PredXGBR: A Machine Learning based Short-term Electrical Load Forecasting Architecture

Preprint · July 2022

CITATIONS
0

READS
103

3 authors, including:

Rifat Zabin
Chittagong University of Engineering & Technology
9 PUBLICATIONS **40** CITATIONS
SEE PROFILE

Tofael Ahmed
Chittagong University of Engineering & Technology
26 PUBLICATIONS **321** CITATIONS
SEE PROFILE

PredXGBR: A Machine Learning based Short-term Electrical Load Forecasting Architecture

Rifat Zabin[1][0000−0002−1672−7875], Labanya Barua[2], and Tofael Ahmed[3][0000−0003−2661−9445]

[1] Chittagong University of Engineering and Technology
u1602105@student.cuet.ac.bd
[2] Chittagong University of Engineering and Technology
u1602106@student.cuet.ac.bd
[3] Department of EEE, Chittagong University of Engineering and Technology
tofael.cuet.ac.bd

Abstract. The increase of consumer end load demand is leading to a path to the smart handling of power sector utility. In recent era, the civilization has reached to such a pinnacle of technology that there is no scope of energy wastage. Consequently, questions arise on power generation sector. To prevent both electricity shortage and wastage, electrical load forecasting becomes the most convenient way out. Artificial Intelligent, Conventional and Probabilistic methods are employed in load forecasting. However the conventional and probabilistic methods are less adaptive to the acute, micro and unusual change of the demand trend. With the recent development of Artificial intelligence, machine learning has become the most popular choice due to its higher accuracy based on time, demand and trend based feature extractions. Even though machine learning based models have got the potential, most of the contemporary research works lack in precise and factual feature extractions which results in lower accuracy and higher convergence time. Thus the proposed model takes into account the extensive features derived from both long and short time lag based auto-correlation. Also, for an accurate prediction from these extracted features two Extreme Gradient Boosting (XGBoost) Regression based models: (i) PredXGBR-1 and (ii) PredXGBR-2 have been proposed with definite short time lag feature to predict hourly load demand. The proposed model is validated with five different historical data record of various zonal area over a twenty years of-2 time span. The average accuracy (R^2) of PredXGBR-1 and PredXGBR-2 are 61.721% and 99.0982% with an average MAPE (error) of 8.095% and 0.9101% respectively.

Keywords: Electrical Load Forecasting · Load Prediction · XGBoost · Definite time lag · Regression .

1 Introduction

Electricity generation according to demand has always been a matter of great concern in the power sector of a country. Generation must ensure the fulfillment of industrial and domestic demand of all over the particular region at the same time restrain the excess generation to prevent power wastage. Development of technology and necessity of green energy utilization have discovered many probability and prospects in the power sector. PV system, wind energy and other renewable sources are being utilized in building up decentralized stand-alone grid stations[1]. These progresses are in vein if system loss in not reduced significantly. Hence naturally, demand prediction draws attention of researchers. Load prediction is not a whole new concept. It has been implemented for a long time in the grid network. Both qualitative and quantitative methods including curve fitting, decomposition, regression, exponential smoothing etc have been studied and applied conventionally over the time. Eventually the statistical techniques turned into probabilistic and heuristic forms involving Auto Regression (AR), Auto Regression Moving Average (ARMA), Auto Regression Integrated Moving Average (ARIMA) model, Support Vector Machine (SVM) and other computer algorithms[2]. All these probabilistic algorithms stated complex multi-variable mathematical models for solution. The more networks added to the central grid, the more non-deterministic polynomial (NP-hard) problems arises which increases complexity. Further studies approaches to the reduction of the complexity and introduces data driven neural networks. Historical data from past two months to two years built forecasting model and reduced complexity until the data-sets increased enormously with the period of time[3].

In the last decade, machine learning approach has reached to the apex in time series prediction technique. Machine Learning is a branch of Artificial Intelligence. It involves the process of accessing some data sets and learn from them in a way similar to which, human brain learns[4]. Machine learning begins with the processing of data set, learning data, extracting feature and finally gain knowledge. The method has been accepted worldwide because of its computational speed, error-less calculation, and feature adaption. Artificial Intelligence is the main objective of ML. [3]. In the passage of time, machine learning has been deployed within many practical applications. Image recognition, hand-writing and language recognition, home automation, IOT based smart waste collection presented great application of machine learning concept[5]. Load forecasting involves labelled data sets for training and hence classified into supervised learning type. Deep learning approaches later served the purpose involving depth of layers. Not only these methods simplified the total process of load prediction, but also speed up the operation and establish robustness of architecture.

In this paper we propose an approach of Extreme Gradient Boosting (XGBoost) regression technique to develop a robust electric load forecasting model with great accuracy. Our goal is to design feature based XGBoost models and to evaluate and cross validate them with 5 different data sets. Make a comparative study on how the different features effect the model performance.

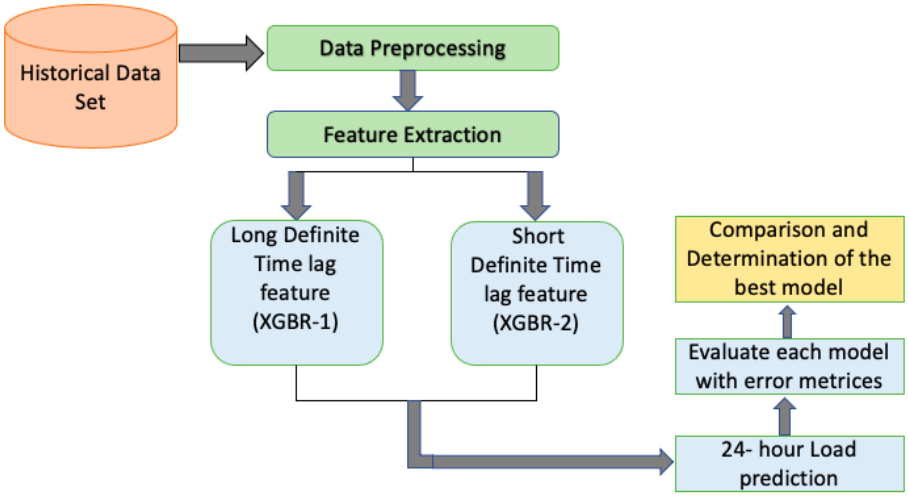


Fig. 1: Load Forecasting Mechanism

The architecture of this paper shows in section 2, works relevant to load prediction using machine learning hence, the main contribution of this study, section 3 presents an elaborated framework on the model, The following sections presents the data preparation, feature extraction and result analysis.

2 Contemporary Research and Authors Contributions

Different approaches of Short-Term Load Forecasting (STLF) have been evidently considered lately as the most effective method for electric load prediction. Many machine learning and deep learning models are derived throughout the passage of time. They have ease the efficient management, economic dispatch and scheduling of generated electrical load[6].
Artificial Neural Network techniques are come out to be most reasonable method of load forecasting. H. Aly proposed six clustering hybrid models [7] consisting of Artificial Neural Network (ANN), Wavelet and Neural Network (WNN) and Kalman Filtering (KF) combination. A regional load forecasting was studied by Saurabh Singh (2017) [18] of NEPOOL region of ISO New England where hourly temperature, humidity and historical electric load data were taken into account. Daniel L. Marino, [8] in one of his studies investigated on conventional LSTM and Sequence to Sequence architecture (S2S) base on LSTM for individual building level forecasting. D.Ageng [9] proposed an hourly load forecasting structure for domestic household merging LSTM and data preparation strategies. A hybrid model was proposed (2021) by Bashir and Haoyong [10] namely Back Propagation Neural network (BPNN) consisting of the combination of Prophet and LSTM. DPSO-LSTM approach was proposed using Discrete Particle Swarm Optimization (DPSO) algorithm by J.Yang.[11] K.Amarasinghe studied on (2017)

classical CNN and bench-marked it against result obtained from other ideal models like LSTM (S2S) [12]. Alhussein [13] proposed a hybrid model namely CNN-LSTM where CNN layers used for feature extraction with LSTM layers for sequence learning.
XGBoost is a recent addition to regression model. Y.Wang[14] applied line regression for trend series and Extreme Gradient Boosting (XGBoost) for fluctuating sub-series data decomposed my VMD and SVMD method.Zheng [15] where a hybrid model was built up involving Similar Day (SD), Empirical Mode Decomposition (EMD) and LSTM combination.

However, most of the contemporary research works related to electric load forecasting are associated with LSTM, RNN, CNN and other statistical algorithms like ARIMA, SVM and so on. To author's knowledge, none of them executed short term definite time lag features and so the characteristics of the trained data record might be random and very much non-linear.
On this aspect our main contributions are:
(i) Designed feature based XGBoost models and evaluated as well as cross validated them with five different data sets.
(ii) Performed a comprehensive comparative study on how different features might impact the model performances.
(iii)Introduced time lag features which improves short term load prediction provided previous 24 hours data available.
(iv)The proposed model performed with an error rate of 1.05% on an average on all the dataset. (v) The operational code has been made publicly available for the ease of further research work.

3 Model Design

Electric load data are unbalanced, non linear and difficult to build up relationships. As stated before, conventional statistical models are insufficient in case of forecasting these type of historical data. XGBoost regression allows a scalable tree boosting algorithm. Our work is mainly associated with this model. The impact of the model drew attention in the kaggle ML competition where most of the winning projects were associated with XGBoost regression and classifier model[17].

XGBoost employs CART (Classification and Regression Tree) to be modified from the residual of each iteration. The principle of CART is a generalized binomial variable called GINI Index. CART facilitate the procedure by allowing splits utilization on the aspect of missing values[16]. Let N be the number of CART, the score by i_{th} sample represented by $f_k(x_i)$.

$$\hat{y}_{xi} = \sum_{k=1}^{N} f_k(x_i), f_k \in \zeta \tag{1}$$

where, $\hat{y}_{xi}$ stands for the final output function, $\zeta = (f_x = w_{q(x)})$ and q represents structure of each tree. Regression tree provides a continuum score as-

sociated with each leaf. This is considered as one of the basic difference between decision and regression tree[17].

By summing up, the minimized objective function hence obtained,

$$\mathcal{L}\{\phi\} = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{2}$$

Where,

$$\Omega(f) = \gamma T + \frac{1}{2}\alpha \parallel \omega \parallel^2 \tag{3}$$

In the above equation (2) and (3), the term 'l' belongs to loss function that determine the difference between the prediction $\hat{y}_i$ and the actual data y_i. The error hence obtained is penalized by the function $Omega(f_k)$. This procedure ultimately smooth the weight of the prediction curve and therefore quality is enhanced by eliminating over-fitting.

The formal objective function we obtain,

$$\widetilde{\mathcal{L}}^{(t)} = \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)\omega_j + \sum_{i\in I_j}(h_i+\alpha)\omega_j^2] + \gamma T \tag{4}$$

I_j stands for all leaf nodes j, g_i and h_i represent the first and second order derivatives. In the equation, ω_j^* is defined as the weight function of j leaf

$$\omega_j^* = -\frac{\sum_i g_i}{\sum_i h_i + \alpha}, where i \in I_j \tag{5}$$

The quality of the regression tree can obtained by the following equation. This is determined over a wide range of objective function, starting from a single leaf and therefore adding branches cumulatively using a greedy algorithm.

$$\widetilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \alpha} + \gamma T \tag{6}$$

Finally, Eq(7) presents the practical formula for evaluation of split candidates.

$$\mathcal{L}_{split} = \frac{1}{2}[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} h_i + \alpha} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} h_i + \alpha} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \alpha}] - \gamma \tag{7}$$

This study is based on XGBoost regression tree algorithm which we make surety to perform extra-ordinarily on time series type forecast. We have two models, one of them works on long term features defined by PredXGBR-1 and the second and the best one works on short term features defined by PredXGBR-2. These characteristics are later discussed in the following sections. However, The two XGBoost models continue iteration until there is presence of residual. It stops at least after 200 runs as soon as the residual goes to null. The model

performs a full iteration for given dataset and thus the parameters are tuned for the subset. The learning process is similar to transfer learning of old parameters. The active function seasonal decomposition is employed here to avoid over-fitting in PredXGBR-1. Each dataset is trained and tested in a particular ratio which is described in later sections.

4 Data Preparation and Feature Extraction

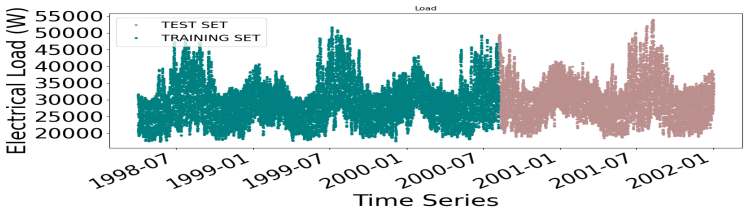
Data selection and preparation Our proposed model has been validated and verified with bunch of dataset. The list of data we employed:

A regional transmission organization which supervise the distribution of wholesale electricity allover the states of Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia and the District of Columbia[]. Long term planning of broad, reliable, efficient and cost- effective interstate electricity wholesale market to ensure coverage for 65 million people. From the historical record, electric load demand from 1998 to 2002 has been extracted.PJM east (PJME) data covers electric load demand within the timeline of 2002-2018 in the eastern region of USA. Likewise PJM West (PJMW) serves the load demand data from 2002 to 2018 over the western region of USA. AEP is known as one of the great investors supplying electricity to almost 11 states over USA. AEP activities involve making strategies and planning through engineering, construction, handling raw materials and renewable energy conversion. Owing almost 38000 Megawatt generation capacity and over 750 KV ultra HV lines, AEP coverage considered as the largest electrical generation company. A complete historical data record between 2004 to 2018 has been employed for the purpose of this study. The DAYTON, Ohio (DPL) power plant is mostly coal-fired electricity generation plants placed in Ohio, meeting the power demand over the state. The plant provides a complete demand record within the timeline of 2004-2018 before it had been reached in an agreement with AES Corp, a global power company in Arlington.

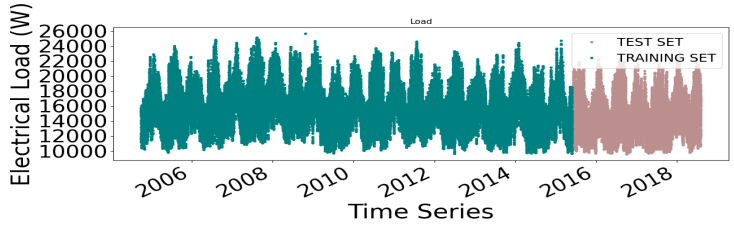
Table 1: Dataset Scheduling

Dataset	Starting Date	Ending Date	Split Ratio	Split Date
PJM	31-12-1998	02-01-2001		07-08-2000
[15 pt] PJM East	31-12-2002	02-01-2018		02-01-2015
PJM West	31-12-2002	02-01-2018	80%	02-01-2015
AEP	31-12-2004	02-01-2018		28-05-2015
DAYTON	31-12-2004	02-01-2018		28-05-2015

The earlier stated data set are the key to our proposed model validation. The records are pre-processed and then trained and tested at a ratio of 80/20 percentage. Fig.2 can clarify the process.



(a)



(b)

Fig. 2: Split up the dataset by pre-processing

Feature Extraction Both of the models are enhanced with unique features. It can be described as below:

Long Time Lag Feature Moving forward to training the model, we initially compose a long term lag feature which can be noted as conventional one. The year, month, day, month of the year, week of the year are the key properties of this feature.

Very Short Term Definite Time Lag Feature When considering this feature, a second model is composed considering the mean and standard deviation value of six hours, twelve hours and 24 hours immediate previous values of load demand.

5 Result Analysis

The collected data set were pre processed and trained by the proposed XGBoost machine learning model.The obtained results are represented and compared in the form of three precision metrics.

R^2 *value* is the measurement of curve fitness indicating the determination of dependant variable by the independent variable. An ideal R^2 value in the scale of 1 should be closest to 1.

$$R^2 = 1 - \frac{\sum(y_i - \widehat{y_i})^2}{\sum(y_i - \overline{y_i})^2} \tag{8}$$

Mean Squared Error (MSE) is the mostly used error function used in machine learning algorithms. Basically MSE is determined by the difference between the predicted value and the ground truth by squaring and followed by taking the average from it.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2 \tag{9}$$

Mean Absolute Percentage Error (MAPE) is slightly different from MSE as it is determined by the absolute value of the difference between predicted value and ground truth, take average and followed by normalized into percentage value. The mathematical expression can be stated as

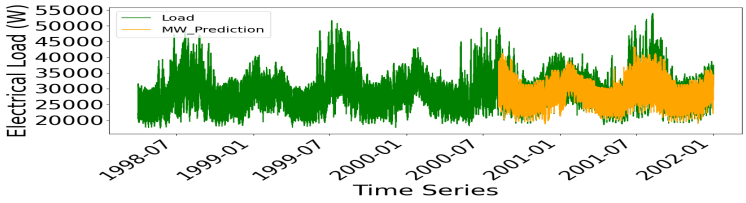
$$MAPE = \frac{1}{n}\sum_{j=1}^{n} \mid \frac{(y_i - \widehat{y_i})}{y_i} \mid \tag{10}$$

Initially model 1 was trained was and tested where PJM interconnection dataset resulted in reasonable R^2 value that is 0.71209. Though the historical record of AEP company showed a very poor value compared to others (0.5762). However the Mean Absolute Percentage Error (MAPE) was calculated worst under the application of PJM East dataset(8.59412) but best under PJM dataset (6.87851). The results of PredXGBR-1 are depicted in Fig.3. The following table-2 presents the corresponding performance metrics results from model 1.

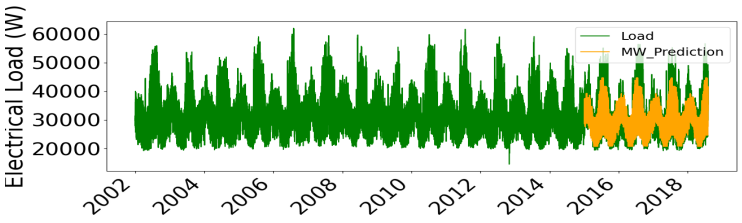
Table 2: Precision Metrics Obtained from Model PredXGBR-1

Model	Dataset	Metrics		
		R^2	MSE	MAPE
PredXGBR-1	PJM	0.71209	9515542.0	6.87851
	PJME	0.581409	443292.8	8.59412
	PJMW	0.59473	335754.53	8.42565
	AEP	0.5762	2575187.3	8.08461
	DAYTON	0.62165	54881.0	8.49580

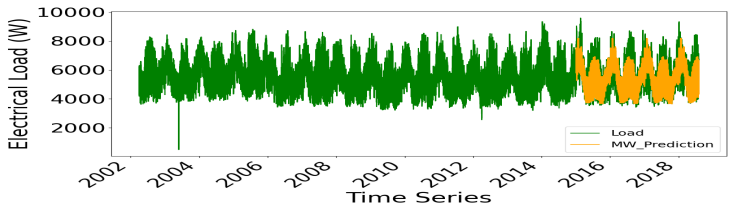
The second model is featured with definite short term lag property. The mean and standard deviation of six hours, twelve hours and twenty-four hours previous value is taken into consideration as feature importance. Consequently the lack of efficiency due to weather condition, peak and off peak hour dependency, seasonal load demand and other factors is minimized by this model. The PJM interconnection organization dataset gives the best outcome in the view of R^2 value (0.991547), whereas PJM West zonal dataset gives comparatively lower outcome though the difference is negligible. On the other hand, AEP generation provides the least MAP error (0.98304) hence considered the effective most dataset, whereas PJM East zone provides comparatively higher error. Again the deviation between these two datsets are negligible. The results obtained from PredXGBR-2 model are depicted in Fig.4.



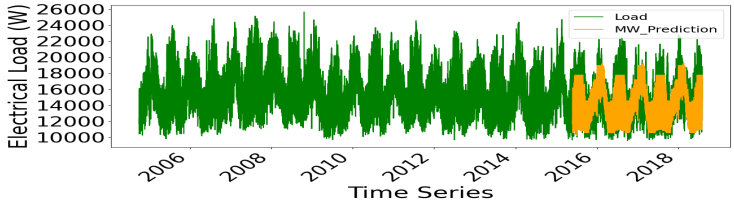
(a) PJM Load



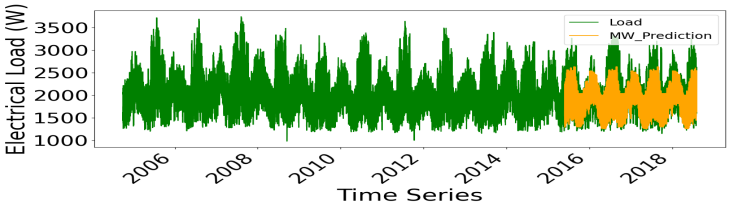
(b) PJME Load



(c) PJMW Load

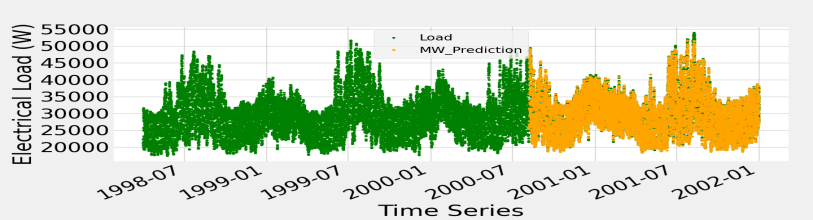


(d) AEP Load

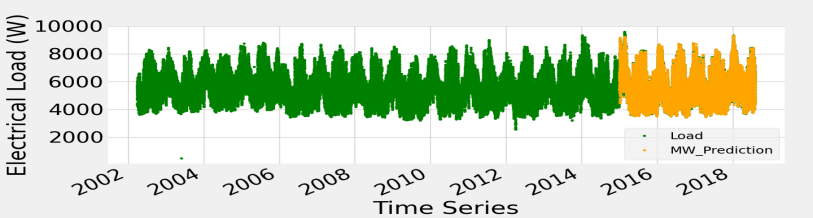


(e) DAYTON Load

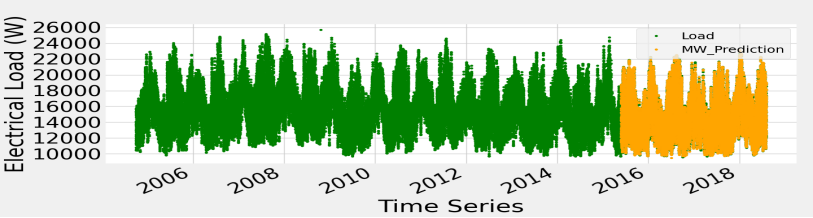
Fig. 3: Electric Load prediction result with PredXGBR-1



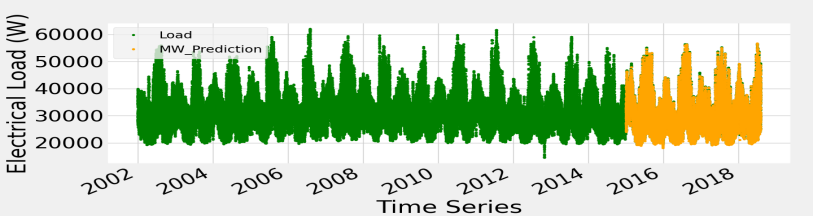
(a) PJM Load



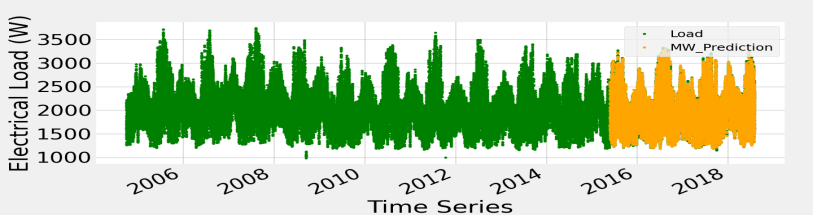
(b) PJME Load



(c) PJMW Load



(d) AEP Load



(e) DAYTON Load

Fig. 4: Electric Load prediction result with PredXGBR-2

The above figures represents the load demand prediction with in the last two or three years of the total record. It is evident that prediction status for model 2 with definite time lag feature provides accurate most output. The following table shows the results of feature 2 model.

Table 3: Precision Metrics Obtained from Model PredXGBR-2

Model	Dataset	Metrics		
		R^2	MSE	MAPE
PredXGBR-2	PJM	0.991547	279381.752	1.0776227
	PJME	0.99114491	368633.416	1.289423
	PJMW	0.98963	10979.769	1.07894
	AEP	0.9912	53421.370	0.98304
	DAYTON	0.99134	1255.176	1.12189

It is evident from the two sets of result that PredXGBR-2 provides far better outcome as it is working based on the very short time lag characteristics. In this way the environmental data including temperature, rain, wind etc are unnecessary, and it is able to adapt with any kind of unexpected change in the demand trend.

6 Conclusion

This paper works on seeking one of the most accurate short-term electric load forecasting model based on Extreme Gradient Boosting (XGBoost) Regression algorithm.. Initially the model extracts a long definite time lag feature which contains days, weeks, years, week of the year, month of the year etc type of data learning characteristics.On the contrary, the proposed model later extracts a "short definite time lag feature" containing the mean and standard deviation of previous few hours demand data. This feature enables the model an explicit training on previous data set. This feature alone can reduce the requirement of many other subsidiary features to be employed. The models PredXGBR-1 and PredXGBR-2 were validated with a wide range of previous data set obtained from five private and regional grid stations of USA with a twenty-years time span. This feature based proposed model is capable of forecasting demand load in hourly basis with about 1.05-1.15% error rate that is, almost 98-99% accuracy.

References

1. Saqib, N., Haque, K.F., Zabin, R., Preonto, S.N.: Analysis of grid integrated pv system as home res with net metering scheme. In: 2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST). pp. 395–399 (2019). https://doi.org/10.1109/ICREST.2019.8644098

2. Singh, A.K., Khatoon, S., Muazzam, M., Chaturvedi, D., et al.: Load forecasting techniques and methodologies: A review. In: 2012 2nd International Conference on Power, Control and Embedded Systems. pp. 1–10. IEEE (2012)
3. Lusis, P., Khalilpour, K.R., Andrew, L., Liebman, A.: Short-term residential load forecasting: Impact of calendar effects and forecast granularity. Applied energy **205**, 654–669 (2017)
4. Zhang, L., Wen, J., Li, Y., Chen, J., Ye, Y., Fu, Y., Livingood, W.: A review of machine learning in building load prediction. Applied Energy **285**, 116452 (2021)
5. Haque, K.F., Zabin, R., Yelamarthi, K., Yanambaka, P., Abdelgawad, A.: An iot based efficient waste collection system with smart bins. In: 2020 IEEE 6th World Forum on Internet of Things (WF-IoT). pp. 1–5 (2020). https://doi.org/10.1109/WF-IoT48130.2020.9221251
6. Chen, K., Chen, K., Wang, Q., He, Z., Hu, J., He, J.: Short-term load forecasting with deep residual networks. IEEE Transactions on Smart Grid **10**(4), 3943–3952 (2018)
7. Aly, H.H.: A proposed intelligent short-term load forecasting hybrid models of ann, wnn and kf based on clustering techniques for smart grid. Electric Power Systems Research **182**, 106191 (2020)
8. Marino, D.L., Amarasinghe, K., Manic, M.: Building energy load forecasting using deep neural networks. In: IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society. pp. 7046–7051. IEEE (2016)
9. Ageng, D., Huang, C.Y., Cheng, R.G.: A short-term household load forecasting framework using lstm and data preparation. IEEE Access **9**, 167911–167919 (2021)
10. Bashir, T., Haoyong, C., Tahir, M.F., Liqiang, Z.: Short term electricity load forecasting using hybrid prophet-lstm model optimized by bpnn. Energy Reports **8**, 1678–1686 (2022)
11. Yang, J., Zhang, X., Bao, Y.: Short-term load forecasting of central china based on dpso-lstm. In: 2021 IEEE 4th International Electrical and Energy Conference (CIEEC). pp. 1–5. IEEE (2021)
12. Amarasinghe, K., Marino, D.L., Manic, M.: Deep neural networks for energy load forecasting. In: 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE). pp. 1483–1488. IEEE (2017)
13. Alhussein, M., Aurangzeb, K., Haider, S.I.: Hybrid cnn-lstm model for short-term individual household load forecasting. IEEE Access **8**, 180544–180557 (2020)
14. Wang, Y., Sun, S., Chen, X., Zeng, X., Kong, Y., Chen, J., Guo, Y., Wang, T.: Short-term load forecasting of industrial customers based on svmd and xgboost. International Journal of Electrical Power Energy Systems **129**, 106830 (2021). https://doi.org/https://doi.org/10.1016/j.ijepes.2021.106830, https://www.sciencedirect.com/science/article/pii/S0142061521000703
15. Zheng, H., Yuan, J., Chen, L.: Short-term load forecasting using emd-lstm neural networks with a xgboost algorithm for feature importance evaluation. Energies **10**(8) (2017), https://www.mdpi.com/1996-1073/10/8/1168
16. Loh, W.Y.: Classification and regression trees. Wiley interdisciplinary reviews: data mining and knowledge discovery **1**(1), 14–23 (2011)
17. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794 (2016)
18. Singh, S., Hussain, S., Bazaz, M.A.: Short term load forecasting using artificial neural network. In: 2017 Fourth International Conference on Image Information Processing (ICIIP). pp. 1–5. IEEE (2017)