Electricity Loss and Fraud Prediction with Deep Learning

Orçun Kitapcı
Department of Computer Engineering
İstanbul Sebahattin Zaim University
Istanbul, Turkey
orcun.kitapci@dedas.com.tr

Alaa Ali Hameed
Department of Computer Engineering
İstanbul Sebahattin Zaim University
Istanbul, Turkey
alaa.hameed@izu.edu.tr

Akhtar Jamil
Department of Computer Engineering
İstanbul Sebahattin Zaim University
Istanbul, Turkey
akhtar.jamil@izu.edu.tr

***Abstract*—In developing countries, energy usage has increased with remaining population, industry, widespread technology and the increasing trend of economy. The main energy source of this increase is electricity. From this perspective, the forecasting of electricity fraud has an important role in the control of this trend to support, run, plan of distribution network's investment. High percentage of fraud in this region damages both the region economy growth and also electricity distribution network. The main source of Fraud usage comes from industry so fraud detection is very hard. So with the correct analysis of daily usage, the usage before theft and last usage of electricity which retrieved from Automatic Meter Reading System (AMRS), we can forecast future theft with Deep Learning. If we use more than one method so we can decide to use one that gives us the best proven.**

Keywords—Electricity Distribution, Lost and Fraud Prediction, Deep Learning, Artificial Intelligence

I. INTRODUCTION

Both electricity loss and fraud may be technical or non-technical. Electricity fraud is generally a non-technical type of fraud, which can be performed in several ways such as intervention in the meter, external intervention to the electricity line, non-payment of invoices, and invoice irregularities [1]. This study will focus on illegal electricity, intervention to the meters or external interventions to the electricity line. With field operation, it is very hard to identify such frauds. So data of the AMRS's which consists of indexes of usage are analyzed with the models of deep learning to predict high possibility theft usage on the power line.

Electricity distribution consists of an interconnected network that distributes electricity at large scale. In some cases, the distribution network spans over many countries. Even a small interruption or outage in the distribution network can cause a disruption at large scale across countries [2]. A simple disruption in a region with high electricity loss and fraud can actually affect the whole system in country and its neighbours. Therefore, it is important to identify the electrical loss and fraud as high outage of electricity effects both production industries and public institutions such as hospitals, schools, sewage treatment, plants.

The literature indicates that some researchers have already work war prediction and identification of electricity theft and loss. Some of them based on the electricity usage data on low tension electric installation [3]. Irregularity detection on low tension electric installations uses ensemble model of neural networks to increase the level of accuracy to identify irregularities. The data retrieved from usage of Light S.A. Company, the Rio de Janeiro in Brazil. The proposal consists of two modules part; one of data mining to obtain the correct subscribers and the other for classification of the subscriber. Each module has 5 neural networks which each of classifies output that indicates whether there is fraud or not. Low tension electricity usages don't decisive to reduce the rate of fraud regionally, however it reduces individual fraud.

On the other hand, the other proposal implements a neural network as previous one but additionally suggest hierarchical model for classification of customers with library of support vector machines (SVM) which is required properly parameters and a training function [4]. When evaluated in terms of feasibility, normally selection of the parameters is very hard because it takes a lot of time for evaluation of instant consumption. With the NN it could be facilitated to predict classification.

Some researchers have proposed methods based on supervised machine learning. For instance, in [5] both gas and electricity consumption data was used. Learning is improved controlling the accuracy with campaign feedback and newly detected technics of fraud usage. The data has no regional characteristics which are not localized and at the same geographical climates also retrieved both gas and electricity consumption. So the learning model could not be used for local consumption and fraud detection.

Since obtaining data on large scale is difficult, therefore the data analysis were performed on a regional basis. Table 1 includes the top 7 distribution companies that are exposed to the most illegal use by distribution region and their fraud rates [6]. Dicle Electricity Distribution Region consists of Diyarbakır, Şanlıurfa, Mardin, Siirt and Şırnak provinces. The region covers most of the Southeastern Anatolian provinces that are developing compared to the west of the country. It is a gateway to countries such as Iraq, Iran, Saudi Arabia and Egypt. Therefore, it has an important position in terms of energy and information transfer to these regions. The Loss and Fraud rate in the region is 51% as of 2020. With this rate, it is seen that there is a long way to go [7].

In addition, the region's other important parameters of energy usage are constantly increasing population and GDP. Population growth has reached 6.5 million from 5 million in 2010 with an annual average increase of 2.5%, and the GDP growth has approached from 36 million TL to 120 million TL with an annual average increase of 19% in 2010 [7].

The development of the region, being a gateway to developing countries, includes important opportunities with its constantly increasing population and GDP. In this respect, loss and fraud prediction is important both to support this development and to provide projection to infrastructure and superstructure works, and to increase the profitability of the company. This will ultimately help the industry to grow and improve the overall economic stability of the country by providing input to economic activities of country.

TABLE 1 DISTRIBUTION REGION LOSS AND FRAUD RATES (%) **[6]**

Distribution Company	Target year 2013	Realized year 2013	Target period (2015)	Realized period (2015)
Dicle Elektrik	71,07	75,41	49,03	-26,38
Vangölü	52,1	65,84	35,94	-29,9
Aras	25,7	27,58	17,73	-9,85
Toroslar	11,8	15,24	10,72	-4,52
Yeşilırmak	9,41	11,47	8,78	-2,69
Akdeniz	8,05	11,32	8,02	-3,3
Boğaziçi	10,76	9,89	9,78	-0,11

In this paper, we performed a detailed comparison of deep learning and traditional machine learning methods for electricity theft detection. With ANN, Random Trees, Logistic Regression, and Linear Support Vector Machines to predict most accurately after classification using training data which is real and on a regional basis. The methodology was validated and its result are detailed in the following section.

The rest of paper consists of five more sections. Section-I explains the methodology, section II explains describes the dataset. The following section IV and V present proposal method experimental results respectively. Finally, the paper is completed with conclusion.

II. METHODOLOGY

There are generally two types of predictions: qualitative and quantitative. Qualitative estimation is mostly based on the interpretation of experts who have experience in the field. Quantitative estimation is based on accepted and proven mathematical functions. Therefore, in this article, a comparison has been made using the quantitative estimation method such as ANN (Artificial Neural Network, Artificial Neural Networks), Random Forest, Random Trees, Logistic Regression, Linear Support Vector Machines. Following sections describe the algorithms used in this study.

A. Artificial Neural Network

Artificial neural networks are a computing technology inspired by the information processing technique of the human brain. The operation of the simple biological nervous system is imitated with ANN. In other words, it is digital modeling of biological neuron cells and the synaptic connection between these cells. Neural Network is a structure established in layers. The first layer is called input and the last layer is called output. The layers in the middle are called "Hidden Layers". Each layer contains a certain number of "Neurons". These neurons are linked to each other by "Synapse". Synapses contain a weigh. These weighs indicate how important the information in the neuron to which they are attached.

Consider that x_0 is an input value and the weight in dendrite (w_0) are multiplied, ($x_0 . w_0$) is transmitted to the nerve cell and this multiplication is done in the nerve cell. After all input and weigh multiplied, all these results are summed. In other words, weighted addition is done. Then, after being summed with a bias (b), the activation function is then transferred to the output. This output can be the final output or the input of another cell. Mathematically, weights and inputs are multiplied finally bias is added. Thus, a simple mathematical model is obtained.

B. Random Forest

Random Forest is a highly effective algorithm developed by Breiman (2001). Random Forest is a collective learning algorithm consisting of many individual training data. Random selection is used with generate random sets to create Random Forest Setup. While in standard trees each node is branched using the best split among all variables, in the Random Forest each node is branched using the best among randomly selected subsets of prediction at that node [8]

C. Random Trees

Random Tree is a classification algorithm that generates a tree by taking randomly selected features on a certain number of nodes in each node. There is no pruning and there is an option that allows prediction of class probabilities based on the data set held [9].

D. Linear Support Vector Machine (LSVM)

Support vector machines are mostly used to separate binary classification data, for example separating each data in a data set into female or male. On the other hand, the data can sometimes belong to more than two classes. In such cases, the basic SVM algorithm becomes dysfunctional. For example, the classification of a data set where certain characteristics of dogs of different breeds are kept based on these characteristics [10]. SVM is based on the principle of inherent risk minimization. SVM can be analysed theoretically using concepts in computational learning theory and can achieve good performance in real world problems. Support vector machines are supervised learning models that select a small number of critical boundary samples called support vectors from each class and create a linear discriminant function that separates them as much as possible [11].

III. DATASET

A. Training Data

As a requirement of deep learning, we will first try to obtain training data. Firstly, training data was obtained by using the data obtained from Automatic Meter Reading

System (AMRS) and enriching the data obtained in SYS (Field Management System). The subscriber information and the date when the fraud was detected was obtained from the SYS system. On the other hand, from the AMRS system, the usage before the date of fraud detection (BF), the usage after the date of fraud detection (AF), the date of the prediction (SYS), the indication that shows it is fraud or not are retrieved for four weeks. Special information such as the Subscriber Number in the data provided here has been changed because it is private as summarized in Table 1. In order to understand the increase between these usages, the rates of this information on the basis of rows are used as input parameters. There is a total of 2,993 rows of training data. Since the data here are training data, there is no subscriber information. The second important column here is the Target column. The training

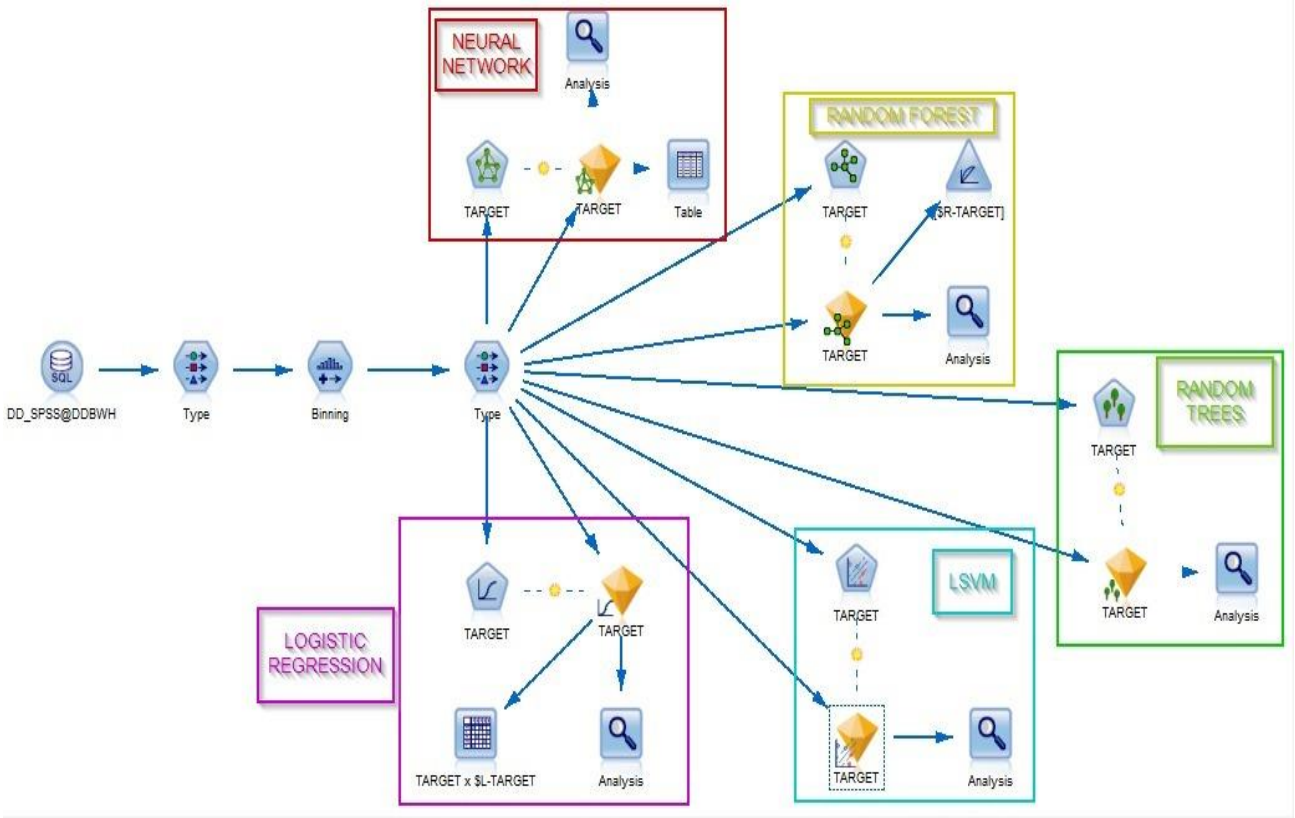

FIGURE 1 PREDICTION MODEL

TABLE 2 SAMPLE CONSUMPTION DATA

No	BF_1	BF_2	BF_3	BF_4	AF_1	AF_2	AF_3	AF_4	SYS_1	SYS_2	SYS_3	SYS_4
1	20,32	20,66	28,07	29,04	21,53	30,53	30,62	38,88	42,07	40,56	37,93	42,93
2	60,9	55,87	58,5	42,62	259,1	170	275,4	269	40,87	25,25	23,65	16,77
3	34,76	34,95	35,29	37,26	36,1	35,73	40,09	45,42	91,74	78,53	77,47	67,41
4	163	165,2	144,7	176,8	72,91	30,28	34,33	33,94	30,03	41,71	29,97	32,2
5	721,3	700,9	962,4	802,9	1322	769,9	620,4	699,3	0	1709	1720	1326

TABLE 3 SAMPLE TRAINING DATA

No	BF_AF_RATE_1	SYS_BF_RATE_1	BF_AF_RATE_2	SYS_BF_RATE_2	BF_AF_RATE_3	SYS_BF_RATE_3	BF_AF_RATE_4	SYS_BF_RATE_4	TARGET
1	0,94	2,07	0,68	1,96	0,92	1,35	0,75	1,48	1
2	0,24	0,67	0,33	0,45	0,21	0,4	0,16	0,39	1
3	0,96	2,64	0,98	2,25	0,88	2,2	0,82	1,81	0
4	2,24	0,18	5,45	0,25	4,22	0,21	5,21	0,18	1
5	0,55	0	0,91	2,44	1,55	1,79	1,15	1,65	0

TABLE 4 SAMPLE TEST DATA

SUBSCRIBER_ID	BF_AF_RATE_1	SYS_BF_RATE_1	BF_AF_RATE_2	SYS_BF_RATE_2	BF_AF_RATE_3	SYS_BF_RATE_3	BF_AF_RATE_4	SYS_BF_RATE_4
10000000	1,16	1,10	2,03	0,94	1,30	0,79	1,54	0,70
11111111	0,37	0,75	0,55	0,85	0,46	0,78	0,89	0,88
22222222	1,00	4,92	0,48	2,96	0,29	5,75	0,26	4,14
33333333	1,15	2,76	0,20	1,78	0,20	2,38	0,19	1,77
44444444	1,93	3,73	0,27	5,43	0,17	5,67	0,20	5,39
55555555	0,92	1,55	0,68	2,53	0,58	1,33	0,65	1,03
66666666	0,82	1,05	0,33	1,69	0,53	1,51	2,37	0,89
77777777	1,20	0,69	0,17	6,26	0,35	3,91	0,83	1,10
77777788	1,00	5,11	0,90	3,21	0,96	2,07	1,18	1,50
99999999	0,00		0,00		4,11	0,01	0,78	0,00

data summarized in Table 3 retrieved also from the AMRS system, including the Target Value information, training data also doesn't contain the subscriber value. The ratio is obtained by dividing each column into one in weekly basis, the columns are summarized in Table 2. If any rate divergent to 1, it indicates that there is probably fraud usage. Because if there are common usages, the consumptions are close to each other so dividing convergent to 1.

TARGET as a boolean field;

$$TARGET = \begin{cases} 1, & There\ is\ Fraud \\ 0, & There\ is\ No\ Fraud \end{cases} \quad (1)$$

B. Prediction Test Data

The data retrieved from the AMRS system, including the subscriber information summarized in Table 4. In this data there is not target column. Each column has its own rate of before usage of fraud, after usage of fraud or consumption on the date of test into one in weekly basis. Each column summarized in Table 4 have similarities with the training data of the table summarized in Table 3 except Target Value and Subscriber ID. Using this dataset in Table 4 prediction will be made and thus the illegal usage of the subscriber will be predicted. The total number of test data is 3.715.

BF/AS	:Before/After Fraud (Pre-Fraud/Post-Fraud)
SYS	:Fraud on System Date
BF_AF	:Ratio of Pre-Fraud to Post-Fraud
SYS_AF	:Ratio of System Date Fraud to Post-Fraud
SYS_BF	:Ratio of System Date Fraud to Pre-Fraud

IV. Proposed Model

In order to detect fraud, we need weekly usage data obtained from AMRS system and fraud data obtained from SYS system to determine whether there is fraud or not. We took pre-fraud and post-fraud data, as well as the ratios between the current usage, and we will get how the usage pattern has changed in these time periods proportionally and linearly. If the use is not fraud, it is expected that the rates before or after the fraud will be convergent to 1, that means, the usage will not change much *(1)*. If there is a fraud, it is expected that there will be no big changes in these ratios and the linear distribution of these rates. The target (2) and (3) equations in which the fraud occurred from the SYS System will enable us to evaluate the realization of these rates and to turn this into a training data. We will be able to obtain the success rate by giving the training data in Table 2 as the input parameter to the Deep Learning methodologies, and then comparing the actual value with the predicted value after the prediction. We will make our prediction by developing 2 projects in SPSS Software. One of them will be our modelling, the other will be the project where our prediction is made with this modelling and the results are obtained.

$$BF_n\_AF_n = \frac{BF_n}{AF_n} \quad (2)$$

$$SYS_n\_AF_n = \frac{SYS_n}{AF_n} \quad (3)$$

A. Modelling

The training data in Table 2 is used as input for our modelling. We classify the rates in our input we use as. We will divide each of our proportions into 36 parts. What is important here is that if our success rate is low, we will increase our success rate by taking our classification lower or higher. Also in here some SPSS Models are used to create prediction model and binning values. After the model are created, it is compiled retrieving data which is already prepared with ETL process on the database preparing a cube and create some binning values to create prediction model.

B. Prediction

After model is ready we can make our prediction on the test data in Table 3 with the classification algorithm we have obtained through modelling. Classified binning values are used to create prediction mode. Each of prediction model of ANN, Random Forest etc. which are created before on the stage of Modelling are used individually to calculate its prediction value of fraud. Also predicted values are viewed, analyzed and confusion matrix calculated by SPSS tools.

V. Experimental Results

Experiments were performed with five different models on our dataset to evaluate their performance. The detailed result of sample data obtained from ANN Model are summarized in Table 5. Because the test data for the other models are similar to Table 5, the test data aren't given for all. In the Test Data, as given in training table, there are rate of the pre-fraud to post-fraud usages, rate of pre-fraud to system date usages, rate of the post-fraud to system date usages for 4 weeks for each. All rows (which means all subscriber in the table) has high probability of fraud usage because the column of Target divergent to 1. This subscriber should be controlled with the field operation to be sure whether there is fraud usage or not. With this operation the test data can be examined. Also comparative table of results that indicates the accuracies are summarized in Table 6 As it can be seen from table, almost all model has the same accuracy except the Random Trees. Accuracy rate are mostly %83. The important amount here is fraud usage prediction. Most of all are very small amount of fraud usage. The main target is to find the fraud usage. So the amount of fraud should be increased. The percentage of accuracy for Random Tree is very low. Because it is a classification algorithm that creates a tree by taking randomly selected features in a certain number of nodes in each node. Our learning and test data are not categorical data, they do not have a tree-pattern structure. Therefore, the accuracy percentage is very low for the Random Tree.

As it can be seen from Table 6, we can evaluate the probability of illegal usage of the relevant subscriber on the basis of Subscriber ID as a percentage in the Target column on the data we predict, and concentrate on the fraud controls of these subscribers. When the weekly consumption chart of these subscribers is examined, the rates of BF_AF and SYS_BF are far away from 1. If these rates can be more less than 1 or greater than 1. This means usage of electricity are not regular. So the irregularity of subscribers' usage can be understood looking at the rate.

VI. Conclusion

In this paper, we exploited different types of machine learning techniques for identification of theft/loss of electricity in various regions in Turkey. The obtained results indicate that machine learning methods were effective. In future, we would like to extend our method on more regions and apply deep learning methods for higher accuracy.

TABLE 5 PREDICTION RESULTS OBTAINED FROM TEST DATA

SUBSCRIBER_ID	BF_AF_RATE_1	SYS_BF_RATE_1	BF_AF_RATE_2	SYS_BF_RATE_2	BF_AF_RATE_3	SYS_BF_RATE_3	BF_AF_RATE_4	SYS_BF_RATE_4	TARGET
222222	1,56	0,00	70,41	0,00	931,80	0,00	1,87	0,00	0,80
333333	3,79	0,04	103,37	0,00	10,08	0,00	2,56	0,00	0,80
444444	1,26	2,33	89,69	0,03	6,30	0,35	2,01	1,00	0,80
555555	0,67	0,02	91,33	0,02	0,93	0,02	197,45	0,01	0,80
666666	0,02	37,38	104,62	0,63	0,52	0,05	0,03	1,48	0,80

TABLE 6 CONFUSION MATRIX OF PREDICTION FOR EACH MODEL (%)

	ANN	
	Predicted No Fraud	Predicted Fraud Exists
Actual No Fraud	2484	3
Actual Fraud Exists	502	3
$Accuracy = \frac{(TP + TN)}{Total}$	83%	

	Random Forest	
	Predicted No Fraud	Predicted Fraud Exists
Actual No Fraud	2484	3
Actual Fraud Exists	497	8
$Accuracy = \frac{(TP + TN)}{Total}$	83%	

	Random Trees	
	Predicted No Fraud	Predicted Fraud Exists
Actual No Fraud	1707	780
Actual Fraud Exists	149	356
$Accuracy = \frac{(TP + TN)}{Total}$	69%	

	Logistic Regression	
	Predicted No Fraud	Predicted Fraud Exists
Actual No Fraud	2484	3
Actual Fraud Exists	497	8
$Accuracy = \frac{(TP + TN)}{Total}$	83%	

	LSVM	
	Predicted No Fraud	Predicted Fraud Exists
Actual No Fraud	2487	0
Actual Fraud Exists	503	2
$Accuracy = \frac{(TP + TN)}{Total}$	83%	

VII. REFERENCES

[1] B. S. Thomas, "Electricity theft: a comparative analysis," *ELSEVIER - ENERGY POLICY*, p. 1, 2013.

[2] Ö. TUTTOKMAĞI and A. KAYGUSUZ, "Büyük Ölçekli Elektrik Kesintilerinin İncelenmesi," *BEU Journal of Science*, no. İnönü Üniversitesi, Elektrik-Elektronik Mühendisliği, Malatya, 2019.

[3] C. Muniz, K. Figueiredo, M. Vellasco, G. Chavez and M. Pacheco, "Irregularity Detection on Low Tension Electric Installations by Neural Network Ensembles," Vols. June 14-19, no. Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, 2009.

[4] S. S. S. R. Depuru, L. Wang, V. Devabhaktuni and P. Nelapati, "A hybrid neural network model and encoding technique for enhanced classification of energy consumption data," no. IEEE Power and Energy Society General Meeting, 2011.

[5] B. Coma-Puig, J. Carmona, R. Gavald`a, S. Alcoverro and V. Martin, "Fraud Detection in Energy Consumption: A Supervised Approach," *IEEE International Conference on Data Science and Advanced Analytics*, 2016.

[6] "Kaçak Elektrik ile Mücadele Üzerine Bir Değerlendirme," 2013. [Online]. Available: https://www.pwc.com.tr/tr/sektorler/enerji-altyapi-madencilik/enerji-spotlights/kacak-elektrik-ile-mucadele-uzerine.html.

[7] Türkiye İstatistik Kurumu, "İstatistik Göstergeler," 15 Mayıs 2019. [Online].

[8] S. Kalmegh, "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News," *International Journal of Innovative Science*, pp. 438-446, 2015.

[9] E. Akçetin and U. Çelik, "İstenmeyen Elektronik Posta (Spam) Tespitinde Karar Ağacı Algoritmalarının Performans Kıyaslaması," *Internet Applications & Management*, 2014.

[10] M. R. Ogiela and L. C. Jain, Computational intelligence paradigms in advanced pattern classification, Berlin: Springer, 2012.

[11] I. H. Witten, E. Frank and M. A. Hall, "Data Mining: Practical," *Machine Learning Tools and Techniques*, 2011.