Pakistan Journal of Science, 71 (4 Suppl.): 2019, Page 252-257
ISSN: 2411-0930

LOAD FORECASTING OF AN EDUCATIONAL INSTITUTION USING MACHINE LEARNING: THE CASE OF NUST, ISLAMABAD

A.Y. Kharal[1,*], A. Mahmood[2] and K. Ullah[1]

[1]USPCAS-E, National University of Sciences and Technology, Islamabad, Pakistan
[2]COMSAT University, Islamabad, Wah Cantt Pakistan
[*]Corresponding author's E-mail: 17eepali@uspcase.nust.edu.pk

ABSTRACT: Forecasting has great importance in the electrical power system. It informs the system in advance about the future load demand and assist the system operator to make decisions regarding spinning reserves, economic dispatch, unit commitment and demand side management etc. Infrastructures of Educational institutions demand power with good quality and reliability. Load demand of Educational institutions differs from other consumers because of sensitive laboratory apparatus, sophisticated instruments and machinery. In order to meet the requirements of Educational institutions, integration of renewable energy sources with national grid require knowledge of accurate demand of energy; a key to building managers and planners. Keeping this in mind we tried to forecast load of an Educational Institute of Pakistan as a case study by exploring decision tree-based algorithm (XGBoost). The results were impressive, and predictions seem near optimal. Previously, this data has never been used in forecasting. However, by training an XGBoost model, short term load forecasting (STLF) was done. The results were determined by calculating MAE, MAPE, R2 and RMSE of predicted and actual values. By efficiently using the renewables during the peak hour, peak load demand can be trimmed.

Keywords: Machine Learning, Load Forecasting, XGBoost, Tree Ensemble, STLF.

INTRODUCTION

Sowing Load forecasting plays a key role in Electrical Power System. With an increase in population the demand for energy is increasing day by day, therefore, load forecasting on hourly basis is important to meet the demand. Where forecasting predicts the future load demand, it also has great importance in the inter-related operations of electrical power system such as spinning reserves, preventive and corrective maintenance economic dispatch, unit commitment and demand side management etc. Apart from these benefits accurate planning can be done beforehand, and scheduled maintenance can be performed in anticipation to the load demand. Hence, a shutdown or blackout can be avoided. Load forecasts also helps in managing fuel inventory and price market of electrical energy. The vast horizon of load forecasting is evident from (Kunwar *et al.*, 2013).

An educational institution needs a high quality of electricity supply. This demand depends on the time of year and environment of the location. The consumption of educational institution is different from other consumer sectors (commercial and residential). The semester, working days, vacations, weekends, special occasions e.g. entrance test and convocations and weather conditions are some factors that make the consumption of universities different. Apart from these, sensitive laboratory apparatus and delicate instruments require consistent electrical power of good quality. There is also a room for distributed energy generation in public sectors especially educational institution in Pakistan. Therefore, to avoid any contingency, unplanned occasion and inaccurate generation and demand-balance, load forecasting is the key. Load forecasting in the past has been done through machine learning and artificial intelligence using time series, neural nets, SVR and other hybrid algorithms. The previous research work focused on distribution level of power grid, whereas, load forecasting of educational institutions has been out of focus. Educational institutions require uninterrupted high-quality power which makes it a priority consumer. To manage the load of educational institutions accurate load forecasting is necessary. This research focuses on the load forecasting of an educational institution of Pakistan to meet their desired requirements using XGBoost as a forecasting tool.

MATERIALS AND METHODS

In past, many techniques have been used to predict or forecast load. In previous decades, load forecasting was done using time series models. In (Papalexopoulos *et al.*, 1994), the authors proposed neural networks and explained the draw backs of time series forecasting such as ARIMA, ARMA.

The electric load and weather conditions are stochastic in nature also due to the integration of Renewable Energy Generation (REG), the generation has

become time and weather dependent (Sun *et al.*, 2016). While time series models do not incorporate the weather conditions, time series technique have become obsolete.

In (Srivastava *et al.*, 2016), authors categorized the forecasting techniques in four groups; deterministic, stochastic and heuristic, knowledge based Expert Systems (ES) and Neural Networks (NN). ANNs have gained a strong ground due to their maturity and excessive research in this field and it has become quite mature in its architecture. Forecasting from multilayer perceptron and fully connected layers gives good results (Sahay *et al.*, 2016).

In (Zheng *et al.*, 2017), the authors used a hybrid technique to forecast the load of New England that was provided by ISO. It used XGBoost based k-means to find the similarity between past days and the forecasting day also termed as similar day method (SD). The input data was then decomposed into various Intrinsic Mode Functions (IMF) by using the Empirical Model Decomposition (EMD) algorithm. The LSTM was then fitted separately to each IMP and residuals respectively. In the end, forecasted values were reconstructed from each LSTM model. The model was compared with classical STLF models (LSTM, SD-LSTM, EMD-LSTM, ARIMA, BPNN, and SVR) and were found better than the previous models.

In (Li *et al.*, 2018), the authors forecasted the electrical load by taking the advantage of fog computing environment. The enterprises were clustered into different groups based on their consumption. The input features were determined by Pearson's correlation matrix. After applying initial tests, a predictive model was selected. If the enterprise data exhibited the periodicity, XGBoost was used and if it showed some randomness and passed the stationarity and white noise test, ARMA model was implemented. The results were compared with ARMA, XGBoost, GBDT, and Random Forest and found that XGBoost-ARMA model performed well on the metrics of MAE and Score.

The selection of the model depends on the data sets and availability of computational power. With advanced models, weather signals such as temperature, humidity, wind speed etc. can be given as inputs simultaneously with the load data therefore accurate results having low errors are achieved. In a neural network-based forecasting model capable of adaptation using online learning was used and tested on data of ten different utilities. Although not famous yet decision tree algorithms are making their way in forecasting, (Khan *et al.*, 2018; Warrior *et al.*, 2017; Lobato *et al.*, 2006).

In (Gulin *et al.*, 2014), a load forecasting model for university building using machine learning algorithms was built after feature engineering using principal component analysis and factor analysis. The data set included load consumption with 15-minute interval spanning more than one year of data. The authors built two models using ANN and Support Vector Regression (SVR) and compared their results with actual electric load. This gave maximum error recorded was up to 10%.

Another paper (Gulin *et al.*, 2014), utilized ANN by developing a data set containing lagged load data together with weather and time data. The input to ANN was given considering two scenarios. It predicted the one hour ahead and a day ahead demand of a university in Zagreb. This technique showed that one hour ahead prediction was accurate than the day ahead because only a single inaccurate prediction affected the predictions that followed it as a result model performance dropped.

In (Papadopoulos *et al.*, 2016), load profiling of nine campuses of the Democritus University of Thrace (DUTh) was conducted. Forecasting was done by incorporating the ANNs and Statistical Analysis. However, the proposed methods provided the basics for an in-depth study that can lead the managers of buildings to understand the complexity of the building behaviors and design their REG accordingly. In (Papadopoulos *et al.*, 2016), predicted the load of a Korean University by using the support vector machines for exponential smoothing of the demand curves. It broke the demand into seven intervals with identical trends and found the optimum coefficients of exponential smoothing model.

This paper (Del *et al.*, 2018), uses ensemble technique using decision trees to forecast the load of a university campus in Spain. Depending on the 48-hour ahead predictions, it also developed a practical application for the Spanish Electricity Market that helped the university to save the revenue by participating as a direct consumer rather than buying energy from a retailer.

This case study explores a tree-based ensembling algorithm XGBoost. Tree based models gives liberty to visualize the path of predictions which further helps in analyzing the predictions as well as data. Electrical load is time-series whereas weather is stochastic. However, it is non-linear therefore; non-linear regression is implemented using the XGBoost. The horizon of forecasting is seven days ahead.

Data collection: The data was collected from a 132 KV grid substation specifically designed to feed National University of Sciences and Technology (NUST). The data of current in amperes was recorded from the grid. The outgoing and incoming voltages were assumed to be constant. The data contained hourly log of ampere values for year 2017. The weather data was obtained from weather station installed in building of Centre for Energy (CES) at NUST.

Table 1, gives the insight into the selection of independent variables. Seven features were selected that were broadly categorized into time and weather data. Features were extracted on trial and error method.

Table 1. Features Used In Data Set.

Time data	1.	Hour of day
	2.	Day of week
	3.	Month of year
Weather data	1.	Relative humidity
	2.	Wind speed
	3.	Air pressure
	4.	Temperature

In **Figure 1**, current in amperes is plotted against total number of hours in a year. The figure is for the training data that is used for training the load forecast model.

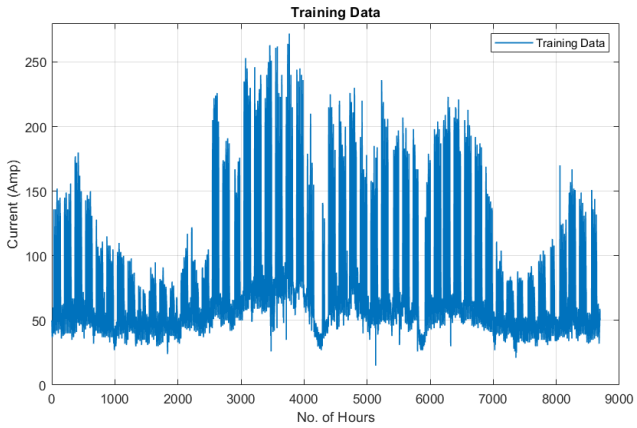

Figure 1. Training Data

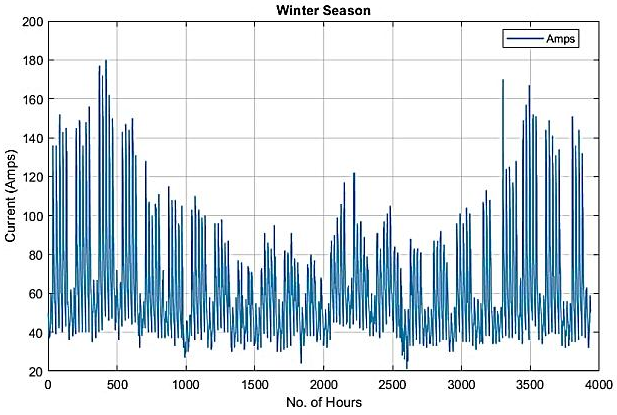

Figure 2. Trend of Demand Current In Winter Season

The data of year 2017 is divided into winter and summer season so that difference of demand in both the seasons is visible. In **Figure 2**, load current of winter season is shown. The maximum current demand in winter season is 180 amps. It is also seen that demand current is relatively less. **Figure 3** shows the demand current of summer season. Current demanded is high as maximum current recorded is 272 amperes. The demand increased as air conditioners and fans come into the system.

Figure 4 is a scatter plot that describes the distribution of load current over the 24-hours of dataset. gradually, and the hostel buildings and street lighting are the only objects of power consumption. The load demand is higher for working hours that is from 9am to 5pm. After 5pm power consumption drops

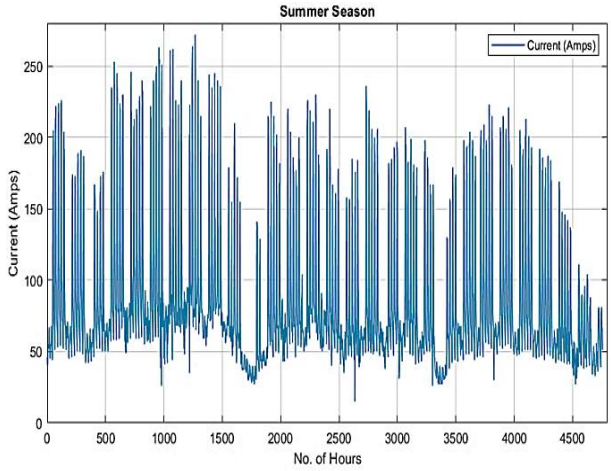

Figure 3. Trend of Demand Current In Summer Season

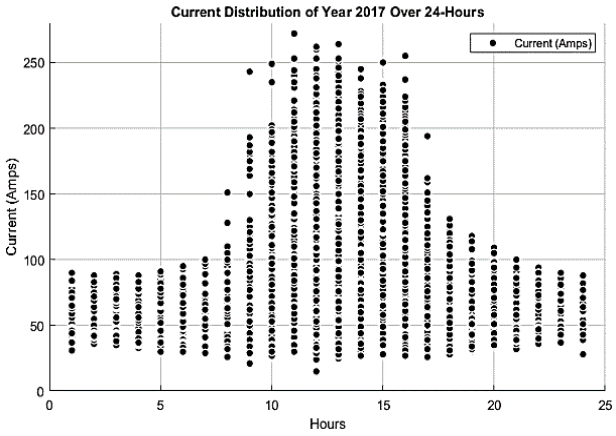

Figure 4. Load Current Distribution Over 24-Hours of Year 2017

Methodology: The aim of this work is also to explore a famous classification algorithm for regression on a new dataset. The model is trained on a dataset comprising of 8696 hours (processed data). Data Preprocessing was done by dropping out the outliers and data points that had missing values.

Table 2 shows the parameters their functions and values that were adjusted to fine tune the model.

Table 2. Parameter of XGBoost, Their Functions and Values.

Sr#	Parameters	Function	Value
1	max_depth	Determines how dense a tree can grow during any boosting round.	15
2	colsample_bytree	Denotes the fraction of columns to be sampled randomly for each tree.	1
3	Subsample	Percentage of data set used per tree.	0.79
4	Min_child_weight	Stops splitting a node once sample size of node goes below a threshold.	1
5	gamma	It describes minimum value of loss required to make a split.	0.8
6	Scale_pos_weight	Controls the imbalanced data.	13

XGBoost: This algorithm is an application of supervised machine learning. XGBoost (Chen and Guestrin, 2016) is a variant of decision tree algorithm and it predicts by ensembling the weak learners that together give the best learner. The predictions of these weak learners are summed up to get the final prediction. Unlike bagged regression tree where trees are built in parallel, here the tree is grown in series or sequential order therefore, fewer trees are required to arrive at a decision which makes it faster and easier to interpret the model.

Decision trees proceed by splitting into nodes. There are many algorithms to find the best split for the tree node. One such algorithm that is also used in XGBoost is the ID3 algorithm. It is defined for the gain. The higher the gain the more favorable is the split. To find the best split, information gain is found for each feature of the data set. With each split more features come into play and more information is added in the tree. Splitting stops when predetermined tree depth or estimators have been satisfied. A flow chart of the algorithm is shown in **Figure 5.**

The complete model can be stated as:

$$\sum_{k=1}^{k} f_k(x_k) \qquad (1)$$

Here, k = number of trees,

y_i =decision from a single tree,

y_i =feature vector

The objective function for the task is as follows,

$$obj = LF + RF \qquad (2)$$

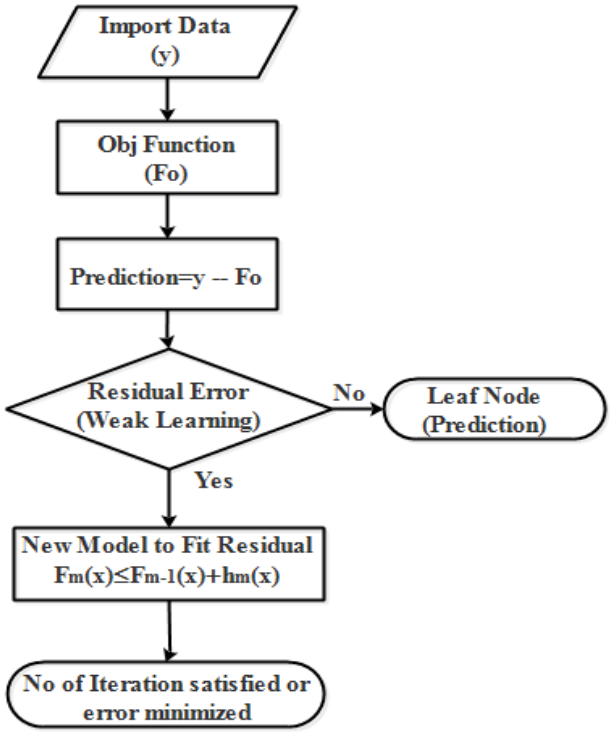

Figure 5. Working of XGBoost

Like many other machine learning algorithms XGBoost uses gradient descent to minimize the loss function whereas regularization is used to avoid the model complexity. From (2) the loss function (LF) is defined as follows:

$$L = \frac{1}{n}\sum_{i-1}^{n}(y_i - y'_i)^2 \qquad (3)$$

Where,

y_i, target variable.

y_i, predicted variable.

The regularization factor (RF) is as follows,

$$\Omega = \gamma T + \frac{1}{2} * \lambda \sum_{j=1}^{T} \omega_j^2 \qquad (4)$$

where,

T=number of leaves,

ω_j =score on the jth leaf of that tree,

γ and λ control the degree of regularization.

It works by building an initial model. This model is called a weak learner. The residual errors of this model are split into two set of values. New model is trained on this splitted data. The residuals of this model are again split in two set of values and so forth. This iterates until the predetermined depth is satisfied. The old

models are ensembled after each split. At each iteration model tries to minimize the error generated by the loss function. The growth of tree stops at a leaf node. The algorithm keeps running until a predefined number of iterations are performed.

RESULTS AND DISCUSSION

We have used Mean Absolute Error (MAE), Mean absolute Percentage Error (MAPE), Coefficient of Determination (R2), Root Mean Square Error (RMSE) for our results.

MAE determines the mean of absolute error between the predicted and actual values. It is calculated as follows:

$$MAE = \frac{1}{n}(\sum_{i=1}^{n} abs(y_i - y'_i)) \quad (5)$$

MAPE is the percentage of MAE. It is calculated as follows:

$$MAPE = \frac{1}{n} * (\sum_{i=1}^{n} abs((y_i - y'_i)/y'_i))) \quad (6)$$

Where,

y_i, actual value of current.

y'_i, the predicted value of current.

n, the total number predicted values.

R2 determines how much the model depicts the actual values. It is measured between 0 and 1. It is calculated as follows:

$$R^2 = 1 - \frac{total\ \text{sum of squares of residuals}}{total\ sum\ of\ \text{squares}} \quad (7)$$

Where,

Total sum of squares of residuals is the sum of squares of the errors between actual and predicted values. Total sum of squares is the sum of the squares of the differences between the dependent variable and its mean.

RMSE is the square root of the mean squared errors. It gives the distance between the predicted and actual values. This shows how close the data is to the best fitted line.

$$RMSE = \sqrt{\frac{(y_i - y'_i)^2}{n}} \quad (8)$$

In **Figure 6,** simulated results were plotted against the real time load. As described earlier, predictions are made for seven days ahead. It can be seen in **Figure 6** that our proposed model traced the target variable. Model performance was recorded in **Table 3**. These predictions were for the first week of January 2018. First five curves were working days and it had high load demand. The next two days were the weekend that had lower power demand.

These load curves shown in **Figure 6**, can be used by the building managers to trim the peaks by integrating the renewable sources. It was observed that the peaks occurred from 12 pm to 1 pm and 2 pm to 3 pm. After 1 pm, one-hour recession was observed in the campus resulting in the lower load demand. There came another smaller peak after the recession period. These three hours can be used as an opportunity for efficiency improvement. The renewables may be utilized to efficiently manage these peaks. The load demand decreased gradually after recession as the campus activities came to the closing hours.

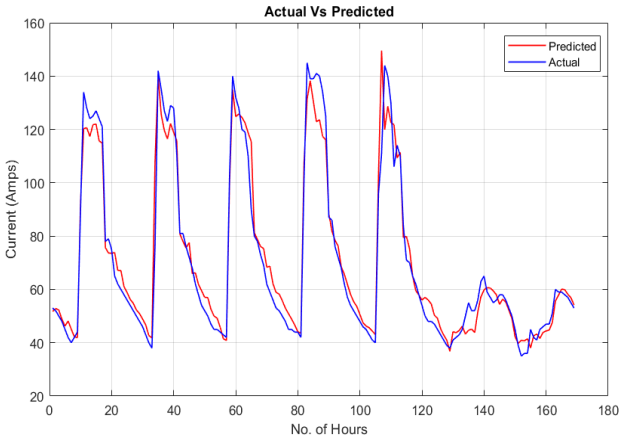

Figure 6. Actual Vs Predicted Load

The results are recorded in **Table 3**

Table3. Simulation Results.

MAE	4.929470
MAPE	7.037071
R2	0.950051
RMSE	7.146219

Conclusion: The implementation of XGBoost for load forecasting of an Educational Institution was done in this case study. Features engineering was done on trial and error method. Data preprocessing was done to refine the data set from outliers. This data was not utilized before therefore a comparison of results could not be drawn. This case study provides a basis for future research in the load forecasting and policy making of education institution. It is also helpful for integrating the renewable resources in the university system. The MAE, MAPE, R2 and RMSE values were achieved by fine tuning the model. The fine-tuned parameters may be utilized for regression of load patterns identical to this research. These results can be improved if we had more historical data, better insight of trends of university events such as

exam days, semester breaks etc. and information of future load to be added in the system. From this research it is pertinent that XGBoost can be used for load forecasting. As compared to Neural Networks, it is relatively a new algorithm as it was built in 2014 and it may need some time to get mature to be used in forecasting. In future, a comparison will be drawn between neural network (trained on this data) and XGBoost. Predictions by building a classifier will be experimented.

REFERENCES

Chen, T. and C. Guestrin (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794.

Del, C., N. Ruiz-Abell, A. Gabaldón and A. Guillamón (2018). Load forecasting for a campus university using ensemble methods based on regression trees. *Energies*, *11*(8).

Gulin, M., M. Vasak, G. Banjac and T. Tomisa (2014). Load forecast of a university building for application in microgrid power flow optimization. *ENERGYCON 2014 - IEEE International Energy Conference*, 1223–1227.

Khan, A.R., S. Razzaq, T. Alquthami, M.R. Moghal, A. Amin and A. Mahmood (2018). Day ahead load forecasting for IESCO using Artificial Neural Network and Bagged Regression Tree. *Proceedings - 2018, IEEE 1st International Conference on Power, Energy and Smart Grid, ICPESG 2018*, 1–6.

Kunwar, N., K. Yash and R. Kumar (2013). Area-load based pricing in DSM through ANN and heuristic scheduling. *IEEE Transactions on Smart Grid*, *4*(3), 1275–1281.

Li, C., X. Zheng, Z. Yang and L. Kuang (2018). Predicting Short-Term Electricity Demand by Combining the Advantages of ARMA and XGBoost in Fog Computing Environment. *Wireless Communications and Mobile Computing*, *2018*.

Lobato, E., A. Ugedo, L. Rouco and F.M. Echavarren (2006). Decision trees applied to spanish power systems applications. *2006 9th International Conference on Probabilistic Methods Applied to Power Systems, PMAPS*, 9–14.

Papadopoulos, T.A., G.T. Giannakopoulos, V.C. Nikolaidis, A.S. Safigianni and I.P. Panapakidis (2016). Study of electricity load profiles in University Campuses: The case study of democritus university of thrace. *IET Conference Publications*, *2016*(CP711), 1–8.

Papalexopoulos, A.D., S. Hao and T.M. Peng (1994). An implementation of a neural network based load forecasting model for the EMS. *IEEE Transactions on Power Systems*, *9*(4), 1956–1962. https://doi.org/10.1109/59.331456

Sahay, K.B., S. Sahu and P. Singh (2016). Short-term load forecasting of Toronto Canada by using different ANN algorithms. *2016 IEEE 6th International Conference on Power Systems, ICPS 2016*. https://doi.org/ 10.1109/ICPES. 2016.7584044

Srivastava, A.K., A.S. Pandey and D. Singh (2016). Short-term load forecasting methods: A review. *International Conference on Emerging Trends in Electrical, Electronics and Sustainable Energy Systems, ICETEESES 2016*, (April 2001), 130–138.

Sun, X., P.B. Luh, K.W. Cheung, W. Guan, L.D. Michel, S.S. Venkata and M.T. Miller (2016). An Efficient Approach to Short-Term Load Forecasting at the Distribution Level. *IEEE Transactions on Power Systems*, *31*(4), 2526–2537.

Warrior, K.P., M. Shrenik and N. Soni (2017). Short-term electrical load forecasting using predictive machine learning models. *2016 IEEE Annual India Conference, INDICON 2016*.

Zheng, H., J. Yuan and L. Chen (2017). Short-Term Load Forecasting Using EMD-LSTM neural networks with a xgboost algorithm for feature importance evaluation. *Energies*, *10*(8).