MACHINE LEARNING BASED ELECTRICITY DEMAND FORECASTING

Zeynep Çamurdan, Murat Can Ganiz
Department of Computer Engineering, Marmara University
Istanbul/Turkey
{zeynep.camurdan, murat.ganiz}@marmara.edu.tr

***Abstract*— In this empirical study we develop forecasting models for electricity demand using publicly available data and three models based on machine learning algorithms. It compares accuracy of these models using different evaluation metrics. The data consist of several measurements and observations related to the electricity market in Turkey from 2011 to 2016. It is available in different time granularities. Our results show that the electricity demand can be forecasted with high accuracy using machine learning algorithms such as linear regression and decision trees and publicly available data.**

Keywords— Electricity demand forecasting; Time Series Analysis; Machine Learning Algorithms

I. INTRODUCTION

Electricity is a necessity in the modern world. Adequate power supply enables better public health and economic growth [1]. The demand of electricity forms the basis for power system planning, power security and supply reliability [2]. Demand planning for electricity consumption is a key success factor for the development of any countries. However, this can only be achieved if the demand is forecasted accurately [3]. With a good electricity demand forecasting, the quantity and quality of electric power generated can fulfill the needs of consumers with the minimum operational cost [4]. In this study, we develop and evaluate prediction models for electricity demand using publicly available data. Three models are developed based on several different machine learning algorithms. The developed models are evaluated using several different evaluation metrics that are commonly used in time series analysis and regression. The data consist of several measurements and observations related to the electricity market in Turkey from 2011 to 2016. It is available in different time granularities Our results show that the electricity demand can be forecasted with high accuracy using machine learning algorithms such as linear regression and decision trees and publicly available data.

II. DATASET

The data subject to our analysis is obtained from daily publication of EPİAŞ company[1].

There are fifty-six features and a class label. Class label shows electricity consumption. Some of the features are related to weather conditions such as temperature, wind, humidity and majority of them are related to the electricity market in Turkey. These are amount of PTF (market clearing amount) (TL/MWh), SMF (system marginal amount) (TL/MWh), previous PTF for a month (TL/MWh), previous SMF for a month (TL/MWh), [2]transaction volume(TL), amount of load forecasting plan (MWh), bilateral agreement (MWh), system sales amount (MWh), KGUP(daily production schedule) (MWh), offered buying amount, offered sales amount, market clearing amount, bilateral settlement volume, public bilateral settlement volume, private sector bilateral settlement volume, GOP (day ahead market) volume, DGP (imbalance power market) volume, total market volume, net bilateral settlement amount, GOP settlement amount, DGP instruction amount, total market amount, portfolio revenue-YPG(TL), Unbalance Amount of YEKDEM (mechanism of supporting renewable energy sources) (MWh), Unbalance Cost of YEKDEM(TL), amount of resources which are wind (MWh), geothermal (MWh), biogas (MWh), dam (MWh), lake type (MWh), canal type (MWh), sun (MWh), biomass (MWh), garbage gas (MWh), river type (MWh), reservoir (MWh), total (MWh).

Firstly, basic statistics of all attributes are calculated for understanding features. Minimum, maximum, mean, median, standard deviation and variance of each features were calculated. In Table 1, basic statistics of the some of the best and worst (in terms of mutual information) features are listed in order to give an idea about the features. As can be seen from this table, features takes varying range of values.

TABLE I. SOME FEATURES VALUES ACCORDING TO MI

	Some of Best Features			Some of Worst Features	
	SSM	Bilateral Agr. Volume(MWh)	Bilateral Agr. Amount	SMF	Temperature
Min	0.00	0.00	$1.65x10^4$	0.00	-7.0
Max	$1.85x10^4$	$1.03x10^5$	$2.89x10^4$	$2.00x10^3$	36.0
Mean	$8.48x10^3$	$5.69x10^4$	$1.90x10^4$	$1.43x10^2$	15.7
Median	$8.27x10^3$	$5.68x10^4$	$1.93x10^4$	$1.50x10^2$	15.0
Std	$2.96x10^3$	$1.08x10^4$	$3.76x10^3$	$6.79x10^1$	7.7
Variance	$8.77x10^6$	$1.18x10^8$	$1.41x10^7$	$4.61x10^3$	60.5

Mutual Information (MI) used for feature extraction. In this way features that are most relevant to the predictive modeling are selected.

[1] https://rapor.epias.com.tr/rapor/

[2] **978-1-5386-0930-9/17/$31.00 ©2017 IEEE**

TABLE II. SOME OF BEST AND WORST VALUES OF MI

Best Features	Mutual Information	Worst Features	Mutual Information
Total market amount	0.729	Wind(mph)	0.003
Total market volume	0.630	Humidity(%)	0.004
SSM(MWh)	0.484	Biogas(MWh)	0.004

Mutual Information has a value between zero and one. If the value is close to zero, it means that there is weak relationship between the two relationships. Otherwise if the value is close to one, it means that there is a strong relationship between the two relationships.

According to the results in the Table II, the mutual information score for total market value is 0.729. That is, there is a close relationship between the total market amount and electricity consumption. And the wind value is 0.003. So, there is weak relationship between the wind and electricity consumption. However, it is important to note in here that although the rest of the data is country wide, weather conditions data is obtained for İstanbul which is by far the biggest city and the industrial capital of Turkey.

Features which are related to electricity consumption are system sales amount, day ahead market's volume, KGUP, YAL (0,1,2,Undelivered), YAT(0,1,2), bilateral agreements (public and private sectors), the balancing power market, total market amount, total market volume, some of sources (canal type, biogas, total), amount and costs of YEKDEM imbalance.

System sales amount is the amount of production increase or consumption decrease offered by market participants [5]. KGUP is the production or consumption values that the balance responsible reports to the system operator at the beginning of the balancing power market [5]. YAL is the amount of instructions given to stabilize the system when there is an electrical charge in the system direction [5]. YAT is the amount of instructions given to balance the system when there is a surplus in the system direction [5]. Bilateral agreements is commercial agreements between the real or legal entities and licensed entities for the purchase and sale of electricity [5]. The day-ahead market is the main arena for trading power. Here, contracts are made between seller and buyer for the delivery of power the following day, the price is set and the trade is agreed [5]. Three main ways for trading electricity; bilateral agreements, day ahead planning, balancing power market. Day ahead planning is the name given to the wholesale electricity market which is established for the purchase and sale of electric energy to be delivered one day later and operated by the Market Operator[5]. YEKDEM imbalance cost is the amount of imbalance created by the reconciliation value of the portfolio under the surveillance [6].

III. RELATED WORK

Similar studies to ours are available. For example in [7], Time Series Analysis Model is applied to the electricity consumption of public transportation in Sofia (Bulgaria). In [8], different forecasting methods—autoregressive integrated moving average (ARIMA), artificial neural network (ANN) and multiple linear regression (MLR)—were utilized to train prediction models of the electricity demand in Thailand. The objective was to compare the performance of these three algorithms and the empirical data used in this study was the historical data regarding the electricity demand (population, gross domestic product: GDP, stock index, revenue from exporting industrial products and electricity consumption) in Thailand from 1986 to 2010. In [9], the authors use an ARIMA model to forecast yearly electricity demand in Tamale.

In our study, data related to Turkey's electricity consumption is used. Additionally, we focus on machine learning methods such as Linear Regression, Decision Tree, Random Forest. A rich set of features which are publicly available is used. Examples of such features are the amount of market settlement, bilateral agreement, total market volume.

IV. METHODOLOGY

Dataset is collected from public web pages mainly from the government web sites reporting electric market. With web scraping method, dataset could be downloaded automatically. Database keeps together datasets which is downloaded. MySQL database system iswas used for this. The admin module controls process of dataset. The dataset can vary in terms of time and some circumstances. Therefore, it should be arrangement in the same category and transfer to the database in this wise.

The dataset communicates with database as data transformation module. Pymysql library of Python was used for processing the dataset. After it has processed in the database then the data preprocessing module has occured. Data preprocessing module is analyzed using pandas library as frame. At the end of this process, time series analysis and prediction module has occured. Decision Tree, Linear Regression and Random Forest models were used with scikit learn library for forecasting.

As a result, the models determine amount of electricity to be produced with under different conditions according to time such as days, months and years.

The reporting module shows the results.

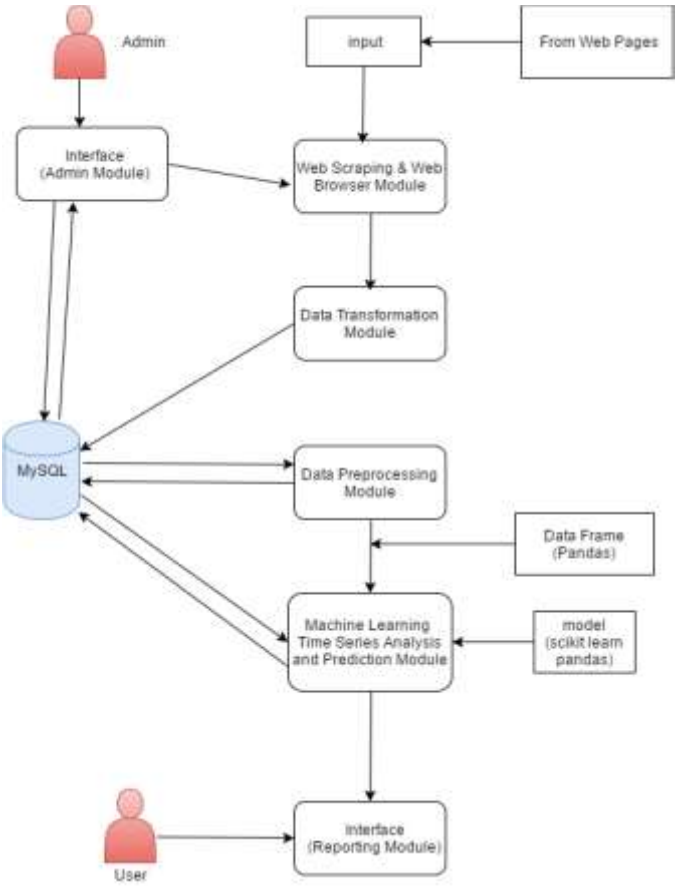

FIG 3. SCHEMA of SYSTEM ARCHITECTURE

V. PLOTS

Scatter plots shows correlation between two attributes. In this project, scatter plots used for understanding relationship between features and class label.

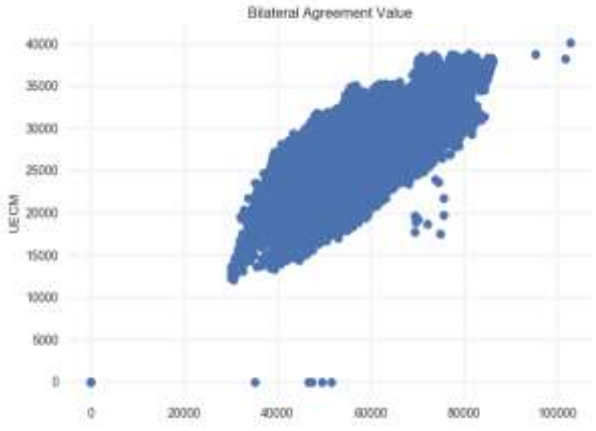

FIG 4. BILATERAL AGR- CLASS CORRELATION

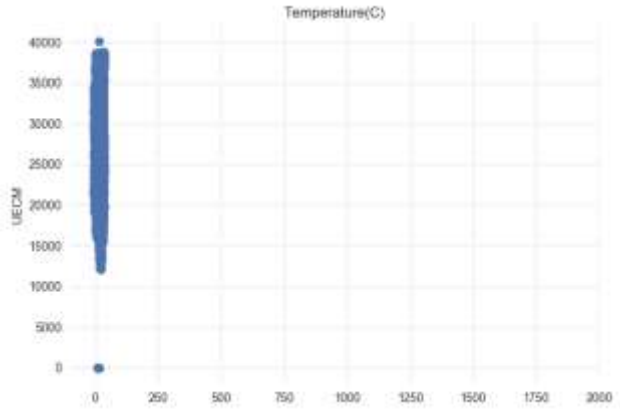

FIG 5. TEMPERATURE-CLASS CORRELATION

There is a positive correlation between bilateral agreement value and class label according to figure 4 and there is no correlation between temperature and class label according to figure 5.

VI. RESULTS

Three different models were used because the data are numeric and the best results are generated with these models. These are decision tree, linear regression and random forest.

A) MODELS

1) Decision Tree

TRAINING	TEST	R^2	MAPE	MAE	MSE
2011–2014	2015-2016	0.97	1.684	450.51	489314.92
2011-2015	2016	0.97	1.860	518.99	438709.92

If the training set is chosen as from 2011 to 2014 and test set is chosen from 2015 to 2016, the result of r square is 0.97, MAPE is 1.684, MAE is 450.51 and MSE is 489314.92. So according to R square, 97% accuracy was reached with decision tree model. According to MAPE, the forecast was made with an 1.7 percentage error. According to MAE, there is an average of 450 differences between the actual value and the forecasting value. According to MSE, the sum of the squares of the error between the actual values and the forecasting values is 489315.

If the training set is chosen as from 2011 to 2015 and test set is chosen 2016, the result of r square is 0.97, MAPE is 1.860, MAE is 518.99 and MSE is 438709.93. So according to R square, 97% accuracy was reached with decision tree model. According to MAPE, the forecast was made with an 1.9 percentage error. According to MAE, there is an average of 519 differences between the actual value and the forecasting value. According to MSE, the sum of the squares of the error between the actual values and the forecasting values is 438710.

When the evaluation metrics are compared, the best result is if the trainingset is chosen as between 2011-2014 and the testset is chosen as between 2015-2016 with decision tree model.

2) Linear Regression

TRAINING	TEST	R^2	MAPE	MAE	MSE
2011–2014	2015-2016	0.98	1.468	394.13	299509.50
2011-2015	2016	0.97	1.908	535.68	405088.18

If the training set is chosen as from 2011 to 2014 and test set is chosen from 2015 to 2016, the result of r square is 0.98, MAPE is 1.468, MAE is 394.13 and MSE is 299509.50. So according to R square, 98% accuracy was reached with linear regression model. According to MAPE, the forecast was made with an 1.5 percentage error. According to MAE, there is an average of 394 differences between the actual value and the forecasting value. According to MSE, the sum of the squares of the error between the actual values and the forecasting values is 299509.

If the training set is chosen as from 2011 to 2015 and test set is chosen 2016, the result of r square is 0.97, MAPE is 1.908, MAE is 535.68 and MSE is 405088.18. So according to R square, 97% accuracy was reached with linear regression model. According to MAPE, the forecast was made with an 1.9 percentage error. According to MAE, there is an average of 536 differences between the actual value and the forecasting value. According to MSE, the sum of the squares of the error between the actual values and the forecasting values is 405088.

When the evaluation metrics are compared, the best result is if the trainingset is chosen as between 2011-2014 and the testset is chosen as between 2015-2016 with linear regression model.

3) Random Forest

TRAINING	TEST	R^2	MAPE	MAE	MSE
2011–2014	2015-2016	0.97	1.394	373.90	390618.82
2011-2015	2016	0.98	1.453	412.48	276778.31

If the training set is chosen as from 2011 to 2014 and test set is chosen from 2015 to 2016, the result of r square is 0.97, MAPE is 1.394, MAE is 373.90 and MSE is 390618.82. So according to R square, 97% accuracy was reached with random forest model. According to MAPE, the forecast was made with an 1.4 percentage error. According to MAE, there is an average of 374 differences between the actual value and the forecasting value. According to MSE, the sum of the squares of the error between the actual values and the forecasting values is 390619.

If the training set is chosen as from 2011 to 2015 and test set is chosen 2016, the result of r square is 0.98, MAPE is 1.453, MAE is 412.48 and MSE is 276768.31. So according to R square, 98% accuracy was reached with linear regression model. According to MAPE, the forecast was made with an 1.4 percentage error. According to MAE, there is an average of 412 differences between the actual value and the forecasting value. According to MSE, the sum of the squares of the error between the actual values and the forecasting values is 2766768.

When the evaluation metrics are compared, the best result of r square is if the trainingset is chosen as between 2011-2015 and the testset is chosen 2016 with random forest model. But according to the other metrics, the best result is if the trainingset is chosen as between 2011-2014 and the testset is chosen as between 2015-2016 with random forest model.

If the models are compared, the best result were achieved by linear regression with 98 percent accuracy rate is if the trainingset is chosen as between 2011-2014 and the testset is chosen as between 2015-2016. And also the best result were achieved by random forest with 98 percent accuracy rate is if the trainingset is chosen as between 2011-2015 and the testset is chosen as 2016.

B) EVALUATION METRICS

An evaluation metric is used to evaluate the effectiveness of information retrieval systems and to justify theoretical and/or pragmatical developments of these systems [10].

Error measurement statistics play a critical role in forecast accuracy.

i) MAPE

The MAPE (Mean Absolute Percent Error) measures the size of the error in percentage terms. It is calculated as the average of the unsigned percentage error [11].

```
100*np.mean(np.abs(ypred[idx]-ytrue[idx])/ytrue[idx])
```

ii) R Square

R square is regression score function. Best possible score is 1.0 and it can be negative. If the r square is 0.97, it means 97 percent accuracy [12].

iii) Mean Absolute Error

Mean Absolute Error(MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes [13].

```
MAE = sum(abs(y-y_pred)) / length(y)
```

iv) Mean Squared Error

Mean Squared Error(MSE) is an average of the squares of the difference between the actual observations and those predicted. The squaring of the errors tends to heavily weight statistical outliers, affecting the accuracy of the results [14].

C) VISUALIZATION

In below, the graphs show actual and predicted values depending on time as year, month, week and day.

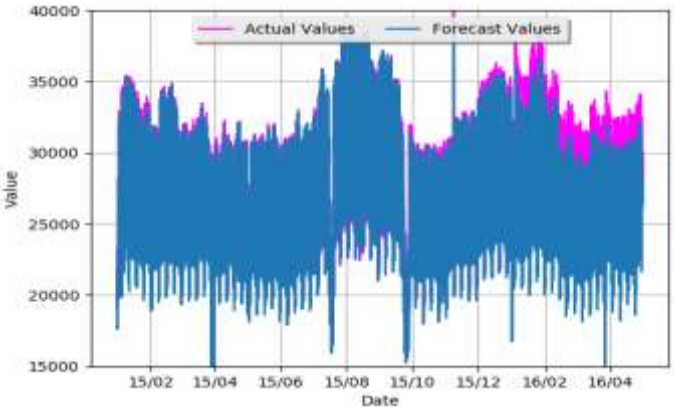

FIG 6. 2015-2016 FORECASTING

This graph shows annual forecasting. It is observed that the actual values and predicted values are very close to each other but generally the predicted values are slightly below the actual values.

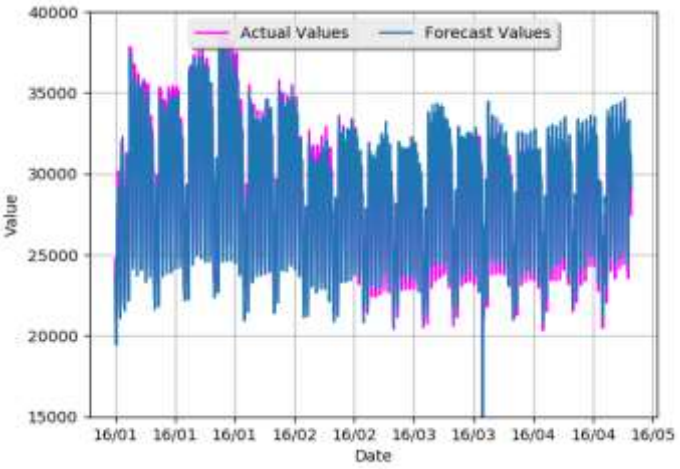

FIG 7. 2016 FORECASTING

This graph shows monthly forecasting. It is observed that actual values are sometimes higher and sometimes lower than predicted but the values are very close to each other.

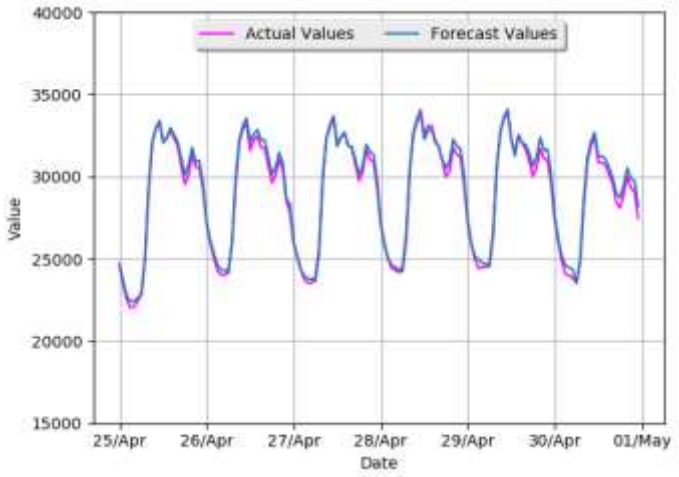

FIG 8. WEEKLY FORECASTING

This graph shows weekly forecasting. It is observed that there is not much difference between the days of the week. Every day there is a similar electricity consumption as the previous day. It is observed that the actual values and predicted values are very close to each other.

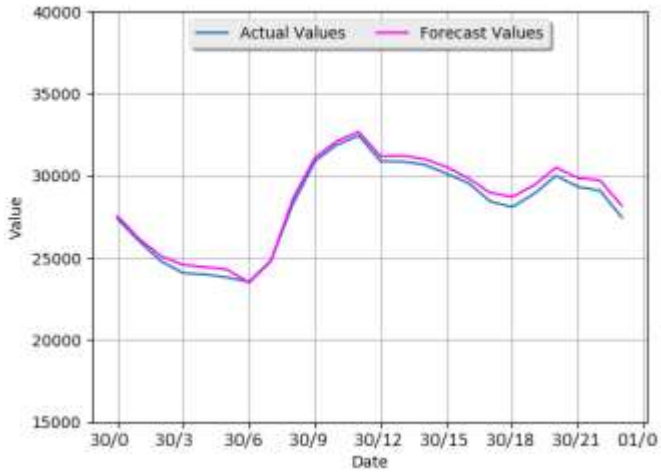

FIG 9. DAILY FORECASTING

This graph shows daily forecasting. It is observed that electricity consumption between 12pm and 6am is the minimum, increase towards noon hours and the most usage is at that time. According to this graph, actual values are sometimes higher and sometimes lower than predicted but the values are very close to each other.

According to the evaluation metrics, accuracy of results of forecasting was 98 percent. It is observed that the values in the all graphs are close to each other and the evaluation results are proved.

The features which are related to Turkey's electricity market have been collected and their effect on electricity consumption have been observed. At first there were 56 different features. It was aimed to increase the electricity demand forecast by extracting features with little or no effect on electricity consumption by performing feature extraction with mutual information, scatter plots methods and statistical results. Thus, the success rate was reached from 86 percent to 98 percent . The most related features with electricity consumption are the total market amount. Total market amount consists of amount of net bilateral agreement , amount of clearing (Sale = Purchase) in the Day Ahead Market, net amount of all instructions (YAL + YAT) in Balancing Power Market.

VII. CONCLUSION

In conclusion, electricity demand was predicted with three different models which are based machine learning algorithm by using some features related to electricity market. Mean absolute percentage error (MAPE), R^2 (R Square), MAE (Mean Absolute Error) and MSE (Mean Squared Error) evalution metrics were used to test the accuracy of the results. According to the R^2, it was observed that up to 98% accuracy was reached with random forest and linear regression models. Otherwise 97% accuracy was reached with decision tree model. And according to the best result of mean absolute percentage error (MAPE), electricity demand was predicted with 1.4 percentage error with random forest model.

REFERENCES

[1] Z. Ismail, F. Jamaluddin and F. Jamaludin. Time Series Regression Model for Forecasting Malaysian Electricity Load Demand. Asian Journal of Mathematics & Statistics, 1: 139-149, 2008

[2] Juan M.Vilar, Ricardo Cao and Germán Aneiros, Forecasting next-day electricity demand and price using nonparametric functional methods, Spain

[3] http://article.sapub.org/, Date accessed: December 12,2017

[4] http://www.mdpi.com/, Date accessed: December 12,2017

[5] http://www.epdk.gov.tr/, Date accessed: April 5,2017

[6] http://enerjienstitusu.com/, Date accessed: April 5,2017

[7] Carmine Tedino, Claudio Guarnacia, Joseph Quartieri, Time Series Analysis and Forecast of the Electricity Consumption of Local Transportation, Italy

[8] Kandananond, K. Forecasting Electricity Demand in Thailand with an Artificial Neural Network Approach. Energies, 4, 1246-1257, 2011

[9] Ghanna, Salifu Katara, Alhassan Faisal, Gideon M. Engmann, A time series analyis of electricity demand in Tamale, International Journal of Statistics and Applications, 2168-5193,2014

[10] Jovan Pehcevski, Benjamin Piwowarski, Evaluation Metrics, France, Latin America, 2009

[11] http://www.forecastpro.com/, Date accessed: April 5,2017

[12] http://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html, Date accessed: March 20,2017

[13] https://www.kaggle.com/wiki/MeanAbsoluteError, Date accessed : March 20,2017

[14] http://www.businessdictionary.com/definition/mean-squared-error.html, Date accessed: April 5,2017