtaining a more realistic view which will be used for fitting the models in the next stages.

![Graphs showing yearly average consumption levels for different clients](image-url)

*Figure 22. Yearly Average Consumption Level 2*
4.2.1.2 Data Transformation

In this step, the data is transformed or converted to another format that is more appropriate for the data analysis methods that will be used. Certain attributes in the dataset were identified and were required to be converted to a factor, to be used appropriately in the machine learning model building and testing phases.

The attributes selected for the factor conversion are:

- Target
- Month
- Distric
- Mode_counter_statue_ELEC
- Mode_reading_remarque_ELEC
- Tarif_Type
- Region
- Client_catg
- Mode_counter_statue_GAZ
- Mode_reading_remarque_GAZ

All the above attributes are considered categorical variables, that have a fixed and known set of possible values, thus the conversion to a factor is required in a classification case since the model will now deal with them as categories rather than numbers.

4.2.1.3 Feature Engineering

Feature engineering is the process of selecting and transforming variables when creating a predictive model. How well machine learning models perform and how accurately they are influenced by such process. It enhances machine learning's prediction capability and aids in revealing hidden patterns in the data. Creating new features in the dataset enhances the effectiveness of machine learning models by assisting in the collection of more information about the data and generating more insight.

This was accomplished through the utilization of statistical features such as the average, variance, standard deviation, median, mode, and range of the values of an attribute. Those features were obtained for different attributes in the invoice dataset including consumption levels 1-4, old index, new index, counter coefficient, counter status, invoice date, and reading remarks, causing the number of attributes to reach 71. The newly obtained features were added as new attributes to the dataset and used for building the machine-learning models.
The datasets have gone through different stages of modifications and tuning prior to their usage in the model-building phase. In the invoice dataset, for example, the newly added statistical features were merged under a unique client ID, however, different counter types resulted in the additional rows of data sharing the same client ID as shown in Table 2 below, thus, the counter type has also been merged within the remaining attributes as illustrated in Table 3, increasing the number of attributes to 139.

Table 2. Dataset Overview Stage 1

<table>
<thead>
<tr>
<th>client_id</th>
<th>counter_type</th>
<th>avg_consom_1</th>
<th>var_consom_1</th>
<th>sd_consom_1</th>
<th>median_consom_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>train_Client_0</td>
<td>ELEC</td>
<td>352.4000000</td>
<td>9.631307e+04</td>
<td>310.343472</td>
<td>267.0</td>
</tr>
<tr>
<td>train_Client_1</td>
<td>ELEC</td>
<td>557.5405405</td>
<td>3.917864e+04</td>
<td>197.935960</td>
<td>520.0</td>
</tr>
<tr>
<td>train_Client_10</td>
<td>ELEC</td>
<td>798.6111111</td>
<td>2.640330e+05</td>
<td>513.841374</td>
<td>655.5</td>
</tr>
<tr>
<td>train_Client_100</td>
<td>ELEC</td>
<td>1.2000000</td>
<td>1.301053e+01</td>
<td>3.607011</td>
<td>0.0</td>
</tr>
<tr>
<td>train_Client_1000</td>
<td>ELEC</td>
<td>663.7142857</td>
<td>5.054914e+04</td>
<td>224.831365</td>
<td>770.0</td>
</tr>
<tr>
<td>train_Client_10000</td>
<td>ELEC</td>
<td>422.0689655</td>
<td>1.568130e+05</td>
<td>395.996206</td>
<td>425.0</td>
</tr>
<tr>
<td>train_Client_100000</td>
<td>GAZ</td>
<td>245.3157895</td>
<td>6.717989e+04</td>
<td>259.190846</td>
<td>70.0</td>
</tr>
</tbody>
</table>

The newly modified invoice dataset was then merged with the original client dataset, increasing the number of attributes by 6 to 145. The number of records became 135,493 which is exactly the number of unique client ID or STEG users in the dataset. However, since the client ID has no value or relationship with the target, the unique combinations of client ID and tariff type were obtained first, and client ID was replaced by tariff type in the resultant dataset as shown in Table 4, increasing the number of records to 197,444. This dataset was then used to train and test the different supervised machine-learning models.

Table 3. Dataset Overview Stage 2

<table>
<thead>
<tr>
<th>client_id</th>
<th>avg_consom_1_ELEC</th>
<th>avg_consom_1_GAZ</th>
<th>var_consom_1_ELEC</th>
<th>var_consom_1_GAZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>train_Client_0</td>
<td>352.4000000</td>
<td>0.000000</td>
<td>9.631307e+04</td>
<td>0.000000e+00</td>
</tr>
<tr>
<td>train_Client_1</td>
<td>557.5405405</td>
<td>0.000000</td>
<td>3.917864e+04</td>
<td>0.000000e+00</td>
</tr>
<tr>
<td>train_Client_10</td>
<td>798.6111111</td>
<td>0.000000</td>
<td>2.640330e+05</td>
<td>0.000000e+00</td>
</tr>
<tr>
<td>train_Client_100</td>
<td>1.2000000</td>
<td>0.000000</td>
<td>1.301053e+01</td>
<td>0.000000e+00</td>
</tr>
<tr>
<td>train_Client_1000</td>
<td>663.7142857</td>
<td>0.000000</td>
<td>5.054914e+04</td>
<td>0.000000e+00</td>
</tr>
<tr>
<td>train_Client_10000</td>
<td>422.0689655</td>
<td>245.315789</td>
<td>1.568130e+05</td>
<td>6.717989e+04</td>
</tr>
</tbody>
</table>

Table 4. Final Dataset Overview

<table>
<thead>
<tr>
<th>tariff_type</th>
<th>district</th>
<th>client_catg</th>
<th>region</th>
<th>avg_consom_1_ELEC</th>
<th>avg_consom_1_GAZ</th>
<th>var_consom_1_ELEC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>60</td>
<td>11</td>
<td>101</td>
<td>352.4000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>2</td>
<td>11</td>
<td>69</td>
<td>11</td>
<td>107</td>
<td>557.5405405</td>
<td>0.000000</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>62</td>
<td>11</td>
<td>301</td>
<td>798.6111111</td>
<td>0.000000</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>69</td>
<td>11</td>
<td>105</td>
<td>1.2000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
<td>62</td>
<td>11</td>
<td>303</td>
<td>663.7142857</td>
<td>0.000000</td>
</tr>
<tr>
<td>6</td>
<td>40</td>
<td>69</td>
<td>11</td>
<td>103</td>
<td>422.0689655</td>
<td>245.315789</td>
</tr>
</tbody>
</table>
4.2.1.4 Feature Selection

Choosing the most relevant attributes that have a significant impact on the target variable in a dataset is a crucial stage in data analytics projects. Several benefits can be achieved by performing the feature selection including Improving the model performance, reducing overfitting and complexity, and achieving faster model training. In this project, all original attributes posed a value to the target variable except for three attributes, which don’t have any correlation with the outcome of the classification model, the client ID, counter number, and counter code. The client ID was replaced by tariff type at the final stages of data pre-processing after obtaining the distinct combinations of client ID and tariff type. Whereas both counter number and counter code were removed from the dataset. Furthermore, as part of model fitting, the model algorithm will check the information gain of each attribute and will act accordingly.
4.3 Modeling

The training data is the foundation for the machine learning model-building process. The model might not be able to recognize the underlying relationships and patterns in the data without a high-quality training dataset, which could result in poor performance and incorrect classifications. Since the obtained final dataset was still imbalanced in terms of the target variable, where fraudulent customers had significantly fewer observations than honest customers.

The down-sampling was essential to be performed to prevent a biased model towards the majority class. It involves reducing the number of samples in the honest customers class, to the number in the fraudulent customer class, ensuring a balanced target-wise dataset as shown in Figure 21 below.

```r
```
# Attach packages
library(groupdata2)

# Using downsample()
rit_full_data_for_model_after_downsample = downsample(rit_full_data_for_model, cat_col = "target")

# check imbalance for testing subset
rit_full_data_for_model_after_downsample$target %>%
  table() %>%
  prop.table()
```

```
## .
## 1 0
## 0.5087337 0.4992663
```

**Figure 23.** Dataset Down-Sampling

The down-sampled dataset was then split into a training dataset and a testing dataset in an 80 / 20 ratio respectively as shown below, which will be used to build and evaluate the performance of the machine learning models.

```r
```
# Split our dataset into 80% training data and 20% test data.
set.seed(54321)
rit_data_validation_split <- sample(nrow(rit_full_data_for_model_after_downsample), nrow(rit_full_data_for_model_after_downsample)*0.80)

rit_data_train <- rit_full_data_for_model_after_downsample[rit_data_validation_split, ]
rit_data_test <- rit_full_data_for_model_after_downsample[-rit_data_validation_split, ]
```

**Figure 24.** Dataset Split
An additional check was required to ensure balanced datasets after the creation of the training and testing datasets, prior to using them for building and evaluating the models, and the results below have shown that both datasets were balanced target-wise.

![Image: Training & Testing Dataset Balance Check]

In this project, and based on the research that was conducted, a total of three supervised machine learning models will be used to classify the customers as fraudulent and honest customers, which are the decision tree, random forest, and support vector machine. Each model will be evaluated based on four performance metrics, accuracy, precision, recall, and F1 score.

The accuracy of the model is obtained by dividing the total correctly classified observations by the total number of observations, the precision is the ratio between the True Positives and all the Positives, the recall is a measure of a model correctly identifying True Positives and the F1 score a performance metric for classification which is a measure of a test’s accuracy. The four metrics are obtained through the following equations:

\[
Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4.1)
\]

\[
Recall = \frac{TP}{TP + FN} \quad (4.2)
\]

\[
Precision = \frac{TP}{TP + FP} \quad (4.3)
\]

\[
F1 Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.4)
\]

Where TP: True Positive, TN: True Negative, FP: False Positive, and FN: False Negative
4.3.1 Decision Tree

The decision tree is used to represent decisions and decision-making visually and explicitly, the data is continuously split according to a certain parameter. It consists of two entities: decision nodes, and leaves. The leaves are the decisions or the final outcomes, whereas the decision nodes are where the data is split. Each node represents either a fraud customer (target class 1) or an honest customer (target class 0).

From the resultant decision tree in Figure 24, the most important attribute which is considered the root node was determined to be the range old index for an ELEC counter type, where a value of less than 7,893 will be classified as an honest customer, else will go through different leaves to determine the outcome. So, for example, if the range old index ELEC has a value larger than 7,893, the minimum value of the old index in an ELEC counter type will be considered, where a value of less than 1 will proceed to another leave considering the account duration, where a value more than 181 months will classify the customer as fraudulent.
The decision tree model was tested through the usage of the testing dataset which comprises 20% of the dataset, where the confusion matrix has shown that among 2,541 fraud customers, the model has classified 1,710. On the other hand, 1,953 honest customers were correctly classified out of 2,571. The accuracy of the model was calculated to be 71.65% since the total correctly classified instances were 3,663 among the 5,112.

![Confusion Matrix](image1)

**Figure 27. Decision Tree Accuracy**

The precision was calculated to be 0.73, whereas the recall of the model was 0.67, and finally, the F1 score was 0.70 as per the below results.

```r
# Recall, Precision & F1 Score Calculation
# Precison = TP / (TP+FP)
# Recall = TP / (TP+FN)
# F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

dt_model_precision = (1710) / (1710+618)
dt_model_precision

dt_model_recall = (1710) / (1710+831)
dt_model_recall

f1_score_dt = 2 * (dt_model_precision * dt_model_recall) / (dt_model_precision + dt_model_recall)
f1_score_dt
```

![R Code](image2)

**Figure 28. Decision Tree Model Precision, Recall & F1 Score**
4.3.2 Random Forest

The next model that was built is the Random Forest, which is a commonly used machine learning algorithm that merges the output of multiple decision trees to reach a single result. Similarly, the random forest model was tested through the usage of the testing dataset which comprises 20% of the dataset, and the obtained accuracy reached 82.73%. The confusion matrix has shown that 2,183 instances were correctly classified as fraudulent customers among 2,708, whereas 2,046 among 2,404 were correctly classified as honest customers as shown in the below output.

![Confusion Matrix and Statistics](image)

**Figure 29. Random Forest Accuracy**

The remaining three evaluation metrics were calculated similarly and found that precision is 0.86, recall is 0.81 and the F1 score is 0.83 as shown in the below output.

```
r
#Recall, Precision & F1 Score Calculation
#Precision = TP / (TP+FP)
#Recall = TP / (TP+FN)
#F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

rf.model.precision = (2183) / (2183 +358 )
r.rf.model.precision

rf.model.recall = (2183) / (2183 + 525)
r.rf.model.recall

f1_score_rf = 2 * (rf.model.precision * rf.model.recall) / (rf.model.precision + rf.model.recall)
f1_score_rf
```

```
[1] 0.8591106
[1] 0.80613
[1] 0.8317775
```

**Figure 30. Random Forest Model Precision, Recall & F1 Score**
4.3.3 Support Vector Machine

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification. The SVM algorithm’s objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, enabling it to quickly classify new data points in the future. This best decision boundary is called a hyperplane. Similarly, the model was built and validated by using the training dataset and testing dataset respectively.

The obtained accuracy was dependent on the cost parameter of the model, which determines the degree at which an SVM should bend with the data. Therefore, for a low cost, the aim is for a smooth decision surface, and for a higher cost, the aim to correctly classify more points. It is referred to as the cost of misclassification. Therefore, changing the cost changes the accuracy of the model. In most runs through R Studio, increasing the cost resulted in higher accuracy. Running the model requires high computing power which consumes time, therefore, only a few runs were performed. Below are some of the accuracies that were calculated based on different costs.

![Figure 31. Support Vector Machine Models' Accuracy Per Cost](image-url)
Based on the different values of cost and accuracy that were obtained, the remaining three evaluation metrics were calculated for each cost value as shown in the table below.

**Table 5. Support Vector Machine Models' Accuracy, Precision, Recall & F1 Score**

<table>
<thead>
<tr>
<th>Model</th>
<th>Cost</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>3</td>
<td>48.63%</td>
<td>0.95</td>
<td>0.49</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>52.27%</td>
<td>0.07</td>
<td>0.68</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>49.45%</td>
<td>0.98</td>
<td>0.50</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>42.00%</td>
<td>0.31</td>
<td>0.37</td>
<td>0.34</td>
</tr>
</tbody>
</table>

The final selection of the SVM model among the above four in terms of performance was selected to be the model with a cost of 0.05, due to the fact that precision, recall, and F1 score metrics were higher compared with the other models, and the accuracy of the model is close compared with the highest accuracy model.

### 4.4 Final Model Evaluation and Selection

The final selection of the model is based on the evaluation metrics’ results that were obtained during the testing phase of the three machine learning models, the decision tree, random forest, and support vector machine. The table below is a summary of the metrics indicating the performance of the three models.

**Table 6. Machine Learning Models' Performance Comparison**

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>71.65%</td>
<td>0.73</td>
<td>0.67</td>
<td>0.70</td>
</tr>
<tr>
<td>Random Forest</td>
<td>82.73%</td>
<td>0.86</td>
<td>0.81</td>
<td>0.83</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>49.45%</td>
<td>0.98</td>
<td>0.50</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Based on the above results, the Random Forest was selected as the final model for this project, to classify the Tunisian Company of Electricity and Gas (STEG) customers as either fraudulent or honest customers. Achieving an accuracy of more than 82%, a precision of 86%, a recall of 81%, and an F1 score of 83% is a good indicator of the performance of the model, which will contribute to decreasing the financial losses caused by electricity theft and fraud through the detection of fraudulent customers.
4.5 Model Deployment

Model deployment, which is the process of making a predictive model available for usage in a production environment, is a crucial phase in CRISP-DM. The ability for stakeholders to use the model's predictions via a real-time Application Programming Interface (API) makes model deployment essential. The deployment can also help in increasing the accuracy of the predictions as the model can be continuously updated with new data. In this project, the Tunisian Company of Electricity and Gas (STEG), which is the company that publicly shared its data, can utilize the model to enhance its operation and processes, through the classification of clients as either honest customers or fraudulent customers, will help in reducing the financial impacts of electricity theft and fraud on the company that used to occur in the past years.
Chapter 5 Conclusion

5.1 Conclusion

Distinguishing between normal and suspicious consumers with conventional meters that are unable to recognize the periodicity and variability of consumption poses some challenges in detecting fraudulent behavior for many studies. In this project, it is no different. The two datasets collected were explored and visualized using different charts to understand the information and patterns embedded in the data, and they had to go through different stages of data preparation and preprocessing including cleansing, data transformation, feature engineering, and selection, to enhance the quality and volume of information provided in the data, and to allow the different models to be appropriately trained to detect the hidden patterns and relationships of different variables and their correlation to the target variable.

Three supervised machine-learning models were built and tested for the classification of customers, which are decision tree, random forest, and support vector machine. A total of four different metrics were used to evaluate the performance of the models, and upon final comparison, the random forest model has been selected since it surpassed the other two models. Random forest achieved 82.73% in terms of accuracy for classifying the correct classes and scored more than 80% in the three remaining metrics, precision, recall, and f1 score. The selected model will contribute to decreasing the financial losses caused by electricity theft and fraud, through the detection and classification of fraudulent customers.
5.2 Recommendations

There are many ways to enhance the performance of the machine learning models, including using more accurate and high-quality data from different sources to accurately capture the patterns of fraudulent customers. Since conventional meters were used in this project, some important patterns were definitely lost during the on-site inspection and collection of data, which could’ve contributed to a more robust model classification. However, with the utilization of different charts as a part of the data visualization, some understanding and insights were obtained on the regions, districts, and electrical consumption patterns of fraudulent customers, which could be utilized for the transition to the implementation of smart meters, which will be the main source of data that fulfill the requirements needed in terms of quality and accuracy for a better customer classification for building the machine learning models.

5.3 Future Work

In terms of future work, more machine-learning models could be built using the same dataset, and different approaches in terms of data visualization, preparation, and preprocessing shall be performed to better understand, how such steps contribute to changing the accuracy of the models and the other evaluation metrics, as well as to generate more insight and patterns that couldn’t be detected in this project due to the resources and time limitations.
Bibliography


