Journal of Global Information Management
Volume 30 • Issue 5

Application of Machine Learning Algorithm in Managing Deviant Consumer Behaviors and Enhancing Public Service.

Shantanu Dubey, Postman, Inc., India
Prashant Salwan, Indian Institute of Management, Indore, India*
Nitin Kumar Agarwal, XALT Analytics Pvt. Ltd., India

ABSTRACT

Consumer-deviant behavior costs global utility firms USD 96 billion yearly, attributable to non-technical losses (NTLs). NTLs affect the operations of power systems by overloading lines and transformers, resulting in voltage imbalances and, thereby, impacting services. They also impact the electricity price paid by the honest customers. Traditional meters constitute 98% of the total electricity meters in India. This paper argues that while traditional meters have their limitation in checking consumer-deviant behavior, this issue can be resolved with ML-based algorithms. These algorithms can predict suspected cases of theft with reasonable certainty, thereby enabling distribution companies to save money and provide consistent and dependable services to honest customers at reasonable costs. The key learning from this paper is that even if data is noisy, it is possible to create a machine learning model to detect NTL with 80% or higher accuracy.

KEYWORDS

Artificial Intelligence, Emerging Economies, Non-Technical Losses, Power Utilities, Traditional Meters

1. INTRODUCTION

Technology interventions like digitalization and Machine Learning (ML) have had a commendable impact on public services (Cheng, Hu, & Wu, 2021) and consumer behavior (Ahmad, Masri, Chong, Fauzi, & Idris, 2020). Technology applications not only help to enhance public services but also reduce deviant consumer behavior (DCB). (Fullerton & Punj, 1997) defines deviant consumer behavior as any behavior which is "against the law, organizational policy or violates the generally accepted norms of conduct." DCB causes financial and physical losses to the organization and emotional harm to the owners and employees (Daunt & Harris, 2012).

Organizations, especially public service organizations, use tactics like communicating with customers to comply with the legal and social norms centering their messaging around, "it's wrong, don't do it." The second tactic that industries such as the retail industry use are evoking fear of punishment. In these tactics, organizations have to proactively demonstrate that customers cannot get away with unethical practices and that they may be caught and punished for their deviant behavior.

DOI: 10.4018/JGIM.292064 *Corresponding Author

The second tactic is defined as the "Deterrence tactic: you will be caught and punished" (Dootson, Lings, Beatson, & Johnston, 2017).

Organizations use hardware like CCTV cameras and non-hardware solutions like analytics, Artificial Intelligence (AI), and ML to capture deterrence tactics. Analytics is used to increase enterprise value by appropriate application in several functional areas, viz data to increase sales, and improve customer service and operations, to name a few (Baker, Al-Gahtani, & Hubona, 2010). It is also finding extensive use in other activities ranging from predicting train tickets and confirmations to checking for water supply leaks and even finding the perfect bride and groom. Governmental services are using analytics to combat crime, improve transparency, and services such as transport, etc. (Raghupathi & Raghupathi, 2014).

Increasingly, energy utility firms are using analytics to optimize power generation and planning (Kim et al., 2016). However, energy theft remains their major concern, adversely affecting the bottom-line and profitability (Dick, 1995). Electricity losses in utility firms are recorded under two heads, namely Technical Losses (TL) and Non-Technical Losses (NTL). Power dissipation in transportation and distribution of power falls under TL. Commercial losses are due to non-billed electricity, defined as non-natural losses and recorded under NTL. Non-billing of consumed electricity happens due to errors in metering or non-legitimate behavior of consumers (Oliveira et al., 2001).NTL reduces the finance available with utility firms for investing in further growth (de Souza Savian et al., 2021).

Emerging economies face the brunt of energy thefts; for instance, Brazil and India record an annual loss of USD 3 billion (Z. Hussain, Memon, Shah, Bhutto, & Aljawarneh, 2016). Extant research has adopted various methods, including AI-based, game theory-based, and state-based models, to capture NTL. ML and deep learning (DL) are constituents of AI-based approaches. ML is the process of training a machine with an algorithm to handle large data efficiently by predictive analysis. On the other hand, DL (Mohammad, Thabtah, & McCluskey, 2012) is based on an artificial neural network (ANN), a human brain model that helps to model irrational functions.

Using AI, utility firms can detect usage patterns, payment history, and other customer information that indicates misconduct (Gunturi & Sarkar, 2021; J. Li & Wang, 2020). For instance, in Brazil, power theft represents up to 40 percent of the distribution of electricity, while India loses approximately 25% of its supply, amounting to INR 200 billion every year (Gunturi & Sarkar, 2021). Hence, the Indian energy utility space calls for an urgent application of AI (Akter et al., 2021) and ML to capture and address deviant consumer behavior.

The focus of our research is to detect Energy Theft (ET) using an ML algorithm. Electricity billing is recorded using electrical meters, which are of two types, traditional/ analog meters, and smart meters. Smart meters are crucial for reducing losses in electricity distribution companies, also known as DISCOMs (Gholami, Nishant, & Emrouznejad, 2021). The smart meter is an automated metering system that requires no manual intervention and reduces meter-reading and data-entry errors and costs. Some DISCOMs that are using smart meters have reported an increase in their per-meter, per-month revenue by Rs 200. (When the national average bill is about Rs 450, assuming average consumption of 90 units per month at Rs 5 per unit) ("EESL," 2021).The use of smart electronic meters is currently the most effective method for reducing NTLs. However, while their installation is advantageous, the expenses are substantial, and new infrastructure for data collection is necessary.

Traditional/analog meters (TM) constitute 98% of India's electricity meters. In TM, the representative of a utility firm has to visit the consumer premises and record the meter reading physically to capture the consumption. In the COVID- 19 situation where physical visits were not possible, utility firms issued bills on estimates, as a result of which they were faced with more service issues and losses.

Characteristics of data decide the effectiveness of an ML solution and the performance of the learning algorithms (Wu et al., 2021).The key research question which we address in this paper is can ML techniques help to viably detect theft in electricity using traditional meter data (which is noisy due to manual reading)and thus eliminate the urgency of switching to smart meters. Smart meters are

good but require a large investment, which is a challenge for the resource-starved utility companies in emerging economies. It will take these companies approximately four to eight years to replace a majority of TMs with smart meters (Alkawsi, Ali, & Baashar, 2021). Hence, ML may be the solution to minimize losses that occur due to data management in traditional meters.

Given the importance of machine learning algorithms, a limited but growing body of literature has examined its application in managing theft and deviant consumer behaviors, thereby enhancing the quality of public services, e.g., (P. K. Jain, Pamula, & Srivastava, 2021; Maamar & Benahmed, 2018; Ravnik, Solina, & Zabkar, 2014; Veale & Brass, 2019).

Prior studies have highlighted the complexities in determining power theft using machine learning, e.g.,(Arif, Javaid, Aldegheishem, & Alrajeh, 2021; Ghaedi, Tabbakh Farizani, & Ghaemi, 2021).In the context of the current study, we pay particular attention to the conceptualization of an ML algorithm, which can identify two attributes: consumption pattern and fraud in the distribution network. Prior studies have mostly focused on using ML for theft detections in smart grids, e.g., (Hasan, Toma, Nahid, Islam, & Kim, 2019; Johncy & Felise, 2020; Nawaz, Akhtar, Shahid, Qureshi, & Mahmood, 2021) and largely ignored its potential impact to detect power theft in traditional meters. For instance, (Glauner, Meira, Valtchev, State, & Bettinger, 2016) highlights the significance of the new model, which is based on supervised ML techniques and actual electricity consumption data. He implemented Support Vector Machine (SVM), Convolution Neural Network (CNN), and Logistic Regression (LR) (S. Hussain et al., 2021) to propose a supervised ML-based electric theft detection approach using the feature engineered-CatBoost algorithm with the SMOTE Tomek algorithm. (Gunturi & Sarkar, 2021) suggest ensemble ML (Srivastava & Eachempati, 2021) models for detecting energy theft in smart grids using customers' consumption patterns. However, no study has considered ML application in managing traditional meters data (Funayama et al., 2021; Yilmaz, Kapoor, Siraj, & Abouyoussef, 2021).

Our study develops the ML technique for traditional meters to reduce energy theft. In light of the ML algorithm and power theft literature, we identify and map the relevant attributes of the "application of ML algorithm in managing deviant consumer behaviors and power theft detections.". To maintain an effective operation, a utility firm needs to focus on power theft detections, technology implementations, and consumer behaviors (Gordon, 1993).

In this study, we specifically explore:

RQ1. How can a large government-run utility company in an emerging market preprocess the noisy data from manual and sometimes irregular readings from traditional meters to derive some insights?
RQ2. How to use the sparse and noisy consumer data collected manually to detect theft cases with reasonable accuracy and business value?
RQ3. How to extract features from this noisy and sparse data to create powerful ML models that can detect theft cases?
RQ4. How to create the ML model that can detect spikes and troughs in consumer data due to theft and separate it from the spikes in the data due to weather-related reasons?
RQ5. Finally, how to develop and fine-tune Random Forest and Neural Network-based to obtain reasonably accurate models?

This paper is the first attempt of its kind to investigate the impact of TM power theft in light of consumer behaviors and public policy. Our findings will contribute to the ML literature by unveiling its different dimensions specified in the context of utility firms. We also propose an ML algorithm based on Random Forest, an ensemble ML classifier-based energy theft detector to recommend means to reduce power theft in traditional/analog meters. The model demonstrates the accuracy and performance of actual data from a large emerging market utility company. Our ML energy theft detector model has been trained and tested with real data obtained from customers of electricity utilities.

The remainder of this paper is organized as follows. Section 2 discusses the literature background on energy theft identification and ML in traditional/analog meters. Section 3 presents the proposed methodology for undertaking the research. Section 4 discusses the application and effects of our ML algorithm for utility firms. Section 5 discusses the results and concludes with a discussion on the theoretical and managerial implications, limitations, and directions for future work.

2. THEORETICAL UNDERPINNINGS

2.1 Problem Identification

Utility firms from emerging economies like Brazil and India and developing economies like Malaysia and Thailand suffer huge losses touching billions of USD due to NTL, specifically electricity theft (Nizar & Dong, 2009). The problem does not stop here, and the honest consumer is left to face the brunt of DCB. The losses are transferred to such consumers as hike-in tariffs (Bhatia & Gulati, 2004). Electricity theft disrupts the power operations by overloading lines and transformers, resulting in voltage imbalances and thus impacting services, including long blackouts.

An Indian state Uttar Pradesh is an apt example wherein during the period 1970 to 2010, 29% of power transmitted was not accounted for (Min & Golden, 2014). There are many motivations for the DCB, including institutional voids in the market system, corrupt employees, protection due to no process of data capture, non-payment of bills by powerful customers, the difficulty for utility firms to zero down on the customer, and the non-enforcement of already weak laws (Yurtseven, 2015).

(Dike, Obiora, Nwokorie, & Dike, 2015) in their studies of Nigerian customers found that their attitude is a significant hindrance given that they see nothing wrong with electricity theft as no one actually gets caught for it. Technology-enabled power consumption tracking will help in capturing electricity theft as well as changing consumer behavior. Technology interventions by public service firms like power utility organizations will help to manage DCB and the loss that utility firms have to face due to NTL.

2.2 Technological Interventions and Consumer Behavior

In recent years, data analytics has attracted the attention of both academia and industry. It has found use in different disciplines, such as engineering, medicine, psychology, agriculture, and the power industry (Watson, 2013). Analytics has witnessed four growth phases. The current (2021) phase is defined as Analytics 4.0, where cognitive technologies such as machine learning have emerged as the keystone for strategic decision-making (Bughin et al., 2017; Insights, 2018). Analytics 4.0 includes the application of AI methodologies and a greater degree of autonomy in the methods' execution, including autonomous machine learning techniques (Subramanian, 2006).

A growing body of research shows that energy savings can be achieved by approaches that aim to modify people's behavior. Behavioral models are necessary to understand what consumers do and why they do. The theory of the economics of criminal behavior "stipulates those offenders are utility-maximizing agents who weigh subjective benefits and costs of offenses so that offenses are committed when the gains are more than cost." This theory also fits the homo-economics viewpoint, which states that economic agents, including corrupt consumers and officials from utilities firms' crime gatekeepers, are also the self-interested individuals who weigh personal gains and costs in their engagement with everyday economic goods or decisions (Jamil & Ahmad, 2019).

Psychology and human resources research have considered the effect of social norms as forms of social control and how peer contexts affect social perceptions of crime(Goldstein, Griskevicius, & Cialdini, 2007). If a customer knows that they are being watched and if they are caught, then the gains will be far less than the loss, they are more likely to avoid DCB (Loroz, 2006). This implies that if a consumer is caught indulging in unethical practices, the rubbing effect will come into play, deterring other consumers in the neighborhood from DCB (Ahmad et al., 2020).

2.3 Related Work

Hardware- and non-hardware-based methods are used in energy theft detection (ETD). Utilities use hardware methods, for instance, additional hardware equipment such as wireless sensors, RF meter readers, communicable distribution transformers meters, and smart meters to detect electricity theft. Non-hardware methods include game-based, and AI, ML-based systems (Glauner et al., 2016).

In non-hardware-based methods, game theory methods use gaming between power thieves and utility firms to gain data and then conduct data analysis to analyze the difference in the power consumption behavior of power thieves and law-abiding users. The challenge in using the game theory method is to know the players' behavior and utility functions of stakeholders; however, these are quite dynamic (J. Li & Wang, 2020). ML methods are used in smart meters to analyze data from smart meters to detect consumption anomalies. The problem with the ML method is that the performance of ML varies in case there is an imbalance in data (Kaur, Pannu, & Malhi, 2019).

In the context of developing economies, loss-making power utility firms have yet not found the right means to achieve optimal prevention efficiency of energy theft (Gunturi & Sarkar, 2021). Of late, ML techniques have gained acceptance to detect ETD as they are cost-effective and not intrusive. ML is categorized into unsupervised techniques (clustering) and supervised techniques (classification). Our work uses supervised techniques to classify a consumer as either theft or no theft. Based on the limitations discussed above, we use the Neural network and Random Forest model to achieve greater predictive accuracy.

2.4 Searching and Classification of Literature

The academic literature is searched using the keywords presented in Table 1. The independent exploration of the ML algorithm is executed on Web of Science and Google scholar using the "and" and "or" operators. Table 1 presents the keywords used in this exploration, and these can be copied and pasted in the advanced search section on www.webofscience.com, www.ebsco.com to see the results. The articles that appear on the execution of this search will vary as the Web of Science, EBSCO, and Google scholar database are actively updated. The literature review was conducted on Web of Science, EBSCO, and Google scholar on 20 May 2021 to identify the key articles for this exploration. The search process is presented in Figure 1. The search resulted in 126 journal papers, all of which have been used in the current study. We first searched for the keywords related to the ML algorithm and power theft, utility firms, deviant consumer behavior (see Table 1). A total of 1,421 documents appeared in this search.

The second stage is executed considering research articles up to May 2021, resulting in 1,511 documents. Researchers, except in the computer science domain, prefer to publish in journals rather than conferences (Derntl, 2014; Hermenegildo, 2012). Therefore, we do not include papers published in conference proceedings and limit the search to articles, articles in the press, and review papers in the third stage. These articles are restricted to the subject area of business management, utility firms & public services, and computer science. Only the articles published in the English language are considered for the study. The above criterion resulted in 261 research articles. We did not find any paper on traditional meters that used ML algorithms. Finally, we only included the articles relevant to the current study and these were 148 in number.

3. RESEARCH METHODOLOGY

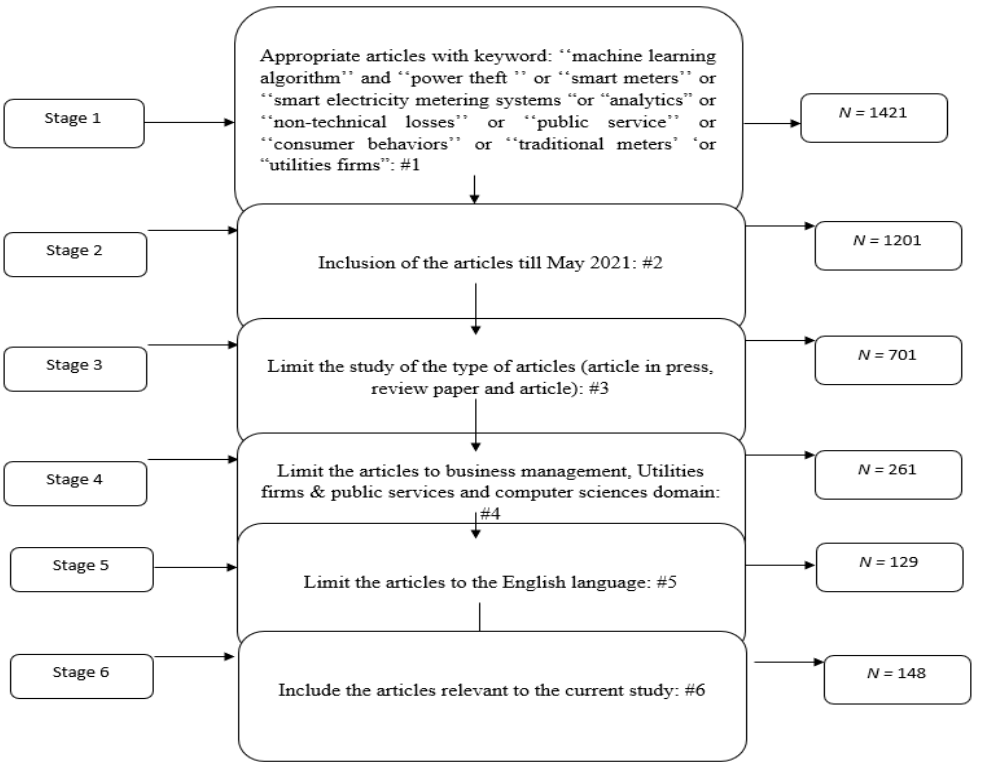
3.1 Context: Power Theft in India

In the context of increasing energy demand, electricity theft has been anathema for energy utility companies in emerging economies. Energy theft is a common problem in India, where energy consumption has been rising steadily, in tune with its population growth. It's impossible to control and solve theft by door-to-door visits for every customer (Oliveira et al., 2001). Electricity can be stolen

Table 1. Keywords used in this study

Keyword	
And	"Machine learning algorithm"
Or	"Power theft"
Or	"Non-technical losses"
Or	"Technical losses"
Or	"Power utilities company"
Or	"Traditional meters"
Or	"Smart meters"
Or Or	"Consumer Behaviors" "Deviant Consumer Behaviors"
Or	"Public service"
Or	"Power grids"
Or	"Non-theft"
Or Or Or Source(s): Author's compilation.	"Artificial intelligence" "Analytics" "Utilities"

Figure 1. Stages of data collection



Source(s): Author's compilation (May 20, 2021.)

in a variety of ways. In India, consumers frequently draw electricity by connecting a line to utility poles. This is theft because the wire is not attached to a meter, and the consumption is not recorded. Even when electricity meters are installed, users frequently defraud electricity utility companies by tampering with meters with magnets to display lower consumption than otherwise (Depuru, Wang, & Devabhaktuni, 2011; Smith, 2004). It is estimated that various forms of electricity theft constitute 20–25% of the generated power in India – an annual cost of INR 20,000 crores (2.7 million USD) (Kumar & Sharma, 2017). Power theft affects the country's GDP hard. According to the World Bank, electricity theft has caused electricity supply losses to exceed 25% of India's supply, 16% in Brazil, and 6% in China (Aryanezhad, 2019). The impact of NTL is also significant in developed countries, and electricity theft is estimated at £173 million every year in the UK. It may be worth up to $6 billion in the USA (Jokar, Arianpoo, & Leung, 2015). In a six-state survey on energy access in India conducted by the New Delhi-based Council on Energy, Environment, and Water (CEEW) in 2018, around 94% of the rural respondents stated that electricity theft is an illegal activity and should be stopped. At the same time, 29% reported that stealing exists in their village (S. Jain, Choksi, & Pindoriya, 2019). Existing policies have failed to curb electricity theft, the major point of focus of the Saubhagya scheme. The Indian government initiative announced in 2017 aimed to electrify all households by the end of 2018.

3.2 Traditional/Analog Vs. Smart Electricity Metering Systems

Traditional meters are electromechanical electricity meters and are also defined as analog meters. The data in analog meters are collected manually and monthly. Smart meters are digital electricity meters that accurately and efficiently measure electricity consumption at 15 minutes intervals and other parameters, such as maximum demand, power factor, and current in various phases (Blazakis, Kapetanakis, & Stavrakakis, 2020).

Smart meters, despite offering several benefits, put additional economic pressures on loss-making utility firms in emerging economies because their initial cost of investment is relatively high. Smart meters have a shorter life expectancy (5 to 7 years), whereas traditional meters have a long life of approximately 20 to 30 years (Weaver & Solutions, 2017). In the case of smart meters, application and maintenance utility firms incur costs in personnel training and equipment development. Consumers in emerging economies are reluctant to use smart meters due to the cost of installment (of meter which is normally passed to them) and concerns about the privacy of their data. Utility firms also face the issue of managing, storing, and analyzing the huge amount of data collected through smart meters (Taft & von Prellwitz, 2012). The cost of replacing an analog meter with smart meters is approximately ten times the cost of analog meters. (Table 2 analog meter cost, Table 3 smart meter cost).

After analyzing the data from Tables 2 & 3, it is clear that replacing analog meters is an economic pressure for both utility firms and customers. One of the main advantages of smart meters is that they can detect NTL. If somehow, this enhanced capability of NTL detection can be added to the traditional meters, then the urgency to make this investment will be greatly reduced. This paper shows that such an ideal situation is possible where applying ML techniques to traditional meters data. This solution would be a boon to the cash-strapped utility firms in India. In our study /experiment, we implement Random Forest with Principal component analysis (PCA) and Neural Network ML-based theft detection model. We use PCA to reduce the dimensionality of data as meter readings are correlated. Fewer dimensions will capture all the nuances of a consumer's consumption and help us keep our model simple, in line with the principle of theory, construction, or evaluation (Akai, 2018).

4. OUR APPROACH

We create a theft detection algorithm by applying ML methodologies to real data (Gholami et al., 2021). The methodology consists of the following steps:

In the following sections, we describe the way we carried out each of these steps:

Table 2. Analog Meter cost

Analog Meter Cost(in Indian Rupee INR)			
S.no		**For Single Phase meter**	**For Three Phase meter**
1	**Meter Cost**	624	1876
2	**Labour cost**	163	222
3	**Total**	778	2098
4	**Meter Reading cost**	INR 6 – 8 per meter reading / per month	

Table 3. Smart meter cost

Smart meter cost (in Indian Rupee INR)			
S.no		**For Single Phase meter**	**For Three Phase meter**
1	**Smart meter**	2500	3500
2	**Installation cost**	500	800
3	**Meter reading cost -network charges:**	20-25 INR per meter/ per month	
4	**Other cost (software etc.)**	2500 per node	250 per node
	Total	INR 7700	8700
	Note	Lump sum rates. The approximate rate per node (Opex + Capex) come around INR 7700 (for 1-ph) and INR 8700 (for 3-ph) considering 7.5 years operation period @30 per node /month	

Source: MPPKVVCL

4.1 Data Sourcing and its Description

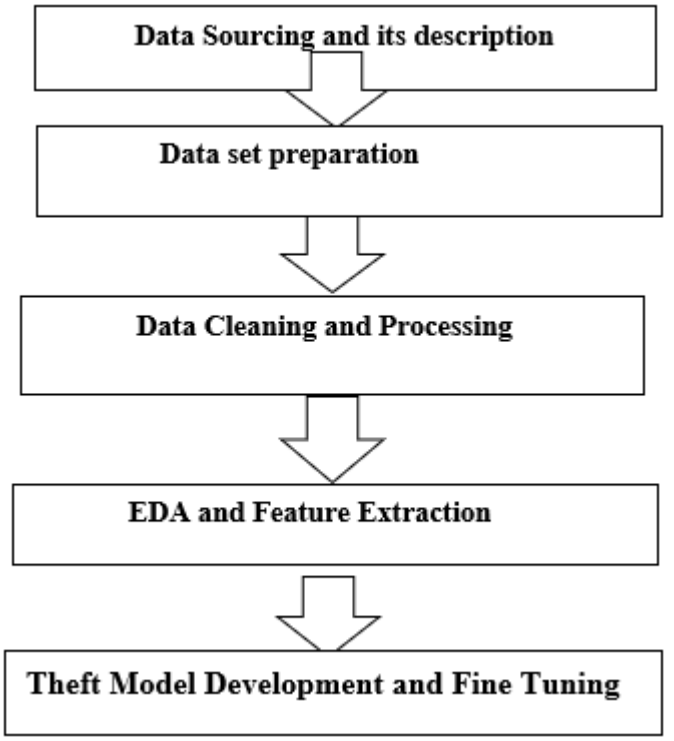
We use real electricity consumption data from Indore, the largest city in India's centrally located province of Madhya Pradesh, as a base for the study. This data is provided by the Madhya Pradesh Paschim Kshetra Vidyut Vitran Company Limited (MPPKVVCL), a western distribution company. For brevity, we will refer to the company as MPPKVVCL, a utility company responsible for distributing electricity in the city of Indore. The consumption data here is based on the manual reading of meters conducted by meter readers, who often work on contracts. The meters are often poorly located in various homes with difficult-to-read conditions.

Moreover, the whole process is also prone to corruption, where either a meter reader may under-report meter reading or report it in a manner to cause benefit to the consumers in order to benefit from the various incentive schemes the company offers. Thus, the consumption data is very noisy. The data we received contained electricity consumption details and other information about 6 14,757 consumers throughout January 2018 till December 2020. Additionally, we received theft data (labeled data) from 900 consumers.

4.1.1 Dataset Preparation

Data preparation is a self-service operation that transforms diverse, raw, and jumbled data into a tidy and uniform perspective. Searching, cleaning, converting, organizing, and collecting data are all part of the process (Lei, 2021). We filtered cases of direct theft and meter tampering irregularities and were left with 398 cases of theft from all the zones. Here, direct theft involves bypassing the electricity

Figure 2. Our approach



meter. In contrast, meter tampering refers to the case where the current is still passing through a meter that has been tampered with to under-report electricity consumption. The company has divided the city of Indore into 33 zones for administrative purposes. These zones also differ significantly in their social and economic parameters and so exhibit different behaviors in terms of electricity theft. Figure 3 shows the distribution of theft in different zones.

Energy theft in Indian cities is concentrated in a few areas. For example, in the Indian capital Delhi, 60% of all energy theft cases are from places such as Najafgarh, Burari, Bawana, and Azadpur. In the metro city of Kolkata, areas like Park Street and Shakespeare Sarani constitute the majority of energy theft (Mahalakshmi, Nikhitha, & Varsha, 2018). This situation finds reflection in Indore too.

We analyzed Indore city zones. Figure 3 shows zones of GPH WEST & DALY COLLEGE have the highest number of theft cases. We filtered the data from these two zones for further study and then modeled it for further analysis. These two zones have 44,177 consumers and collectively report 99 cases of electricity theft.

We can see from figure 3 that the Daily College zone has 25,712 consumers, and the GPH WEST zone has 18,465 consumers.

4.2 Data Cleaning and Processing

This is an essential step of ML, and its goals are to transform the raw data into an understandable format and improve prediction accuracy. Data cleaning is essential because raw data is hardly clean or complete. The data cleaning step identifies the missing values and inconsistencies in the dataset. The missing value is defined as the value not present in the cell of a particular column. These missing values are either discarded or imputed as missing data. (Dogan & Birant, 2021). Challenges like

Figure 3. Power theft in different zones

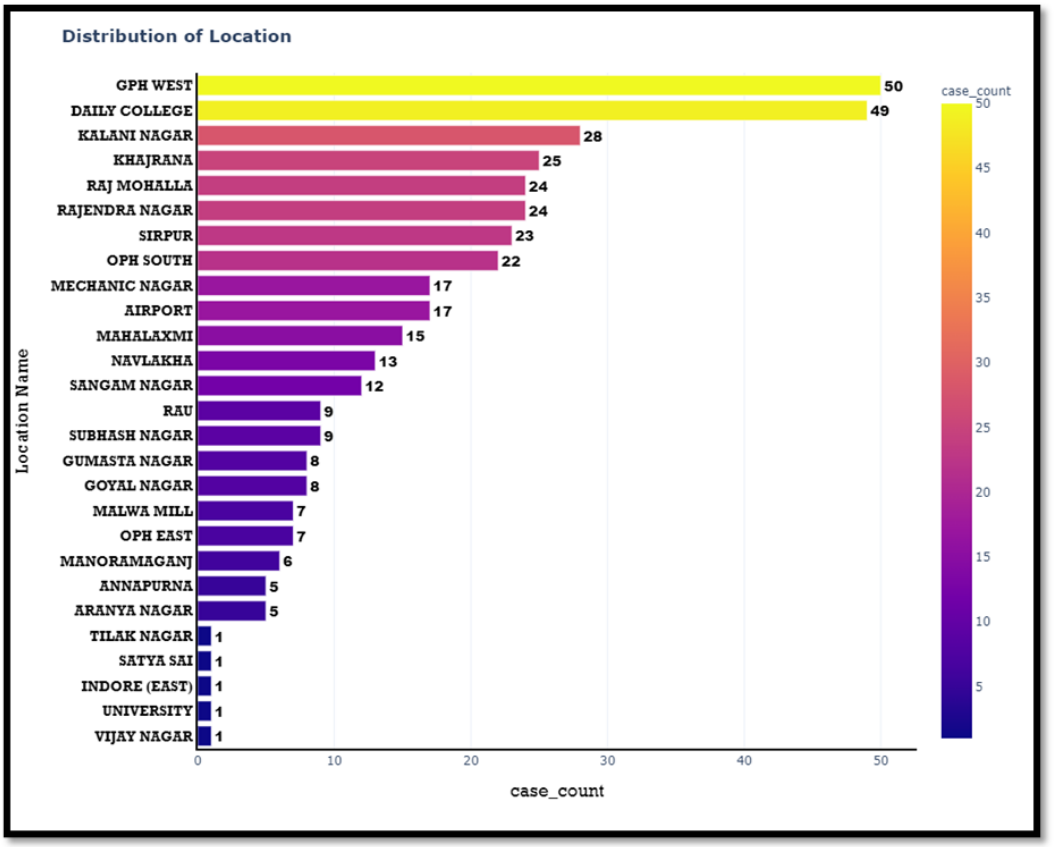


Table 4. location and count

Location Code	Count
3424502 (Daly College)	25,712
3424407(GPH West)	18,465

missing values for some of the variables, seasonality of data, and theft data preparations are dealt with in the data preprocessing.

We handled each of the challenges in the following manner:

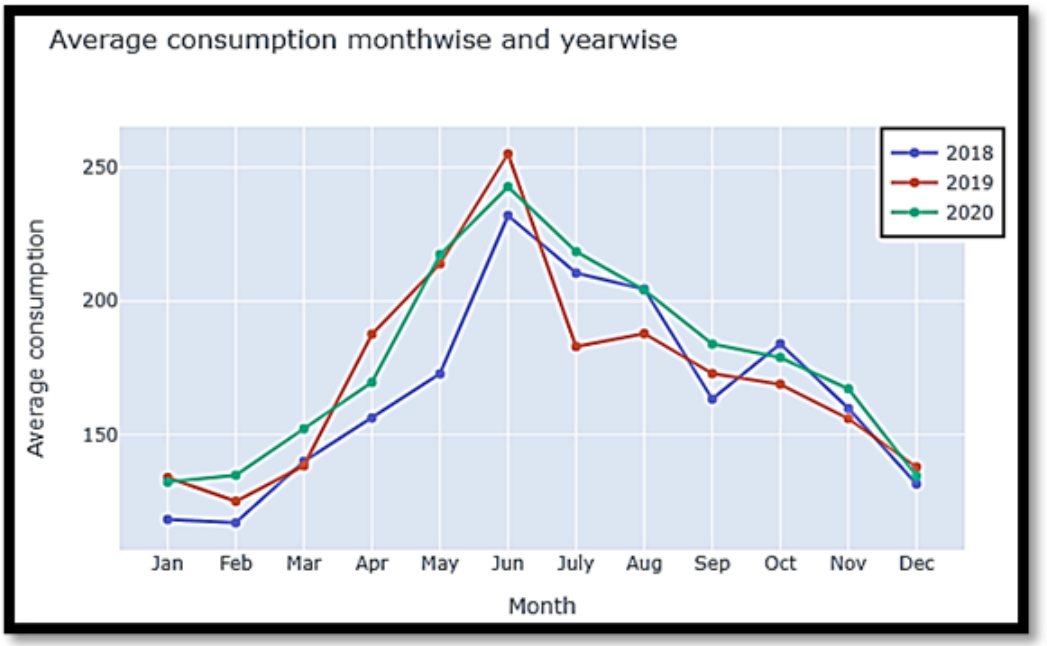
4.2.1 Imputation of Missing Values

Missing values in the meter readings occur as it is a manual meter reading process, and this could occur either because the reader could not visit the premises (like lockdowns due to Covid) or found them locked during their visit. In such cases, the average of the readings dropped the rows in which the group and/or reader were missing. Dropped cases where the reading type was PFL (Premises Found Locked), and also, we dropped cases where the meter was replaced more than two times. For treatment of missing values, we placed missing values by mean of their previous and next value in total unit columns. Finally, we were left with 40,813 consumers from 44,177 consumer data.

4.2.2 Seasonality of Data

Seasonality is a time series feature in which the data undergoes regular and predictable changes that repeat each calendar year. Seasonal refers to any predictable fluctuation or pattern that recurs or repeats over one year (Stolwijk, Straatman, & Zielhuis, 1999).

Figure 4. Seasonality of Data



The graph in Figure 4 shows the monthly average consumption in 3 years. There is seasonality in the data where the consumption is higher in May, June, and July as these are the summer months. It is usually low in the winter months of December and January as people don't require extensive heating given the tropical climate. We found seasonality to be a double-edged sword in using electricity consumption data to detect theft. While, on the one hand, lack of seasonality in a consumer's electricity consumption data may indicate theft, it could also hide the fluctuations in energy consumption occurring due to theft. Therefore, we decided to remove seasonality and trend from each user's time series of consumption separately from the model data of 197 consumers and then create additional features to take care of variations due to seasonality and otherwise. These features will be described in the subsequent sections.

4.3 Data Set Preparations (Theft, Non-Theft)

Data set preparation is an important step in the ML process. It is a set of procedures that helps make our data more suitable for ML. The data preparation also includes establishing the right data collection mechanism (Corbeil, Williams, & Labute, 2012). For theft data, we assume that, after the theft is detected, the consumer turns into a non-theft case from the month of theft detection.

This assumption may not always be true, as some consumers can redeploy the theft mechanism soon after company inspectors leave. However, in the absence of any further confirmatory data, we decided to stay with this assumption. This assumption requires us to estimate the theft reading for

this consumer. We estimate this by using the consumption data before theft and taking a month-wise average of theft months to estimate non-theft months. We then forecast the consumption observations after the date of Panchnama (A written account of a transaction that occurred between two or more than two individuals as narrated by the persons who were witnesses. (Yadav & Mohania, 2017)) to get a whole series of theft data for 99 consumers (who have stolen in the past). We randomly select 100 non-theft data and create our knowledge dataset by taking all the theft cases about two zones thus estimated. We then choose an equal number of non-theft cases randomly from the remaining data.

4.4 Exploratory Data Analysis (EDA) and Feature Extraction

Exploratory Data Analysis (EDA) is a critical process for conducting preliminary investigations on data to discover patterns, detect anomalies, test hypotheses, and validate assumptions using summary statistics and graphical representations. (Abukmeil, Ferrari, Genovese, Piuri, & Scotti, 2021; Hakak et al., 2021; Q. Li et al., 2021).

Figure 5 describes the statistical dispersion of monthly consumption readings. For example, the first reading describes the distribution in the consumption data for January 2018. We have total data points for about 44177 consumers. The mean consumption is 118 units, while the standard deviation is 158 units. Mean being much higher than the median of 88 indicates few consumers with very high consumptions, with the highest consumption being 13228 units. There are about 11000 consumers with consumption of more than 144 units. We also notice that all these readings vary significantly when we compare with June 2018. It is important to note here that this is an aggregate behavior, and consumer behavior can be very different varying from one individual to another. For example, for some homes, the consumption in summer months may not go up simply because either they travel away from their homes during summer months or lack the gadgets like ACs, fans, refrigerators, etc., which cause consumption to go up during summer months.

4.5 Feature Extraction

Our first set of features is obtained by decomposing consumption readings into trends, seasonality, and residuals and keeping residuals as feature values. Additionally, we create additional features (Table 5) drawing on the current practice of company officials as well as to the account of seasonality and other variations in the data:

Moreover, the tendency to steal occurs in a similar neighborhood. If a consumer is caught stealing electricity, there are increased chances that more consumers are doing the same in the same neighborhood (Higgins, Ricketts, & Wolfe, 2008). Our database did not have data about the consumers' location in a manner that allowed us to compute their closeness. But we did have the data about the meter reader who goes and take the manual reading. Using the assumption that the same meter reader must be assigned to the entire neighborhood, we decided to use this information as the proxy for that neighborhood. We calculate a variable which counts the number of theft cases in the same meter reader code and assigns it to the consumer. This is our final feature set, which we denote by (c). Thus, our entire feature set available is (a) + (b) + (c). We scale these feature values using Standard Scalar, which scales by subtracting the population's mean from the observation and then dividing the difference by the population's standard deviation.

Table 6 contains the p-values of these variables tested against the null hypothesis that they are individually not significant in determining whether a particular consumer can be suspected of theft. We can see that they all are insignificant in determining that except for the last variable of the group reader flag.

However, we will see later that some of these features play an essential role in determining suspicious cases.

Figure 5. Statistical dispersion of monthly consumption readings

	count	mean	std	min	25%	50%	75%	max
TOT_UNITS_2_JAN_2018	44177.0	118.192838	158.351299	0.0	52.00	88.0	144.00	13228.0
TOT_UNITS_2_FEB_2018	44177.0	116.990221	183.792539	0.0	50.00	86.0	142.00	20000.0
TOT_UNITS_2_MAR_2018	44177.0	139.956312	159.456091	0.0	64.00	105.0	172.00	8916.0
TOT_UNITS_2_APR_2018	44177.0	156.219356	172.633247	0.0	76.50	120.5	189.50	12211.0
TOT_UNITS_2_MAY_2018	44177.0	172.749689	217.183358	0.0	81.00	130.0	207.00	18711.0
TOT_UNITS_2_JUN_2018	44177.0	232.001494	355.229028	0.0	100.00	168.0	274.00	32456.0
TOT_UNITS_2_JUL_2018	44177.0	210.442606	240.635493	0.0	96.00	157.0	257.00	17400.0
TOT_UNITS_2_AUG_2018	44177.0	204.482853	231.666746	0.0	96.00	154.0	254.00	19200.0
TOT_UNITS_2_SEP_2018	44177.0	163.290332	193.630016	0.0	80.00	124.0	200.00	13320.0
TOT_UNITS_2_OCT_2018	44177.0	183.964959	218.118160	0.0	86.00	138.0	225.00	12538.0
TOT_UNITS_2_NOV_2018	44177.0	159.755508	196.500909	0.0	74.00	121.0	199.00	18934.0
TOT_UNITS_2_DEC_2018	44177.0	131.508953	160.749203	0.0	58.00	100.0	159.00	10000.0
TOT_UNITS_2_JAN_2019	44177.0	133.842973	412.103591	0.0	50.00	92.0	160.00	78333.0
TOT_UNITS_2_FEB_2019	44177.0	125.032189	179.548683	0.0	50.00	91.0	151.00	20139.0
TOT_UNITS_2_MAR_2019	44177.0	138.289676	171.297251	0.0	57.00	102.0	172.00	8879.0
TOT_UNITS_2_APR_2019	44177.0	187.569323	204.977179	0.0	80.00	143.0	239.00	11976.0
TOT_UNITS_2_MAY_2019	44177.0	213.930462	234.525941	0.0	86.00	162.0	269.00	11400.0
TOT_UNITS_2_JUN_2019	44177.0	255.079725	273.530076	0.0	100.00	194.0	323.00	11972.0
TOT_UNITS_2_JUL_2019	44177.0	182.942662	226.604726	0.0	42.00	137.0	250.00	12034.0
TOT_UNITS_2_AUG_2019	44177.0	187.816262	197.701136	0.0	83.00	145.0	239.00	7794.0
TOT_UNITS_2_SEP_2019	44177.0	172.823936	169.072339	0.0	83.00	136.0	219.00	8772.0
TOT_UNITS_2_OCT_2019	44177.0	168.829934	174.790814	0.0	81.00	135.0	210.00	10588.0
TOT_UNITS_2_NOV_2019	44177.0	155.981800	165.339471	0.0	78.00	126.0	190.00	12959.0
TOT_UNITS_2_DEC_2019	44177.0	137.819250	139.394230	0.0	67.00	111.0	167.00	7200.0
TOT_UNITS_2_JAN_2020	44177.0	132.196030	161.433095	0.0	72.00	104.0	150.00	19125.0
TOT_UNITS_2_FEB_2020	44177.0	134.767730	151.662401	0.0	87.00	100.0	154.00	14359.0
TOT_UNITS_2_MAR_2020	44177.0	152.177128	176.470083	0.0	100.00	106.0	172.00	19022.0
TOT_UNITS_2_APR_2020	44177.0	169.545895	167.079462	0.0	100.00	121.0	197.00	11332.0
TOT_UNITS_2_MAY_2020	44177.0	217.410934	216.452044	0.0	100.00	163.8	271.65	9214.0
TOT_UNITS_2_JUN_2020	44177.0	242.791240	269.657818	0.0	100.00	174.2	299.66	10048.0
TOT_UNITS_2_JUL_2020	44177.0	218.508439	276.403431	0.0	96.00	150.0	254.73	11438.0
TOT_UNITS_2_AUG_2020	44177.0	204.126188	253.560020	0.0	97.00	151.0	248.00	18967.0
TOT_UNITS_2_SEP_2020	44177.0	183.905563	204.435769	0.0	97.00	144.0	226.49	9072.0
TOT_UNITS_2_OCT_2020	44177.0	178.870939	188.168130	0.0	97.02	141.0	212.00	9930.0
TOT_UNITS_2_NOV_2020	44177.0	167.125514	178.053581	0.0	90.00	130.0	202.00	7488.0
TOT_UNITS_2_DEC_2020	44177.0	134.478454	156.033137	0.0	73.00	101.0	159.37	7696.0

4.6 Theft Model Development and Fine-tuning

Having extracted the features, we now develop the model for detecting theft. The model development consists of the following steps:

4.6.1 Features selection
4.6.2 Creating training and test datasets
4.6.3 Train a model and test its performance
4.6.3.1 Random Forest
4.6.3.2 Neural Network

Table 5. Additional features to capture seasonality and other variations

S. No	Feature	Description
1	Current 6 Months compared with Previous 6 months	Sum of ratio of (previous 6 months average consumption - current 6 months average consumption)/ current 6 months average consumption
2	Current Month compared with Previous 6 months	Sum of ratio of (previous 6 months average consumption - current month consumption)/ current month consumption.
3	Current Month compared with Previous month	Sum of ratio of (previous month consumption – current month consumption)/ current month consumption
4	Current Month compared with same month Previous Year	Sum of ratio of (previous year month consumption – current year month consumption)/ current year month consumption
5	Count of Zero	Count of zero in three years of consumption for each consumer
6	Current Year compared with Previous Year	Sum of ratio of (previous 12 months average consumption - current 12 months average consumption)/ current 12 months average consumption
7	Upper outlier count	Upper outlier count is how many times consumption falls more than above the upper quartile.
8	Lower outlier count	Lower outlier count is how many times consumption falls more than below the lower quartile
9	Abnormal High	If Current month consumption greater than or equal to 3*(average of last six-month consumption) then count it as abnormal high
10	Abnormal Low	If Current month consumption less than or equal to (average of last six-month consumption)/3 then count it as abnormal low
11	Last 6-month comparison with initial 6 months consumption	Sum of ratio of (first 6 months average consumption - last 6 months average consumption)/ last 6 months average consumption
12	Coefficient of Variation	Standard deviation of all the 3 years / Average consumption of all the 3 years
13	Group Reader Flag	If theft found in a group and reader in theft data, we flag those group reader as 1 otherwise 0

Table 6. P values

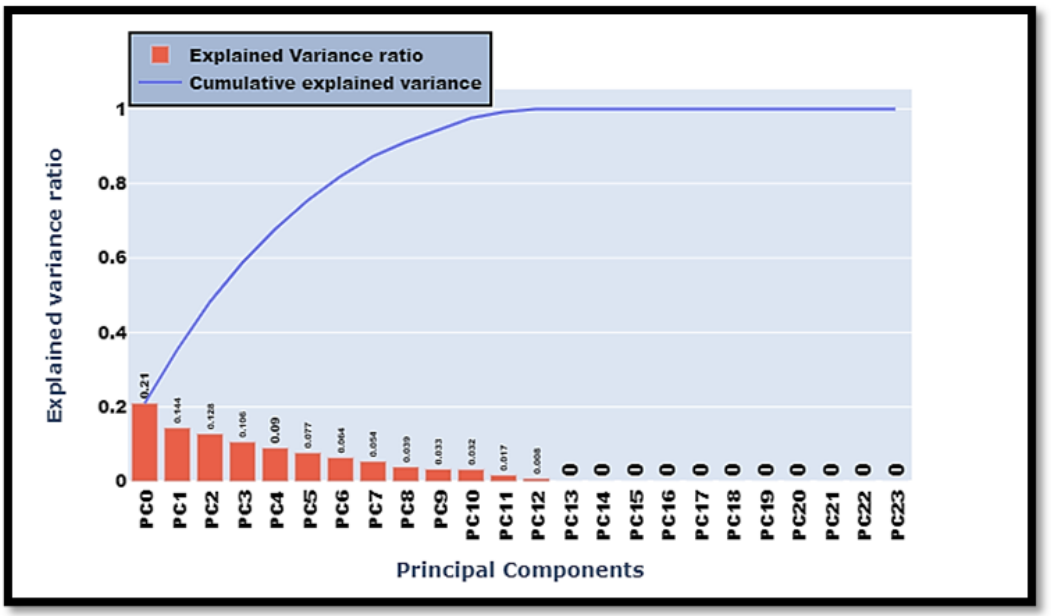
Features	P-Value	Significant (Yes/No)
Current 6 Months compared with Previous 6 months	0.91	No
Current Month compared with Previous 6 months	0.36	No
Current Month compared with Previous month	0.3	No
Current Month compared with same month Previous Year	0.49	No
Current Year compared with Previous Year	0.65	No
Upper outlier count	0.49	No
Lower outlier count	0.76	No
Abnormal High	0.07	No
Abnormal Low	0.06	No
Last 6-month comparison with initial 6 months consumption	0.15	No
Coefficient of Variation	0.81	No
Group Reader Flag	0	Yes

4.6.1 Feature Selection

Since we have many features corresponding to the meter reading (36 in total), which are also likely to be correlated, we reduce the number of elements to make our model simple as per Occam's Razor. Here, we use the Principal Component Analysis (PCA) to select the features. The graph in Figure 6 shows the principal components and their explained variance.

From the graph, we can see that only 12 principal components explain 99% of the variance. Thus, we choose the 12 components (from PC0 to PC11). The total explained variance is 0.991.

Figure 6. PCA



4.6.2 Creating Training and Testing Dataset

After creating our knowledge dataset and following the ML practice, we split the dataset using the train test split procedure. This is essential to conduct an unbiased evaluation of prediction performance. The data set is split into two subsets, and the first subset, which is the training dataset, is used to fit the model. The second subset, referred to as the test dataset, is used to predict the value of the dependent variable and compare it to the already known value to determine how well our trained model is performing. The splitting ratio we use is 80/20 (80% training data and 20% testing data).

4.6.3 Training of the Model and Testing its Performance

There are many machine learning classification algorithms available to train a classification model. Some of the widely known ones are Logistics Regression, Naïve Bayes, Support Vector Machines, Decision Trees, Random Forest, and Neural Networks. We choose Random Forest and Neural Network to train our classification model. Random Forest is known to work very well in a broad class of practical problems with a lot of noise in the data. Neural Networks help with creating additional features from the data which could have been missed. Now, we describe their usage in more detail.

4.6.3.1 Random Forest

"Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. The more significant number of trees in the forest leads to higher accuracy and prevents overfitting." Antony Christopher

On applying this algorithm to our data, we achieve an accuracy of 100% on training and 88% on the testing dataset. We make our feature selection after applying baseline random forest. Random Forest also gives us the list of features that play an essential role in the classification. The list is shown in figure 7.

As we can see, the essential features based on the Gini index with an importance of more than 0.05 are Group Reader Flag, PC8, PC5, PC3, and PC7.

Moreover, we optimize the hyper parameters for Random Forest using Grid search CV. The final optimal parameters are:

o n_estimators=100
o max_depth=4
o min_samples_leaf=20,
o min_samples_split=50,
o max_features=4

After fine-tuning, the "Random Forest with PCA" model gives an accuracy of 100% on training & 91% on the testing dataset. Moreover, we compute the Out-of-bag score, AUC score, and sensitivity.

Cross-validation score and Out-of-bag score are computed as the number of correctly predicted rows from the out-of-bag sample. The score is 90%, which is good. AUC (Area under the Curve) score is a metric for capturing the prediction accuracy of the model. The score is 91%, which is good.

The sensitivity of a test is its ability to determine the theft cases correctly. The sensitivity is 95%, which is good. The specificity of a test is its ability to determine the non-theft cases correctly. Our specificity is 79%. Cross-validation is a resampling procedure used to evaluate models on a limited data sample. The cross-validation score using K-fold is 90%, Stratified K-Fold is 88%, and Leave One Out is 90%.

4.6.3.2 Neural Network

It is for us to design a neural network based on our understanding of the problem at hand. After several trials and errors, we use the final neural network model of 23 nodes at the input layer. The first hidden layer has seven nodes and uses the rectified linear activation function, called the Re LU activation function. The second hidden layer has five nodes and uses the Re LU activation function. The output layer has one node and uses the sigmoid activation function. We used cross-entropy as the loss argument. This loss is for a binary classification problem and is defined in Keras as binary cross-entropy. The optimizer is the efficient stochastic gradient descent algorithm. This is a popular version of gradient descent because it automatically tunes itself and gives good results in a wide range of problems. Epoch is one pass through all the rows in the training dataset. We could fit the model for 1000 epochs. A batch is one or more samples considered by the model within an epoch before weights are updated. We have a fit model for batch size 1.

The "Neural Network" model gives an accuracy of 100% on training & 75% on the testing dataset. Moreover, similar to the random forest conducted before, we estimate several more performance parameters. The AUC (Area under the Curve) score is a metric for capturing the prediction accuracy in the model. The score is 76%.

We conduct a sensitivity test to check the ability of our model to determine the theft cases correctly. The sensitivity is 67%. and the specificity is 85%.

Figure 7. List of features

features	importance
grp_reader_Flag	0.361575
8	0.063296
5	0.051378
3	0.048426
7	0.046146
4	0.040014
0	0.032124
Current Year is 50% lower than Pre Yr	0.030877
2	0.030153
6	0.029796
11	0.028840
Current 6 Months is 50% lower than Pre 6 months	0.028549
Coeff_of_Variation	0.028480
9	0.025170
1	0.023863
10	0.022936
Abnormal High	0.020927
Current Month is 50% lower than Pre month	0.019843
Current Month is 50% lower than same month Pre Yr	0.019804
Current Month is 50% lower than Pre 6 months	0.018020
Last 6 month comparision with intial 6 months ...	0.014646
Abnormal Low	0.009548
Lower outlier count	0.003050
Upper outlier count	0.002540
Count of Zero	0.000000

Figure 8. ROC curve for random forest based algorithm

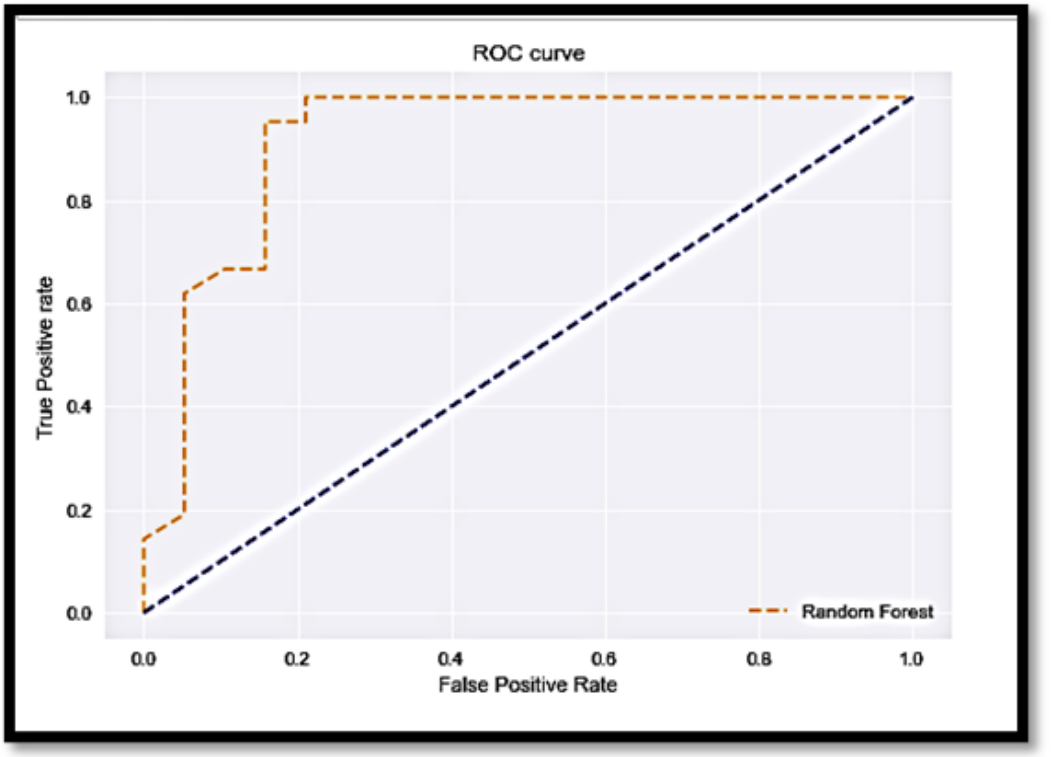
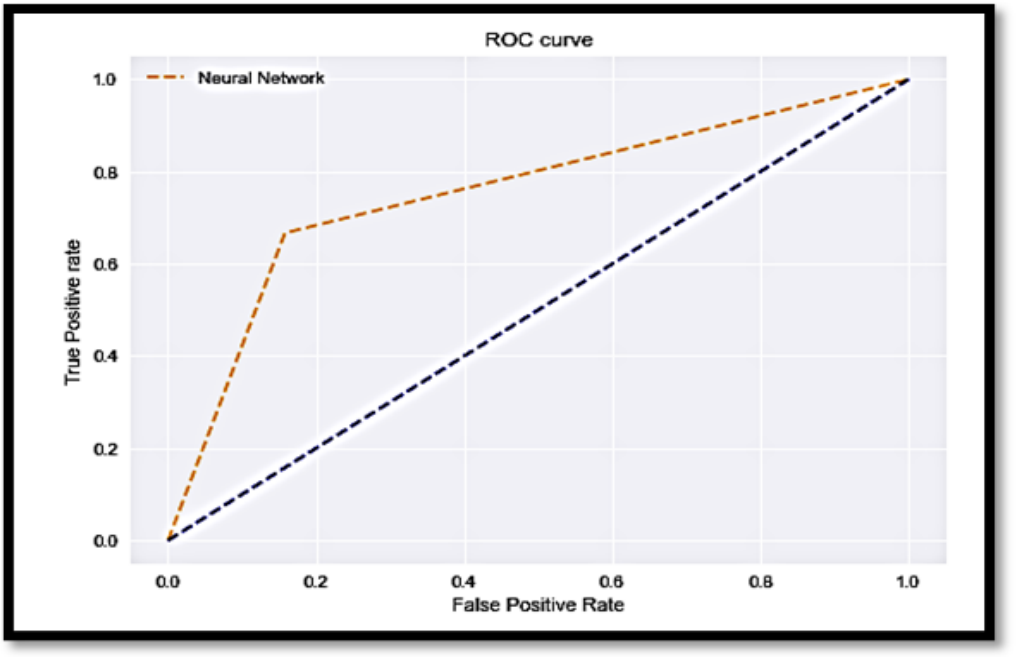


Figure 9. ROC curve for neural network based algorithm



5. CONCLUSION, LIMITATIONS AND FUTURE SCOPE OF RESEARCH

As described in section 4.2, the data was very noisy due to infrequent and irregular manual readings and possible incentives for corrupt behavior for the reader. We performed imputation of missing values in the data by using means of the adjacent values.

Moreover, data had seasonality due to the normal behavior of electricity consumers in India, where the consumption is more in the hot summer months and less in the less intense winter months. We use time series decomposition to remove seasonality as it would interfere with detecting theft, as described in section 4.2.

Section 4.5 described various ways to create features that will then serve as inputs for the machine learning models. We used two sets of features: one offset from the time series decompositions and the other emerging from current practices, where we used current practice as our basis to create valuable features.

The key challenge we faced in this research was with regards to our key premise, which was that theft would generate spikes and troughs in the meter consumption data, and which we would be able to detect. However, these spikes and troughs could also be attributed to the more normal behavior exhibited by the consumers under various scenarios, for instance, weather conditions (as we discussed earlier), people moving in or out, or the purchase of new appliances like refrigerators, ACs and so on. Our main contribution to literature is the creation of the ML model, which can distinguish between the crests and troughs generated due to these two types of consumer behavior. The accuracy of the model indicates that we are successful in achieving our aim.

The main limitation of this research is the lack of field verification of the theft prediction generated by our model. The same model creation exercise needs to be conducted for different geographies to validate the approach.

5.1 Impact for Utility Firms

Energy theft results in uncertainty on real consumption at the grid and at the distribution transformer (DTR) level. The use of smart meters has the potential to increase billing and collection efficiencies. However, in the near foreseeable future, the penetration of smart meters will be limited due to costs, consumer behavior, manufacturing constraints, technology adoption, and communication infrastructure limitation.

Utilities are struggling to identify theft in electricity consumption without incurring a large amount of expenditure and adopting a non-intrusive approach. Given the customer base of 16.8 million consumers in Madhya Pradesh, about 98 percent of consumers use traditional meters. It is practically impossible to identify theft by manual analysis. Besides, there are several parameters, which can go wrong in case theft is being committed. However, variation in a single parameter cannot be interpreted as theft with certainty, so we need the intersection of data points indicating theft to be doubly sure. This is not possible by manual analysis, and hence, there is a need to deploy ML techniques to identify theft. The historical data provided by the distribution company itself shows the minimal incidence of identified theft cases despite high NTL losses (up to 40%) in some of the zones of Indore city.

The ML technique deployed with high accuracy, sensitivity, and specificity gives us the confidence that the predicted suspected cases have a high probability of true positives. This will reduce workforce cost in identifying the theft, and lead to higher satisfaction, the better realization of revenue, and increased compliance amongst consumers.

We have used the trained ML model in the city areas to help identify the theft cases with encouraging results. Despite the limitation in the granularity of data from traditional meters, we could achieve about 20 percent success in identifying theft cases from the limited number of cases identified by the model. This is even better than the rule-based analysis deployed to identify theft in about 0.2 million smart meters in the city.

Though the federal government is now stressing to convert all traditional meters with prepaid smart meters, it will still take a long time to achieve this goal. Thus, until they can be replaced, the ML technique can be effectively used by the utilities in developing countries to minimize electricity theft. Based on the success achieved, utilities in MP are keen to use the ML technique on a large scale, and as the model gets trained further, the capability to identify theft will increase manifold.

5.2. Impact on ML Literature

This work applies the ML techniques to real-world data, which is very noisy and sparse in terms of theft cases. It demonstrates that such techniques can be very useful in helping simplify the complicated, challenging, and costly task of reducing NTL in a third-world country with an antiquated metering system. We used several techniques to deal with noisy data and then decomposed them as time series. The results were compared to data obtained using the smart metering system, which costs a lot more to install and run. Thus, we have amply demonstrated the power of machine learning in helping organizations to re-evaluate the kind of investments they need to make to improve and modernize their business processes.

Our model is superior in terms of handling large time-series data and accurate classification. It can be efficiently applied by the utility companies using the real electricity consumption data to identify the electricity thieves and, thus, overcome the major revenue losses they face. With this model, utilities will be able to monitor deviant consumer behavior and reduce their NTL.

The limitation of this research is that we have developed the model for a limited number of customers, mainly due to the lack of sufficient theft data from all the zones under study. This gap can be bridged by further research. There is also a need to further study how this model developed on the basis of data in one region or zone can predict theft in other zones. This study does not consider the change in consumer behavior after the model is implemented. Arresting deviant consumer behavior and increased public service quality due to technology intervention are two areas for future research.

REFERENCES

Abukmeil, M., Ferrari, S., Genovese, A., Piuri, V., & Scotti, F. (2021). A survey of unsupervised generative models for exploratory data analysis and representation learning. *ACM Computing Surveys*, *54*(5), 1–40. doi:10.1145/3450963

Ahmad, A. H., Masri, R., Chong, C. V., Fauzi, R. U. A., & Idris, I. (2020). Evolution of Technology and Consumer Behavior: The Unavoidable Impacts. *EVOLUTION*, *7*(11), 2020.

Akai, M. (2018). Is Occam's Razor Meaningful for Selecting Significant Outcome Items and to Narrow Down Question Numbers in a Psychometric Scale? The Journal of Rheumatology. doi:10.3899/jrheum.180264

Akter, S., Dwivedi, Y. K., Biswas, K., Michael, K., Bandara, R. J., & Sajib, S. (2021). Addressing Algorithmic Bias in AI-Driven Customer Management. *Journal of Global Information Management*, *29*(6), 1–27. doi:10.4018/JGIM.20211101.oa3

Alkawsi, G., Ali, N., & Baashar, Y. (2021). The Moderating Role of Personal Innovativeness and Users Experience in Accepting the Smart Meter Technology. *Applied Sciences (Basel, Switzerland)*, *11*(8), 3297. doi:10.3390/app11083297

Arif, A., Javaid, N., Aldegheishem, A., & Alrajeh, N. (2021). Big data analytics for identifying electricity theft using machine learning approaches in microgrids for smart communities. *Concurrency and Computation*, e6316.

Aryanezhad, M. (2019). A novel approach to detection and prevention of electricity pilferage over power distribution network. *International Journal of Electrical Power & Energy Systems*, *111*, 191–200. doi:10.1016/j.ijepes.2019.04.005

Baker, E. W., Al-Gahtani, S. S., & Hubona, G. S. (2010). Cultural impacts on acceptance and adoption of information technology in a developing country. *Journal of Global Information Management*, *18*(3), 35–58. doi:10.4018/jgim.2010070102

Bhatia, B., & Gulati, M. P. (2004). *Reforming the power sector: Controlling electricity theft and improving revenue*. Academic Press.

Blazakis, K. V., Kapetanakis, T. N., & Stavrakakis, G. S. (2020). Effective Electricity Theft Detection in Power Distribution Grids Using an Adaptive Neuro Fuzzy Inference System. *Energies*, *13*(12), 3110.

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlstrom, P., & Trench, M. et al. (2017). Artificial intelligence: The next digital frontier? Academic Press.

Cheng, L.-C., Hu, H.-W., & Wu, C.-C. (2021). Spammer Group Detection Using Machine Learning Technology for Observation of New Spammer Behavioral Features. *Journal of Global Information Management*, *29*(2), 61–76.

Corbeil, C. R., Williams, C. I., & Labute, P. (2012). Variability in docking success rates due to dataset preparation. *Journal of Computer-Aided Molecular Design*, *26*(6), 775–786. doi:10.1007/s10822-012-9570-1 PMID:22566074

Daunt, K. L., & Harris, L. C. (2012). Motives of dysfunctional customer behavior: An empirical study. *Journal of Services Marketing*, *26*(4), 293–308. doi:10.1108/08876041211237587

de Souza Savian, F., Siluk, J. C. M., Garlet, T. B., do Nascimento, F. M., Pinheiro, J. R., & Vale, Z. (2021). Non-technical losses: A systematic contemporary article review. *Renewable & Sustainable Energy Reviews*, *147*, 111205. doi:10.1016/j.rser.2021.111205

Depuru, S. S. S. R., Wang, L., & Devabhaktuni, V. (2011). Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft. *Energy Policy*, *39*(2), 1007–1015. doi:10.1016/j.enpol.2010.11.037

Derntl, M. (2014). Basics of research paper writing and publishing. *International Journal of Technology Enhanced Learning*, *6*(2), 105–123. doi:10.1504/IJTEL.2014.066856

Dick, A. (1995). *Theft of electricity-how UK electricity companies detect and deter*. Academic Press.

Dike, D. O., Obiora, U. A., Nwokorie, E. C., & Dike, B. C. (2015). Minimizing household electricity theft in Nigeria using GSM based prepaid meter. *American Journal of Engineering Research*, 2320-0936.

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, *166*, 114060. doi:10.1016/j.eswa.2020.114060

Dootson, P., Lings, I., Beatson, A., & Johnston, K. A. (2017). Deterring deviant consumer behaviour: When 'it's wrong, don't do it'doesn't work. *Journal of Marketing Management*, *33*(15-16), 1355–1383. doi:10.1080/0267257X.2017.1364285

EESL. (2021). Retrieved from https://eeslindia.org/en/home/

Fullerton, R. A., & Punj, G. (1997). *What is consumer misbehavior?* ACR North American Advances.

Funayama, Y., Nakamura, K., Tohashi, K., Matsumoto, T., Sato, A., Kobayashi, S., & Watanobe, Y. (2021). Automatic analog meter reading for plant inspection using a deep neural network. *Artificial Life and Robotics*, *26*(2), 176–186. doi:10.1007/s10015-020-00662-y

Ghaedi, H., Tabbakh Farizani, S. R. K., & Ghaemi, R. (2021). Improving power theft detection using efficient clustering and ensemble classification. *International Journal of Electrical & Computer Engineering, 11*(5).

Gholami, R., Nishant, R., & Emrouznejad, A. (2021). Modeling Residential Energy Consumption: An application of IT-based solutions and big data analytics for sustainability. *Journal of Global Information Management*, *29*(2), 166–193. doi:10.4018/JGIM.2021030109

Glauner, P., Meira, J. A., Valtchev, P., State, R., & Bettinger, F. (2016). *The challenge of non-technical loss detection using artificial intelligence: A survey.* arXiv preprint arXiv:1606.00626.

Goldstein, N. J., Griskevicius, V., & Cialdini, R. B. (2007). Invoking social norms: A social psychology perspective on improving hotels' linen-reuse programs. *The Cornell Hotel and Restaurant Administration Quarterly*, *48*(2), 145–150. doi:10.1177/0010880407299542

Gordon, S. (1993). Standardization of information systems and technology at multinational companies. *Journal of Global Information Management*, *1*(3), 5–15. doi:10.4018/jgim.1993070101

Gunturi, S. K., & Sarkar, D. (2021). Ensemble machine learning models for the detection of energy theft. *Electric Power Systems Research*, *192*, 106904. doi:10.1016/j.epsr.2020.106904

Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, *117*, 47–58. doi:10.1016/j.future.2020.11.022

Hasan, M., Toma, R. N., Nahid, A.-A., Islam, M., & Kim, J.-M. (2019). Electricity theft detection in smart grid systems: A CNN-LSTM based approach. *Energies*, *12*(17), 3310. doi:10.3390/en12173310

Hermenegildo, M. V. (2012). *Conferences vs. journals in CS, what to do? Evolutionary ways forward and the ICLP/TPLP model*. Leibniz-Zentrum für Informatik.

Higgins, G. E., Ricketts, M. L., & Wolfe, S. E. (2008). Identity theft complaints: Exploring the state-level correlates. *Journal of Financial Crime*. Advance online publication. doi:10.1108/13590790810882883

Hussain, S., Mustafa, M. W., Jumani, T. A., Baloch, S. K., Alotaibi, H., Khan, I., & Khan, A. (2021). A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection. *Energy Reports*, *7*, 4425–4436. doi:10.1016/j.egyr.2021.07.008

Hussain, Z., Memon, S., Shah, R., Bhutto, Z. A., & Aljawarneh, M. (2016). Methods and techniques of electricity thieving in pakistan. *Journal of Power and Energy Engineering*, *4*(09), 1–10. doi:10.4236/jpee.2016.49001

Insights, D. (2018). *State of AI in the enterprise*. Deloitte.

Jain, P. K., Pamula, R., & Srivastava, G. (2021). A systematic literature review on machine learning applications for consumer sentiment analysis using online reviews. *Computer Science Review*, *41*, 100413. doi:10.1016/j.cosrev.2021.100413

Jain, S., Choksi, K. A., & Pindoriya, N. M. (2019). Rule-based classification of energy theft and anomalies in consumers load demand profile. *IET Smart Grid*, *2*(4), 612–624. doi:10.1049/iet-stg.2019.0081

Jamil, F., & Ahmad, E. (2019). Policy considerations for limiting electricity theft in the developing countries. *Energy Policy*, *129*, 452–458. doi:10.1016/j.enpol.2019.02.035

Johncy, G., & Felise, A. A. (2020). *An efficient power theft detection using mean-shift clustering and deep learning in smart grid.* Paper presented at the IOP Conference Series: Materials Science and Engineering.

Jokar, P., Arianpoo, N., & Leung, V. C. (2015). Electricity theft detection in AMI using customers' consumption patterns. *IEEE Transactions on Smart Grid*, *7*(1), 216–226. doi:10.1109/TSG.2015.2425222

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys*, *52*(4), 1–36. doi:10.1145/3343440

Kim, H. C., Wallington, T. J., Arsenault, R., Bae, C., Ahn, S., & Lee, J. (2016). Cradle-to-gate emissions from a commercial electric vehicle Li-ion battery: A comparative analysis. *Environmental Science & Technology*, *50*(14), 7715–7722. doi:10.1021/acs.est.6b00830 PMID:27303957

Kumar, A., & Sharma, A. (2017). Systematic Literature Review on Opinion Mining of Big Data for Government Intelligence. *Webology, 14*(2).

Lei, C. (2021). Data Preparation for Deep Learning. In *Deep Learning and Practice with MindSpore* (pp. 329–362). Springer. doi:10.1007/978-981-16-2233-5_14

Li, J., & Wang, F. (2020). Non-technical loss detection in power grids with statistical profile images based on semi-supervised learning. *Sensors (Basel)*, *20*(1), 236. doi:10.3390/s20010236 PMID:31906158

Li, Q., Lin, H., Tang, C., Wei, X., Peng, Z., Ma, X., & Chen, T. (2021). Exploring the" Double-Edged Sword" Effect of Auto-Insight Recommendation in Exploratory Data Analysis. *CEUR Workshop Proceedings*.

Loroz, P. S. (2006). *The generation gap: A Baby Boomer vs. Gen Y comparison of religiosity, consumer values, and advertising appeal effectiveness*. ACR North American Advances.

Maamar, A., & Benahmed, K. (2018). Machine learning techniques for energy theft detection in AMI. *Proceedings of the 2018 International Conference on Software Engineering and Information Management*. doi:10.1145/3178461.3178484

Mahalakshmi, H., Nikhitha, J., & Varsha, B. (2018). Implementing Anti-theft Systems for ATM and Vehicles. *Perspectives in Communication, Embedded-systems and Signal-processing-PiCES*, *1*(12), 196–201.

Min, B., & Golden, M. (2014). Electoral cycles in electricity losses in India. *Energy Policy*, *65*, 619–625. doi:10.1016/j.enpol.2013.09.060

Mohammad, R. M., Thabtah, F., & McCluskey, L. (2012). *An assessment of features related to phishing websites using an automated technique.* Paper presented at the 2012 International Conference for Internet Technology and Secured Transactions.

Nawaz, R., Akhtar, R., Shahid, M. A., Qureshi, I. M., & Mahmood, M. H. (2021). Machine learning based false data injection in smart grid. *International Journal of Electrical Power & Energy Systems*, *130*, 106819. doi:10.1016/j.ijepes.2021.106819

Nizar, A., & Dong, Z. (2009). *Identification and detection of electricity customer behaviour irregularities.* Paper presented at the 2009 IEEE/PES Power Systems Conference and Exposition. doi:10.1109/PSCE.2009.4840253

Oliveira, C. d., Kagan, N., Meffe, A., Jonathan, S., Caparroz, S., & Cavaretti, J. (2001). *A new method for the computation of technical losses in electrical power distribution systems.* Paper presented at the 16th International Conference and Exhibition on Electricity Distribution, 2001. Part 1: Contributions. CIRED. doi:10.1049/cp:20010889

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems*, *2*(1), 1–10. doi:10.1186/2047-2501-2-3 PMID:25825667

Ravnik, R., Solina, F., & Zabkar, V. (2014). *Modelling in-store consumer behaviour using machine learning and digital signage audience measurement data.* Paper presented at the International Workshop on Video Analytics for Audience Measurement in Retail and Digital Signage. doi:10.1007/978-3-319-12811-5_9

Smith, T. B. (2004). Electricity theft: A comparative analysis. *Energy Policy*, *32*(18), 2067–2076. doi:10.1016/S0301-4215(03)00182-4

Srivastava, P. R., & Eachempati, P. (2021). Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi-Criteria Decision-Making Approach. *Journal of Global Information Management*, *29*(6), 1–29. doi:10.4018/JGIM.20211101.oa23

Stolwijk, A., Straatman, H., & Zielhuis, G. (1999). Studying seasonality by using sine and cosine functions in regression analysis. *Journal of Epidemiology and Community Health*, *53*(4), 235–238. doi:10.1136/jech.53.4.235 PMID:10396550

Subramanian, R. (2006). India and information technology: A historical & critical perspective. *Journal of Global Information Technology Management*, *9*(4), 28–46. doi:10.1080/1097198X.2006.10856431

Taft, J., & von Prellwitz, L. (2012). Utility data management & intelligence. *Cisco White Paper, 372*.

Veale, M., & Brass, I. (2019). Administration by algorithm? Public management meets public sector machine learning. *Public management meets public sector machine learning*.

Watson, H. J. (2013). All about analytics. *International Journal of Business Intelligence Research*, *4*(1), 13–28. doi:10.4018/jbir.2013010102

Weaver, K., & Solutions, S. (2017). *Smart meter deployments result in a cyber attack surface of "unprecedented scale*. Smart Grid Awareness.

Wu, Z., Zang, C., Wu, C.-H., Deng, Z., Shao, X., & Liu, W. (2021). Improving Customer Value Index and Consumption Forecasts Using a Weighted RFM Model and Machine Learning Algorithms. *Journal of Global Information Management*, *30*(3), 1–23.

Yadav, R. K., & Mohania, S. (2017). Claim settlement of Pradhan Mantri Suraksha Bima Yojana under Pradhan Mantri Jan Dhan Yojana. *World Scientific News*, *65*, 123–134.

Yilmaz, I., Kapoor, K., Siraj, A., & Abouyoussef, M. (2021). Privacy Protection of Grid Users Data with Blockchain and Adversarial Machine Learning. *Proceedings of the 2021 ACM Workshop on Secure and Trustworthy Cyber-Physical Systems*. doi:10.1145/3445969.3450431

Yurtseven, Ç. (2015). The causes of electricity theft: An econometric analysis of the case of Turkey. *Utilities Policy*, *37*, 70–78. doi:10.1016/j.jup.2015.06.008

Prashant Salwan is Professor of Strategy and International Business at Indian Institute of Management Indore India. He is an alumnus of London School of Economics and Political Science UK and a British Chevening Scholar and a Fulbright Scholar.

Nitin Kumar Agarwal is Co-Founder and Chief Data Science Officer at Xalt Analytics. He brings along around 25+ Years of experience in data science and operations research out of which approx. 12 + years his experience lies in building data-driven innovative solutions using data science tools and techniques to drive business value to the customer.