National College of Ireland

Non-Technical Electricity Loss:

Predicting and Defining correlation of Electricity Theft Determinants Using Machine Learning Algorithms

MSc Research Project
FINTECH

Steven Kawala
X15018121

School of Computing
National College of Ireland

Supervisor: Victor Del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Steven Kawala
Student ID:	XI5018121
Programme:	MSc FINTECH **Year:** 2020
Module:	MSc RESEARCH PROJECT
Supervisor:	Victor Del Rosal
Submission Due Date:	18th August 2020
Project Title:	Non-Technical Electricity Loss: Predicting and Defining correlation of Electricity Theft Determinants Using Machine Learning Algorithms
Word Count:	6444 **Page Count** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Steven Kawala

Date: 18/08/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Non-Technical Electricity Loss:
Predicting and Defining correlation of Electricity Theft Determinants Using Machine Learning Algorithms

Steven Kawala
X15018121

Abstract

Of the issue related with non-specialized irregularities in electricity usages, different strategies have been put in place for effective administration of non-specialized peculiarities in the electricity industry. The effective and best strategy implemented so far to diminish non-specialized peculiarities and revenue losses is the utilization of Smart advanced utility meters. This strategy makes deceitful exercises increasingly difficult, and it is simple to identify when such deceitful exercise happens. However, this strategy is not extensively utilized in most countries because of the cost associated with the procurement and installation of the smart meters. This research paper looks at how well can Artificial Intelligent Algorithms be used to predict power theft considering the social and economic determinants behind electricity theft.The research proposes the use of seven models on local area power utilization in China, to improve constant precision on the recognition of nontechnical inconsistencies and save revenue loss from utility companies. The models will distinguish and anticipate malevolent power utilization in real-time and chronicled data with anomalous utilization patterns will be associated with electricity theft. All models were evaluated based not only on the accuracy of the model but also sensitivity, specificity, and the AUC results. The analysis of the results did not only look at the exactness deciding the exhibition of the models to judge the performance of the model but also looked at the proportion of right theft forecast. The analysis is successful in predicting theft of electricity and a clear comparison of the models gave a rank to a promising model that supports the researcher.

Keywords: Feature Selection, Electricity theft, Theft determinant, SVM, SOM, kNN, Logistic regression, Decision Tree, Random Forest, Naïve Bayes.

Introduction

Power extortion is a worldwide issue and it is costing utility companies a great deal of cash in revenue loss (Li et al., 2019).Commonly, the theft of power is related with, illegal electrical reroutes and meter altering (Sardar and Ahmad, 2016, Zhang et al., 2018, (McLaughlin et al., 2013). The power theft consequences, aside from income and conservative misfortune can likewise bring safety issue worries to the public for instance, causing electrical fire[1]. Even though, in the reason for power theft catastrophes occur yet the fundamental destinations for the act is not to cause fire but the aggressors need to pay not exactly the standard expense of

[1] https://www.bchydro.com/news/conservation/2011/smart

their power utilization, causing revenue loss for the utility companies. According to Jiang et al., (2014), Zhang et al., (2018), utility companies suffer two kinds of power losses namely, technical, and non-technical losses. Nontechnical losses (NTLs) starting these power thefts are the most significant worries of any electricity company (Nizar, Dong and Wang, 2008). As an example, Li et al., (2019), pointed out that the USA alone losses $4.5 billion consistently and utility companies overall lose a gauge of more than 20 billion every year in revenue because of power theft. In further research, India reported $16.2 billion loss every year and in 2012 India GDP dropped by 1.5% as a result of electricity theft(Hu, Yang, Huang and Cheng, 2020).The USA uncovered $6 billion per year economical loss because of power theft (Singh, Bose and Joshi, 2019).Canada reported CAD$100 million loss every year[2].This is one of the fundamental reasons why this domain area remains a research domain globally to save the utility companies from revenue loss. These losses affect utility companies in the way quality of power supply is maintained, increases electricity generation load and implementation of electricity tariff on sincere customers. In developing countries, power companies use field auditors to explore the pernicious use of electricity at domestic unit level. This is dull and increases overhead expense of the company.

1.1.Motivation Overview

Although electricity theft domain has been researched vigorously to improve customer safety and detect anomaly, Andrysiak, Saganowski and Kiedrowski, (2017) , it is fascinating to observe that most of the researchers have focused more on NTLs and reduction of these losses without considering the social and economic determinants behind the theft of electricity. In developing countries there are more determinant factors compared with developed countries. In developed countries, tariff, income, population, agricultural load, urbanization, and temperature are the major determinants. In developing countries, additional factors like unemployment, literacy, corruption, politics, and correction efficiency rule the customer and employee behavior towards electricity theft (Saini, 2017).

The author sees a gap from the motivation overview and wishes to explore and contribute to the knowledge of academia. Recently there has been a rise in the number of electricity theft in Asian countries, Tiong et al., (2012) which include India and China. The focus in this paper will be on China, where most of customers who steal electricity are associated with Bitcoin mining. From the knowledge of the literature reviewed, table 1 below summarizes some of the models researched in the past, where SVM outperformed the other techniques.

[2] https://www.bchydro.com/news/conservation/2011/smart

Table 1: Summary of Electricity Theft detection Techniques (Related Work)

Technique Applied	Year	Researcher
ANN	1998	Galvan et al., (1998)
Statistical Based Outlier	2002	Bolton and Hand, (2002)
Rough Sets	2004	Cabral, Pinto, Gontijo and Filho, (2004)
Decision Tree	2004	Cabral, Pinto, Gontijo and Filho, (2004)
KDD	2006	Nizar, Dong and Zhao, (2006)
ELM	2008	Nizar, Dong and Wang, (2008)
SVM	2009	Marrakchi, Agina and Elghali, (2009)
SVM	2010	Nagi et al., 2010)
OS-ELM & BPNN	2012	Tiong et al., 2012)
LTSM & CNN	2019	Hasan et al., (2019)

The exploration of this paper analyzes the feasibility of applying different algorithms to predict electricity theft and define the electricity theft determinants in China. This will be done by identifying non-specialized irregularities. The approach includes the use of, Decision tree, Random forest, Naïve Bayes, SVM, kNN, Logistic Regression and Kohonen artificial neural Network(KANN) predictive models for the evaluation of the abnormal consumption patterns to determine continuous exactness on the distinguishing proof of a nontechnical irregularity.

Even though SVM and Random forest strategies radiate an impression of being creative, a different school of thought suggest that there is yet a necessity for more exploration to be carried out for the results of the frameworks to be in every way trusted. These contributory conflicts make this space an area more interesting for extra examination. The study helps future researchers to apprehend and extend the knowledge of building a robust predictive model that will withstand huge volume of dataset companies are generating in this digital era. The research has contributed by:

1) Use of public data to conduct a set of experiments to predict electricity theft and compare how each model performs under the same parameters.
2) Use of geographical region to define the major determinants behind electricity theft.

1.2. Research Question

The proposed research question is picked subject to what Jokar, Arianpoo and Leung, (2016) recently contributed to literature and further supported by Singh, Bose, and Joshi, (2019) as a locale for exploration. The research question for this paper looks at:

How well can Artificial Intelligent Algorithms be used to predict power theft considering the social and economic determinants behind electricity theft?

The focus of the research will have six areas to examine and each area is partitioned into smaller realms as indicated in figure 1 below:

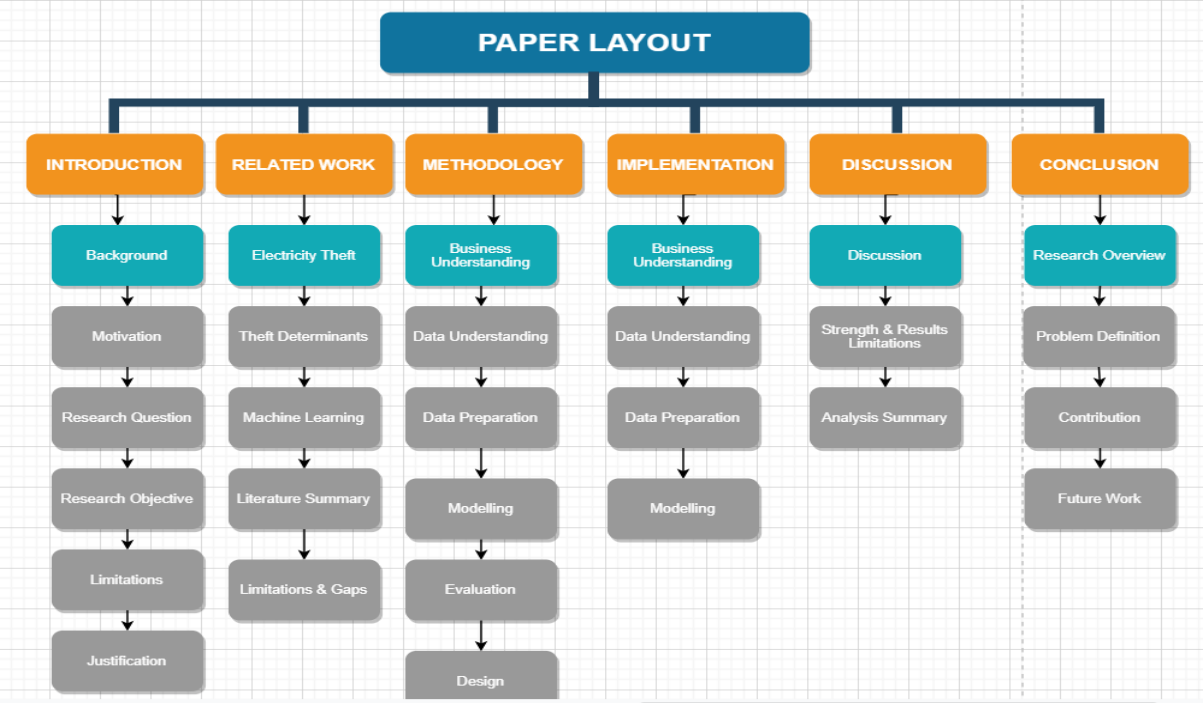


Figure 1: Paper layout (Source: Author's own)

1.3. Research Paper Layout

- Related Work: This area will create knowledge and contribute to the previous researchers in the electricity theft domain and help the author of the research to establish the research gap and form the nitty-gritty of the research question.
- Methodology: This section covers all CRISP-DM steps for research including how data has been prepared and the statistical techniques being used in the paper.
- Implementation and evaluation: After identifying the research methodology, this area will describe in detail all methodology steps including a description of how the models are developed and evaluation of each model results.
- Conclusion and future work summarize the state of art success, problem definition and a contribution of knowledge for future research.

1.4. Objective of the Research:

The objective of the investigation attempts to react to the proposed examination question above in two sections:

- To develop a model that can recognize and envision noxious power usage constantly, mainly to suggest a novel model that can look at the limitations of already tried and researched models in this domain (SVM, Random Forest, ANN)
- To define the social and economic determinants behind the electricity theft and predicate results of the previous researches in the electricity domain.

1.5. Research Limitations

The author feels necessary at this point to put down the clampdowns of the research in determining the success of the research:

- Timeline set up for completion of academic research projects.
- The limitation of the research content fueled by word count restriction (6000).
- Use of small dataset size, 21684 obs by 1024 variables due to limitation on open source dataset in the electricity domain.

1.6. Justification

Most of utility companies are still hooked up in rule-based systems in detecting and predicting theft and the author of the research thinks, it is high time that utility companies should fully embrace the machine learning based solutions considering the following:

- The rule-based approach cannot detect correlations and it is hard to process constant information streams that are basic for the advanced domain.
- Machine learning solutions can work on huge datasets with numerous factors in trying to discover shrouded verifiable connections in information and the probability of false activities.
- Most interestingly, machine learning solution-based system has faster data processing power because of streamlined steps[3]

2.Related Work

2.1. Electricity Theft

According to Seger, (2005) electricity theft is categorized into honest theft and dishonest theft. Whatever category theft falls into, it is not acceptable. However in Pakistan power manager have come to the point of accepting power theft as the norm of the business because of the magnitude of the practice(Smith, 2004).In order to facilitate the detection of electricity theft, an international association has been established(International Utilities Revenue Protection Association) to protect the electricity utility companies(Smith, 2004).Table 2 below pictures the NTLs summary per year of selected countries.

Country	Non-Technical Loss/Year
Jamaica	$46 Million -
India	$17 Billion -
USA	$6 Billion/year +
United Kingdom	$173 million +
Canada	$100 million -

+ ***Cost*** ***-Loss***

Table 2: Yearly Electricity NTLs (Alazab, 2019)

2.2. Determinants

Different researchers have used different methods to show the relationship between electricity theft and its determinants. Table 3 below summarises the results of the researches.

[3] https://bigdata-madesimple.com/top-5-free-data-mining-tools-to-try-for-your-business/

Study Method	Year	Researcher	Findings
Correlations Analysis	2004	T.B Smith	Time
3 Stage Least Square method	2015	C.Yurtseven	Population, price, temperature Agricultural production
FGLS model	2016	Gaur and Gupta	Poverty, corruption, urbanization, populism
Survey	2018	O. Yakubu, C. N. Babu & O. Adjei	Price, corruption, poor governance, poor quality
Theoretical model	2019	Jamil and Ahmad	Benefits of stealing
Random Forest model	2019	Razavi and Leury	Crime rate, electricity consumption per capita

Table 3: Determinants Source**:** Briseño and Rojas, (2020)

Briseño and Rojas, (2020) pointed out that econometric models provide evidence of the impact of socio-economic factors on the electricity theft. The results of the models confirm that population, price, and temperature were the significant factors that drive the theft of electricity. In the same school of thought Gaur and Gupta, (2016) added other determinants which were associated with electricity theft in India like, poverty, urbanization, and corruption.

In his contribution to the methods listed in the earlier sections above, Briseño and Rojas, (2020) identified the determinants by using least square method which revealed that an increase of determinant variable for example, population, increases the theft of electricity. There is an assumption that other variables in the experiment remained the same for the population to have an impact on the theft of electricity. Using population for example as a determinant, Briseño and Rojas, (2020) assumed that every persona in the population used the electricity without limitation of the demographics of the population.

2.3. Machine Learning

Apart from determinants of electricity theft, researchers have shown keen interest in trying different models to detect electricity theft. In this domain Yeckle and Tang, (2018) contributed the use of Outlier Detection and k means clustering Algorithms as methods of detecting electricity theft in customer consumption. His novel contribution was on the generation of types of electricity theft on which the models were tested and compared their performance to enhance the security of Advanced Metering Infrastructure. During this experiment, the density of the data instances was taken care off by using k-nearest neighbour.

Yeckle and Tang, (2018)'s experiment did not represent a fair amount of the population of Irish homes and businesses who had access to electricity. The dataset only had 90 instances by 24 components. Based on the results obtained from the experiment, Yeckle and Tang, (2018) observed that the use of outlier methods produces excellent results in the detection of electricity thefts and could be used to enhance the security of advanced metering infrastructure.

Depuru, Wang and Devabhaktuni, (2017) proposed utilization of SVM data classifier for power theft detection. Jokar, Arianpoo and Leung, (2016) used SVM classifier to distinguish anomalous practices and power theft attacks with zero usage results. However, Depuru, Wang and Devabhaktuni, (2017) preposition suggested that, geographical locations, annual cycle and customers classification should be considered when choosing data to train SVM model.

Depuru, Wang and Devabhaktuni, (2017) experiment results indicated that one classified group of customers were prone to theft, but the evaluation of the models was not clear because of the style of the presentation of the results. Using raw dataset for classification problems can be challenged with class imbalance when it comes to evaluating the accuracy of the model. This common problem was not clearly addressed by Depuru, Wang and Devabhaktuni, (2017). However, Hasan et al., (2019) proposed the generation of synthetic data to address class imbalance problem. Another area which is not clearly reflected in Depuru, Wang and Devabhaktuni, (2017) experiment is the data mining technique carried out before the application of the models. It is important for a data scientist to carry out data pre-processing as the first step of the analysis to avoid bias in the results. We can assume that the dataset Depuru, (2017) used had no missing gaps contrary to the behaviour of raw data which is always noisy and inconsistent. Hasan et al., (2019) proposed the generation of synthetic data to address class imbalance problem.

However, Hasan et al., (2019) did a comparison of SVM performance with his proposal combining convolutional neural network (CNN) and long short-term memory (LSTM) architecture. According to Hasan et al., (2019), CNN technique automates feature extraction and the classification process and uses human visual cortex for object recognition. It was evident that the use of synthetic dataset improved CCN classification of fraud users. Compared with other methods, SVM and Logistic regression, CNN had better precision, recall and accuracy (Hasan et al., 2019). In another comparison analysis, Wang and Ahn, (2020) used closest neighbor strategy, neural network and SVM. In this experiment closest neighbor calculations outflanked the two other techniques. Although SVM was not proficient in irregularity location and had low identification accuracy, still it can be used to solve any complex issues if kernel function is used correctly. According to Zhang et al., (2019) the performance of a classification technique relies on elimination of unnecessary features through application of optimisation calculations,

2.3. Literature Gaps

The literature reviewed shows that SVM has been a popular successful technique that has outperformed on average other techniques. SVM proved to be a reliable technique of detecting electricity theft (Ahmed et al., 2008). SVM has nonlinear partitioning hypersurfaces that give it high separation and it gives great speculation capacity to concealed information characterization (Ahmed et al., 2008). However, Alazab, (2019) described the SVM technique as shallow and Zhang et al., (2018) had the same school of thought that the techniques had low accuracy in determining electricity theft .Recent study has shown that most of the electricity theft detective algorithms produce poor performance because of the assorted variety and power utilization behavior inconsistencies, which are practically difficult to completely see, by NTL utilization only(Hu, Yang, Huang and Cheng, 2020).

To explore the gaps the paper proposes a feasibility application of predictive algorithms of electricity theft using correlated electricity determinants variables through use of Supervised Machine Learning Algorithms. The author wishes to use already existing algorithms to evaluate their performance and compare the current results with the previous results as an attempt to answer the research question.

3.Research Methodology

The methodology adopts CRISP-DM approach of data mining described by Chapman et al., (2000) as opposed to KDD and SEMMA. The former (CRISP-DM) covers, understanding of the business, understanding of the dataset, data preparation, modelling, evaluation, and deployment. Figure 2 illustrate the CRISP-DM approach and likewise figure 3 illustrate the KDD approach.

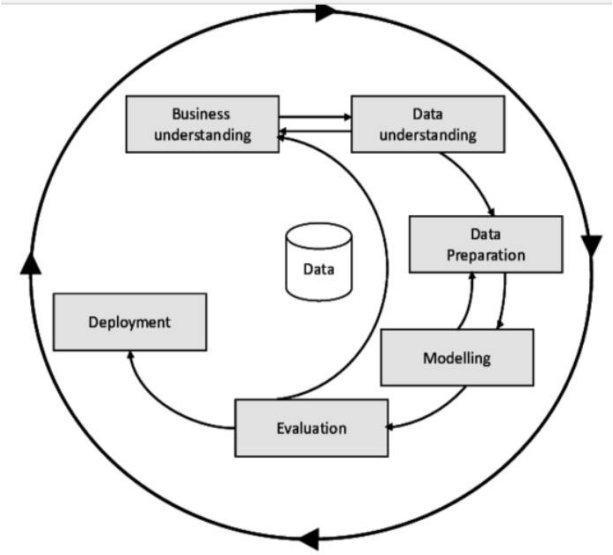


Figure 2:CRISP-DM sourced from(Chapman et al., 2000)

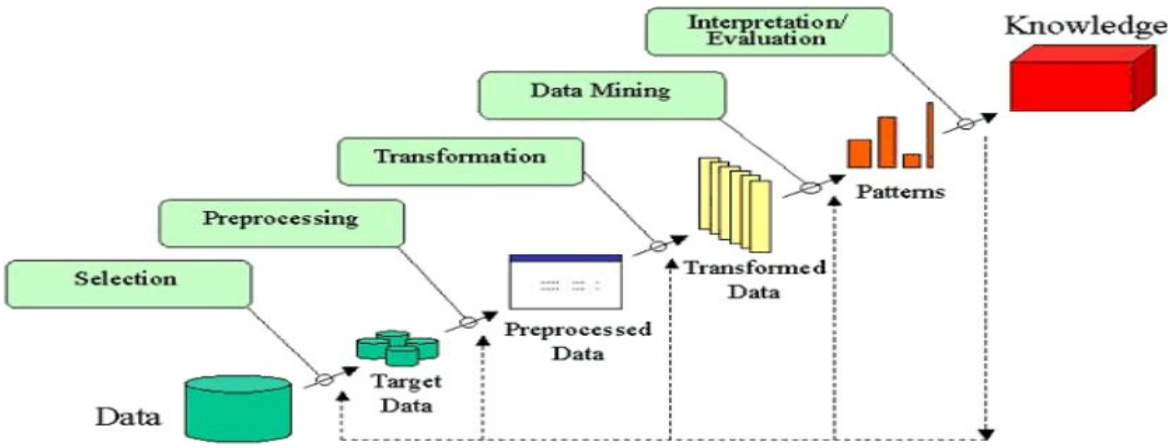


Figure 3: KDD methodology (Azevedo and Santos, 2008)

Table 4 below shows a detailed overview of the tasks to be carried out on each stage of CRISP-DM

CRISP-DM TASKS

usiness Understanding	Data Understanding	Data Preperation	Modeling	Evaluation	Deployment
Determine Business Objectives Background Business objectives Success criteria **Determining Data Mining Goals** Data mining success criteria **Techniques** Assessment of Techniques.	**Collect Data** Initial collecting of data from source **Data Description** Describing data and identifying attributes **Explore Data** Data exploration- Self Organising Map and histograms **Data quality** Verifying quality of data	**Data Set** Data Set Description **Select Data** Rationale for inclusion and exclusion-Feature selection. **Clean Data** Data cleaning-missing data **Data Normalisation** Scale values to 0 and 1 **Data split 70:30**	**Select a modeling technique.** Modeling Technique and assumptions **Build Model** Model description Parameter settings in R **Assess Model** Model assessment	**Evaluate Results** Assessment of data mining results **Review Process** Review of process	**Review Project** There will be no experience documentation

Table 4: CRISP-DM TASKS, source: modified original from (Wirth and Hipp, 2000)

The entire process of model development will be done in 'R' language. 'R' language is the language that combines S3, S4 and R5 OOP system.'R' has been chosen in this research in comparison with the other open source data analytics tools(Orange and Knime) simply because it is free and has a lot of statistical analysis packages ready to use[4].

3.1. Business understanding

The scope of the research focuses on electricity theft and the determinants behind the theft. To explore the pain points of the problem being investigated, a research question has been formulated: predicting electricity theft using Supervised Machine Learning Algorithms and defining correlation of electricity theft determinants. Through machine learning techniques,

[4] https://bigdata-madesimple.com/top-5-free-data-mining-tools-to-try-for-your-business/

six models will be trained on the same dataset and accuracy rate will be one of the performance determination measurement for all the models.
In this research, Business Understanding will cover the following scope:

- Business objectives
- Data mining goals
- Project planning.

3.2. Data Understanding

The data set is obtained from SGCC (Sgcc.com.cn ,2018). The dataset does not reveal any knowledge at this Data Extraction and Analysis stage (Idreos, Papaemmanouil and Chaudhuri, 2015). The Data Understanding stage will get an insight of the variables and their relationships in an unstructured manner and also get the areas that will be interesting to investigate further (Martinez, Martinez and Solka, 2010).In order to get the relevant data to focus on, the research will use both manual and automated methods like graphs and then add graphical summary for visualization**.** The full scope of this section will cover:

- Collection of initial data
- Data description
- Data exploration

3.3. Data Preparation

Data preparation stage involves transformation of data for example, making sure that there is one observation per row and one variable per column. This will be done through data wrangling, feature selection and data normalization. The result of this exercise brings a training dataset for the models. This stage is important in data analysis because ,it will dissect the dataset to a level that can be controlled easily.There after the data is split into a ratio of 70:30. The former being training set 70% and the latter being test set 30%.

3.3.1. Instance Selection

In machine learning problems it is an automatic requirement to classify instance (López, Carrasco, Martínez and Kittler, 2010).

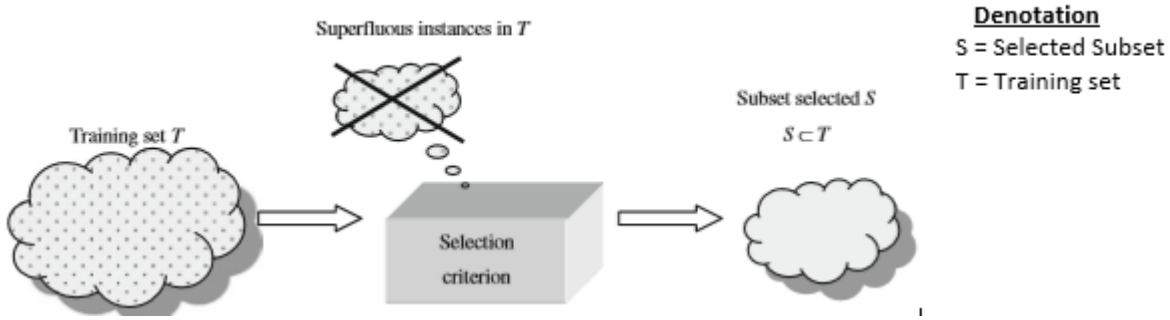


Figure 4: Instance selection Source: (López, Carrasco, Martínez and Kittler, 2010)

Figure 4 demonstrates how the classifier gets '**T**'- training set by using supervised classification to be given a class. The whole reason for this selection process is to get S ⊂ T so that S has no obsolete instances.

3.3.2. Data Normalization

This is a pre-processing phase where dynamic range of feature values are ranged into specific range. This step will be carried out before building any model to make sure that feature values are on the same scale by rescaling the numeric variables within the range of 0 to 1 (Patro and Sahu; 2015).Standard score will be used to convert scale of all the parameters with zero mean using the formula:

$$z_i = \frac{x_i - \overline{x}}{S} \qquad \text{.....(3.1)}$$

In this formula mean sample is represented by $\bar{x}$ and standard deviation of the sample is represented as 'S'.

4.Modeling

Naïve Bayes is a quite popular and widely used method Viaene, Derrig and Dedene, (2004), known to provide good performance with categorical variables than numerical variables and small training data. It gives a prediction of the likelihood of a place that a given data point has within a specific class. This technique works well in practice and is adaptable (Bhowmik, 2011). The biggest flow of this technique is on the problem of making impracticable assumptions which might be unattainable practically (Kaviani and Dhotre, 2017).

$$c^* = \underset{c}{\operatorname{argmax}}\, Pr(c \mid x_1, \ldots, x_m). \qquad \text{.....(3.2)}$$

Denotation

Attribute Values = $(x_1, x_2, \cdots x_n)$

Classification Variable = C

Value of C = c

Figure 5:Naïve Bayes Classifier,Source: Kaviani and Dhotre, (2017)

The goal for Naïve Bayes Classifier,is to develop a classifier with class label of the training set.The class label is selected by labeling new instances (Kaviani and Dhotre, 2017).
Random Forest – Breiman, (2001) has categorised random forest under supervised learning algorithm. This technique is known to provide high accuracy level and suitable for classification and regression problems. However, it cannot be recommended where relationship descriptions are of essence and delays real time predictions (Donges,2019).

10-fold cross validation will be used to train the model for improved prediction on accuracy. This translate that the evaluation and training of the model will be done 10 times and all iterations will be used to determine the accuracy.

5.Evaluation

This section describes some of the limitations of the previously researched models as described in related work section. Zheng et al., (2018) described the following limitations currently challenging the predictive electricity models:

1)Specificity of devices required.

2)Low accuracy detection.

3)Manual involvement on feature extraction.

To see if some of these limitations have been addressed, a measure of accuracy level is done on each model using confusion matrix. Confusion matrix will evaluate the performance of the models and calculate measurable factors which will aid comparison of classification accuracy (Table 5).

	Positive	Negative
Positive	**TP**	**FP**
Negative	**FN**	**TN**

Table 5: Predictive Matrix

5.1. Process Review

The figure 7 pictures a complete architecture of the whole process from extracting data from SGCC repository passing all stages to the model results.

The process started with dataset description including where it has been sourced and identification of variables. Any issues on the unprocessed dataset has been managed through data pre-processing techniques. which involved data feature selection, data normalization and data mapping before developing the models.

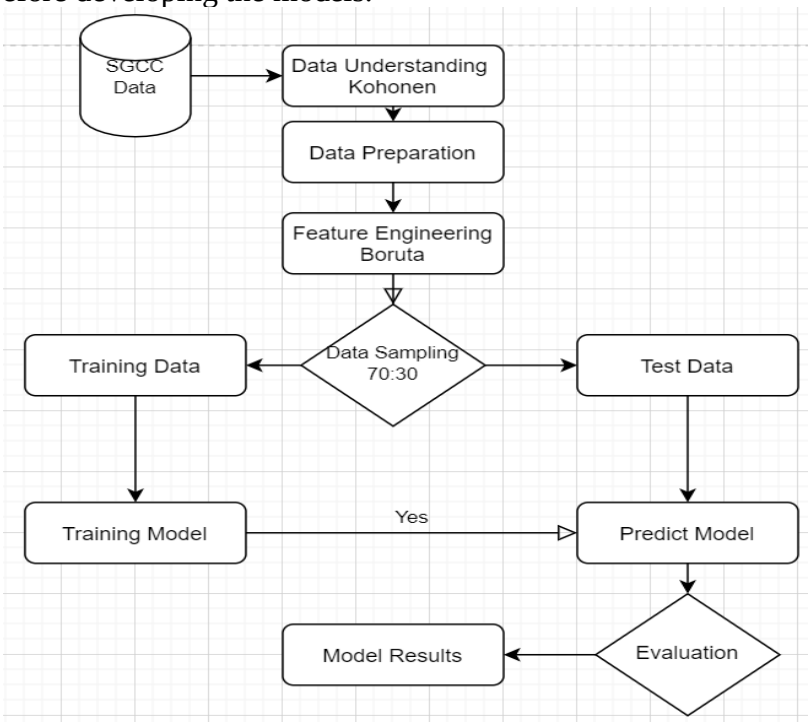


Figure 7: Process Architecture (Own source)

6.Deployment

A summary report is generated through confusion matrix to show a comparison of the model focusing on the sensitivity, specificity, and accuracy rate. There will be no further action needed after the model with high performance is selected.

7.Implementation

The CRISP-DM methodology was chosen as opposed to KDD and SEMMA to describe the process model hierarchic to answer the research question. To this regard CRISP-DM is regarded as well structured that makes it easy to understand this research. Table 6 below summarises the different stages of each method.

CRISP-DM	KDD	SEMMA
Business Understanding	Pre KDD	n/a
Data Understanding	Selection	Sample
	Pre processing	Explore
Data preparation	Transformation	Modify
Modeling	Data Mining	Model
Evaluation	Evaluation	Assessment
Deployment	Post KDD	n/a

Table 6: Comparative Summary of stages

7.1. Business Understanding

The essential objective here as delineated in the past part is to guarantee that the objectives are met by the stages in relation to the data mining goals for our business case. The response variable has partitioned our dataset into two classes, fraudulent and genuine electricity consumption with the view of determining anomaly. The data will be examined to identify human interpretable patterns that will help to describe the behavior of the dataset. The success of the interpretation will depend on the following:

1. Dataset exploratory analysis
2. Data cleaning and missing value imputation.
3. Training of all the models.
4. Score the results significance to the research.
5. Interpret the results of the models to answer the research question

7.2. Data Understanding

The data gathered from State Grid Corporation of China was made available in anonymised files containing 42,372 observations against 1,035 variables covering from 01/01/2014 to 31/10/2016.Only one period is examined (01/01/2016 to 31/10/2016) containing 21,684 observations and 1,024 variables. The dataset shows electricity consumption in kWh per day.

The table 7 below. summarizes the meta data of the dataset.

Table 7: Meta Data

Variable	Behaviour	Description
Meter ID	Numeric	Unique Meter identifier number
Risk Class	Categorical	Dependent variable 1= Risk, 0=No Risk
Normal customers	n/a	19732
Suspected Customers	n/a	1952
Date(interval)	Continuous	Consumption from Jan. 1, 2014 to Oct. 31, 2016 in kWh.

The flagged class is dependent variable identified by ‘0’ or ‘1’. Fraud risk is ‘1’and no risk is ‘0’.
Table 8 gives a summary of descriptive statistics of mean, standard deviation, maximum value, minimum value, and quartile range of each numeric variable.

no. unique	% missing	min	first quartile	median	third quartile	max	mean	sd
6525	0	-3.63829	-0.682168	-0.0001436272	0.6618652	3.812359	-0.009488019	1.004048

Table 8: Descriptive Statistics Summary

The distribution of the data is normal as the mean is close to ‘0’ and the standard deviation is 1 (1.0040488). Figure 8 pictures the distribution of the dataset.

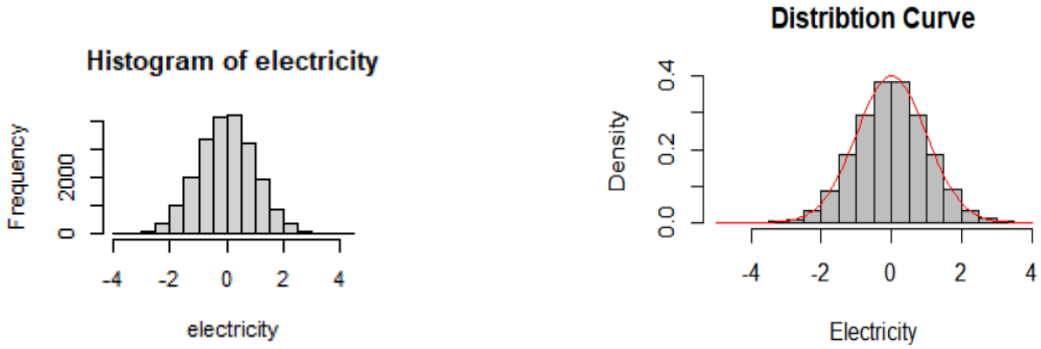


Figure 8:Standard Normal Distribution

After data extraction, the dataset is analyzed to explore the missing values. The missing data was dealt with, using functionality in “R” (***na. omit),*** to omit all missing values. The missing values were identified in the dataset as “0s”and blanks that resulted from uncompletion of the exercise by the participating customers due to different reasons. Some of the reasons for discontinuation from customer side included, customer relocation and death of the participating household owner. This process was part of the data cleaning process where only data from participants who completed the whole survey was used for the analysis. All repetition of customers, customers do not present for the full research period and customers who joined after the exercise had already started were removed from the dataset leaving only 4969 observation out of 21,684 identified as eligible for the research. The missing values were not replaced by mean of the variables because they were related to a meter identifier not variable. Therefore, by replacing them with zero or mean would have destroyed the credibility of the electricity consumption.

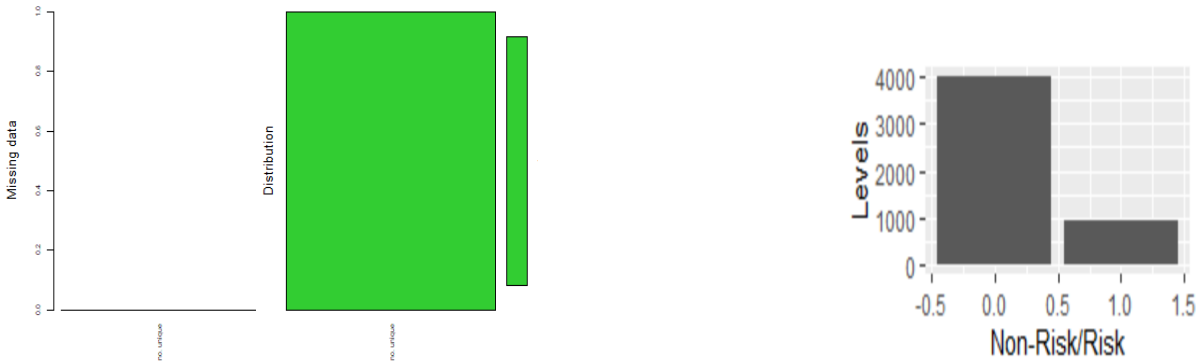


Figure 9: Missing Values Visualisation AND Variable Levels

The varable levels show the data inbalance which is dealt with before model development (Figure 9)

7.3. Data Preparation

7.3.1. Data normalisation

The reason why the data is meant to be normalised is to scale the values to "0" and "1" so that there are no extreme low or high values when using self-organising maps for visualisation of the dataset.

Boruta was selected for feature selection and SOM was chosen for its ability to visualise data in 2D and mapping the data points. The models trained include, Naïve Bayes, Random forest, Decision tree, kNN, Kohonen artificial neural network, Support Vector Machine (SVM). A summary of results for each technique is analysed in evaluation section.

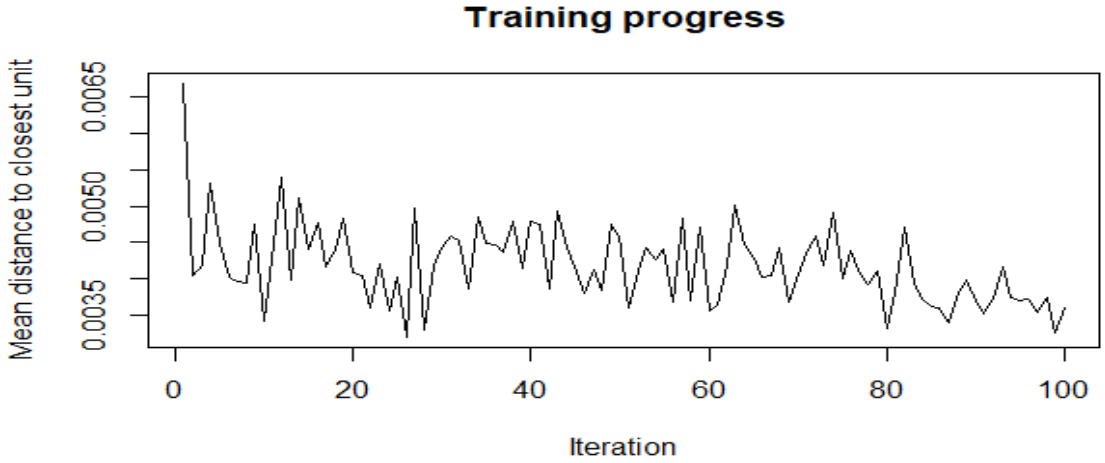


Figure 10: Data Visualisation using Kohonen artificial neural Network

The highest mean distance (MD) to closest node is 0.0065 and default iteration is 100.However the highest point MD flactuates and drops down to 20 indicating that even if the iteration is reduced from 100 to 50 the output will be the same.

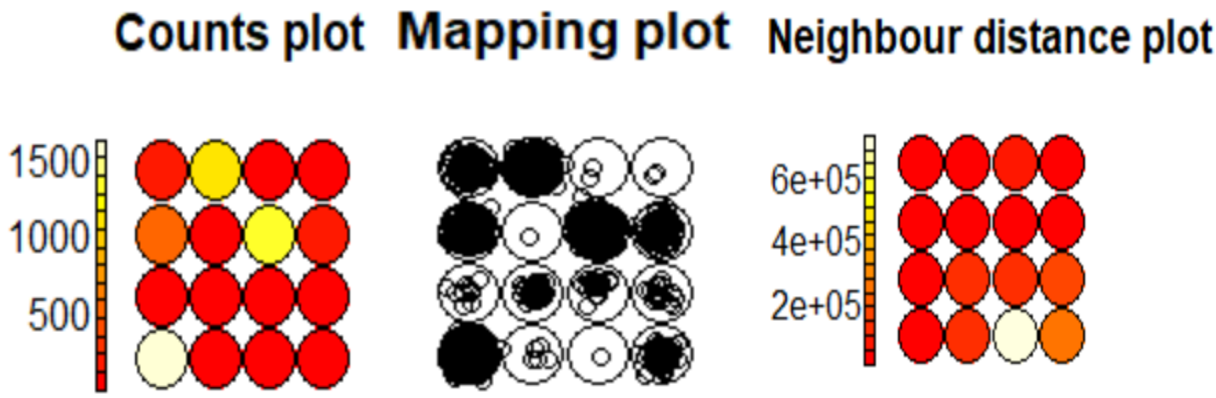


Figure 11: Data points visualization(SOM)

From the count plot it shows that only one node has the highest count of data points (white circle=1500 data points).It is also shown in the mapping plot the number of points in each node.The ducker the node the more points it has.The distance between neighbours for the lighter red circles(cream white=6e+05 distance) in the neighbour distance plot is bigger than in the bright red circles.

7.3.2.Feature Selection

As part of data cleaning methods, Boruta feature selection technique was used, and after 99 iterations 289 attributes were confirmed important,3 tentative attributes were left, and 1 attribute confirmed unimportant (Meter ID). Only unimportant attribute was dropped. Figure 16 displays the level of importance of the variables.

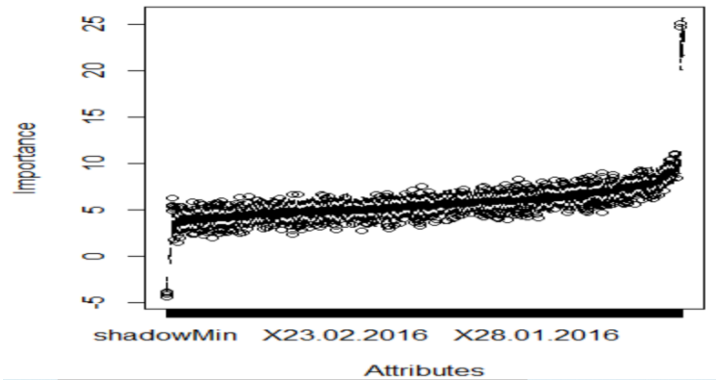


Table 16: Boruta plot

7.3.3.Data Split

Data is splitted into 70% training set and 30% testing set for all the models: Decision tree, Random forest, Naïve Bayes, SVM, kNN and Kohonen artificial neural Network(KANN).

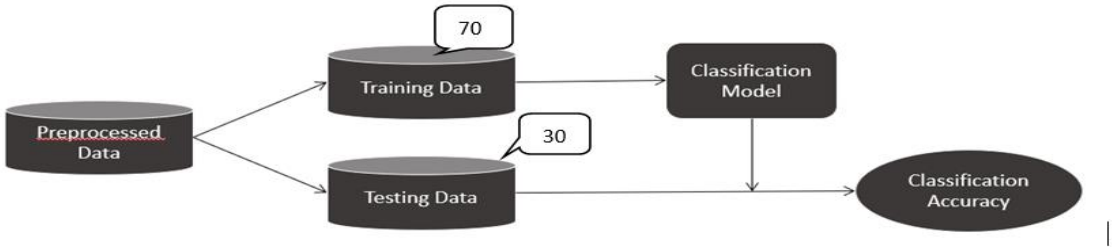


Figure 12:Data Split 70:30

The model will be fitted on the training dataset, implicitly to minimize error. If the fitted model provides a good prediction on the training dataset, then the model is tested on test dataset. If the model predicts good on the test dataset, then the level of confidence on the model will be high.

7.3.4.Evaluation and Results Analysis

Evaluation is focused on the six models; Decision tree, Random forest, Naïve Bayes, SVM, kNN and Kohonen artificial neural Network(KANN).The mathematical calculation for performance of each model were evaluated using the confusion matrics shown below in figure 13.

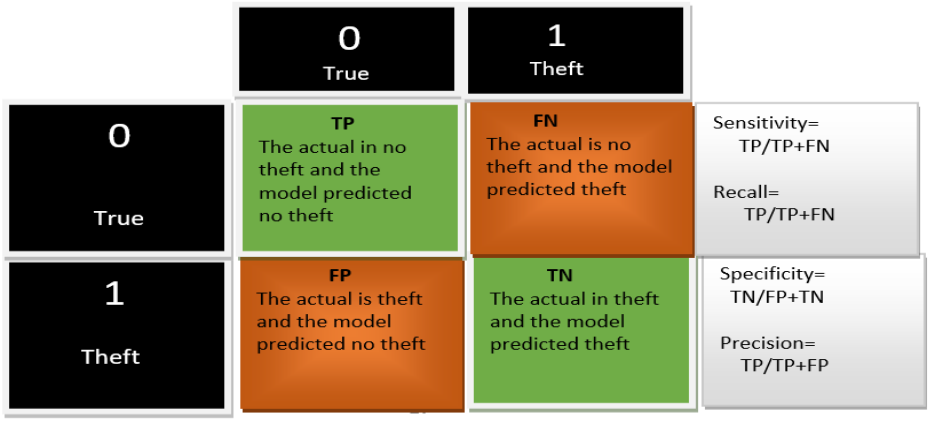


Figure 13: Confusion Matrix Evaluation

Based on our business case sensitivity can not be compromised.It will be dangerous to predict non theft when infact there is theft than predicting theft where there is no theft.The focus is on the true negatives(detected theft).

8.Models

8.1.Method 1 KANN

The first method test KANN performance. Table 9 summarizes the performance based on 1476 instances. The model had accuracy of 76% and achieved sensitivity of 82% in predicting 1071 out of 1476 observations as genuine consumption and mistakenly identified 122 as theft, almost half of the actual theft figure which is dangerous according to the business case of this study and specificity of 27% as electricity theft. The model failed to reject the null Hypothesis.

	Actual	
Predicted	0	1
0	1071	237
1	122	46

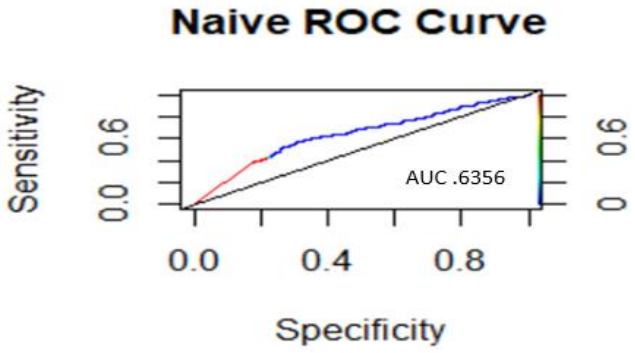
SOM Model Performance	Rate%
Accuracy	76%
Sensitivity	82%
Specificity	27%
Recall	82%

Table 9:SOM Performance Evaluation

The high recall rate of 82% indicates that the model was complete and the majority of electricity theft were identified.

8.2.Method 2 Naïve Bayes

The second method tests Naïve Bayes performance. Table 10 summarizes the performance based on 1476 instances. The model had accuracy of 71% and achieved sensitivity of 77% in predicting 926 out of 1476 observations as genuine consumption and specificity of 43% as electricity theft. The high recall rate of 77% is lower than KANN but it is not bad interms of indicating the completeness of the model and identification of the majority of electricity theft .267 is the number of mistaken evaluation of theft by the model compared to the actual figure of 161.This is good, based on the business case than identifying less than the actual theft.



Confusion Matrix

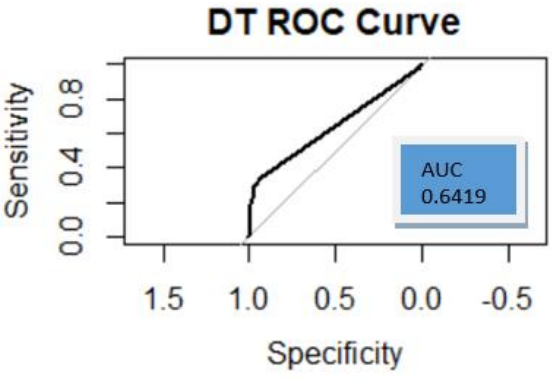
Actual_pred	0	1
0	926	161
1	267	122

Naïve Model Performance	Rate %
Accuracy	71%
Sensitivity	77%
Specificity	43%
Recall	77%

Table 10:Naïve Bayes Performance Evaluation

8.3.Method 3 Decision Tree

The third method tests Decision Tree performance. Table 11 summarizes the performance based on 1476 instances. The model had accuracy of 95% and achieved sensitivity of 95% in predicting 1135 out of 1476 observations as genuine consumption and specificity of 31% as electricity theft. The high recall rate of 95% is close to 100 indicating the highest completion percentage of the model and almost every theft has been identified by the model. The model though failed to reject the null Hypothesis, by mistakenly identifying 58 cases as theft.



Confusion Matrix

Prediction	Reference 0	1
0	1135	193
1	58	90

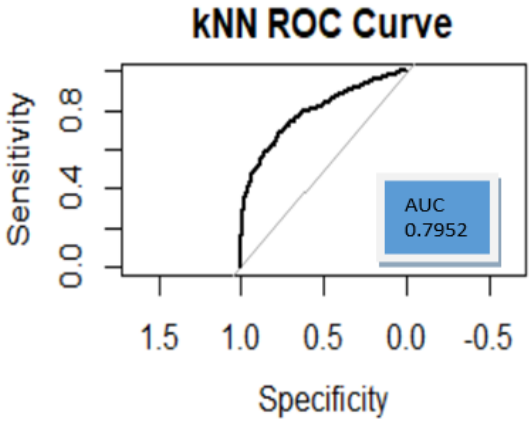
Decision Tree Performance	Rate%
Accuracy	95%
Sensitivity	95%
Specificity	31%
Recall	95%

Table 11:Decision Tree Performance Evaluation

8.4.Method 4 kNN

The fourth method tests kNN performance. Table 12 summarizes the performance based on 1476 instances. The model had accuracy of 86% and achieved sensitivity of 97% in predicting 1163 out of 1476 observations as cleared consumptions (genuine consumption) and specificity of 37% as electricity theft. The high recall rate of 97% is close to 100

indicating the highest completion percentage of the model and almost every theft has been identified by the model. The model failure to reject the null Hypothesis is not good indication. Only 30 cases are mistakenly identified as theft.



Confusion Matrix

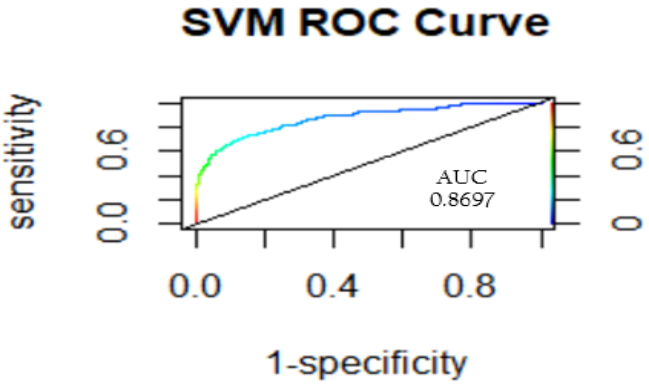
	Reference	
Prediction	0	1
0	1163	178
1	30	105

K-fold Performance	Rate%
Accuracy	86%
Sensitivity	97%
Specificity	37%
Recall	97%

Table 12:KNN Performance Evaluation

8.5.Method 5 SVM

The fifth method tests SVM performance. Table13 summarizes the performance based on 1476 instances. The model had accuracy of 83% and achieved sensitivity of 98% in predicting 1170 out of 1476 observations as cleared consumption (genuine consumption) and specificity of 18% as electricity theft. The high recall rate of 98% is close to 100 indicating the highest completion percentage of the model and almost every theft has been identified by the model. The model failed to reject the null h Hypothesis only 23cases are mistakenly picked as theft.



Confusion Matrix

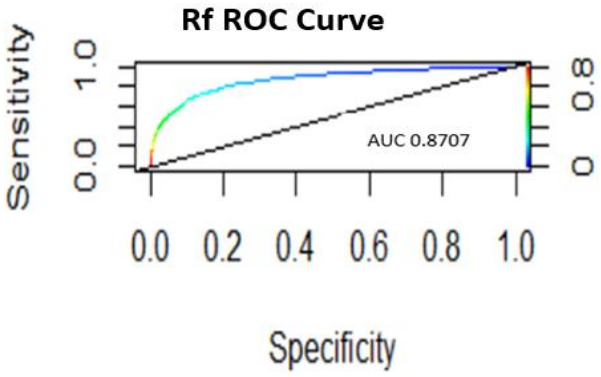
test_pred	0	1
0	1170	232
1	23	51

SVM Performance	Rate%
Accuracy	83%
Sensitivity	98%
Specificity	18%
Recall	98%

Table 13:SVM Performance Evaluation

8.6. Method 6 Random Forest

The sixth method tests Random Forest performance. Table 14 summarizes the performance based on 1476 instances. The model had accuracy of 86% and achieved sensitivity of 97% in predicting 1158 out of 1476 observations as cleared consumptions (genuine consumption) and specificity of 42% as electricity theft. However 23 cases were mistakenly identified as theft which is not correct.The high recall rate of 97% is close to 100 indicating the highest completion percentage of the model and almost every theft has been identified by the model.



Confusion Matrix

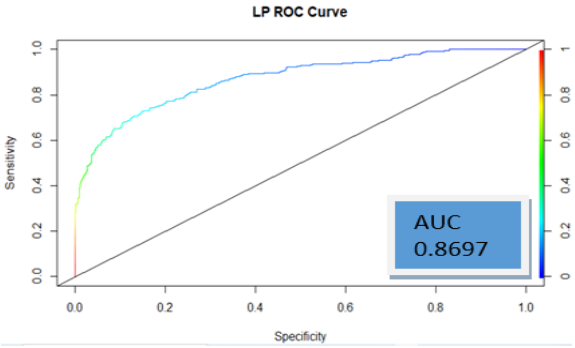
	Reference	
Prediction	0	1
0	1158	165
1	35	118

RF Performance	Rate%
Accuracy	86%
Sensitivity	97%
Specificity	42%
Recall	97%

Table 14: Random Forest Performance Evaluation

8.7. Method 7 Logistic Regression

The seventh method tests Logistic Regression performance. Table 15 summarizes the performance based on 1476 instances. The model had accuracy of 88% and achieved sensitivity of 89% in predicting 1162 out of 1476 observations as genuine consumption and specificity of 19% as electricity theft.



Confusion Matrix

p	0	1
0	1162	150
1	31	133

Logistic Regression Performance	Rate%
Accuracy	88%
Sensitivity	89%
Specificity	19%
Recall	89%

Table 15:Logistic Regression Performance Evaluation

SVM has been rated as reliable model for electricity theft detection in previous researches, however in contrast this research has shown that SVM performance is low compared to Decision tree and Random forest. These outcomes bolster the hypothesis in the way of thinking of the authors where the research question is persuaded from, - the limitations SVM has in predicting electricity theft. For better comparison and evaluation of the results, few more additional techniques have been looked at. Table 16 summarises the results of all models trained.

Statistical Measurement	Decision Tree	Random Forest	Naïve Bayes	SVM	kNN	Logistic Regression	SOM
Accuracy	95%	86%	71%	83%	86%	88%	76%
Sensitivity	95%	97%	77%	98%	97%	89%	82%
Specificity	31%	42%	43%	18%	37%	19%	27%
Recall	95%	97%	77%	98%	97%	89%	82%
AUC	0.6419	0.8707	0.6356	0.8697	0.7952	0.8697	n/a

Table :16 Performance Result summary

Based on the results from each model tested in this research , it is not just the exactness deciding the exhibition of the models yet in addition the proportion of right theft forecast is considered, similar to affectability of the models which shows how well the model has distinguished the real theft against the absolute guaranteed theft. While explicitness has indicated the genuine negatives of how well the models have recognized the reasonable cases. In theory high area under curve (AUC) indicates the correctness of the accuracy of the model and more chances of identifying the true positives. The results of the 7 experiments show that despite Decision Tree having higher accuracy rate than every other model, it has low AUC results. This indicates that Decision tree has low chances of identifying the true positives and less correct accuracy prediction.

As apparent from Table 16, the results of Naïve Bayes came rearward in examination with the other methods. Although kNN did deliver higher precision it appears to gain nothing from the training data. Be that as it may, the prescient estimation of the experiments of all the 7 models in this exploration has been dictated by affectability and explicitness and by the commonness of the condition explicitly for the business case under examination. Performance of a suitable model for this research has high accuracy, sensitivity, and AUC. Therefore, in the authors 'opinion, random forest will be the suitable technique for predicting the electricity theft on this dataset.

9.Discussion

The performance evaluation based on the accuracy shows that Decision tree was the best fit for the prediction of electricity theft in this research. In contrast the research performance evaluation further looked at accuracy and area under curve of all the trained models, which has revealed that random forest has promising results. However, there are these limitations to consider when choosing Random Forest as a promising methodology.

There is no unwavering quality on variable significance score on categorical variables with various levels when using Random Forest, albeit partial permutation appears to determine the issue. Random forest tends to build tree nodes based on random variable omission choices when one variable is of higher priority than the other.

The prediction results obtained are the average of recently watched labels in the training data. This conduct creates a problem in circumstances where the forecast inputs contrast in their conveyances and this is hard for Random Forest because it cannot extrapolate.

Despite these limitations, every machine learning algorithm has the room to improve the results for better decision making as suggested in the future work section.

10.Conclusion

The reason for this examination is to check how well Artificial Intelligent Algorithms can be used to predict power theft considering the social and economic determinants behind electricity theft. The research has answered the research question by training seven models and compare how well they performed in predicting the power theft. At the underlying stage, literature review was led to sum up existing condition of craftsmanship on the domain. The gaps identified were utilized as inspiration for the thesis.
The analysis of the results did not only look at the exactness deciding the exhibition of the models to judge the performance of the model but also looked at the proportion of right theft forecast, similar to affectability of the models which shows how well the model has distinguished the real theft against the absolute guaranteed theft. Be that as it may, the prescient estimation of the trial of all the 7 models in this exploration was dictated by the following criteria: affectability, explicitness and by the commonness of the condition explicitly for the business case under examination.
Based on the results obtained from each model, there are indications that Artificial Intelligent Algorithms can be used to predict power theft by looking at the behavior patterns of the dataset that models expose. Therefore, the evaluation of the performance of the models can be based not only on the accuracy of the model but also sensitivity, specificity and the AUC results as discussed in the evaluation section of the research. In view of this, the technique of using Random Forest for electricity theft prediction demonstrated promising. Therefore, the author suggests that, utility companies can be saved from loss of revenue through NTL if they embrace machine learning-based solution than rule-based solution in predicting anomaly.

10.1 Future Work

As for future work AdaBoost can be applied to all 7 models to see if they can give the same performance on real time predictions of other classification problems like in health, insurance, and financial industry where fraud has also risen. It is worth mentioning in this section that the clampdown of this research was on the size of data, future work should try work on this size of dataset using leave-one-out (LOO) and see the improvements that can bring on the performance of the models. LOO is useful for an exceptionally little dataset and works on the idea of training whole dataset and leaving one case for training.

11.References

Azevedo, A. and Santos, M., 2008. *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*,

Ahmed, S., Tiong, S., Yap, K., Mohammad, A. and Nagi, J., 2008. *Non-Technical Loss Analysis for Detection of Electricity Theft using Support Vector Machines*,

Alazab, M., 2019. *Deep Learning Applications for Cyber Security*. Springer Nature Switzerland AG, p.75.

Andrysiak, T., Saganowski, Ł. and Kiedrowski, P., 2017. *Anomaly Detection in Smart Metering Infrastructure with The Use of Time Series Analysis*.

Bhowmik, R., 2011. *Detecting Auto Insurance Fraud by Data Mining Techniques*,

Briseño, H. and Rojas, O., 2020. FACTORS ASSOCIATED WITH ELECTRICITY THEFT IN MEXICO. *International Journal of Energy Economics and Policy*, 10(3), pp.250-254.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., &
Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide.

Cabral, J., Pinto, J., Gontijo, E. and Filho, J., 2004. *Fraud Detection in Electrical Energy Consumers using Rough Sets*,

Bolton, R. and Hand, D., 2002. *Statistical Fraud Detection: A Review*, 17(3).

Sgcc.com.cn. 2018. *STATE GRID Corporation of China*. [online] Available at: <http://www.sgcc.com.cn/> [Accessed 5 July 2020].

Depuru, S., Wang, L. and Devabhaktuni, V., 2017. *Support Vector Machine Based Data Classification for Detection of Electricity Theft*,

Gaur, V., Gupta, E. (2016), The determinants of electricity theft: An empirical analysis of Indian states. Energy Policy, 93, 127-136

Galvan, J., Alices, A., Munoz, A., Czernichow, T. and Sanz-Bobi, M., 1998. *System for Detection of Abnormalities and Fraud in Customer Consumption*,

Hasan, M., Toma, R., Nahid, A., Islam, M. and Kim, J., 2019. Electricity Theft Detection in Smart Grid Systems: A CNN-LSTM Based Approach. *Energies*, 12(17), p.3310.

Hu, W., Yang, Y., Huang, X. and Cheng, Z., 2020. *Understanding Electricity-Theft Behavior via Multi-Source Data*,

Idreos, S., Papaemmanouil, O. and Chaudhuri, S., 2015. *Overview of Data Exploration Techniques*,

Jiang, R., Lu, R., Wang, Y., Luo, J., Shen, C. and Shen, X., 2014. Energy-theft detection issues for advanced metering infrastructure in smart grid. *Tsinghua Science and Technology*, 19(2), pp.105-120.

Jokar, P., Arianpoo, N. and Leung, V., 2016. Electricity Theft Detection in AMI Using Customers' Consumption Patterns. *IEEE Transactions on Smart Grid*, 7(1), pp.216-226.

Kaviani, P. and Dhotre, M., 2017. SHORT SURVEY ON NAIVE BAYES ALGORITHM. *International Journal of Advance Engineering and Research Development*, 4(11).

Li, S., Han, Y., Yao, X., Yingchen, S., Wang, J. and Zhao, Q., 2019. Electricity Theft Detection in Power Grids with Deep Learning and Random Forests. *Journal of Electrical and Computer Engineering*, 2019, pp.1-12.

Martinez, W., Martinez, A. and Solka, J., 2010. *Exploratory Data Analysis With MATLAB*. 2nd ed. Chapman Hall.

Marrakchi, O., Agina, A. and Elghali, A., 2009. *Evaluating SVM and BPNN Classifiers for Remote Sensing Data*, 4(5), pp.600-605.

McLaughlin, S., Holbert, B., Fawaz, A., Berthier, R. and Zonouz, S., 2013. A Multi-Sensor Energy Theft Detection Framework for Advanced Metering Infrastructures. *IEEE Journal on Selected Areas in Communications*, 31(7), pp.1319-1330.

Nizar, A., Dong, Z. and Wang, Y., 2008. Power Utility Nontechnical Loss Analysis with Extreme Learning Machine Method. *IEEE Transactions on Power Systems*, 23(3), pp.946-955.

Nagi, J., Yap, K., Tiong, S., Ahmed, S. and Mohamad, M., 2010. Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines. *IEEE Transactions on Power Delivery*, 25(2), pp.1162-1171.

Nizar, A., Dong, Z. and Zhao, J., 2006. *Load profiling and Data Mining Techniques in Electricity Deregulated Mark*,

Olvera-López, J., Carrasco-Ochoa, J., Martínez-Trinidad, J. and Kittler, J., 2010. A review of instance selection methods. *Artificial Intelligence Review*, 34(2), pp.133-143

Patro, S. G. K. and Sahu, K. K. (2015). A technical analysis of financial forecasting, International Journal of Computer Sciences and Engineering.

Razavi, R., Fleury, M. (2019), Socio-economic predictors of electricity theft in developing countries: An Indian case study. Energy for Sustainable Development, 49, 1-10

Saini, S., 2017. Social and behavioural aspects of electricity theft: An explorative review,

Sardar, S. and Ahmad, S., 2016. *Detecting and Minimizing Electricity Theft: A Review*,

Seger, K., 2005. *Revenue Protection: Combating Utility Theft & Fraud*. PennWell Corporation.

Smith, T., 2004. Electricity theft: a comparative analysis. *Energy Policy*, 32(18), pp.2067-2076.

Tiong, S., Nagi, J., Koh, J., Nagi, F. and Yap, K., 2012. *Comparison of Supervised Learning Techniques for Non-Technical Loss Detection in Power Utility*,

Viaene, S., Derrig, R. and Dedene, G., 2004. A case study of applying boosting Naive Bayes to claim fraud diagnosis. *IEEE Transactions on Knowledge and Data Engineering*, 16(5), pp.612-620.

Wang, X. and Ahn, S., 2020. Real-time prediction and anomaly detection of electrical load in a residential community. *Applied Energy*, 259, p.114145.

Wirth, R. and Hipp, J., 2000. *CRISP-DM: Towards a Standard Process Model for Data Mining*,

Yeckle, J. and Tang, B., 2018. *Detection of Electricity Theft in Customer Consumption using Outlier Detection Algorithms*,

Zhang, R., Nie, F., Li, X. and Wei, X. (2019). Feature selection with multi-view data: A survey. Information Fusion, 50: 158-167

Zheng, Z., Yang, Y., Niu, X., Dai, H. and Zhou, Y., 2018. Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids. *IEEE Transactions on Industrial Informatics*, 14(4), pp.1606-1615.