Machine Learning Based Residential Electricity Theft Detection

Talacheeru Harshavardhan
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
harshavardhan.talacheeru@gmail.com

C.Sivamurugan
Assistant Professor,
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
c.sivamurugan@klu.ac.in

Tanguturi Shantan
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
shantantanguturi@gmail.com

Ummadisetty Sasi Kumar
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
ummadisettysasikumar@gmail.com

Thanniru Rohin
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
thannirurohin@gmail.com

Abstract- **This study examined the indiscriminate theft of electricity, which is classified as a non-technical loss and affects both customers and electric distribution companies. It can have major repercussions, such as fires and blackouts. The goal of the study was to identify the most effective machine learning prediction model for electrical energy theft. A real-time dataset released by the E&D Corporation of India served as the source of statistics on the electricity use of 155 home clients. Feature extraction was the technique employed to enhance the detection of energy theft. There were eight machine learning models examined. Thus, 93.5% was the accuracy indicator for the SVM model, 93% for K-Nearest Neighbors, 96.7% for Random Forest, 96.7% for Logistic Regression, 93.4% for Naive Bayes, 93% for Decision Tree, 83.8% for Kmeans, and 93% for Gradient Booster. It is determined showed the Random Forest plus Logistic Regression model yields the highest results, with an accuracy of 96.7%.**

Keywords: Machine learning, load shape dictionary, random forest, logistic regression, and electricity theft detection.

I. INTRODUCTION

A variety of methods are used in machine learning-based home electricity theft detection to pinpoint instances of illicit electricity use. Usually, this is done to guard against energy theft, guarantee accurate invoicing, and preserve the integrity of the electrical infrastructure. Using smart meters, tracking voltage and current levels, evaluating usage trends, and applying data analytics to spot abnormalities are some techniques for identifying electricity theft. In addition to costing utility companies money, illegal electricity diversion puts community safety in danger and puts stress on the infrastructure supporting power distribution. The importance of effective energy management has increased due to the world's growing energy consumption and the requirement for sustainable energy resources. Utility companies encounter a significant obstacle in the form of domestic energy theft, a widespread problem that not only results in significant financial losses but also presents risks to public safety and inefficiencies in operations. The inaccuracies and delays in traditional theft detection systems make sophisticated technological interventions necessary. Within this framework, machine learning (ML) presents itself as a potent instrument for revolutionizing the field of electricity theft detection. Machine learning algorithms are a promising way to find abnormalities in electricity consumption that might be signs of theft because of their capacity to recognize complex patterns within big and complicated datasets. This study explores the field of machine learning (ML)-based household electricity theft detection with the goal of creating advanced models that can precisely identify and stop theft in real-time.This study analyzes the immense potential of predictive analytics by utilizing machine learning approaches, such as deep learning structures like neural networks and supervised learning algorithms like decision trees and support vector machines. Voltage swings, unusual usage patterns, and consumption patterns are just a few of the features that are thoroughly examined in this research of several elements taken from historical electricity usage data. Additionally, the study looks into how anomaly detection techniques might be included to identify minute abnormalities that could slip past traditional detection systems. The effective adoption of this measure would have significant ramifications not only for utility companies looking to reduce revenue losses and improve operational efficiency, but also for consumers who will benefit from more equitable billing procedures and a more secure electricity supply. This research endeavors to advance the field of electricity theft detection, contributing valuable insights that bridge the gap between theoretical advancements and practical applications, thereby paving the way for a more sustainable and equitable energy future.

II. LITERATURE REVIEW

The study Decision Tree based Electricity Theft Detection in Smart Grid focuses on the detection of electricity theft using machine learning algorithms. The authors analyze electricity usage data from 114 single-family apartment complexes using gradient boosting, random forest, and decision tree approaches. Finding non-technical loss, or energy theft due to various reasons, is the goal.

The object To improve the performance of fraud detection models, a helpful feature-engineering framework for electricity theft detection in smart grids makes use of clustering and feature engineering techniques.

The study analyzed demand data from over 4000 households in six different attack scenarios using five machine learning techniques. The study found that GBM was the most successful machine learning algorithm for identifying fraud after comparing its efficacy with that of other algorithms.

In article Electricity theft detection using Empirical mode decomposition and K-nearest neighbor the significance of electricity in modern life is undeniable, with its crucial role in various sectors. Addressing electricity theft, particularly in countries like Pakistan, is essential for economic stability. Implementing smart grids, such as through smart meters, offers a promising solution to detect and reduce power losses from theft. Utilizing thirteen features like Mean, Standard Deviation, and others, the Fine KNN classifier achieved an accuracy of 91.0%.

The article "Electricity Theft Detecting Based on Density-Clustering Method" claims that unusual conduct by electricity users, especially electricity theft, has resulted in large financial losses for power companies globally. The methods based on clustering do not require unlabeled data. These methods take patterns from a vast amount of user features and use them to find outlier patterns. The confusion matrix divides the whole dataset into four categories: true positive (TP), false positive (FP), false negative (NP), and true negative (TN). TP, FP, FN, and TN.

The authors of the paper Real-time power theft monitoring and detection system with double connected data capture system utilized a GSM (Global System for Mobile Communications) module to show how useful this technique can be for identifying electrical theft. To detect electricity theft, the smart meter that tracks energy use needs to be connected to the GSM module.

In the work Machine learning for identifying theft of electricity. SVM, K-Nearest Neighbors, Random Forest, Logistic Regression, and Naive Bayes are the five types of algorithm models that were tested. SVM's maximum accuracy of 81%. consequences of non-technical losses, such as fires, blackouts, and power theft.

Study on Weighted Naive Bayes Classification and Recognition of User Stealing Detection In order to assist supervisors in monitoring user behaviors and reducing the number of users participating in power theft behaviors, this paper suggests a naive Bayes method. Users of electricity theft are identified by applying feature weighting in line with the preprocessing strategy of screening the attributes of data related to electricity theft.

III. METHODOLOGY

Developing a machine learning-based residential electricity theft detection system involves several key steps. First, collect a comprehensive dataset comprising legitimate and fraudulent electricity consumption patterns, ensuring it encompasses diverse demographics and usage scenarios. Preprocess the data, cleaning outliers and handling missing values, while also engineering relevant features such as usage patterns. Next, employ various machine learning algorithms, including decision trees, random forests, Logistic regression, Naive Bias, K-Nearest Neighbor to train the model on the prepared dataset. Utilize techniques like cross-validation and hyper parameter tuning to optimize the model's performance. Implement anomaly detection methods to identify Finally, integrate the trained model into the existing electricity grid infrastructure, enabling real-time monitoring and alerts for potential theft, thereby enhancing the overall efficiency and reliability of the residential electricity supply. unusual usage patterns, indicative of theft. Validate the model rigorously on unseen data to assess its accuracy, sensitivity, and specificity.

	A	B	C	D	E	F	G	H	I	J
1	id	date	enegry_m	energy_m	energy_m	energy_co	energy_st	energy_su	energy_m	flag
2	1	15-12-201	0.485	0.432045	0.868	22	0.239146	9.505	0.072	0
3	2	16-12-201	0.1415	0.296167	1.116	48	0.281471	14.216	0.031	0
4	3	17-12-201	0.1015	0.189813	0.685	48	0.188405	9.111	0.064	0
5	4	18-12-201	0.114	0.218979	0.676	48	0.202919	10.511	0.065	0
6	5	19-12-201	0.191	0.325979	0.788	48	0.259205	15.647	0.066	0
7	6	20-12-201	0.218	0.3575	1.077	48	0.287597	17.16	0.066	0
8	7	21-12-201	0.1305	0.235083	0.705	48	0.22207	11.284	0.066	0
9	8	22-12-201	0.089	0.221354	1.094	48	0.267239	10.625	0.062	0
10	9	23-12-201	0.1605	0.291125	0.749	48	0.249076	13.974	0.065	0
11	10	24-12-201	0.107	0.169	0.613	47	0.150685	7.943	0.065	0
12	11	25-12-201	0.2175	0.339188	0.866	48	0.263101	16.281	0.069	0
13	12	26-12-201	0.1495	0.261708	0.838	48	0.244793	12.562	0.066	0

Figure 1 : Dataset of 12 Residents with different attributes

IV. PROPOSED SYSTEM

Our proposed system for residential electricity theft detection integrates cutting-edge machine learning techniques to create an efficient and accurate solution. The system consists of three main stages: data collection, feature engineering, and machine learning model implementation. Firstly, a comprehensive dataset of residential electricity usage patterns is gathered, ensuring diversity and representativeness. During feature engineering, relevant features such as consumption behavior, voltage irregularities, and temporal usage patterns are extracted to enhance the model's predictive power. Then, a range of machine learning methods, including logistic regression, random forests, decision trees, and K-nearest neighbor, are employed to produce trustworthy predictive models. These models are trained and validated using advanced techniques such as cross-validation. The system also has anomaly detection tools to find unusual behavior that might be a sign of electricity theft. By putting the recommended system into place, utility companies can significantly improve their ability to detect and prevent residential electricity theft. This will lead to a decrease in energy losses, fair billing practices, and overall improvements in operational efficiency.

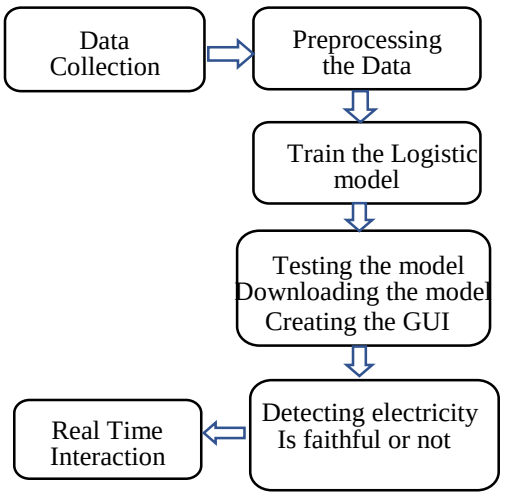

Figure.2: Machine Learning Based Electricity theft detection

A.Data Collection:

Block diagram for Electricity theft detection is shown in Figure 2. The primary step in the Electricity theft detection is data collection. Collected a data of 155 residents with 9 different attributes i.e date, energy_median,energy_mean,energy_max,energy_count, energy_std, energy_sum, energy_min, flag.

B.Preprocessing the Data:

The collected residential electricity dataset undergo preprocessing to normalize the data and remove the missing values from the dataset. Preprocessing techniques may include feature engineering, normalization.

C.Training the Logistic Regression model:

After the dataset was used to train the models, random forest and logistic regression showed the highest accuracy. For this reason, the data are trained using the logistic regression model.

D.Testing the model and downloading the model and creating GUI.:

To find cases of electricity theft, a tested and trained logistic regression model is used. The trained model is downloaded and tested with new data to see if the streamlit-created GUI can reliably process the data.

E.Detecting Electricity is faithful or not:

Logistic model predicts the output based on the different data points, weather the electricity is theft or not and provide the output in GUI.

F.Real-time Interaction:

The final step in this is to predict the electricity is theft or not for the real time data.

V. RESULT AND DISCUSSION

Eight different machine learning models for the electricity theft detection system were compared after the model was trained using a dataset consisting of 155 unique residents with 9 attributes. Random forest and logistic regression demonstrated an amazing 96.77% accuracy after training. Nevertheless, k-Nearest Neighbor, naive Bias, Decision Tree, and SVM demonstrated a respectable accuracy of 93.55% after training on the same dataset. Despite its lack of accuracy compared to other architectures, Logistic Regression remains a viable choice. After training on the same dataset, k-means proved to be the least accurate method, scoring 84%. Logistic regression is an effective tool for identifying instances of electricity theft using real-time data. In summary, the detection of electricity theft is more accurately predicted by logistic regression.

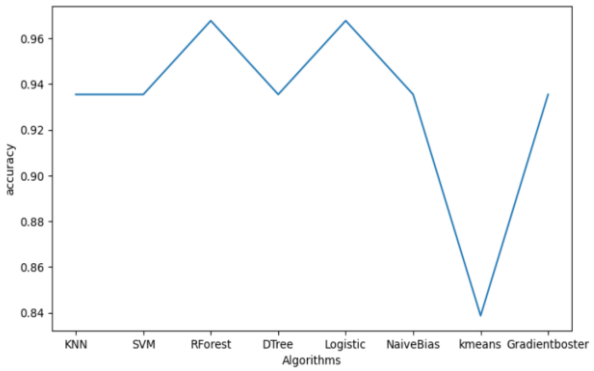

Figure.3: Various machine learning models' accuracy after being trained on household data

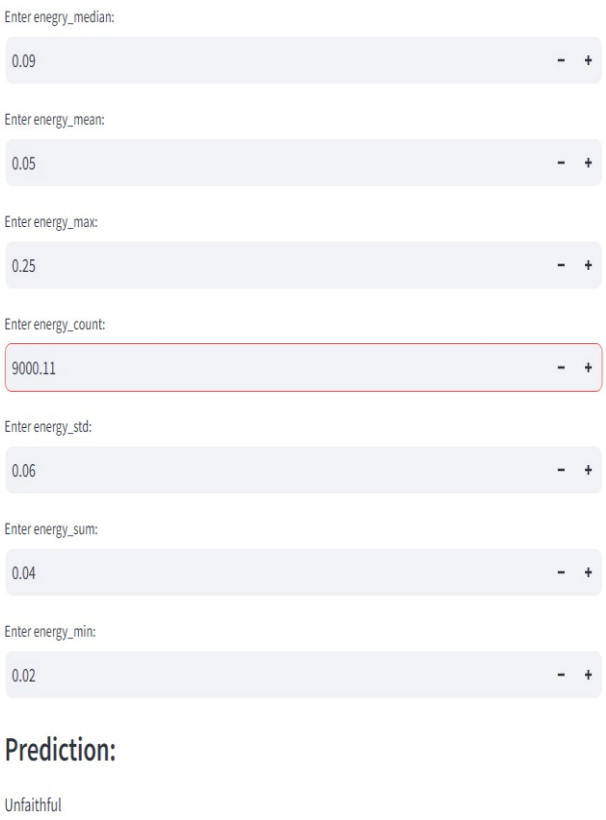

Figure.4,5: Represents the output prediction of electricity theft detection.

Model accuracy is

$$Accuracy = \frac{TP+FN}{TP+TN+FP+FN} \quad (1)$$

Where, TP=true positive rate, TN=true negative rate, FP = false positive rate and FN=false negative Rate. Performance testing flow is shown in Figure 7, which briefs the accuracy and loss performance for validating the model. Validation loss must be small as possible. Number of epochs vs. loss graph Training loss, validation loss.

Validation Accuracy = 1-Validation Loss. Validation loss should be small as possible. If the validation loss is larger than the training loss, this tells overfitting. However, some degree of overfitting can be ignored.

Validation loss >> training loss : Overfitting.

Validation loss > training loss: some overfitting

Validation loss < <training loss:some underfitting Validation loss

< training loss : underfitting.

Validation loss == training loss : perfectfitting

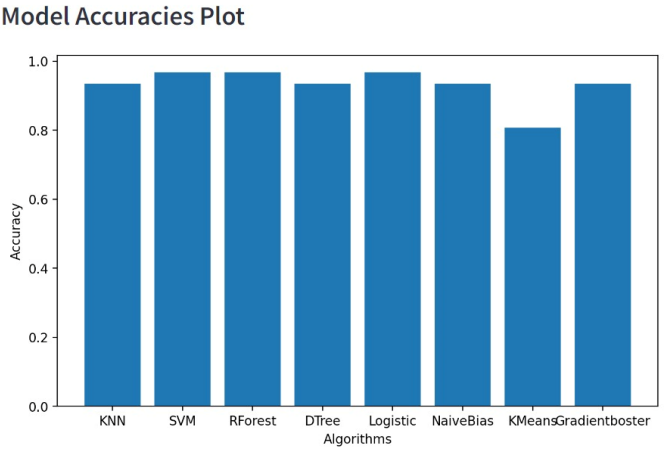

Figure 6: Bar chart representation of accuracy of all machine learning models.

Figure 4,5 presents all the attributes of the dataset and forecasts electricity theft based on those values. If all the values are high and energy_min is low, which indicates that no theft has taken place, then electricity theft is detected. The output prints as faithful if there is no theft, and as unfaithful if theft is discovered. Figure 3 displays the diagrammatic representation of accuracy for each of the various machine learning models. Figure 3 displays the accuracy of each machine learning model graphically, and Figure 6 displays the accuracy of each model as a bar chart.

VI. CONCLUSION

To summarize, after using various algorithms we concluded that the Random Forest is simple and effective model which is worked good with the large data and Logistic regression is a commonly used classification algorithm that is easy to interpret and works well with linearly separable data and it worked good with this data. Decision Tree is a tree-based model that is simple to understand and interpret, and can handle both categorical and numerical data. However, it may suffer from overfitting, particularly when dealing with noisy data and its worked somewhat good only with this data. We used Logistic regression model for predicting the electricity theft.

VII. REFERENCES

[1] Sumair Aziz Department of Electronics Engineering, University of Engineering and Technology, Taxila Taxila, Pakistan, sumair.aziz@uettaxila.edu.pk

[2] Ivan Petrlik, Faculty of Industrial and Systems Engineering, National University Federico Villarreal, Lima, Perú

[3] Soroush Omidvar Tehrani, Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran, omidvar@mail.um.ac.ir

[4] Mohammad Hossein Yaghmaee Moghaddam Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran, yaghmaee@ieee.org

[5] Mohsen Asadi Department of Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran, mohsen.asadi@um.ac.ir

[6] Rouzbeh Razavi Department of Management and Information Systems, Kent State University, USA, rrazavi@kent.edu

[7] Amin Gharipour School of Information and Communication Technology, Griffith University, Australia.
[8] Martin Fleury School of Computer Science and Electronic Engineering, University of Essex, UK
[9] Pedro Lezama Faculty of Industrial and Systems Engineering, National University Federico Villarreal, Lima, Perú.
[10] Ciro Rodriguez Faculty of Electronic Engineering and Informatics, National University Federico Villarreal, Lima, Peru.
[11] Ricardo Inquilla Faculty of Engineering, National University of Cañete, Cañete, Peru.
[12] Julissa Elizabeth Reyna-González Faculty of Industrial and Systems Engineering, National University Hermilio Valdizán, Huánuco, Peru.
[13] Roberto Esparza Faculty of Industrial and Systems Engineering, National University Federico Villarreal, Lima, Peru.
[14] 13. Christina Juliane Informatics Engineering Department, STMIK AMIK Bandung, Indonesia martiti@stmik-amikbandung.ac.idW. Rahmaniar, A. Ma'Arif and T. -L. Lin, "Touchless HeadControl (THC):