Fraud Detection in Customers' Electricity Consumption in Nigeria Using Support Vector Machine and C4.5 Decision Tree Algorithms

Sarumi, J.A. (PhD)
Department of Computer Science
Lagos State Polytechnic
Ikorodu, Lagos Nigeria
E-mails: jerrytechnologies@yahoo.co.uk; sarumi.j@mylaspotech.edu.ng
Phone: +2348023408122

ABSTRACT

Over the years, Electricity theft has been estimated to cost billions of Naira per year in Nigeria. To reduce electricity theft, electric utilities are leveraging data collected by using data analytics to identify abnormal consumption trends and possible fraud. In this study, use of data analytics in detecting electricity theft, and a metric that leverages this threat model in order to evaluate and compare anomaly detectors. Data mining technology have helped several industries and sectors in improving their various forms of technology, this study therefore employ machine learning algorithms for the classification of fraud detection in the electricity consumption of costumers. In this project, Support Vector Machine (SVM) and C4.5 Decision tree classification algorithms were employed for fraud detection using customer electricity consumption data. SVM and C4.5 achieved 63.4% and 65.9% accuracy respectively. Thus, C4.5 outperformed SVM based on the dataset experimented in this project.

Keywords: Active databases, distributed processing, intelligent processing, optimization, query processor

1. INTRODUCTION

The legal purpose of the electricity distribution authority is to expand and sustain a proficient organized and economical system of electricity supply throughout the Federation. In the early 1960s, the Niger Dam Authorities (NDA) and Electricity Corporation amalgamated to form the Electricity Corporation of Nigeria (ECN). In the late 2000s, the corporation became a public limited company (NEPA plc), and then later the name was changed again from NEPA plc to the Power Holding Company of Nigeria (PHCN). For several decades, despite consistent perceived cash investment by the federal government to improve the quality of power distribution, poor nationwide power outages have been the norm instead of the exception. Generally, the tariff has been criticized as being too low compared to the cost of generating power. The federal government of Nigeria has increased the tariff to attract foreign investors since 1st July 2010, and most likely for an increase in tariff coming 1st of April 2020 in order to meet the growing concern for foreign investors into the electricity sector.

Major issues within the Nigerian power sector, principally concerning power outages and unreliable service, compelled the Nigerian government to take radical action. It enacted the Electric Power Sector Reform Act of 2005, which called for unbundling the national power utility company into a series of 18 successor companies: six generation companies, 12 distribution companies covering all 36 Nigerian states, and a national power transmission company. The act stipulated that ownership of these companies be granted to the Bureau of Public Enterprises (the privatization arm of the federal government) and the Ministry of Finance Incorporated. This unbundling paved the way for an ambitious privatization program to be carried out by the Bureau of Public Enterprises in Nigeria.

Utility companies provide an essential service to developed societies, supplying electricity, gas and water to homes, businesses and factories. The necessary infrastructure that guarantees services includes from the kilometers of pipes or lines that transport the energy to the millions of meters that monitors consumption of individual customers. An important problem that these companies face is the imbalance between the energy billed with respect to the energy provided, called energy losses. Non-technical losses are a widely used, somewhat euphemistic name including fraud and meter malfunctions among other. Methods for committing fraud include splicing the pipes to bypass the meter, tampering with the meter to stop it or to slow it down, and simply connecting to the distribution network without even having a contract with the company or a meter. On the other hand, an accidental malfunction of the meter also results in net energy loss for the company. For brevity, along this paper, we will use the term fraud to refer to both forms of non-technical loss, intentional fraud and meter malfunction: The company actions and revenue after the detection will be very different, but both are issues that the company wants to detect and fix as early as possible, and the detection schemes are essentially the same.

Most cases of fraud involving meter tampering or malfunctioning can be detected by direct inspection by a trained technician. However, it is extremely expensive to send technicians to inspect a large number of meters. Therefore, companies usually perform a pre-selection of a subset of meters to be directly verified by technicians in a given period of time and area, a concept that we call a campaign. Every customer visit has a cost, so in order to be worthwhile campaigns need to have relatively high precision (i.e., percentage of problems detected with respect to the number of meters verified). Company gains, on the other hand, are directly proportionally to campaign recall (i.e., fraction of the existing fraud that is detected), so campaign design is all about the classical precision-recall tradeoff. Traditionally, campaigns are based on simple sets of rules indicating fraud (e.g., abrupt decrease of consumption, or no consumption during a long period of time). These rules can be used to detect the fraudulent/irregular customers, but achieve a low success rate, not much higher than selecting customers for the campaign at random.

This can be explained by many other reasons besides the fraud (e.g. customer spends a long convalescence in the hospital, or the house is a second residence that does not follow the consumption patterns). It is very natural to try to exploit data from the past to design better campaigns, including statistical and machine-learning based techniques. This study proposes an approach to detect fraud and irregularities in Nigeria electricity distribution companies. This study also proposes the detection of fraud as a classification problem, where several techniques over the set of historic cases of fraud are applied. Remarkably, the system continuously evolves its knowledge by incorporating the results of each campaign in order to improve model quality and obtain more efficient campaigns. The development of this study is based on real datasets from the electricity distribution company; the approach proposes a significant step towards automating the detection of fraud and irregularities in the consumption of electric energy, when compared to the traditional rule-based approach that nowadays utility companies apply. This study proposes machine learning algorithm that can aid the detection of fraud in the consumption of electricity by customers.

1.2 Statement of the problem

In Nigeria, electricity Power, kWh (Kilowatt-hour) meters record readings are used to bill customers for the amount of electricity consumed. In most cases, inspections are carried out at random, while some targeted raids are undertaken based on information reported by the public or meter readers. With the application of an intelligent detection system using machine learning techniques, customer installation inspections are guided to a reduced group of likely fraudulent customers. Therefore, an intelligent model to guide the inspections of electricity meters is proposed in this study to improve the supply and consumption strategies of electricity consumption in Nigeria by reducing non-technical loses in the supply distribution.

1.3 Aim and Objectives

This study detects fraud in electricity Power supply in Nigeria using machine learning algorithms.

The objectives of this study are to:
1. Develop a classification algorithm using SVM (support vector machine) and Decision tree.
2. Compare the output performances in (i) in terms of Accuracy, Sensitivity, Specificity and Precision.

1.4 Significance of the Study

This study analyzes a fraud detection that will provide information to Nigerian Electricity Power for efficient detection and classification of activities in order to increase effectiveness of their onsite operation. With the implementation of the proposed model, operational costs due to onsite inspection in monitoring activities will be significantly reduced. This will also reduce the number of inspections carried out at random, resulting in higher fraud detection hit rate.

LITERATURE REVIEW

2.1 Nigeria Power and Electricity Distribution Company Limited

By 1950, in order to integrate electricity power development and make it effective, the then colonial government passed the ECN ordinance No. 15 of 1950. With this ordinance in place, the electricity department and all those undertakings which were controlled came under one body.The statutory function of the authority is to develop and maintain an efficient co-ordinate and economical system of electricity supply throughout the Federation. The decree further states that the monopoly of all commercial electric supply shall be enjoyed by NEPA to the exclusion of all other organizations. This however, does not prevent privy individuals who wish to buy and run thermal plants for domestic use from doing so. NEPA, from 1989, has since gained another status-that of quasi-commercialization. By this, NEPA has been granted partial autonomy and by implication, it is to feed itself. The total generating capacity of the six major power stations is 3,450 megawatts. By 1970, the military government appointed a Canadian Consultant firm "Showment Ltd" to look into the technical details of the merger. The report was submitted to the government in November 1971. By Decree No. 24 the ECN were merged to become the NEPA with effect from 1 April 1972. The actual merger did not take place until 6 January 1973 when the first general manager was appointed. The day-to-day running of the authority is the responsibility of the managing director.

In the early 1960s, the Niger Dam Authorities (NDA) and Electricity Corporation amalgamated to form the Electricity Corporation of Nigeria (ECN). Immediately after the end of the 1967-1970 Nigerian civil wars, the management of ECN changed its name to the National Electric Power Authority, or NEPA. In the late 2000s, the company became a public limited company (NEPA plc), and then later the name was changed again from NEPA plc to the Power Holding Company of Nigeria (PHCN). Generally, the tariff has been criticized as being too low compared to the cost of generating power. The federal government of Nigeria has increased the tariff to attract foreign investors since 1 July 2010 in order to meet the growing concern for foreign investors into the electricity sector.

cisdi Journal

Computing, Information Systems, Development Informatics & Allied Research Journal
Vol. 12 No. 1, 2021 - www.cisdijournal.net

Privatization

Major issues within the Nigerian power sector, principally concerning power outages and unreliable service, compelled the Nigerian government to take radical action. It enacted the Electric Power Sector Reform Act of 2005, which called for unbundling the national power utility company into a series of 18 successor companies: six generation companies, 12 distribution companies covering all 36 Nigerian states, and a national power transmission company. The act stipulated that ownership of these companies be granted to the Bureau of Public Enterprises (the privatization arm of the federal government) and the Ministry of Finance Incorporated. This unbundling paved the way for an ambitious privatization program to be carried out by the Bureau of Public Enterprises in Nigeria.

In 2007, the Bureau of Public Enterprises hired CPCS Transcom Limited, an international consulting firm based in Ottawa, Ontario, Canada to provide advice about the best ways to move forward with the privatization of the country's 11 distribution companies and the 6 generation companies. In 2010, CPCS was consulted again in order to provide advice on the Nigerian government's privatization program. On 30 September 2013, following the privatization process initiated by the Goodluck Jonathan regime, PHCN ceased to exist. In its stead, the Nigerian Electricity Regulatory Commission (NERC) was formed. The independent regulatory agency, as provided in the Electric Power Sector Reform Act of 2005 was tasked with monitoring and regulating the Nigerian electricity industry, with issuing licenses to market participants, and with ensuring compliance with market rules and operating guidelines.

PHCN Successor Companies

Following the 2013 divestiture of the federal government from PHCN, the company was divided into separate companies or entities called Local Electric Distribution Companies or Local Distribution Companies (LDC).

Each company will be responsible for handling electricity distribution in each state or region.
- Abuja Electricity Distribution Company plc
- Benin Electricity Distribution Company plc
- Eko Electricity Distribution Company plc
- Enugu Electricity Distribution Company plc
- Ibadan Electricity Distribution Company plc
- Ikeja Electricity Distribution Company plc
- Jos Electricity Distribution Company plc
- Kano Electricity Distribution Company plc
- Kaduna Electricity Distribution Company plc
- Port Harcourt Electricity Distribution Company plc
- Yola Electricity Distribution Company plc

Generation (6 companies):
- Afam Power plc
- Egbin Power plc
- Kainji Hydro-Electric plc
- Sapele Power plc
- Shiroro Hydro-Electric plc
- Ughelli Power plc

Transmission (1 company):
- Transmission Company of Nigeria

2.2   Data Mining

The construction of a data warehouse, which involves data cleaning and data integration, can be viewed as an important pre-processing step for data mining. However, a data warehouse is not a requirement for data mining. Building a large data warehouse that consolidates data from multiple sources, resolves data integrity problems, and loads the data into a database, can be an enormous task, sometimes taking years and costing millions of dollars. If a data warehouse is not available, the data to be mined can be extracted from one or more operational or transactional databases, or data marts. Alternatively, the data mining database could be a logical or a physical subset of a data warehouse. Data mining uses the data warehouse as the source of information for Knowledge Data Discovery (KDD) systems through an amalgam of artificial intelligence and statistics-related techniques to find associations, sequences, classifications, clusters, and forecasts.

Typically the data warehouse architecture has three components:

    i.   Data acquisition software (back-end) which extracts data from legacy systems and external sources, consolidates and summarizes the data, and loads them into the data warehouse.

    ii.   The data warehouse itself contains the data and associated database software. It is often referred to as the "target database."

    iii.   The client (front-end) software, which allows users and applications (such as DSS and EIS) to access and analyze data in the warehouse.

These three components may reside on different platforms, or two or three of them may be on the same platform. Regardless of the platform combination, all three components are required. Data mining offers value across a broad spectrum of industries and can be used as a vehicle to increase profits by reducing costs and/or raising revenue. A few of the common ways in which data mining can accomplish those objectives are:

    i.   lowering costs at the beginning of the product life cycle during research and development;

    ii.   determining the proper bounds for statistical process control methods in automated manufacturing processes;

Data mining is performed by people such as the project leader, who has the overall responsibility for planning, coordinating, executing, and deploying the data mining project. The data mining client, who is the business domain expert that requests the project and utilizes the results, but generally does not possess the technical skills needed to participate in the execution of the more technical phases of the data mining project such as data preparation and modeling. The data mining analyst, who thoroughly understands, from a business perspective, what the client wants to accomplish and assists in translating those business objectives into technical requirements to be used in the subsequent development of the data mining model(s). The data mining engineer, who develops, interprets and evaluates the data mining model(s) in light of the business objectives and business success criteria.

Data mining engineering is performed in consultation with the data mining client and the data mining analyst in order to assist in achieving business ends. The IT analyst, who provides access to the hardware, software and data needed to complete the data mining project successfully. It is important to note that data mining is a technology that needs to co-exist harmoniously with other technologies in the organization. In addition, the data to be mined could be coming from virtually any existing system, database, or data warehouse in the organization. Depending on the scale and scope of the project, multiple individuals may assume each of the various roles. For example, a large project would likely need several data mining analysts and data mining engineers; eliminating expensive mailings to customers who are unlikely to respond to an offer during a marketing campaign; facilitating one-to-one marketing and mass customization opportunities in customer relationship management.

Many organizations use data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers. By determining characteristics of good customers (profiling), a company can target prospects with similar characteristics. By profiling customers who bought a particular product a firm can focus attention on similar customers who have not bought that product (cross-selling). Profiling also enables a company to act to retain customers who are at risk for leaving (reducing churn or attrition), because it is usually far less expensive to retain a customer than acquire a new one. However, profiling introduces issues of privacy.

Examples of other industries where data mining can make a contribution include:

i.	Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services.
ii.	Insurance companies and stock exchanges are interested in applying data mining to reduce fraud.
iii.	Medical applications use data mining to predict the effectiveness of surgical procedures, medical tests, or medications.
iv.	Financial firms use data mining to determine market and industry characteristics as well as to predict individual company and stock performance.
v.	Retailers make use of data mining to decide which products to stock in particular stores (and even how to place them within a store), as well as to assess the effectiveness of promotions and coupons.
vi.	Pharmaceutical firms mine large databases for chemical compounds and genetic material to discover substances that might be candidates for development as agents for the treatments of disease.

Data mining uses the classical statistical procedures such as logistic regression, discriminant analysis, and cluster analysis, machine learning techniques such as neural networks, decision trees, and genetic algorithms. In the continuum of data analysis techniques, the disciplines of statistics and of machine learning often overlap.

Machine learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience. Machine learning aims to provide increasing levels of automation in the knowledge engineering process, replacing much time-consuming human activity with automatic techniques that improve accuracy or efficiency by discovering and exploiting regularities in training data. Although machine learning algorithms are central to the data mining process, it is important to note that the process also involves other important steps, including:

i.	Training data is data that is used to estimate or train a model. Training data are existing data that's pre-classified in that the outcomes are already known.
ii.	building and maintaining the database,
iii.	data formatting and cleansing,
iv.	data visualization and summarization,
v.	the use of human expert knowledge to formulate the inputs to the learning algorithm and to evaluate the empirical regularities it discovers,
vi.	Determining how to deploy the results.

2.3 Types of Data Mining

Data mining techniques are divided into two kinds as supervised and non supervised techniques. "In supervised modeling, whether for the prediction of an event or for a continuous numeric outcome, the availability of a training dataset with historical data is required. Models learnt from past cases" . Supervised techniques such as classification models depend on a label class which must exist in the database under study.

This type of mining is used in several fields like direct marketing, Credit/loan approval, Medical diagnosis if a tumor is cancerous or benign, fraud detection if a transaction is fraudulent, etc. In the non-supervised learning such as clustering models, there is no class label. Clustering is a method of grouping data that share similar trend and patterns, applied in Insurance, city planning to identify groups of houses according to their house type, value, and geographical location.

2.4 Data Mining Methodology
The data mining methodology consists of the following steps.
  i.   Data cleaning to remove noise and inconsistent data.
  ii.   Data integration where multiple data sources may be combined.
  iii.   Data selection where data relevant to the analysis task are retrieved from the database.
  iv.   Data transformation where data are transformed into forms appropriate for mining by performing summary. Data mining (an essential process where intelligent methods are applied in order to extract data patterns).
  v.   Pattern evaluation to identify the truly interesting patterns representing knowledge based on some interesting measures.
  vi.   Knowledge presentation where visualization and knowledge representation techniques are used to present the mined knowledge to the user

2.5 Classification
Classification models are linear and non-linear. "Linear models are analytical models that assume linear relationships among the coefficients of the variables being studied. Furthermore, defined as "Approximation of a discriminant function or regression function using a hyper plane. It can be globally optimized using simple techniques, but does not adequately model many real-world problems". The other type of classification prediction models is the non-linear models defined as "Non-Linear Predictive Models are analytical model that does not assume linear relationships among the coefficients of the variables being studied" . Linear models in spite of their computational simplicity, stability they have an obvious potential weakness. "The actual process may not be linear, and such an assumption introduces uncorrectable bias into the predictions.

2.5.1 Support Vector Machines
The first selected mining technique is SVM that can be used as linear and non linear. "In machine learning, SVM is supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Moreover a support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite- dimensional space which can be used for classification regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class since in general the larger the margin the lower the generalization error of the classifier. Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space, presumably making the separation easier in that space.

To keep the computational load reasonable, the mapping used by the SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function K(x,y) selected to suit the problem. The hyperplanes in the higher dimensional space are defined as the set of points whose inner product with a vector in that space is constant in addition to performing linear classification. SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

2.5.2 Neural Network

The second selected mining technique is ANN which is applicable when working with the nonlinear problems. This technique used successfully in business fraud detection, and defined as "artificial neural networks (ANNs) are widely used for fraud detection", a neural network (NN), in the case of artificial neurons called artificial neural network (ANN) or simulated neural network (SNN), is an interconnected group of natural or artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach to computation. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network. In more practical terms, neural networks are non-linear statistical data modeling or decision making tools". Furthermore, "The inspiration for neural networks was the recognition that complex learning systems in animal brains consisted of closely interconnected sets of neurons. Although a particular neuron may be relatively simple in structure, dense networks of interconnected neurons could perform complicated learning tasks such as classification and pattern recognition".

2.5.3   K-Nearest-Neighbor

The last selected mining method is KNN, which also supports non-linear problem. This section reviews some point about KNN. "The k-nearest-neighbor method was first described in the early 1950s. The method is labor intensive when given large training sets, and did not gain popularity until the 1960s when increased computing power became available. It has since been widely used in the area of pattern recognition" (Han and Kamber, 2005). "In pattern recognition, the knearest neighbor algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally, and all computations is deferred until classification. The k-nearest neighbor algorithm is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of its nearest neighbor."

2.6 Methods For Preprocessing Noisy Data

Pre-processing due to the data being fine-grained, it is possible for some of the data to be inaccurate or missing. It is widely accepted that effective processing of data not only depends on the quality of the selected algorithms, but also the quality of the data. Therefore, a variety of pre-processing techniques were used to ensure the quality of the data. In this section, we discuss the two main pre-processing techniques utilized on the raw energy consumption data before the data are analyzed: feature selection and data aggregation. For faster access to the data in the files, an information retrieval program was used that compresses and indexes the raw energy consumption data.

Logistic Regression

Logistic regression is a predictive modelling technique where there is a correlation between the probability of a result and its predictor variables where π_i is the probability of the outcome, β_1, …, β_k are coefficients, X_1, …, X_{ik} are predictor variables. The β coefficients are transformed into odds ratios with the degree of importance of predictors well known. The Hosmer-Lemeshow statistic is accepted extensively in assessing the goodness of fit of developed models in logistic regression with a dichotomous outcome.

In a specific manner, logistic regression details the linear function of observed attributes for endogenous variables as the fitted probability of event. The fitted probability, logit (πi), is defined as logit, the logarithm of odds, which is the natural log of the probability of success and failure. Logistic regression has been used by many authors to create predictive models in healthcare, preparing landslide susceptible maps, prediction of fraud, among others.

2.7 Discriminant analysis
Discriminant analysis (DA) is a generalized linear modelling technique used in machine learning and pattern recognition for the linear combination or separation of categories of objects. DA is also applied in determining the variable that distinguishes two or more categories that helps in prediction. The normality of the explanatory variables in DA is assumed and applied in the prediction. A discriminant function is used to dichotomize the two groups.

2.8 Related Works
Lekini, Ndzana and Oumarou (2017) proposed an Evaluation of Customer Behaviour Irregularities in Cameroon Electricity Network using Support Vector Machine for a Non-Technical Losses (NTLs) in the Cameroonians electricity network are approximately 30 to 40% of production and are estimated at several billion CFA francs per year for National Electricity Company (ENEO); Hence the importance of finding effective solutions to fight against these losses. The purpose of this work was to develop a tool for the fraud detection for Cameroon National Electricity Company (ENEO) using support vector machines which consisted in data preprocessing base on the load profile, development of a model for classification, parameter optimization and detection of customer's irregularities and prediction.

Moslehi, Haeri, and Moini (2017) proposed an Analyzed and Investigated the Use of Electronic Payment Tools in Iran using Data Mining Techniques. The aim of the present work is to survey and analyze the use of electronic payment instruments in banks across the country using the statistics and information retrieved from the Central Bank and the data mining techniques. For this purpose, initially, according to the volume of the transactions carried out and using the K-Means algorithm, a label was dedicated to any record. Then the hidden patterns of the E-payment instrument transactions were detected using the CART algorithm. The results obtained enable bank administrators to balance their future policies in the field of E-payment and in the bank and customers' interest direction based on the detected patterns, and provide higher quality services to their customers.

The analysis and clustering of residential customers energy behavioral demand using smart meter data was examined by Haben, Singleton, and Grindrod, (2016). They presented an in-depth analysis of customer smart meter data to better understand peak demand and major sources of variability in their behavior. They presented four key time periods in which the data should be analyzed and use this to form relevant attributes for our clustering. They presented a mixture model based clustering where they discovered 10 distinct behavior groups describing customers based on their demand and their variability. Finally, using an existing bootstrapping technique we show that the clustering is reliable. To the authors knowledge this is the first time in the power systems literature that the sample robustness of the clustering has been tested.

Soma (2012), presented a worked on Modeling, Detection, and Prevention of Electricity Theft for Enhanced Performance and Security of Power Grid, proposed algorithms is enhanced in two directions - reducing the overall computation time, and including more real-time parameters using High Performance Computers (HPC). The encoding and classification algorithms are parallelized (in both Task Parallel and Data Parallel approaches). On the other hand, impact of Time-Based Pricing (TBP) and Distributed Generation (DG) on illegal consumers as well as the algorithms used for detection of illegal consumers are analyzed. Economics involved in terms of losses due to illegal consumption of electricity is also explained. Moslehi, Haeri, and Moini (2017) proposed, analyzed and Investigated the Use of

Electronic Payment Tools in Iran using Data Mining Techniques. This purpose, initially, according to the volume of the transactions carried out and using the K-Means algorithm, a label was dedicated to any record. Then the hidden patterns of the E-payment instrument transactions were detected using the CART algorithm. The results obtained enable bank administrators to balance their future policies in the field of E-payment and in the bank and customers' interest direction based on the detected patterns, and provide higher quality services to their customers.

3. METHODOLOGY

3.1 Project Approach

This study is a case study and focused on the electricity distribution company of Nigeria. This study aimed at collecting information from respondents on the causes of non-technical electricity losses in the Nigeria Power with the view of proposing a suitable fraud detection model to address this challenge. The study design refers to the arrangement of conditions for collection and analysis of data in a manner that aims to combine relevance to the research purpose with economy in the procedure. In addition Kothari (2004) observed that research design is a blue print which facilitates the smooth sailing of the various research operations, thereby making research as efficient as possible hence yielding maximum information with minimal expenditure of effort, time and money. The developed system consists of stages: the processing of the dataset and Classification. The dataset loads as an input, it obtains the relevant information in the dataset, the features of the dataset are classified by classification module using the SVM and decision tree algorithm Classifiers the results are displayed and compared in terms of the performance evaluation used in the work.

(i)   Use SVM for classification of the dataset.
(ii)   Use Decision Tree for classification of the dataset.
(iii)   Analyze the results in terms of accuracy, specificity, sensitivity, precision and computational time.

The overall system design of the proposed work is shown in the figure 3.1.
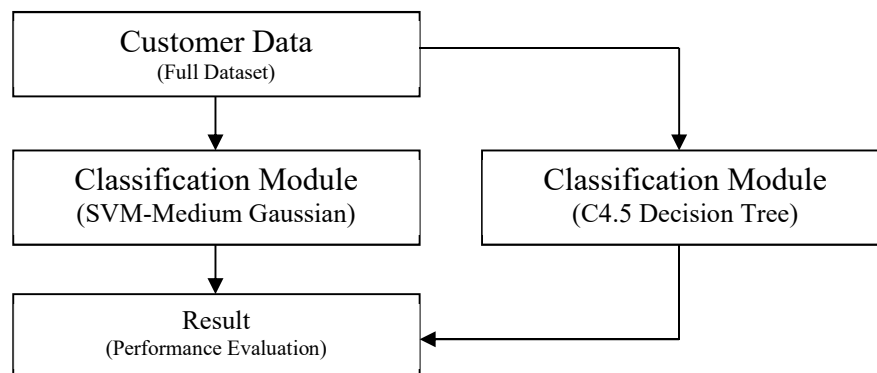
Figure 3.1: Framework of Proposed Approach

- Step 1- Process the customer dataset.
- Step 2- Use Support Vector Machine Medium Gaussian, as a Classification algorithm to carry out and simplify the performance of the dataset, so as to predict Customers irregularities in Electricity network level.
- Step 3- Use Simple Tree Decision Tree, as a Classification algorithm to carry out and simplify the performance of the dataset, so as to predict Customers irregularities in Electricity network level.
- Step 4- Compare the classification performance.

3.2 Experimental Dataset

To achieve desired goal, the proposed system uses electricity distribution customer datasets from Ikeja Electricity Distribution Company, Akowonjo Branch, Lagos Nigeria.

Customer: the customer dataset contains 200 instances and 5 attributes.

Table 3.1: Attributes and Instances

Instances	83
attributes	15

3.2.1 Dataset

Dataset was obtained from Akowonjo Office of the IKJELECTRIC Nigeria with a total instance of 83 and 15 attributes definition, with one response variable being the class grade factor. The attributes definitions are given in table 3.2 below:

Table 3.2: Attribute Descriptions

ATTRIBUTES DECRIPTORS		
Attributes	Data type	Description
Customer's Name	Male or female(Nominal)	Customer's Name of account
Active Terrain	Nominal	Consumption Purpose
Meter Number	Numeric	Meter Reading
Balance	Numeric	Customer's Account balance
Supply Reference	Numeric	Customer's Supply Number
Consumption	Numeric	Customer Consumption Rate
Monthly Charges	Numeric	Customer Charges
Status	Nominal	Customer Account Status

Figure 3.2: Customer Consumption Data from IKELECTRIC

3.2.2 Classification

Classification models have great influence to the final performances. Different classifiers have different performance on any given data set. The classification module uses the concept of Support Vector Machine (SVM); it is a machine learning method that can effectively deal with small-scale sample problems as well as with samples of large dimensions, SVM is confirmed as relatively the best and highly accurate classifier. SVM Polynomial kernel looks into the combinations of the given features of input samples to determine their similarity, regression analysis with polynomial in the number of parameters to be learned is put into consideration, these features corresponds to logical conjunctions of input features.

Decision tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital). The term Classification And Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by Breiman et al. in 1984. Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split.

3.3 Experimental Setup
The proposed system is implemented in MATLAB 2015.

3.3.1 MATLAB
Millions of scientists worldwide use MATLAB in analyzing and designing the systems and products transforming our world. MATLAB is a way of expressing computational analysis, with its easy way of visualizing and gaining insight of data. It has built-in graphics with desktop environment for experimentation, exploration and discovery. MATLAB tools have rigorous capabilities tested and designed to work. MATLAB is a collection of machine learning algorithms for solving data mining tasks.

3.3.2 System Configuration
This study has a system configuration of; Intel core duo processor, 1.1 Ghz speed, 4GB RAM, 20GB hard disk for its hardware configuration. A Windows 7 and later operating System, MATLAB 2015A operating tools and Customer Consumption Data from IKELECTRIC dataset for its software configuration.

3.3.3 Performance Evaluation Metrics
The performance evaluation metrics of classifier is evaluated in terms of classification accuracy, time, sensitivity, specificity, and precision.

The terms are defined below:
Sensitivity= TP/ (TP+FN) %
Specificity = TN/ (TN+FN) %
Accuracy = (TP+TN)/ (TP+TN+FP+FN) %
Precision: TP/ (TP+FP)

Where:
TP (True Positives) = correctly classified positive cases,
TN (True Negative) = correctly classified negative cases,
FP (False Positives) = incorrectly classified negative cases,
FN (False Negative) = incorrectly classified positive cases.

Sensitivity (true positive fraction) is the probability that a diagnostic test is positive, given that the person has the disease. Specificity (true negative fraction) is the probability that a diagnostic test is negative, given that the person does not have the disease. Accuracy is the probability that a diagnostic test is correctly performed. The developed system examined the customer electricity consumption in Nigeria using a data mining approach. This was passed through a classification approach of the stated method which follows the sequence of data collection, data filtering and normalization, data classification. The experimental setup were carried out and developed with the MATLAB programming (MATLAB 2015A) various functions were developed and link to a graphic user interface for user interactivity and responsiveness. The developed systems made use of various component environments in MATLAB to develop and output result of the data mining task namely. Due to the large capacity of MATLAB and its robustness the following are the minimum requirements the software application can be compiled on.

3.4 Data Analysis
Data was analyzed and evaluated using content analysis. The data collected was summarized according to the study objectives being causes of non- technical losses of commercial power at Electricity distribution and data preparation for preprocessing was analyzed using Microsoft Excel and Matlab.

4. RESULTS AND DISCUSSION

4.1 Experimental Results

In this study, the detection framework proposed in chapter three uses historical customer billing data of electricity power distribution customers' and transforms the data into the required format for the Support Vector Machine and Decision tree Simple Tree by data preprocessing and classification. Nigeria power consumption customers are represented by their consumption profiles over a period of time. These profiles are characterized by means of patterns, which significantly represent their general behavior and it is possible to evaluate the similarity measure between each customer and their consumption patterns. This creates a global similarity measure between a normal and fraud customers as a whole. The identification, detection and prediction are undertaken by the Support Vector Classification, which is the intelligent "classification engine".

4.2 Development of the Experiment

The overall developed system at run time is show in the figure 4.3 below. The customer's churn modeling is presented with the results obtained at each point of execution

    i.   Loading of Dataset
    ii.   Data Classifcation
    iii.   Training of with 10 fold hold out
    iv.   Classification of the holdout data (Test Dataset).

The collected dataset has a total observation of 83 customers in a Nigeria electricity distribution company with fifteen attributes taken. This fifteen attributes are significant attributes so as to be able to discover a well generalized knowledge from it. The dataset was loaded on MATLAB environment and the output is shown in Figure 4.1 below with 15 Attribute and 83 Instances.

Figure 4.1: Loaded Electricity Consumption Data

The modeling technique for classification of the consumption data is selected and used. There are several solutions for data mining problems. In this study, classification algorithm is used for applying methods. The data used includes 83 records of customers related to the performance of electricity consumption in the case study for the payment of electricity charges, province, and months during the year 2019. The main purpose of this work was to identify the factors affecting the irregularities in power consumptions in Nigeria according to the average amount of transactions. To do this, classification algorithm was applied to the given data. Two objective variables were required in the classification algorithm implementation. Using this algorithm requires a categorical target variable. In this study, the target variables are "attributes and instances". These variables had continuous numeric values. Figure 4.2 below shows the classification response and predictors table for the 10-folds cross validation holdouts.

Figure 4.2: Response Table for the Classification of the Data

Classification is one of the most important areas in machine learning, due to the huge variety of problems that can be stated as different or specific classification tasks. This study shows how there is a large amount of applications and problems in different aspects of classification systems, which can be solved successfully with classification algorithms.
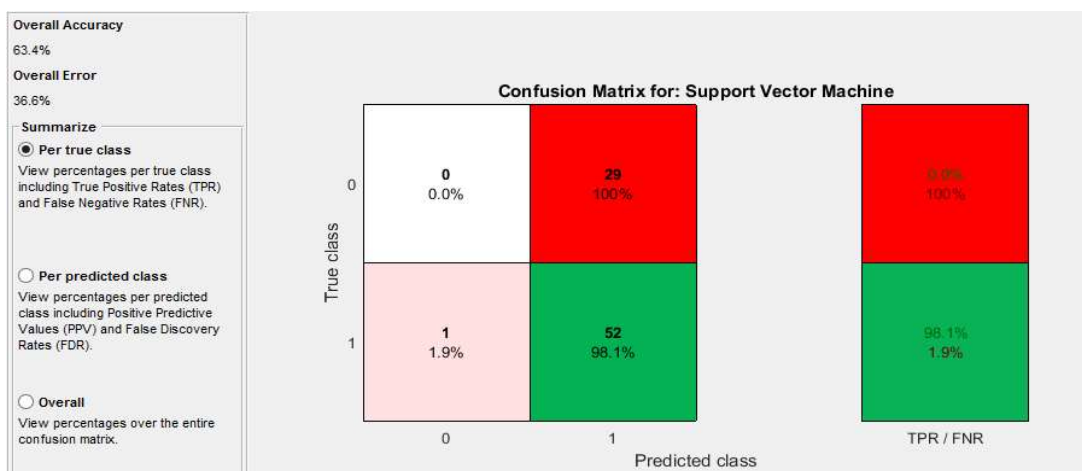

Figure 4.3: Overall Accuracy of the Classification of Power Consumption Using SVM

The improvement of techniques for classification is, without a doubt, the most important research line in the area, which will produce very promising results in applications in the near future. In this sense, some problems that are currently tackled as classification problems, exploiting the continuity of the data, could be tackled in the future as special cases of classification. For the future intelligent electrical network will be stated as classification problems and tackled with some of the algorithms described in this review (or improvements of them).figure 4.3 below shows the result output for the classification of power consumption in Nigeria using SVM. SVM achieved 63.4% overall accuracy.

TP= 52; TN=1; FP=29; FN=0
Accuracy rate = 0.6341
Percentage accuracy rate: 63.41%
Sensitivity Rate = 0.9811
Percentage Sensitivity Rate = 98.11
Specificity Rate = 0.34
Percentage Specificity Rate =34.5
Precision Rate = 0.6420
Percentage precision: 64.2

The scattered plot creates the x,y with the circles at the location specified by the vectors x and y. the graph is a bubble plot type. In this study the scattered plot shows the grouped sample data for the simple plot of one variable against another variable.
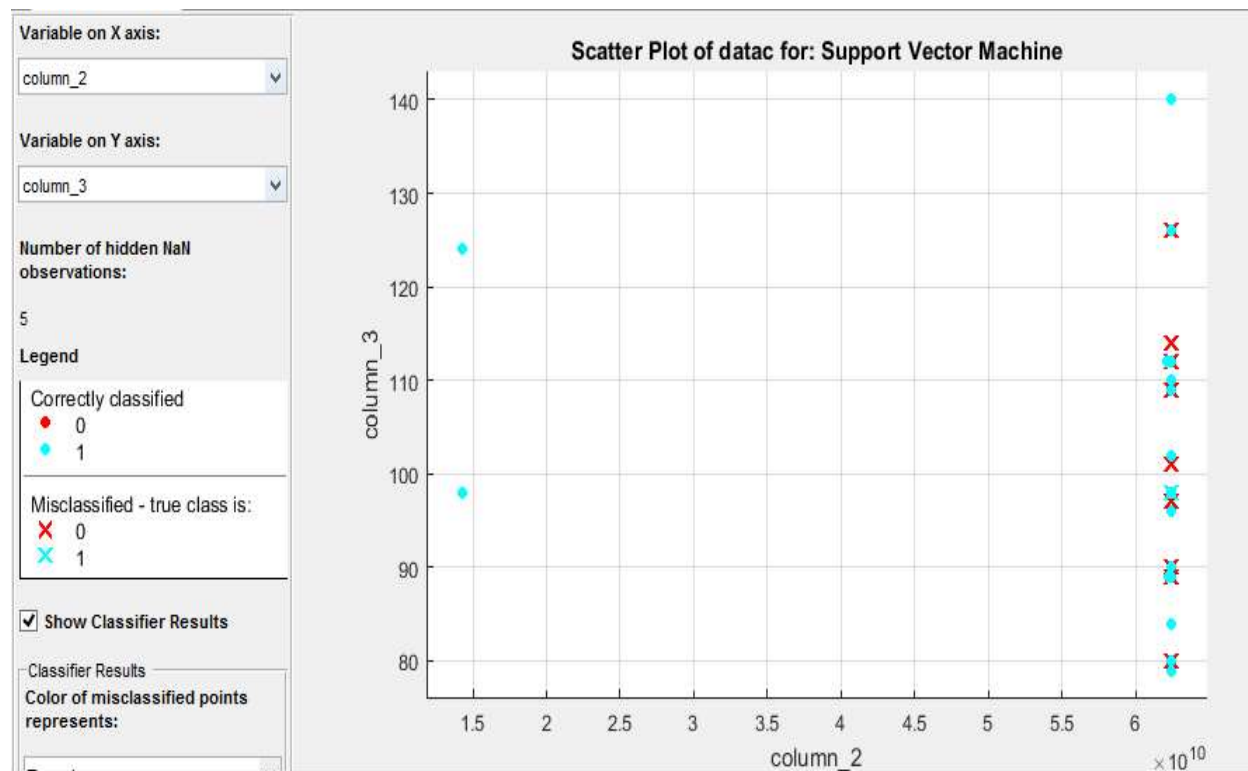

Figure 4.4: The Scattered Plot for the Classification Using SVM

This study presents the ROC curve that represents the specificity and sensitivity of the two classes of variables. The figure 4.5
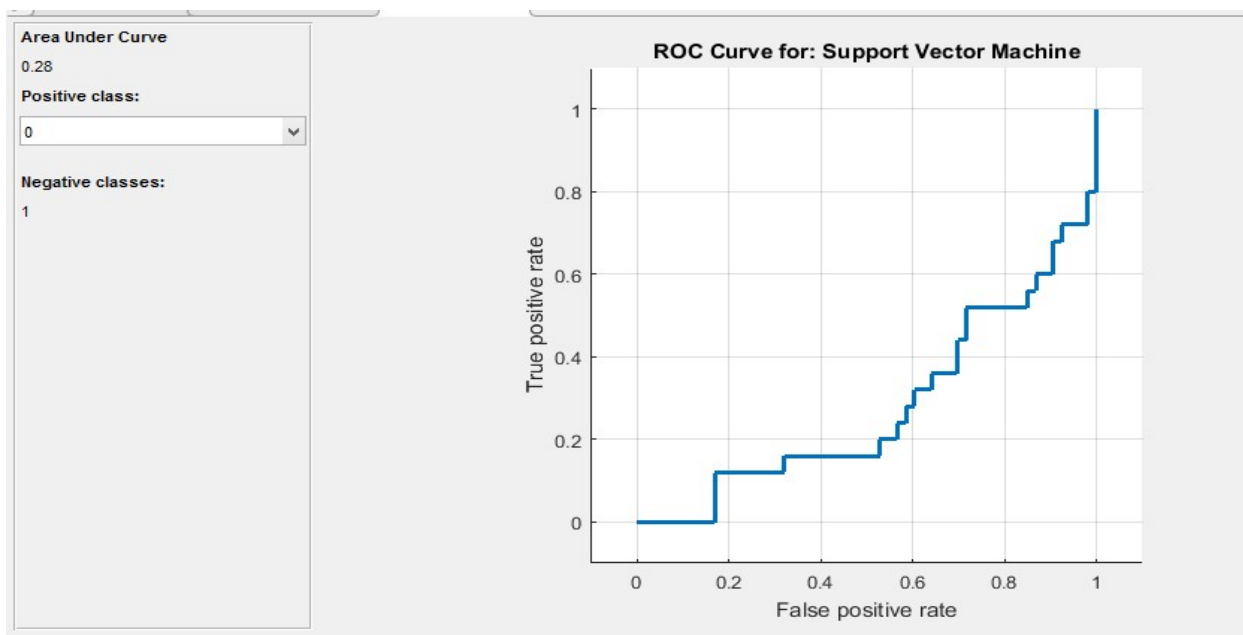

Figure 4.5: The ROC Curve Using SVM

Decision tree classification algorithm classification method using Simple Tree is used in this study, the result is shown in figure 4.6 below, and its result outperformed the SVM. Decision tree have proven to be a better method in achieving the aim of this study.
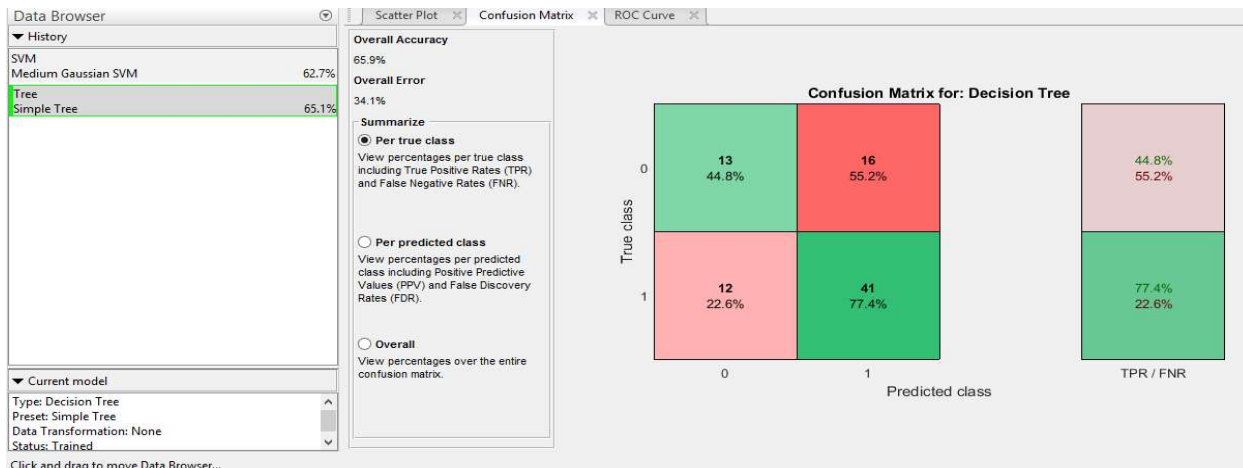

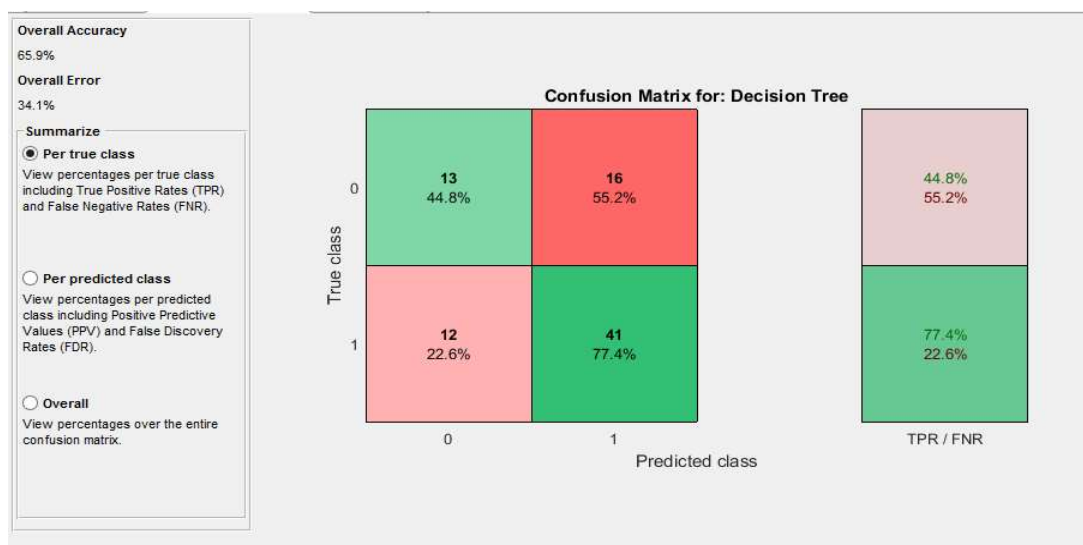**Figure 4.6: Classification Result for Decision Tree**

Figure 4.7: Confusion Matrix for Decision Tree

Accuracy rate = 0.659
Sensitivity Rate = 0.8039
Percentage Sensitivity Rate = 80.39
Specificity Rate = 0.52
Percentage Specificity Rate = 52.0
Precision Rate = 0.7735
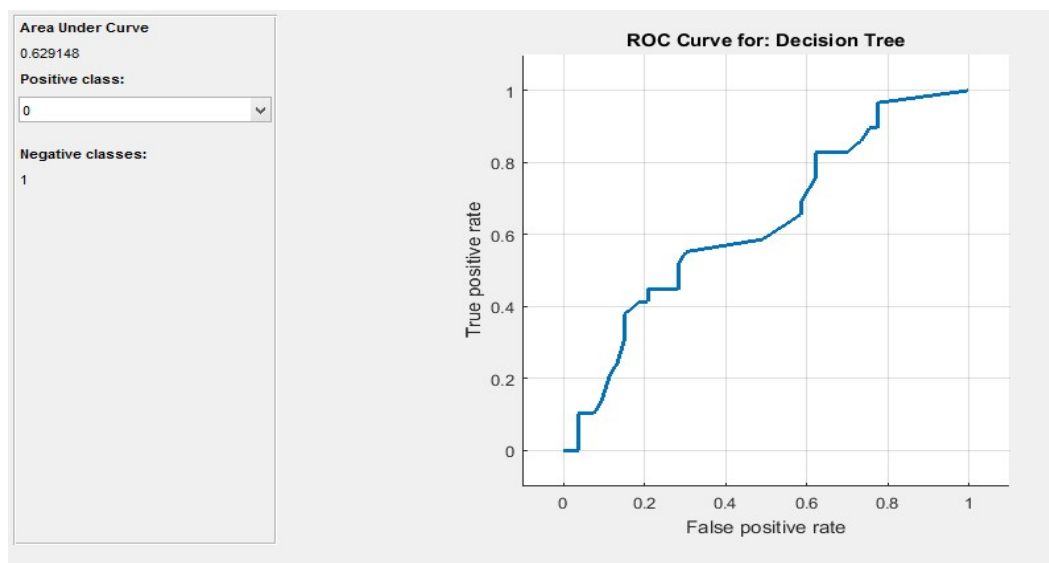*Percentage precision: 77.35*

Figure 4.8: ROC Curve for the Decision Tree Classifier

In this study, Decision tree has proven its efficiency in the classification of power consumption in the electricity distribution sector. Yet SVM is considered the best in terms of sensitivity A comparative

Table 4.1.

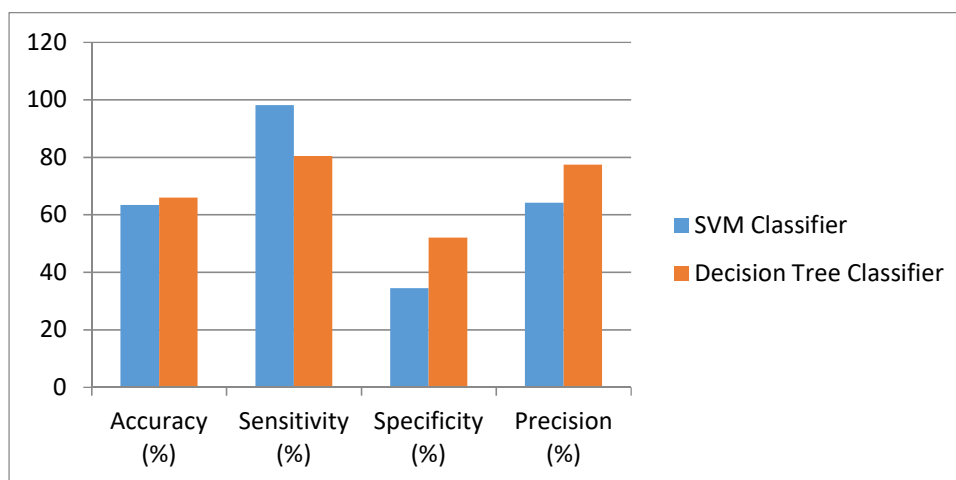| Performance Metrics	SVM Classifier	Decision Tree Classifier
Accuracy (%)	63.4	65.9
Sensitivity (%)	98.1	80.39
Specificity (%)	34.5	52
Precision (%)	64.2	77.35

Figure 4.9: Performance Metrics

5. CONCLUSION AND RECOMMENDATION

5.1 Conclusion

In this study a machine learning approach is carried out using classification algorithms to carry out the irregularities affection the electricity power consumption in Nigeria. The study used efficient classification algorithms SVM and Decision tree. Decision tree was able to achieve 65% accuracy and SVM achieved 62% performance accuracy. The SVM sensitivity proves to be high and can still be considered highly efficient. The results obtain shows that the power consumption sector subject has to do more in customer reviews and satisfaction so as not lose their revenue generation.

Hence this study has helped to improve its management of power supply and revenue protection. Despite the outcome of the study is favorable, working on larger data set is vital before using the SVM classifier for prediction to be put into practice in the real world context. The concept of data mining has also been actualized during the course of the project has the case study dataset was subject to clustering techniques. The classification with Decision tree behaved quite good as it was able to reach a 65% accuracy which really shows confidence in its knowledge discovery from the dataset used.

5.2 Recommendation

It is recommended that larger dataset involving more geographical location can be considered and other classification algorithms can be used for customers management and fraud detection.

REFERENCES

1. Abdelmalik, A.A. (2014). Power Transformer Life Management; relevance to Nigerian power industry.Department of Physics Ahmadu Bello University Zaria, Nigeria.
2. Afful-Dadzie, Eric, Stephen. N., And Zuzana, K.O. (2014). Patterns and Trends in the Concept of Green Economy: A Text Mining Approach." In Modern Trends and Techniques in Computer Science, pp. 143-154. Springer International Publishing, 2014.
3. Amadi, D. (2012).The Structure of the Nigerian Electricity Industry, Major On-going Projects; Opportunities and Challenges. A paper presented to the public and private development centre, procurement monitors training, 19- 21 July, 2012.
4. Arunagiri, T. and Agarwal, O. (2005).Electricity Distribution Management System using GIS, Institute of Remote Sensing, Chennai.
5. Caliş, A., Boyaci, A., and Baynal, K. (2015). Data mining application in banking sector with clustering and classification methods. In Industrial Engineering and Operations Management (IEOM), 2015 International Conference on (pp. 1-8). IEEE.
6. Claudius, A. (2014). NIGERIA Electricity Industry: Issues, Challenges And Solutions. Media & Corporate Affairs Covenant University, Covenant University 38th Public Lecture . Vol. 3, No. 2. pp.1-40.
7. Dare-Alao, D. (2013).Geospatial Modelling of Electricity distribution network.Geospatial World Weekly 22 April 2013.
8. Degeneff, B.C.(2004). Power Transformers in The Electrical Engineering Handbook. Chen W. (Ed.), Elsevier Academic Press. 715-720.
9. Depuru, S.R., Wang, L., Devabhaktuni, V., and Nelapati, P. (2011). A hybrid neural network model and encoding technique for enhanced classification of energy consumption data, Proc. IEEE PES General Meeting, Detroit, MI, July 2011, pp. 1–8.
10. Farokhi, S., Teimourpour, B., Shekarriz, F., and Masoudi, M. (2016). A new application of clustering for segmentation of banks'e-payment services based on profitability. Uncertain Supply Chain Management, vol. 4, no. 1, pp. 55-60.
11. Farquad, M. A. H., Ravi, V., and Raju, S. B. (2014). Churn prediction using comprehensible support vector machine: An analytical CRM application. Applied Soft Computing, vol. 19, pp. 31-40. Elsevier B.V. Retrieved from http://dx.doi.org/10.1016/j.asoc.2014.01.031.
12. Gharipour, A., Sameti, M., and Yousefian, A. (2010). A Comparative Approximate Economic Behavior Analysis of Support Vector Machines and Neural Networks Models. Iranian Economic Review, Vol. 15, No. 26, pp. 17-40.
13. Hu, Y. C., and Liao, P. C. (2011). Finding critical criteria of evaluating electronic service quality of Internet banking using fuzzy multiple-criteria decision making. Applied Soft Computing, vol. 11, no. 4, pp. 3764-3770.
14. Kareem, A.(2006), Maintenance of Electricity Power System and Equipment, NSE Proc.Of the National Engineering Conference, pp 34-42.
15. Kayode, O., Benjamin, C.M., Seiichi, O. and Tetsuo. T. (2018). Estimating Residential Electricity Consumption in Nigeria to Support Energy Transitions. Mdpi journal of sustainability.

16. Khashei, M., Rezvan, M. T., Hamadani, A. Z., and Bijari, M. (2013). A bi-level neural-based fuzzy classification approach for credit scoring problems. Complexity, vol. 18, no. 6, pp. 46–57.

17. Khobzi, H., Akhondzadeh-Noughabi, E., and MinaeiBidgoli, B. (2014). A new application of rfm clustering for guild segmentation to mine the pattern of using banks's e-payment services. Journal of Global Marketing, vol. 27, no. 3, pp. 178-190.

18. Lekini, N.C., Ndzana, B., and Oumarou, H. (2017). Evaluation of Customer Behavior Irregularities in Cameroon Electricity Network using Support Vector Machine. American Journal of Engineering and Applied Sciences. Vol. 10, No.1, pp. 32-42.

19. Manga, A.(2012).Engineering department PHCN, PZ, Zaria, Kaduna State.

20. Mohammed, K.A. (2011). Welcome Address at the Summit On Energy And Sustainable Economic Growth, Energy Commission of Nigeria Organized workshop at Ladi Kwali Hall, Sheraton Hotel And Towers, Abuja, 9 - 10 March, 2011.

21. Moslehi, Haeri, and Moini (2017) Analyzing and Investigating the Use of Electronic Payment Tools in Iran using Data Mining Techniques. Journal of AI and Data Mining. Pp. 1-22.

22. Muhammed, S. (2005). Checking the Odds, NEPA Review Magazine, Vol.4

23. Musa, D.A. (2009). The Economics of Electric Power Supply in Nigeria. Retrieved at: gamji.com/article9000/news9075.htm

24. Nagi, J., K.S. Yap, S.K. Tiong and S.K. Ahmed, 2008a. Detection of abnormalities and electricity theft using genetic support vector machines. Proceedings of the IEEE Region 10th Conference, Nov. 18-21, IEEE Xplore Press, Hyderabad, India, pp: 1-6. DOI: 10.1109/TENCON.2008.4766403

25. Nagi, J., K.S. Yap, S.K. Tiong, A.M. Mohammad and S.K. Ahmed, 2008b. Non-technical loss analysis for detection of electricity theft using support vector machines. Proceedings of the 2nd IEEE International Power and Energy Conference, Dec. 1- 3, IEEE Xplore Press, Johor Bahru, Malaysia, pp: 907-912. DOI: 10.1109/PECON.2008.4762604

26. Nagi, J., K.S. Yap, S.K. Tiong, S.K. Ahmed and F. Nagi, 2010b. Improving SVM-based nontechnical loss detection in power utility using the fuzzy inference system. IEEE Trans. Power Delivery, 26: 1284-1285. DOI: 10.1109/TPWRD.2010.2055670

27. Nagi, J., K.S. Yap, S.K. Tiong, S.K. Ahmed and M. Mohamad 2010a. Nontechnical loss detection for metered customers in power utility using support vector machines. IEEE Trans. Power Delivery, 25: 1162-1171. DOI: 10.1109/TPWRD.2009.2030890

28. Nigerian Electricity Regulatory Commission. Enugu Disco Tariffs. 2017. Available online: http://nerc.gov.ng/nercdocs/tariffs/enugu-tariff.pdf (accessed on 27 February 2018).

29. Nigerian Electricity Regulatory Commission. How Much Do I Pay for Electricity? 2017. Available online: http://nerc.gov.ng/index.php/home/consumers/how-much-do-i-pay-for-electricity (accessed on 3 October 2017).

30. Nwachukwu, O.B. (2009). Consumers Expectation on Service Delivery of PHCN: A Study in Lagos and Ibadan Metropolis. Unpublished MBA Thesis, Owerri: Imo State University.

31. Odiah, R.I., (2011). Challenges surrounding power investment in Nigeria: Insights and perspectives from owners and developers, Paper presented at the Summit on Energy And Sustainable Economic Growth, Energy Commission of Nigeria Organized workshop at Ladi Kwali Hall, Sheraton Hotel And Towers, Abuja.

32. Okonedo, C. (2011), Evaluation of National Electric Power Authority (NEPA) Utility Information Infrastructure Towards Steady Power Supply by the End of 2010: A case Study of Bida NEPAll, Proceedings of the Technical Session of the 37th Annual General Conference and Meeting of Nigerian Institution of Surveyors, Owerri, Imo State, Nigeria, pp. 63-67.

CISDI Journal

Computing, Information Systems, Development Informatics & Allied Research Journal
Vol. 12 No. 1, 2021 - www.cisdijournal.net

33. Olatoye, S.O. (1979). Maintenance of Electric power system and equipment, Nigerian Society of Engineers, Proceedings of the National Engineering Conference on Engineering Maintenance (19th -21st September 1979), pp 44-77.
34. Oluseyi, P.O., Akinbulire, T.O., and Awosope, C.O.A., (2007). Solving the technical problems facing electrical energy development in Nigeria, UNILAG fair and conference, 2007 (Poster Presentation).
35. Onohaebi, O.S., and Lawal, Y.O. (2010). Poor Maintenance Culture; The Bane To Electric Power Generation In Nigeria, Journal Of Economics And Engineering, ISSN: 2078-0346, May, 2010 pp28-33.
36. Sumathi, S., and Sivanandam, S. N. (2006). Introduction to data mining principles. Introduction to Data Mining and its Applications, pp. 1-20.
37. Sunday, S.E.(2017). Electricity: Over 50 % Customers Unmetered 4yrs after Privatisation. Daily Trust Newspapers. 3 October 2017. Available online: https://www.dailytrust.com.ng/electricity-over-50-customers-unmetered- 4yrs-after-privatisation.html (accessed on 3 October 2017).
38. Wijaya, T.K., Ganu, T., Chakraborty, D., Aberer, K., Seetharam, D.P. (2014). Consumer segmentation and knowledge extraction from smart meter and survey data, in: SIAM International Conference on Data Mining (SDM14).
39. Haben, S., Singleton, C. and Grindrod, P. (2016) Analysis and clustering of residential customers energy be havioral demand using smart meter data. IEEE Transactions on Smart Grid, Vol.7,   No.   1, pp. 136-144. ISSN 19493053 doi: https://doi.org/10.1109/TSG.2015.240978.