National College of Ireland

Predicting Electricity Theft Based on the Impact of Socio-Economic Indicators

MSc Research Project
Data Analytics

Varun Rajendra Ghorpade
Student ID: x18104169

School of Computing
National College of Ireland

Supervisor: Dr. Paul Stynes

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Varun Rajendra Ghorpade
Student ID:	x18104169
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Paul Stynes
Submission Due Date:	12/08/2019
Project Title:	Predicting Electricity Theft Based on the Impact of Socio-Economic Indicators
Word Count:	6952
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	10th August 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

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

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

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

Predicting Electricity Theft Based on the Impact of Socio-Economic Indicators

Varun Rajendra Ghorpade
x18104169

Abstract

The main focus of this research is to find out the socio-economic indicators, which are the most significant in predicting electricity theft and to find out there predictive power as the determinants of electricity theft. Much of efforts, by the governments across the world, has been made to control the electricity losses, be it the technical or non-technical losses. Though many solutions are imparted to overcome the technical losses, such as smart metering, feeder renovations, etc., the main concern remains the same, non-technical losses (NTL). Electricity theft or pilferage of electricity is considered to be the most impacting in increasing aggregate technical and commercial (AT&C) and transmission and distribution (T&D) losses. There were many pieces of research in past which attempted to predict the electricity theft by detecting the anomalies in the usage pattern of the customers and many of these researches suggest that only looking at the technical aspect won't provide a wholesome solution. In this research to find out the root cause of electricity theft, KDD methodology of data mining is used. The raw data for socio-economic indicators and T&D losses data from different sources are merged and various regression techniques have been used such as OLS, ridge regression, SVR, CARTs(decision tree and RF). Majority models after evaluation suggested that the most significant factor which is impacting electricity theft is unemployment(male) and RF turned out to be the most efficient model in predicting electricity theft.

Keywords Electricity theft, socio-economic, India, smart meters, machine learning, Maharashtra, T&D losses

1 Introduction

Electricity is a form of energy which is considered to be the most crucial source in energizing the infrastructure of any country, which also makes it crucial from the economic point of view (Jamil; 2013). However, some harmful practices, such as electricity theft hamper the overall growth of a nation and can cause many repercussions. Authors like Messinis and Hatziargyriou (2018), Jamil and Ahmad (2019) and Ahmad (2017) explained the impact of electricity theft on the economy of various countries and how the developing countries India, Pakistan, Brazil and Turkey are struggling to overcome the issue of electricity theft. This situation motivates the research undertaken to find out

the root level cause of electricity theft and what factors affect it the most. Apart from the economic implications, (Depuru et al.; 2011) explained the various adverse effects of electricity theft which includes the overestimation of the demand of energy by the utility providers, resulting in damaging the home appliances of the customers and how it puts an extra burden of the tariff on the benign customers. One more aspect which motivates this research is electrocution. (Munro and Munro; 2008) and (Kumar et al.; 2017) explain the electrocution is a severe issue as it includes human lives. These research describe the electrocution as death caused due to the passage of a high voltage current through the body cardiac arrest or brain damage. As per the National Crime Records Bureau (NCRB), Government of India, 2015 data there was a total of 1373 deaths recorded in Maharashtra due to electrocution (Kumar et al.; 2017). This raises the question of the literacy of the state. Are the people not educated or literate enough to figure out the adversities of stealing electricity? Why do people intend to steal the electricity when there are many regulations of government which ensure a fair amount for energy distribution? These kinds of question need to be answered, and they can be answered by identifying the factors of the society which are responsible for electricity theft. That is why socio-economical indicators play a key role in eradicating such malfunctions in society.

Authors like Depuru et al. (2013), Jokar et al. (2016), Ahmad (2017) and Nallathambi (2017) inspired this research to implement the supervised machine learning techniques in detecting electricity theft. However, the data used in this research are the customer level data, mainly the consumption and billing data. They are mainly focused on detecting anomalies and abnormal consumption patterns. Studies by Min and Golden (2014), Yurtseven (2015), Faria et al. (2016), Gaur and Gupta (2016), Jamil (2018) and Razavi and Fleury (2019) made use of socio-economic indicators to predict the electricity theft. The methodologies and techniques used in these research are studied thoroughly in section.. and based on which the machine learning techniques are implemented. Out of these studies like Min and Golden (2014), Gaur and Gupta (2016), and Razavi and Fleury (2019) implemented OLS regression to find out the significant variables in predicting electricity theft. In all these studies, the OLS model was found to be at the bottom of the performance matrix. In this research, it is tried to implement various techniques like K-fold cross-validation, parameter tuning, and log transformation of data to improve the performance of OLS with respect to other techniques like Ridge regression. Along with regression models, this research also implemented the classification and regression trees like Decision Tree(DT), and Random Forest(RF) are used to predict the electricity theft as suggested by the studies likeBabar et al. (2015), Nallathambi (2017), Razavi and Fleury (2019), Brosnan (2018). Evaluation techniques such as RMSE, MSE, R^2, accuracy, kappa, and detection rate are used to evaluate the performance of above mentioned models.

Keeping all this into consideration, the research question for this research is framed as ***"can the regression-based models significantly predict electricity theft based on socio-economic indicators?"***

The main focus of this research is to present such a predictive model which will be efficient in predicting the electricity theft for the districts of State of Maharashtra, India. Along with this, the research also tried to address the following objectives:

- To design the models which will identify the variables that are best suitable in predicting the electricity theft accurately.

- To investigate and find out the intricacies of the data and implement the solutions to make it work for the machine learning models.
- To investigate if the gender-based data is decisive in predicting electricity theft.
- To implement robust methodologies with several experiments to improve the performance of any given model.

This research is divided into six sections, which give an insight into the approach adopted to address the research question and objectives. Section 2 focuses on the related work done in previous literature in the domain. This section is divided into further six subsections. Section 3 discusses the methodology adopted to carry forward the research and to achieve the objectives. Each step taken in this research is briefly discussed in the section. Section 4 provides the specifications which are used to carry the research steps discussed in Section 3. Section 5 presents the evaluation of models and a discussion of the results. Section 6 summarizes the findings of the research and discusses the future scope for the same.

2 Related Work

This section provides an insight into the past studies carried out in predicting electricity theft using various techniques. The research undertaken, tried to adopt a novel approach with the expertise of this literature and also tried to improve the existing prediction models.

2.1 Rationale Behind Using Transmission and Distribution (T&D) Losses As Electricity Theft

As explained by Sharma et al. (2016), Gaur and Gupta (2016), Smith (2004) and Lewis (2015), T&D losses are those losses which take place due to both technical and non-technical factors such as failing of transformers, non-payment of the bills by the users and hooking the electricity wires. T&D losses are chosen to be the proxy for electricity theft for this research. Gaur and Gupta (2016) argued that though in a country like India, technical losses are very much evident due to ignorance towards the infrastructure, the losses incurred due to commercial losses are huge. In the highly populated countries like India, where the places are usually crowded, it becomes very easy to tamper with the distribution line and the transformers. Hooking the wires are the most common way of stealing the power/electricity from the grid. These NTLs are very difficult to identify, which makes it even more difficult in estimating the exact losses(Kumar et al.; 2017). In one of the studies by Lewis (2015), it is stated that the rationale behind using the T&D losses data as the proxy to electricity theft is that the data provided by many utility providers do not separately specifies the commercial losses, rather the data is an aggregation of both the losses. Therefore, the data made available by MAHADISCOM makes this research more precise and focuses on non-technical losses as they are separately categorized.

Production function approach was used by Lewis (2015) to estimate the T&D losses in Jamaica. As per the analysis, the Hotel industry in the country was consuming maximum energy, which was creating an imbalance in the transmission and distribution. This

approach suggests that the use of energy for any particular industry can impact and motivate the theft. In another research by Kumar et al. (2017) the T&D losses are used as a proxy to electricity theft. This research came up with the four types of electricity theft, meter tampering, tapping the electrical connections, non-payment of the bills, and corruption in the distribution system.

2.2 Detecting Electricity Theft Using Supervised Learning

Many studies (Razavi et al. (2019), Nallathambi (2017), Ahmad (2017), Jokar et al. (2016) and Depuru et al. (2013)), by now have used a wide range of machine learning techniques in identifying the consumption pattern and abnormalities in the usage of electricity which contributes to T&D losses. Majority Techniques that have used belong to Supervised learning. Tampering the electric lines or meters is the most common ways of stealing electricity (Kumar et al.; 2017). Following listed are some researches carried out over a period to identify the electricity theft with the help of Advanced Metering Infrastructure (AMI) or Smart Meters (SM) using supervised learning.

SVM and Rule Engine (RE) were implemented by Depuru et al. (2013) in identifying electricity theft using the customers usage pattern. Anomalies detected in the usage pattern and then by implementing the parallel algorithm technique, in which the results for both the models were compared and re-evaluated if they didnt match. 5-fold cross-validation resulted in better performance of SVM with an accuracy of 89%, and also it turned out to be faster than RE.

Advanced Metering Infrastructure (AMI) was advocated as the best solution to overcome the problem of electricity theft by Jokar et al. (2016). In this research, multi-class SVM was implemented in Consumption Pattern-based Energy Theft Detector (CPBETD), which helped in capturing the usage pattern of the consumption for the users. The model was evaluated using the detection rate and false-positive rate. This model predicted the anomalies in the patter with the detection rate of 94% and accuracy of 95%. This clearly suggests the Multi-class SVM implemented in this research outperformed the SVM implemented in the study by Depuru et al. (2013).

Ahmad (2017) in his research, in finding out the abnormal patterns in the consumption, used SVM-Radial model, Artificial Neural Network(ANN) and Optimum Path Forest(OPT). In this research, SVM was not optimized as it was done by Depuru et al. (2013) and Jokar et al. (2016), resulting in poor performance, wherein, OPT outperformed both SVM and ANN. The seasonal analysis was also carried out in this study, which indicated the period with maximum anomalies. One exciting factor this research pointed out is that the education/literacy level amongst the employees who read the meter and prepare the bill can be a deciding factor as they can make mistakes with the number causing losses in the revenue of the utility provider.

Electricity theft was also detected using CARTs such as Decision Tree(DT) and Random Forest(RF). In research by Nallathambi (2017), the customers were classified using CARTs in the categories based on their pattern of consumption. Decision tree was implemented as a regression analysis, and the dataset was grouped using the standard deviation(SD). The variables with the maximum SD are considered in the model. This study made a point that to overcome the overfitting issue in DT, RF can be used. ROC curve was used as the evaluation method in this and RF outperformed DT with accuracy as high as 95.78%. Some other studies (Cody et al. (2015) and Razavi et al. (2019)) also

used CARTs to detect electricity theft.

2.3 Electricity Theft Detection Using Unsupervised learning

There are many studies in which supervised learning like, classification and regression, approaches have been adopted, but there are very few with the unsupervised learning approach. Few of them are listed below.

Unsupervised learning, in the domain of electricity theft detection, came into the picture in 2017 when Viegas and Vieira (2017) used clustering to divide the data into the groups based on a similar pattern of usage by the customer resulting in the creation of new data. The new data is then tested along with the previous data(historical), and the patterns were extracted. Fuzzy C-means(FCM), Gustafson-Kessel, K-means clustering techniques were compared with the one-class SVM. The models were evaluated using AUC, TPR, FPR, and TPR-FPR. Results demonstrated that the FCM outperformed all other techniques. The author made a point that the one-class SVM is the most efficient and popular, wherein, in the previous study by Jokar et al. (2016), it was proved that the multi-class SVM is a better performer. The drawback of this which can be pointed out is that the interpretation of the results cannot be made easily, making it much more complex than that of supervised techniques.

Singh et al. (2017) studied the consumption data obtained from Advanced Metering Infrastructure (AMI) for 5000 costumers over Ireland. This research performed principle component analysis and evaluated the models based on the receiver operating curve(ROC) and TPR. TPR is considered as detection rate(DR), and the ROC curve was plotted for TPR and FPR. There were three applied, A1, A2, and A3, out of which A3 outperformed the other models with a DR of 90.87% which can be considered as a very high DR. However, if it has been compared to the multi-class SVM, the study by Jokar et al. (2016), detected the electricity theft with the DR of 94% which is surely better than PCA.

2.4 Various Issues Predicted Through Socio-economic Indicators

In the context of socio-economical influence on electricity theft, there is a scarcity of research work (Messinis and Hatziargyriou; 2018). To address the research question (section 1), it becomes very important to explore and gather knowledge regarding how the socio-economic indicators can be used for prediction of various issues of society. What type of machine learning techniques and models are suitable in employing the indicators in the predictive analysis? Also, how the results can be interpreted for these indicators? To answer all these question, literature by Catlett et al. (2019), Brosnan (2018), Hashemian et al. (2017) and Babar et al. (2015) are studied.

Research by Babar et al. (2015) is based on the prediction of disease outbreak in Pakistan and socio-economic factors affecting it. In this Babar et al. (2015) implemented DT, RF, Naive Bayes, Logistic Regression, and Bayes Net to detect factors of disease-outbreak. The models were evaluated using F-measure and precision. This study explains DT as a better model for the prediction that RF, which contradicts with the studies by Razavi and Fleury (2019) and Nallathambi (2017). So, the research undertaken tried to

find out if really DT can be a better prediction model than RF, where socio-economic indicators are used as input.

In research by Brosnan (2018) and Hashemian et al. (2017), OLS along with various other models were implemented using socio-economic indicators as predictors. In both these studies, OLS found to be at the bottom of the performance table. So, the research undertaken in this paper tried to improve the OLS regression model using several methods.

Catlett et al. (2019) is one of the recent studies in which the socio-economic indicators are used as predictors of crime in Switzerland. Methods like ARIMA, RF, RETree, and Zero were implemented and were evaluated using MAE, MAPE, ME, and RMSE.

2.5 Socio-Economic Indicators As The Determinants of Electricity Theft

Section 2.2 gives an insight into how different machine learning algorithms are used to detect electricity theft over a period of time using the consumer-level data. Now, this section discusses the literature which is more focused on the area level data i.e., the socio-economic indicators data to understand how these indicators are significant in answering the research question framed in Section 1 and how the machine learning techniques can help in achieving the objectives of this research.

Min and Golden (2014) carried out one of the first research in which the socio-economic indicators were considered to identify the NTLs and predicting electricity theft. Only the OLS regression model was implemented in this research to identify the theft. The p-value was used to evaluate the strength of the model. The factor which was found to the most influential in identifying the theft was the electoral cycle. The biggest drawback of the study is that it only implemented OLS regression and not tried to get a more accurate prediction using other models. So, the research undertaken in this paper tried to overcome the drawback by implementing various models and optimization techniques to yield better results.

In a research conducted by Yurtseven (2015), demonstrated the use of Instrumental Variable Generalized Method-of-Moments (IV-GMM), 3-stage-least-square (3SLS) in finding out the determinants of theft in Turkey. As per the output of the research, education and per capita income turned out to be the most significant indicators of electricity theft. The models were evaluated using only R^2 value. One very constructive suggestion which the author kept forward in the study is that the tariffs of electricity should be levied on the basis of the customers income bracket. This will certainly bring some balance in society and can definitely reduce electricity theft in the region. The one drawback of this study is that it uses only R^2 as the evolution parameter for the models. R^2 cannot always make a correct judgment regarding the model-fit and its efficiency. Even models with low R^2 can be proved to be efficient in predicting the outcome[1]. To overcome this drawback, the study undertaken in this research has evaluated the models with various evolution techniques.

NTLs for the cities in Brazil were predicted by Faria et al. (2016). Generalized additive model(GAM) was implemented in this research to find out the spatial-temporal patterns in NTLs. Various maps indicating the variations in the NTLs were plotted, and Markov Chain was used to find out those variations. Also, this research made an attempt to find

[1]https://statisticsbyjim.com/regression/interpret-r-squared-regression/

out the most influential factors which can possibly be the determinants of electricity theft. GAM model predicted the NTLs with the accuracy of 78.2%. In the study, variables like the number of households and number of residents were taken into consideration, and these factors are also part of the research undertaken in this very paper.

Gaur and Gupta (2016) implemented OLS regression and Feasible Generalized Least Square regression (FGLS) models to predict the electricity theft by passing socio-economic indicators as input. FGLS outperformed OLS based on R^2 value. This experiment backs the results of Brosnan (2018) and Hashemian et al. (2017) as OLS model lags in these research as well. In this, the most significant factors were urban population, literacy rate, and industrialization. The research undertaken, too, have all these indicators as input.

The studies like Jamil (2018), Gaur and Gupta (2016), Faria et al. (2016), Yurtseven (2015) and Min and Golden (2014), the country level has been used. However, in the most recent by Razavi and Fleury (2019), district-level data has been used to predict the electricity theft on the socio-economic indicators of the region. This research deployed various regression-based models like OLS, CARTs, SVM, and GBM to develop a model which can accurately predict electricity theft. This study addressed the issue of multicollinearity amongst the variables with the help of Variance inflation factors (VIF), which are calculated and variables with the values much higher than 5, are removed from the model. The models were estimated on the basis of R-squared(R^2) and accuracy. The RF model outperformed all the other models with explaining 87% variability in the T%D losses with variation in the socio-economic indicators. The one drawback of this research is that no feature selection technique was used which the research undertaken in this paper tried to implement using Boruta algorithm. In this algorithm, several iterations of the data take place, and every iteration looks for the significance of each variable. In the end, the Boruta classifies the variables in three categories, namely, confirmed important, confirmed unimportant, tentative, and rejected features(Shaheen and Iqbal; 2018).

2.6 Conclusion

After studying the related work mentioned in this section, it can be said that the machine learning approach to the problem really brought some revolutionary changes in the domain of electricity theft detection. Evolution of techniques and results can be observed. Initially there were research (Min and Golden (2014) and Yurtseven (2015)) that adopted statistical modeling for detection of electricity theft ,wherein, in later phase there were research(Faria et al. (2016), Gaur and Gupta (2016), Jamil (2018) and Razavi and Fleury (2019)) that go on to use much advanced machine learning techniques like SVM, OLS, RF, DT, NN, etc. All the literature studied, motivated this research in adopting a machine learning approach that tries to discover those socio-economic indicators which are significant in predicting electricity theft.

3 Methodology

The methodology adopted in this research is based on the lines of KDD(Knowledge Discovery of Databases)(Fayyad et al.; 1996). Also, this research follows a path which has been constructed in such away that the question raised in Section 1 is answered in the most honest manner. Figure 1 gives an insight into the methodology adopted in the research. The steps followed in this research are as follows:

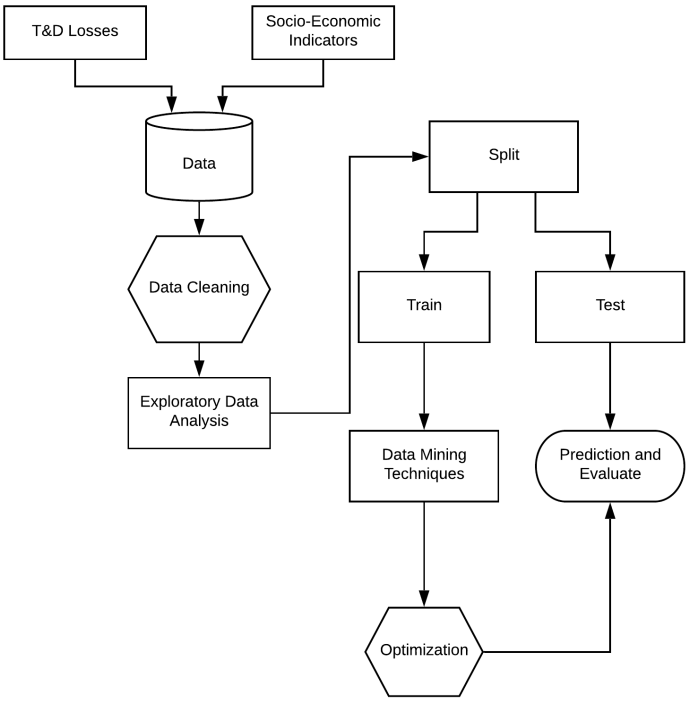

Figure 1: Research Flow

- **Step 1:** The data for the socio-economic indicators is gathered from Open Government Data(OGD) Platform India[2] in CSV format and the data for T&D losses is taken from Maharashtra State Electricity Board portal[3].

- **Step 2:** Then the merging of the data has been done on the basis of the district and the block names.

- **Step 3:** Next, the data has been cleaned and transformed in which the missing values were checked and addressed, the values with 0 have been removed, and commas between the values were removed.

  As this project intends to use regression models, it is essential that all the attributes of the data must contain numeric values. The columns, State/UTs_Name, Level(village/CDBlock), and Total/Rural/Urban were removed as they contained character values and deemed to be less important in the research. Columns such as State/ UTs_Code, District_Code, CD Block_Code, Town/Village_Code, Ward_Code, EB_Code were are also removed as they are no use for further analysis.

  The summary of all the columns was checked and found that there were values with 0 in the data. 0 values were checked and found that there were many blocks in districts for which the data is not captured, hence removed the rows with the 0 values in them. Apart from this, no missing values were found in the data.

- **Step 4:** Exploratory data analysis is done on the cleaned data. Firstly, it is checked if the data is normally distributed or not. After the normality test, it was found

[2]https://data.gov.in/catalog/villagetown-wise-primary-census-abstract-2011-maharashtra?filters%5Bfield_catalog_reference%5D=166483&format=json&offset=0&limit=6&sort%5Bcreated%5D=desc

[3]https://www.mahadiscom.in/consumer/distribution-atc-losses/#1554095676205-f296c751-2f6a

that the data is highly skewed(positively), which may not fit the model. To overcome this issue, log transformation of the data has been done, which reduced the skewness of the data. Feature selection was then made on the transformed data. Boruta algorithm is used for feature selection purpose. Out of 87 variables, 42 were confirmed, 12 were tentative, and rest were rejected by the Boruta algorithm. On the basis of previous research(Section 2.5), 11 variables were chosen in accordance with Boruta. The chosen variables are listed in Table 1 with their importance score. Multicollinearity and addressing class imbalance has been taken to understand and transform the data. Next, the multicollinearity is checked and addressed. In this project, to address the multicollinearity amongst the variables variance inflation factor(VIF) is used. Variance inflation factor gives the score for the variable, which is in the maximum variance of the regression coefficient. To perform this, first, a linear model was built with all the variables which were confirmed by Boruta and then Variance inflation factor is performed. Boruta already takes care of multicollinearity and throws the importance score as per the relation between variables. As in step 3, the rows with the values 0 in it are removed, it becomes important to check the imbalance in the data. After checking, it was found that the data is imbalanced.

Table 1: Most important Variables - Boruta

Variables	**meanImp**	**medianImp**	**Status**
No.of.Households	4.420738	0.93939394	Confirmed
Total.Population.Person	3.771017	0.7979798	Confirmed
Total.Population.Male	4.271348	0.81818182	Confirmed
Total.Population.Female	4.194578	0.81818198	Confirmed
Literates.Population.Person	4.669783	0.8989899	Confirmed
Literates.Population.Male	4.102508	0.85858586	Confirmed
Literates.Population.Female	4.542475	0.95959596	Confirmed
Main.Agricultural.Labourers.Population.Person	5.848452	1.00000000	Confirmed
Main.Household.Industries.Population.Person	3.772565	0.56565657	Tentative
Main.Other.Workers.Population.Person	7.160648	1.00000000	Confirmed
Non.Working.Population.Person	3.998022	0.78787879	Confirmed

- **Step 5:** Once the data is explored, various stages of experiments were carried under which the various machine learning and data mining techniques were used. First, the data for every model was split with a simple split(hold-out) method. The first experiment which was carried was for OLS regression model to find out the significant variables in predicting the electricity theft, which addresses one of the objectives(Section 1) of this research. In the second experiment, the OLS model was tried to improve using the log-transformed data and by tuning the parameters, which also included 10-fold cross-validation.

  The next experiment was done with the Ridge regression model with hyperparameter tuning in which it was tried to see if the Ridge works better with the original data. Next, the Ridge was implemented with log-transformed data to check the improvement and get better results, which gives more accurate information regarding the variables.

In the fifth experiment, SVR was implemented on log-transformed data. The best values for parameters cost and elsilon were selected under hyperparameter optimization. Radius Basis Function (RBF) kernel was used in the SVR model. RBF helps SVR in finding the fit and mapping the data on the plane.

In the next experiment, DT and RF regression was performed. In which the data split of 75:25 is used.

Binning approach is adopted to build classification models and check whether the data is suitable enough for classification as well. However, this cannot be considered as the solution, or the ultimate result of this research as an expert opinion for binning the T&D losses is required(please refer configuration manual for the code). In this research, T&D losses higher than 21.7% were considered are categorized as "theft" and values below that are binned as "no theft." On the basis of this following experiments were done.

Experiments six to nine were carried out in the direction of improving the performance of the decision tree(DT) in predicting electricity theft with the socio-economic indicators. In the first experiment, the log-transformed data is used with a simple split of 75:25. The next experiment included control function with cross-validation in which the parameters such as minsplit, minbucket, cp, maxcompete maxsurrogate and maxdepth were tuned to get better performance. In the next experiment, train control was used with cross-validation to enhance the performance of the DT. Finally, as the data is imbalanced and with limited attributes, SMOTE(Synthetic Minority Over-sampling Technique) was applied on the train data to boost the performance of DT.

The last set of experiment, in order to find the most suitable model to achieve the objectives of this research was carried out with random forest. The first experiment in this was done with hyperparameter ntree = 500 and the second one with ntree = 100 along with SMOTE.

- **Step 6:** Lastly, the models implemented were evaluated using various evaluation techniques which are chosen as per the studies discussed in Section 2. RMSE value is used to evaluate both the OLS models and SVR. The R^2 value is used to evaluate both the OLS models, the Ridge models, and the RF model. Apart from this, RMSE values were used to evaluate SVR, OLS, and ridge regression models. Accuracy is used to evaluate the DT and RF models. Also, Kappa and detection rate are used to evaluate DT and RF models.

4 Implementation

To implement the methodology described in Section 3, the following procedure is carried out and tried to achieve the objectives of this study.

- The data for this research is gathered from the sources mentioned in Section 3. The data for socio-economic indicators have been downloaded directly in CSV format. There were separate files for each district which were combined using builtin function in R like list.file(), file.path() and do.call(). The data for T&D losses was in PDF format and extracted in Google colab using Python packages such as, 'PyPDF2' to import pdf file from the source, 'tabula' package to import the data

in a tabular form. Function 'PyPDF2.PdfFileReader' is used to read the pdf file from the source and keep it in a data frame(pfr in this case) and function 'tabula.read_pdf()' is used to read the file from the data frame(pfr) in a table form. Lastly, function 'tabula.convert_into' is used to save the extracted table in the desired format(CSV in this case). The extracted files are then read in R studio and merged using merge() function in R.

- The data is then cleaned in R Studio with R version 3.5.1 and Microsoft excel. The columns which were not required were removed with the R programming which left the data with 87 attributes which contained meaningful data. The 0 values were identified using function summary() in R, sample of which is shown in 4. The rows with 0 values were then removed in Microsoft Excel. Using function str() is was made sure that all the data types are correct as per the requirement. R function is.na() is used to detect the missing values in the data.

Table 2: Summary of the data

Variable	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**
No.of.Households	0	109	215	1074	411	148561
Total.Population.Person	0	510	1032	5098	2015	658928
Literates.Population.Person	0	339	688	3514	1351	485775

- After cleaning the data, exploratory analysis on that data is performed. The very first thing which was checked is if the data is normally distributed or not. To check the distribution, R function ggplot() is used. The R packages needed for this function are 'dplyr', 'ggpubr' and 'ggplot2'. From the plots obtained, it can be inferred that the data is highly skewed(positively).To overcome this issue, log transformation of the data is used. R function log1p() is used to transform the data.

- This project adopted Boruta as the feature selection technique. For this Boruta package in R is used.

- The multicollinearity amongst the variables is checked using VIF for which the 'car' package in R is installed. This package provides function vif() using which the score for each variable is calculated.

- Before implementing the models, the data prepared in the above steps is split into train and test sets. Initially, the data has been divided using 75:25 ratio using function sample() of package "dplyr". Then the first model, OLS regression, was implemented is. Function lm() in R is used to build the model. Summary () is then used to get the entire summary of the model. The T&D losses were then predicted using function predict(). "Transport" package has been used to do hyperparameter optimization using function trainControl() in the second experiment.

- For Ridge regression, the "glmnet" package in R is used. It comes with function glmnet() which trains the model with hyperparameter 'lambda.' Before this model.matrix() is used to remove the intercepts and to store the response column. Function log1p() is again used in the second experiment of ridge regression for log transformation. Also function cv.glmnet() is used get an optimal value of lambda.

- To implement the SVR model, package "e1071" was installed. It provides the function SVM() using which the model has been built. Function tune() which have an inbuilt 10-fold cross-validation method, is used for hyperparameter(cost and elsilon) tuning. Function plot() is used to plot the prediction and the optimized model, which gives the value of mean square error.

- R packages "rapart", "rpart.plot" and "RColorBrewer" has been used to perform regression using a decision tree. Library "rattle" is used for plotting the decision tree. Functions rpart.control() and trainControl() were used for hyperparameter tuning in second and thrid experiment of DT. In the last experiment for DT, the class imbalance issue is addressed using function SMOTE() for which the package "DMwR" is installed. Function predict() is used to predict the response(T&D losses).

- For RF, package "randomforest" is installed. In the second part of this experiment, SMOTE() is applied to take care of the imbalanced data.

- For evaluation purpose, the packages which were used are "Metrics" for calculating RMSE for SVM, "Caret" to get the confusion matrix for DT and RF. A function has been created using the equation (1) to calculate the RMSE for OLS and Ridge regression.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} |y - y_t|}{n}} \tag{1}$$

Where,

- y = Predicted T&D Loss.
- y_t = Actual T&D Loss.
- n = Sample size

5 Evaluation

In order to arrive at a conclusion whether the implemented models can provide the results which are expected as per the set objectives(Section 1), it is important to evaluate them using various techniques. In this section, all the experiments are listed with the evaluation of the models.

- **OLS with simple split**
  The first phase of machine learning in this research starts with the OLS regression with a simple split of data in 70:30 ratio. The first thing which can be observed is the p-value. The model's p-value, in this case, is 2.2e-12 which as per the assumptions

of linear regression is less than 0.05 making the model statistically significant. The other statistical measures obtained were:
Residual standard error = 1.42,
R-squared = 0.3491, Adjusted R-squared = 0.3392,
F-statistic = 15.06, RMSE = 1.49.

Table 3: Comparison of the OLS models for predicting T&D loss rate in percentage

Variables	t-**value _simple_ols**	p-**value _simple_ols**	t-**value _Log_ols**	p-**value _Log_ols**
No.of. households	-0.536	0.59222	-6.871	0.00089
Total.Population. Person	0.659	0.51008	3.081	0.002152
Total.Population.Male	-0.632	0.52780	-3.142	0.001759
Total.Population.Female	-0.581	0.54109	-2.442	0.014892
Literates.Population. Person	-0.56318	0.57385	-1.365	0.172879
Literates.Population.Male	0.56853	0.56997	1.830	0.067690
Literates.Population. Female	-0.43779	0.61514	-0.033	0.973644
Main.Agricultural. Labourers.Population.Person	-1.161	0.24583	-0.118	0.906051
Main.Household. Industries.Population.Person	0.309	0.75736	-1.764	0.078197
Main.Other.Workers. Population.Person	-1.799	0.07240	1.937	0.053217
Non.Working. Population.Person	2.788	0.00541	-2.851	0.004496 **
Non.Working. Population.Male	-0.936	0.34946	3.416	0.000676***
Non.Working. Population.Female	-0.841	0.47798	2.606	0.009369 **

- **OLS with parameter tuning and log transformation**
  To improve the results of the OLS regression, this research then adopted the log-transformed data and the parameter tuning technique. For tuning the model trainControl package in R has been used, which makes use of k-fold cross-validation to tune the parameters. In this case, 10-fold cross-validation has been implemented. There were significant improvements observed in the OLS model after the transformation and tuning. The measures are as follows:
  Residual standard error: 0.3086,
  R-squared = 0.647, Adjusted R-squared = 0.6354.
  F-statistic = 88.85, RMSE = 0.903105.

- **Ridge Regression with tuning hyperparameter, Lambda** Ridge regression is implemented using glmnet() function and design matrix called model.matrix. Hyperparameter lambda is used to tune the model. A design matrix is created, which

is then passed to cv.glmnet() function, which finds the optimal value of lambda. The lowest point in the curve shown in Figure 2 represents the optimal value of

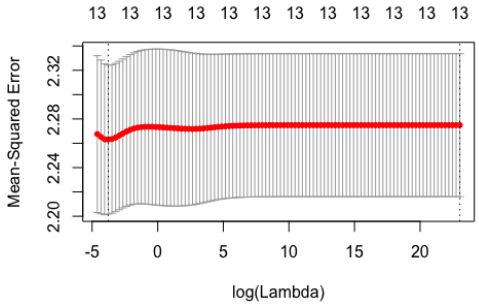

Figure 2: Finding the optimal value of Lambda

lambda as log(lambda). In this experiment, the log value of lambda observed was 0.3053. With $R^2 = 0.4385$ and RMSE = 0.821036

- **Ridge regression with log-transformed data** After performing ridge regression on original data, the model is tested on log-transformed data. With the log-transformed data, the log lambda value is further optimized to 0.1, and with this, the R^2 value also improved to 0.7211 and also the RMSE is reduced to 0.718359

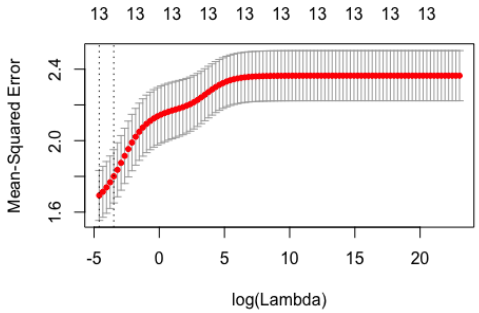

Figure 3: Finding the optimal value of Lambda with Log transformed data

Improvement in the ridge model can be marked from Figure 2 to Figure 3 where it can clearly be seen that the the mean-squared error is reducing with the log transformed data.

- **SVR with hyperparameter tuning and Cross validation** Once the simple linear regression(SLR) models were implemented, the SVR model is used. SVR can be seen as an optimization problem as it works on the principle of maximal margin, which reduces the error in the model by huge value. In this research, the SVR technique used Radius Basis Function (RBF) kernel to build the model. The parameter optimization is done in the SVR using tune() function in R in which maximum allowable error and cost parameters are varied. The RMSE of 0.68 is observed for SVR model. Figure 4 gives an idea of the mean-squared error. The side axis depicts the error. The lighter the color, the lesser is the error. Figure

5 represents the prediction plot of SVR in which the white circles are the actual values, and the red ones are the predicted values.

- **Decision Tree and Random Forest regression models:** The RMSE for the DT regression was found to be as high as 1.293 and the variance percentage explained by RF model was 81.89% with the ntree = 500, which can be considered as a descent value.

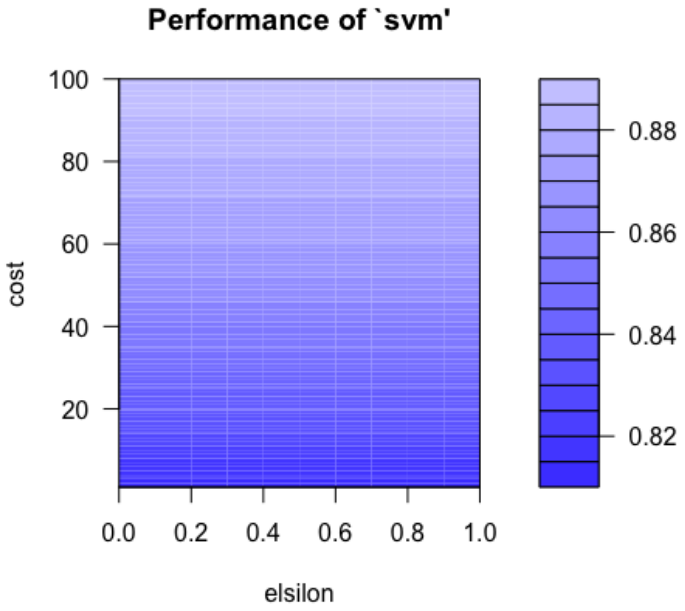

Figure 4: Performance of SVM

- **Decision Tree with simple split:**When trained with the training data with simple split, the DT model yields following result:
  Accuracy = 56.25%
  Kappa = 0.1757
  Detection Rate = 23.33%

- **Decision Tree with rpart.control():** rpart.control() is incorporated with cross validation technique in it. For this research, 10-fold cross validation was applied for DT and following results were obtained:
  Accuracy = 57.55%
  Kappa = 0.3254
  Detection Rate = 36.12%

- **Decision Tree with trainControl() :**Using the trainControl() function, in which parameters like complexity parameter(cp = 0.2) minisplit = 5, minibucket = 5 and maxdepth = 10 were tuned. The results obtained after hyperparameter tuning are as follows:
  Accuracy = 59.01%
  Kappa = 0.4962
  Detection Rate = 48.09%

- **Decision Tree with SMOTE:** To apply the SMOTE function is applied on the optimized parameter model of DT and the results are as follows:
  Accuracy = 65.671%
  Kappa = 0.6062
  Detection Rate = 53.09%

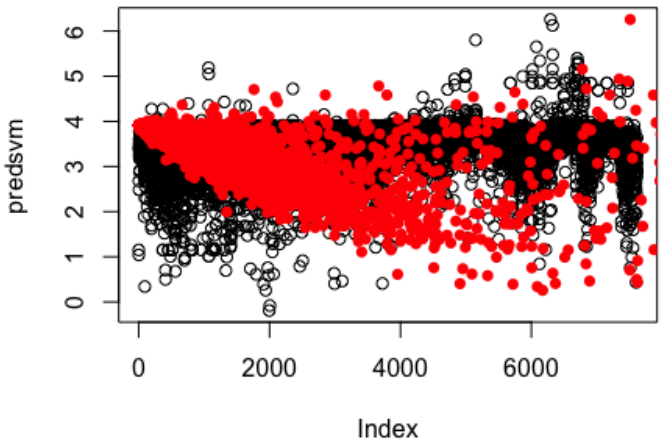

Figure 5: True Response (white) vs. prediction (red) using SVR.

So it can be clearly be inferred that with hyperparameter optimization and SMOTE the performance of DT can be enhanced.

- **Random Forest with Original data:** The random forest does not need hyperparameter optimization. So, first, the RF is implemented on the original data which is unbalanced and obtained the following results:
  Accuracy = 66.25%
  Kappa = 0.3757
  Detection Rate = 43.33%

- **Random Forest with SMOTE:** To overcome the issue of unbalanced data in the RF model, SMOTE function is used, and the results are as follows:
  Accuracy = 89.97%
  Kappa = 0.846
  Detection Rate = 78.90%

5.1 Discussion

In Section 5, all the models were evaluated in several phases of experiments. The first model in this research which implemented was OLS regression. The OLS regression is carried in two phases, one with a simple split and the other one with log-transformed data and parameter tuning. In the first experiment, the RMSE value came up to 1.49, which is considered to be a very high error rate. However, when log-transformed data

along with hyperparameter tuning is applied, the RMSE value dropped down to 0.90, which is very much in the considerable range. Even the adjust-R^2 value and F-statistic value improved with the second experiment. However, if it is compared with the recent study (Razavi and Fleury; 2019) in which the adjust-R^2 improved to 0.83, the improved adjust-R^2 of 0.6354 seems very low. This result explains that the input in the OLS model can explain the variance only up to 63%. As per Table 3, OLS model the socio-economic indicator which is more significant in detecting the electricity theft is unemployment in the male population(Non.Working.Population.Person) based on t-value and p-value.

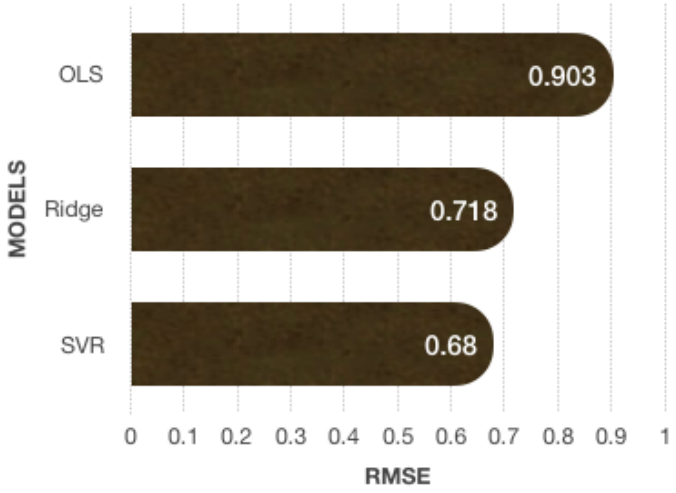

Figure 6: RMSE of Models

Ridge regression was implemented, keeping in mind the shortcomings of OLS regression like multicollinearity, high residual error. Here it can be observed that ridge regression with original data and optimal lambda(hyperparameter) value the R^2 comes to 0.4385 and with log-transformed data, it goes up to 0.7211, which is better than the R^2 value of OLS regression model. Although this research improved the OLS regression model with several techniques and experiments, it still not able to achieve a larger value of R^2 and lowest value of RMSE as it did in the recent study by Razavi and Fleury (2019). This implies that the OLS model can explain 63% variance in the T&D loose with the socio-economic indicators as the input, wherein, Ridge can explain up to the 72% variance.

On the other hand the SVR out performed OLS and Ridge with accuracy and RMSE value of 0.68. As mentioned earlier, the SVR works on the same principle as SVM, it will not wrong to say that the results obtained for it are in trend with the contemporary studies(Razavi and Fleury; 2019).

Table 4: Summary of the data

Models	**Most Significant Variable**
OLS	Non.Working.Population.Male
Ridge Regression	Literates.Population.Person
SVR	No.of.Households
Decision Tree	Non.Working.Population.Male
Random forest	Non.Working.Population.Male

Although the RF model achieved the highest variance of 81.89%, it still not good enough as compared to the state of art model(Razavi and Fleury; 2019), where the author was able to achieve the variance of 89% with RF regression model.

An array of experiments were then performed using the decision tree to enhance its efficiency(Figure 6). Binning approach has been included on an experiment basis. The final results suggest that the decision tree can predict the electricity theft with more accuracy and detection rate for much-balanced data as the SMOTE is performed on training data. The accuracy DT achieved is of 65.67%. Even for the RF, the balanced data produced some exceptional results with the accuracy of 89.97% and Detection rate as high as 78.90%.

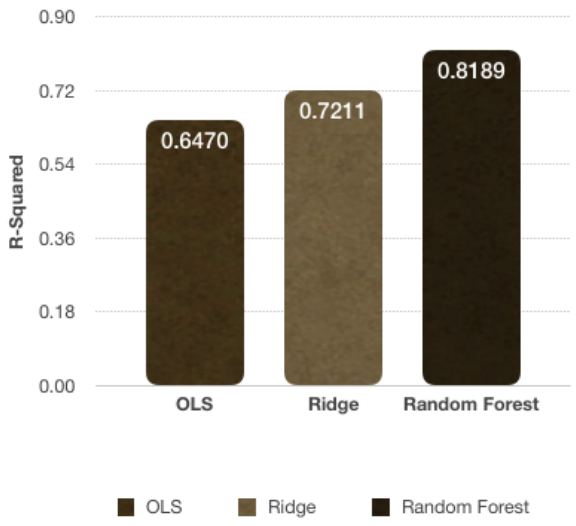

Figure 7: R^2 values for Regression Models

Also, dependence plot is used in this project which has made use of RF model to represent the variation in the response i.e., T&D losses with respect to the change in the significant variables(please refer configuration manual for the plot).

So as per Figure 7, it can be said that the RF regression can explain the variability in the T&D losses better than other regression models. As RF seems to be the most efficient model, the companies that supply electricity can make use of it to construct a road map and can improve their financial as well as supply and demand estimates.

6 Conclusion and Future Work

As the research question framed in Section 1, the study implemented the regression models to investigate the socio-economic indicators in predicting electricity theft. It was observed that RF regression could be more significant in explaining electricity theft with the change in electricity theft than other models. The objective of this project was to investigate if the gender-specific data can create any difference in predicting the electricity theft and it was found that OLS, DT, and RF identified the unemployment among the male population more effective in terms of increase in electricity theft(Table 4). Another objective of this research was to improve the models on given data, and this research was able to achieve that as various experiments using the transformed data are performed, and the models showed the trend of improvement. Although the models were improved, the limitation in the data has effected the outputs. The trend of outputs are in line with the state of art model(Razavi and Fleury; 2019) and other literature, for instance, RF model with highest variance, OLS regression sits at the bottom of the performance matrix(Brosnan (2018), Hashemian et al. (2017), Gaur and Gupta (2016) ,and Razavi and Fleury (2019)). However, the efficiency based on the evaluation matrix, is not achieved.

One limitation which was observed during this research was the limitations of the data. Though with different techniques, this research tried to prepare the data in such a way that the models trained should achieve the maximum level of efficiency, but there are few things which cannot be ignored such as unavailability of data for multiple blocks of districts. In future, by using more sampling methods on the data, the desired efficiency can be achieved. For future study, this research intends to include spatial-temporal data, which may help in marking the regions where electricity theft is dominant. Apart from this, an honest effort has been made in achieving the objectives of the research so that it can be beneficial for the utility providers to frame their financial planning and will also help the authorities in focusing on the areas of the society which are promoting the harmful practices like electricity theft.

Acknowledgement

I want to thank Dr. Paul Stynes and Sachin Sharma for their guidance throughout the research.

References

Ahmad, T. (2017). Non-technical loss analysis and prevention using smart meters, *Renewable and Sustainable Energy Reviews* **72**: 573 – 589.
URL: *http://www.sciencedirect.com/science/article/pii/S1364032117300990*

Babar, Z., Mannan, A., Kamiran, F. and Karim, A. (2015). Understanding the impact of socio-economic and environmental factors for disease outbreak in developing countries, *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 124–131.

Brosnan, S. (2018). The socioeconomic determinants of crime in ireland from 2003-2012, *The Economic and Social Review* **49**: 127–143.

Catlett, C., Cesario, E., Talia, D. and Vinci, A. (2019). Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments, *Pervasive and Mobile Computing* **53**: 62 – 74.
URL: *http://www.sciencedirect.com/science/article/pii/S157411921830542X*

Cody, C., Ford, V. and Siraj, A. (2015). Decision tree learning for fraud detection in consumer energy consumption, *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, pp. 1175–1179.

Depuru, S. S. S. R., Wang, L. and Devabhaktuni, V. (2011). Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft, *Energy Policy* **39**(2): 1007 – 1015. Special Section on Offshore wind power planning, economics and environment.
URL: *http://www.sciencedirect.com/science/article/pii/S030142151000861X*

Depuru, S. S. S. R., Wang, L., Devabhaktuni, V. and Green, R. C. (2013). High performance computing for detection of electricity theft, *International Journal of Electrical Power Energy Systems* **47**: 21 – 30.
URL: *http://www.sciencedirect.com/science/article/pii/S0142061512005947*

Faria, L., Melo, J. and Padilha, A. (2016). Spatial-temporal estimation for nontechnical losses, *2016 IEEE Power and Energy Society General Meeting (PESGM)*.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *Open Journal Systems* **17**: 37 – 51.

Gaur, V. and Gupta, E. (2016). The determinants of electricity theft: An empirical analysis of indian states, *Energy Policy* **93**: 127 – 136.
URL: *http://www.sciencedirect.com/science/article/pii/S0301421516300878*

Hashemian, B., Massaro, E., Bojic, I., Murillo Arias, J., Sobolevsky, S. and Ratti, C. (2017). Socioeconomic characterization of regions through the lens of individual financial transactions., *PLoS ONE* **12**(11): 1 – 20.

Jamil, F. (2013). On the electricity shortage, price and electricity theft nexus, *Energy Policy* **54**: 267–272.
URL: *http://www.sciencedirect.com/science/article/pii/S0301421512010166*

Jamil, F. (2018). Electricity theft among residential consumers in rawalpindi and islamabad, *Energy Policy* **123**: 147 – 154.
URL: *http://www.sciencedirect.com/science/article/pii/S0301421518302398*

Jamil, F. and Ahmad, E. (2019). Policy considerations for limiting electricity theft in the developing countries, *Energy Policy* **129**: 452–458.
URL: *http://www.sciencedirect.com/science/article/pii/S0301421519301181*

Jokar, P., Arianpoo, N. and Leung, V. C. M. (2016). Electricity theft detection in ami using customers consumption patterns, *IEEE Transactions on Smart Grid* **7**(1): 216–226.

Kumar, S., Prasad, J. and Samikannu, R. (2017). Overview, issues and prevention of energy theft in smart grids and virtual power plants in indian context, *Energy Policy* **110**: 365 – 374.
URL: *http://www.sciencedirect.com/science/article/pii/S0301421517305360*

Lewis, F. B. (2015). Costly throw-ups: Electricity theft and power disruptions, *The Electricity Journal* **28**(7): 118 – 135.
URL: *http://www.sciencedirect.com/science/article/pii/S1040619015001633*

Messinis, G. M. and Hatziargyriou, N. D. (2018). Review of non-technical loss detection methods, *Electric Power Systems Research* **158**: 250 – 266.
URL: *http://www.sciencedirect.com/science/article/pii/S0378779618300051*

Min, B. and Golden, M. (2014). Electoral cycles in electricity losses in india, *Energy Policy* **65**: 619 – 625.
URL: *http://www.sciencedirect.com/science/article/pii/S0301421513009841*

Munro, R. and Munro, H. M. (2008). 10 - injuries associated with physical agents, *in* R. Munro and H. M. Munro (eds), *Animal Abuse and Unlawful Killing*, W.B. Saunders, Edinburgh, pp. 70 – 74.
URL: *http://www.sciencedirect.com/science/article/pii/B9780702028786500192*

Nallathambi, S. (2017). Prediction of electricity consumption based on dt and rf: An application on usa country power consumption, *2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE)*, pp. 1–7.

Razavi, R. and Fleury, M. (2019). Socio-economic predictors of electricity theft in developing countries: An indian case study, *Energy for Sustainable Development* **49**: 1 – 10.
URL: *http://www.sciencedirect.com/science/article/pii/S0973082618311177*

Razavi, R., Gharipour, A., Fleury, M. and Akpan, I. J. (2019). A practical feature-engineering framework for electricity theft detection in smart grids, *Applied Energy* **238**: 481–494.
URL: *http://www.sciencedirect.com/science/article/pii/S0306261919300753*

Shaheen, A. and Iqbal, J. (2018). Spatial distribution and mobility assessment of carcinogenic heavy metals in soil profiles using geostatistics and random forest, boruta algorithm, *Sustainability* **10**(3).
URL: *https://www.mdpi.com/2071-1050/10/3/799*

Sharma, T., Pandey, K., Punia, D. and Rao, J. (2016). Of pilferers and poachers: Combating electricity theft in india, *Energy Research Social Science* **11**: 40–52.
URL: *http://www.sciencedirect.com/science/article/pii/S2214629615300293*

Singh, S. K., Bose, R. and Joshi, A. (2017). Pca based electricity theft detection in advanced metering infrastructure, *2017 7th International Conference on Power Systems (ICPS)*, pp. 441–445.

Smith, T. B. (2004). Electricity theft: a comparative analysis, energy policy, **32**: 2067–2076.
URL: *http://www.sciencedirect.com/science/article/pii/S0301421503001824*

Viegas, J. L. and Vieira, S. M. (2017). Clustering-based novelty detection to uncover electricity theft, *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6.

Yurtseven, C. (2015). The causes of electricity theft: An econometric analysis of the case of turkey, *Utilities Policy* **37**: 70 – 78.
URL: *http://www.sciencedirect.com/science/article/pii/S0957178715000429*