An Improved Logistic Regression Algorithm For Tamilnadu Electricity Theft Prediction And Analysis

V.Kayalvizhy¹, Dr. A Banumathi²

¹Research Scholar, PG & Research Dept. of Computer Science, Govt. Arts College (A), Karur–639005, Tamil Nadu, India.

²Research Advisor, PG & Research Dept. of Computer Science, Govt. Arts College (A), Karur–639005, Tamil Nadu, India.

Corresponding Author: V.Kayalvizhy, Dr. A Banumathi

(Affiliated to Bharathidasan University, Tiruchirappalli)

Abstract

In India, there are several different types of power theft, the most of which take place in metropolitan areas. Technical loss and non-technical loss are the two categories of losses that can occur in an organization. Technical loss is an inescapable concern. Companies that produce energy are suffering enormous losses as a result of the theft of energy owing to non-technical losses. This is a consequence of the fact that energy is being stolen. Theft of electricity is currently a significant problem all around the world. A top-down approach strategy based on logistic regression should be proposed as a means of preventing large-scale theft of energy. The monitoring of the transport and distribution divisions can be accomplished by any one of the aforementioned ways. The structure that predominates places an emphasis on openly catering to transmission and distribution levels. At this time, provide a method for doing logistic regression classifier analysis in order to detect power theft from the producer to the consumer. This includes power generation, transmission, distribution, and consumption. A number of contemporary technologies, such as smart meters, enhanced metering infrastructure, Internet of Things-based smart metering, etc., have been developed to fight this issue. In spite of the finding of potential solutions, power thefts continue to be a problem across the country, leading to an increase in the amount of money lost by the government. The end product is a model and a pattern of the daily power usage of electricity customers who purchase consumer goods. After that, measurements of how much electricity they use each day are kept an eye on.

Keyword: Logistic Regression, IoT Smart Meter, Technical loss, non-technical loss.

1. Introduction:
There are three main components of the power source system named as power station, transmission line, and distribution system. Electricity is typically generated at power plants located in promising areas far from consumers. The hub is then sent over long distances for loading using conductors known as transmission lines. The electricity has circulated through a distribution network to a large number of huge as well as small consumers.

The primary categories for power supply systems are either alternating current (AC) or direct current (DC). System (ii) Overhead or subterranean system. At this time, a three-phase, three-wire alternating current (AC) system is the method that is generally anticipated to be used for the generation and transmission of energy for commercial purposes. In contrast, the power supply is alternating current with three phases and four wires. System. One that is concealed is more elegant than one that is displayed aloft.

As a consequence of this, the transmission and distribution of energy in our country frequently makes use of systems that are suspended in the air. The enormous network system of conductors that extends all the way from the power station to the end user may be broken up into two distinct components: the transmission system and the distribution system. The primary exchange and the secondary exchange make up each network region. ii) two types of distribution: main and secondary.

The massive network of conductors that connects the power station and the consumer can be broken up into two distinct parts: the transmission system and the distribution system. The distribution system is responsible for delivering the electricity that was generated at the power plant to the end user. Each component may be further broken down into primary and secondary trade, as well as primary and secondary distribution. [1]

**Overview of TNEB**

On July 1, 1957, an instance of administrative meddling known as the Tamil Nadu Electricity Board (TNEB) came into existence. It is necessary for TNEB to fulfil its duties as the owner of the distribution authority and the transaction utility.

**2.1 Power Generation:**

The TNEB now has a capacity of 630 MU, an installed capacity of 10,098 MW, and an annual preliminary generation of 156 MW (MW) (million units). As of the first of April in the year 2020, the Tamil Nadu Power Generation and Distribution Corporation has a production capacity of 16,034.58 MW. The normal (renewable) connection capacity is 15,871.29 MW, and it creates power using water, wind, sun, bio production, and cogeneration. Other forms of electricity generation include cogeneration. [3]

**2.2 Transmission and Distribution**
The TNEB serves 185.82 million clients in the state of Tamilnadu. The distribution network has a total of 1,73,053 distribution transformers along with 1148 substations, 1,54,104 circuit kilometers of Extra High Tension (EHT) and High Tension (HT) lines, and 5.02 lakh kilometers of Low Tension (LT) lines. As of the 31st of March in 2007, there were a total of 63,956 households, cities, and towns that were wired for electricity. In addition to that, there were 18,02 lakh agricultural pump assemblies and 10,55,705 houses that were electrified. The role that the Tamil Nadu Electricity Commission plays in the enhancement of the economy of the state through the electrification of large numbers of villages, the electrification of large numbers of agricultural pump projects, and the extension of electricity services to poor, backward, and oppressed areas of the community is discussed. A great number of businesses are aware of the rise in the amount of money spent by governments.
Figure 1 Architecture of Power Generation to Distribution [2]

2. Generating station:

A power generation station is a power manufacturing plant that uses a three-phase transformer that runs parallel to one another. An average production voltage of 11 kV. The power produced at this voltage ranges from 132 kV to 220 kV and up to 400 kV. Conducting electricity at high voltage has several advantages, including conductive material storage, higher transmission capacity, and low sine loss.

1. Primary Transmission:

A three-phase, three-wire system supplies the high voltage power supply (for example, 132 kV), which is given by the system. Overhead transportation networks for the outskirts of the city. This is the principal method by which information is sent. The electromagnetic force A system that
consists of three phases and three wires is used to transport power at a high voltage (for example, 132 kV). Overhead transportation networks for the outskirts of the city. This is the principal method by which information is sent.

2. Secondary Transmission:
After the suburbs, the principal transmission line is the overhead system, and this line often terminates at the receiving station that is situated on the outskirts of the city. Using a three-phase, three-wire overhead system, the voltage is lowered from 132 kilovolts (kV) to 33 kilovolts (kV) at the receiving station so that it may be distributed to the other stations that are strategically placed all over the city. This kind of transaction is known as a secondary exchange.

4. Primary Distribution:
The voltage is lowered from 33 kV to 11 kV when the secondary transmission line comes to an end at the substation. Following its voltage reduction, the 11 kV line will go alongside the major arteries that run through the city. The main distribution is produced as a result.

5. Secondary Distribution:
   i. Power is supplied to the distribution substation from the primary distribution (PD) line.
   ii. These substations are situated in close proximity to the consumer districts, and they bring the voltage down to a maximum of 400 V, with a voltage of 230 V existing between each phase and the neutral.
   iii. The domestic lighting load for a three-phase system is linked between any one phase and neutral.
   iv. On the other hand, motor loads that require 3 phase 400V are connected directly in 3 phase lines.
   v. Minimizes corona loss and interference in the connection circuit.
   vi. The transmission of high voltage d.c. does not incur any costs, particularly those associated with conductive loss cable casing.
   vii. In d.c. transactions, stability issues and synchronization issues.

As the next phase, electricity is sent from the primary distribution (PD) to the distribution substation so that it can be further distributed. The voltage level is lowered at these substations because of their proximity to the areas with the highest power demand. The voltage between any phase and neutral is 230 volts, and it goes up to 400 volts. The home lighting load for three phases is linked between all of the phases and the neutral whereas three-phase, 400-volt motor loads are connected directly in the three-phase lines. Corona loss and interferences in connecting circuits are reduced as a result of this feature.

Figure:2 Power Generation Three phase system

http://www.webology.org
3. ELECTRICAL POWER LOSSES
During the generation, transmission, and distribution of energy from its sources at the generation station to the distribution point, two different kinds of electrical power were lost (Fig.3).

3.1 TECHNICAL NON-TECHNICAL LOSSES AND POWER OUTAGES:
In an electrical system, there are two types of losses: (1) Technical losses are caused by process failures due to the physical properties of the system mechanisms and apparently happen in the custom of dissipating energy in transformers, transmission lines and distribution. A technical loss may be determined as a percentage and measures the amount of energy that is not billed to customers in relation to the total amount produced/bought/distributed to customers.

(2) Non-technical or business losses arising from actions outside the electricity grid e.g.: electricity theft, non-payment of bills, faulty meters, incorrect meter readings, estimation of unmeasured energy, poor record keeping, etc. [4] [5, 6] and [7] non-technical losses are often not easily calculated. However, if the technical loss is known, it is possible to calculate a nontechnical loss. NTL, also called trade losses, estimates the amount of energy that is not charged to customers for non-technical reasons. It is usually expressed as a percentage, computed as \[ NTL = (CL - TL) / CL \]. It includes fraud, theft and process problems (meters, human errors and system). TL stands for technical losses due to defects in physical distribution processes. for example, electrical impedance. CL stands for commercial loss.[8][9]

4. Proposed scheme:
This part provides a visual representation of the execution of the suggested plan, which includes a design of the architecture for data collecting on the Smart Grid (as shown in Figure) (4). Clearly This architecture illustrates how a smart grid may be constructed out of a large number of interconnected components. One example of each of these is the power industry, which includes power generating, transmission lines, distribution stations, and customers. The utility server will initially do a comparison between the entire amount of power that has been transferred and the total amount of power that has been received at many levels, including transmission (stage 1) and...
distribution (stage 2). These planning procedures are carried out on a real-time basis, and the server is responsible for carrying out the integrated sampling function.

![Diagram of Electrical Losses]

Figure:3 Electrical power losses in electric power System

Based on the results of these computations, the server that processes requests for information decides whether or not theft has happened at a particular power distribution to consumer level. If the server determines that the value of the total power transmitted has been calculated to be greater than the limit of the power received (which takes into account primary station, secondary station, primary distribution, secondary distribution, and D&T losses), it takes into account that the relevant level of theft has been confirmed after a manual inspection. After this, it is determined that there was no act of theft at the corresponding level of transmission, and the inquiry into the correct procedures is moved on to the subsequent level. The following equation is used to determine PT losses, which stand for power transmission losses.

\[ PT = PD + TDL \]

Stage 1 is the level at which electricity is transmitted, Stage 2 is the level at which power is received, and TD is the variable loss factor during power transformation. Where stage 1 and stage 2 represent as power transmission and distribution of these, Stage 1 is the level at which electricity is transmitted. Stage 2 is the level at which power is received. It is a challenging task to detect theft at T&D Stage 1 and end consumer Stage 2, as it is essential that the electricity board has a unique consumer in which some consumer theft of electricity is used by the theft connection without the use of a physically tampered with energy meter. Detecting theft at these stages requires the electricity board to have a unique consumer. Another kind is the electrical manipulation of measured data, as well as the theft of power by hooking. When this is done, the values of energy consumption that are caused by these kinds of thefts will be decreased, and end users' energy consumption will also be lowered as a result. This results in a lower overall value for the monthly electricity bill. Customers who defraud the system by stealing power are known as fraudulent consumers. The work that is being offered here suggests a categorization approach to identify customers who have fallen for this intricate power theft scheme. The suggested categorization
scheme at the consumer level is one of the most used and widespread classification methods, known as LR.

\[
PT = PD - TDL \quad \ldots \quad (1) \\
PD = CUR - L \quad \ldots \quad (2)
\]

where PT stands for power transmitted and the first level of the hierarchy is the location where power is transferred. Subtracting the values of power distribution and TD loss are the first steps in

Figure: 4 System Architecture

Power Transmission = Power Distribution – TD Loss
Power Distribution = Consumer usage Reading – Loss

http://www.webology.org
the process of determining the deviation values for power transmission. PD is an abbreviation for "Power Distribution," and the level 2 of the hierarchical structure is the location where electricity is dispersed after having received power from level 1.

5. Logistics Regression

Logistic regression is a typical technique that statisticians use to determine the likelihood that a given instance belongs to a specific class. LR is an abbreviation for the classification technique that divides the qualities into classes based on the relevant characteristics of each class. During the LR Classification process, the training set is divided into several categories. One of these categories is an energy saving dataset, which includes data on voltage, current, and power. Let's put together a model using the logistic regression. [11] One must determine the link between the value that was expected and the actual value. Creating separate sets of variables to use for training and testing the dataset Eighty percent of the data is utilized for training the model, twenty percent of it is used to evaluate the performance of our model, and after carrying out LR classification, we can forecast which customers are likely to be harmful.

The values of the parameters are sent to the LR Classifier in the capacity of input. The LR classifier is used to perform the aforementioned operations on the parameters, and it computes the pattern value, which is comprised of the users' anticipated levels of power consumption. After that, contrast the actual value with the one that was anticipated. If there was any kind of divergence between what was intended and what really happened. LR has finally reached the dishonest customers. [12]

![LR Classification Process](image)

**5.1 Sigmoid function:**

Statisticians came up with the logistic function, which is also known as the Sigmoid function, in order to characterize the qualities of population increase in ecology. These properties include growing swiftly and reaching a maximum at the carrying capacity of the ecosystem. It is a curve
in the shape of a S that can translate any real-valued number onto a value between 0 and 1, but never exactly at those boundaries. This mapping can be done for every integer.

\[
\frac{1}{1+e^x}
\]

Natural logarithms use e as their basis, and the value x that you wish to modify using the logistic function is what you pass into the function. The representation of the logistic regression equation is quite similar to that of the linear regression equation. The output value that is being modelled is of a binary type, which is the main distinction between the two.

\[
\hat{y} = \frac{e^{\beta_0 + \beta_1 x_1}}{1 + e^{\beta_0 + \beta_1 x_1}}
\]

\[\beta_0\] is the intercept term
\[\beta_1\] is the coefficient for \(x_1\)

- is the anticipated result, which has an actual value somewhere between 0 and 1. In order to convert this to a binary output of 0 or 1, either this would need to be rounded to an integer value or a cutoff point would have to be supplied in order to designate the point at which classes are separated.

It is necessary to use your training data in order to make an estimate of the logistic regression algorithm's coefficients, also known as beta values. In order to do this, maximum-likelihood estimation is utilized.

Although it does involve assumptions about the way in which your data is distributed, maximum-likelihood estimation is a popular learning method that many different kinds of machine learning algorithms employ (more on this when we talk about preparing your data).

A model would be created using the best coefficients if it predicted a value extremely close to 1 (for example, a malicious customer) for the default class and a value very close to 0 (for example, a normal customer) for the other class. The maximum-likelihood method of logistic regression is based on the premise that a search technique should look for values for the coefficients (values of beta) that would result in the smallest amount of deviation between the probabilities predicted by the model and those found in the data (e.g., probability of 1 if the data is the primary class).
5.2 Prepare Data for Logistic Regression

The assumptions that are made by logistic regression regarding the distribution of your data and the correlations between the variables are quite similar to the assumptions that are made by linear regression.

There has been a significant amount of research put into establishing these assumptions, and the terminology that is employed is quite probabilistic and statistical. In the end, the primary focus of machine learning initiatives in predictive modelling is on creating correct predictions rather than evaluating the findings. As a result, you should disregard some of the presumptions as long as the model is solid and functions effectively.

- **Binary Output Variable**: Logistic regression is designed to analyses situations that include binary (or two-class) classifications. It will make a prediction on the likelihood of an instance belonging to the default class, which may be categorized as either 0 or 1.

- **Remove Noise**: Since logistic regression presupposes that there is no mistake in the outcome variable (y), you should think about eliminating outliers and cases that might have been misclassified from your training data.

- **Gaussian Distribution**: A linear method is required for logistic regression (with a non-linear transform on output). It does make the assumption that there is a linear relationship between the variables that are input and the outcome. A more accurate model could be the result of transforming the data of your input variables in a way that exposes this linear connection more clearly. Eliminate Inputs That Are Highly Correlated Just like linear regression, this model might overfit if you have numerous inputs that are highly correlated with one another. To exclude highly correlated inputs, you might want to calculate the pairwise correlations that exist between all of the inputs first.

- **Fail to Converge**: It is conceivable for the process of estimating the anticipated likelihood that learns the coefficients to fail to converge on the correct answer. This may occur if your data has a large number of inputs that are highly associated with one another, or if your data is particularly sparse (e.g., lots of zeros in your input data).

![Example Data Table]

Algorithm 1: Formation of logistic Regression

Input: Features
Output: Logistic regression (LR)
1. Collect the data
2. Extract the features
3. Split data for Train (80%) and Test (20%)
4. Find the class label value:
   5. \( df['Tsum'] = df['TQ1'] + df['TQ2'] + df['TQ3'] + df['TQ4'] \)
   6. \( df['Taverage'] = df['Tsum'] / 4 \)
   7. \( df['Rsum'] = df['RV1'] + df['RV2'] + df['RV3'] + df['RV4'] \)
   8. \( df['Raverage'] = df['Rsum'] / 4 \)
   9. \( df['RV1-TQ1'] = df['RV1'] - df['TQ1'] \)
   10. \( df['RV2-TQ2'] = df['RV2'] - df['TQ2'] \)
   11. \( df['RV3-TQ3'] = df['RV3'] - df['TQ3'] \)
   12. \( df['RV4-TQ4'] = df['RV4'] - df['TQ4'] \)
   13. \( df['Rsum-Tsum'] = df['Rsum'] - df['Tsum'] \)
   14. \( df['Raverage-Taverage'] = df['Raverage'] - df['Taverage'] \)
   15. \( df['classlabel'] = 0 \) # add a class column with 0 as default value
   16. # find all rows that fulfills your conditions and set class to 1
   17. \( df.loc((df['Raverage'] - df['Taverage']) > 0.00) \)
   18. # if difference is more than 0.00 of Raverage-Taverage
   19. \( (df['Raverage'] > 0.00), \)
   20. \( 'classlabel' = 1 \) # then set class to 1
21. Calculate the following metrics from the confusion matrix.
22. \( TP = confusion[1, 1] \) # True Positives
23. \( TN = confusion[0, 0] \) # True Negatives
24. \( FP = confusion[0, 1] \) # False Positives
25. \( FN = confusion[1, 0] \) # False Negatives
26. Find the classification accuracy:
27. \( Accuracy = TP + TN + TP + TN + FP + FN \)

### 5.3 Model Evaluation Procedures

In general, we avoid training and testing a model with the same data because this can result in overfitting. Models that are too well suited to training data will perform poorly when applied outside of the sample data.

1. Train/Test Split
2. K-Fold Cross Validation
Train/Test Split

The data set is split into two pieces for the purpose of this methodology: a training set and a test set. The model will be trained using the specialized training set. Additionally, we are able to test the accuracy of the model throughout the training set; however, we should not evaluate the models based just on this metric. After training has been completed, the testing set is simply used to test the model and determine how accurate it is. During the training process, the model is never shown any data samples that come from the test set. The reliability of the test data provides a clearer picture of how well the model will work with the newly acquired information. The scikit-learn package gives us access to a method that is called for the purpose of segmenting the data into train and test sets. From the model selection, generate the train test split. Module. In the beginning, we will separate the data into a training set and a test set. After that, we will construct the model of logistic regression utilizing the train. After that, we will make our predictions using the test set. At last, we compute the performance of the model by employing the evaluation measure known as Classification Accuracy and obtaining an accuracy score of 0.906, which is equivalent to 90.6 percent.
Model Evaluation Metrics

A criterion that is used to evaluate a model's performance or correctness is referred to as an evaluation measure for a module. The train dataset and the test dataset are both generated with the use of a method called k-fold cross-validation. The parameters of the proposed model are trained using the train dataset, and the test dataset is used to do the evaluation.

6. Classification Accuracy

Accuracy in classification is the model evaluation metric that is utilized by far the most frequently in relation to classification issues. The proportion of accurate predictions is the measure of a classification's accuracy. Classification is a useful method of measurement; but, when the distribution of classes is uneven, it might create the impression that the level of precision is quite high. Accuracy is evaluated using a distinct technique that can be found in the metrics module of Scikit-learn and is referred to as the accuracy score. In addition to this, the accuracy estimator has been incorporated into cross_val_score in the form of a parameter. The value of the score is used to determine how accurate the categorization is.

When attempting to measure the accuracy of a classification, it is recommended to use the same number of samples for each class.

However, if the same model were to be applied to a dataset with a different class distribution, for example one in which 60 percent of the samples belong to class A and 40 percent to class B, the test accuracy score would decrease to 60 percent.

6.1 Confusion Matrix

The performance of a classification model on a test dataset in which the actual values are known can be described using a confusion matrix, the definition of which is open to interpretation. A perplexing matrix has a lot of room for interpretation and may be applied to the estimation of a variety of different metrics. The Confusion matrix may be performed on the test data set using a technique that is provided by Scikit-learn. In order to calculate the confusion matrix, the method known as confusion matrix has to have access to both the actual values of the response class and the predicted values.

\[
\begin{bmatrix}
68 & 0 \\
7 & 0
\end{bmatrix}
\]

Fig — Confusion Matrix

Because there are only two possible answer classes in our problem, we are dealing with what is known as a binary classification problem. Because of this, the matrix of confusion is a two-by-two
grid. Depending on the specific implementation, the confusion matrix can be interpreted in a variety of ways. The matrix that is displayed above is ambiguous enough for us to comprehend everything.

## 6.2 Performance Measurement

### 1. Confusion Matrix

- True Positives (TP): correct prediction as a legitimate customer.
- True Negatives (TN): correct predictor as a malicious customer.
- False Positives (FP): Wrong Prediction as Genuine customer (Type I Error)
- False Negatives (FN): wrong prediction as a malicious customer (Type II error)

![Confusion Matrix Diagram](image)

Metrics computed from the confusion matrix

The first step is to analyze the resulting confusion matrix. True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

# [row, column]

\[\begin{align*}
TP &= \text{confusion} [1, 1] \# \text{True Positives} \\
TN &= \text{confusion} [0, 0] \# \text{True Negatives} \\
FP &= \text{confusion} [0, 1] \# \text{False Positives} \\
FN &= \text{confusion} [1, 0] \# \text{False Negatives}
\end{align*}\]

We can calculate the following metrics from the confusion matrix.
2. Classification Accuracy

The link between a precise forecast and the total number of predictions is what we mean when we talk about classification accuracy. Or, to put it another way, how many times does the classifier get it right?

\[
\text{Accuracy} = \frac{\text{No. of Correct Predictions}}{\text{Total No. of Predictions}}
\]

3. Accuracy using Confusion Matrix:

The confusion matrix allows us to do the calculation necessary to determine the accuracy. In order to determine the accuracy while employing the confusion matrix, the following equation is utilized:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

It is also possible to determine the accuracy by utilizing the process known as accuracy score. It's possible for us to note that the accuracy is 0.795.

```python
print ((TP + TN) / float (TP + TN + FP + FN))
0.9066666666666666
0.9066666666666666
```

4. Specificity using Confusion Matrix:

The ratio of the number of negative predictions that were accurate to the total number of predictions that were incorrect is referred to as specificity. This impacts the specificity of the classifier when it comes to making predictions about positive cases.

\[
\text{Specificity} = \frac{\text{No. of Correct Negative Predictions}}{\text{Total No. of Negative Predictions}}
\]

The specificity can be calculated by means of the confusion matrix as follows.

\[
\text{Specificity} = \frac{TN}{TN + FP}
\]

```python
specificity = TN / (TN + FP)
print(specificity)
```
7. ROC curve

The performance of a binary classifier, or a classifier with only two potential output classes, may often be evaluated using a ROC curve, which is a common method. The graph illustrates the ratio of the True Positive Rate (Recall) to the False Positive Rate (also interpreted as 1-specific). We are able to draw the ROC curve by making use of a function that is provided by Scikit-learn and is referred to as roc curve. This method finds the false positive and true positive rates for a given set of thresholds. In the following fashion, we are able to draw the curve.

```
evaluate_threshold(0.5)
```

Sensitivity: 0.0
Specificity: 1.0

```
evaluate_threshold(0.3)
```

Sensitivity: 0.0
Specificity: 1.0

ROC curve is a reliable indicator in measuring the performance of a classifier. It can also be extended to classification problems with three or more classes using the “one versus all” approach.
AUC (Area Under the Curve)
The proportion of the ROC plot that falls inside the area indicated by the AUC, also known as the Area Under the Curve. The area under the curve (AUC) is a helpful metric to utilize as a single number overview of classifier performance.

The percentage of the ROC trace that falls below the curve is the AUC, often known as the Area Under the Curve. The ACU is helpful since it can summaries the performance of the classifier into a single number.

The roc auc score function in Scikit-learn may be used to find the area under the curve (AUC) score.

0.7668067226890756

8. Comparative Analysis:
The figure below illustrates the comparative analysis of NB, SVM, LDA, KNN, LR in terms of accuracy. Figure --- shows comparative classifier analysis based on dominance parameters. Clearly demonstrates that NB's performance is stronger than others.
9. Discussion:

Theft of energy at several levels can result in annual economic losses that are in the billions of rupees for a country. Theft of power, unregistered meters, and improper charge management are additional contributing factors to the terrible situation the nation is currently facing. This article discusses the strategies and procedures that may be utilized to steal electricity in order to conduct a study of home time series. The government has utilized millions of energy meters in the past, but now it plans to employ the prepaid meter system to cut down on electricity losses and theft, and research on smart meters is still going on. About ten consumption data are chosen for statistical analysis in order to determine the difference in monthly energy consumption. This allows one to identify the dishonest consumer, the consumer, the rate of theft, the units consumed, and the minimal losses that occur in order to differentiate the various types of consumers who steal.

10. Conclusion

Theft of electricity is a type of criminal activity that obstructs the community's development toward its economic goals. Instead of identifying and combating theft in the non-technical electrical infrastructure sector, the government and utilities are entirely focused on developing new power plants. As part of this research, a comprehensive survey was carried out to uncover instances of theft and loss on the part of consumers. Different customers have resorted to a variety of deceptive practices in order to cover up the fact that they steal power from their houses. Testing and balancing that is not done correctly. Because the smart meter system is a two-way communication mechanism between electricity companies and consumer sites, it can monitor the theft of electricity from consumers' premises and prevent the theft of electricity from sites that do not have technology. This makes smart metering systems the ideal solution.

REFERENCES:
2. http://www.industrial-electronics.com/11_Electric-power-  
11. Decision Tree and SVM-based Data Analytics for Theft Detection in Smart Grid Anish Jindal, Student Member, IEEE, Amit Dua, Student Member, IEEE, Kuljeet Kaur, Student Member, IEEE, Mukesh Singh, Member, IEEE, Neeraj Kumar, Member, IEEE and S. Mishra, Senior Member, IEEE  
13 https://www.kaggle.com/vishalansh/power-consumption-data-visualization  
14 https://pythonda.ta.com/comparing-machine-learning-methods/  
15 https://dibyendudeb.com/comparing-machine-learning-algorithms/  
16 https://machinelearningmastery.com/compare-machine-learning-algorithms-python-scikit-learn/  
18. https://www.kaggle.com/cmchauhan/we4tech-power-consumption-india