Knowledge Transactions on Applied Machine Learning, Vol 01, Issue 04, Sep (2023)
ISSN (Online): 2972-3817

Exploring the Gradient Boosting and LSTM for Power Distribution-based Time Series Analysis

M. Sudha [1*], S. Mehertaj [2]

Research Scholar, Department of mathematics, Amet university, Chennai, India[1]
Assistant Professor, Department of mathematics, Amet university, Chennai, India[2]
seedinmenew@yahoo.com [1*], taj.meher@gmail.com[2]

ABSTRACT

Electrical energy production is an important function in a country that has abundant demand for society. Therefore, power supply forecasting is very significant to keep the balance of power consumption and production and should not be in short supply at any cost. Thus, Forecasting and modeling time series analysis of power production becomes important in prediction. We explore the synergy between Gradient Boosting and Long Short-Term Memory (LSTM) networks in the context of time series analysis. Specifically, we investigate how XGBoost and LightGBM, well-known gradient boosting frameworks, complement LSTM, a recurrent neural network architecture. By combining these methods, we aim to uncover deeper insights and elevate predictive potential for time series forecasting. In this paper, we examine the machine learning methods of forecasting the power generation of 33-year time series data. In the conclusion, we put forward a comparative study between the outcomes attained from the use of LSTM and Gradient-boosting tree-based algorithms XGBoost and LightGBM. The results show that the XGBoost model outclasses the other models with a low error value in forecasting power generation.

Keywords: **XGBoost, LightGBM, LSTM, RMSE, Power Production, Time series Analysis.**

1 INTRODUCTION

XGBoost and LightGBM, two widely recognized gradient-boosting frameworks, can be employed with great effectiveness in the field of time series forecasting.". XGBoost and LightGBM depend on the specific characteristics of the dataset we have selected power distribution time series and the requirements of the machine learning task at hand. It's often a good idea to try both frameworks and compare their performance on our data and this article tries to determine which one works better for our use case. Time series data is helpful in the analysis and forecasting of many real-

https://doi.org/10.59567/ktAML.V1.04.01

time applications like finance, medicine, astronomy, weather forecasting business development etc. Time series analysis is extremely useful in observing data and evaluating the changes in the period.

XGBoost is often used for structured data, where the features are well-defined and have clear relationships. It is commonly applied in scenarios such as predicting customer churn, fraud detection, and credit risk assessment. XGBoost's ability to handle class imbalances through weight adjustments makes it well-suited for tasks where the target classes are not evenly distributed, such as fraud detection or rare disease prediction. It provides built-in feature importance scores, making it useful for identifying which features impact predictions most. This can be crucial for understanding the factors driving a model's decisions. It can handle medium-sized datasets effectively and is known for its robust performance even without extensive hyperparameter tuning [1]. While not as inherently interpretable as linear models, XGBoost models are often easier to interpret than other complex ensemble methods, which can be valuable in domains where interpretability is required. Previous research work might focus on utilizing the efficiency and speed of LightGBM for detecting anomalies in time series data. This could involve incorporating time-related features and optimizing hyperparameters to effectively identify rare events. With the ability of LightGBM to handle high-dimensional data efficiently, recent studies could explore its performance in detecting outliers in datasets with a large number of features, such as in computer vision or genomics [2]. The field of cyber security [3] often involves detecting anomalies in network traffic. Recent research could investigate using LightGBM to analyze network patterns and identify unusual or suspicious behaviors. LightGBM's capability to handle imbalanced datasets makes it well-suited for fraud detection in financial transactions. Recent work [4] focuses on enhancing fraud detection accuracy using LightGBM-based techniques. Monitoring sensor data from industrial equipment and identifying potential anomalies or malfunctions. In health care, recent research explores the use of LightGBM for detecting medical anomalies from various sources, such as patient records, diagnostic images, or wearable devices.

The introduction provides an overview of the article's focus on anomaly detection using LightGBM. It outlines the significance of the investigation and its aims. The related works section discusses prior research in anomaly detection, highlighting gaps. The material and methods section presents the dataset, and preprocessing steps, and introduces LightGBM. The proposed investigation outlines the specific approach and any modifications made. Experiment results detail the setup, metrics, and performance visualization. The discussion interprets results, contrasts with existing methods, and acknowledges limitations. The conclusion summarizes findings, highlights contributions, and suggests future directions.

2 RELATED WORK

This section describes the review of time series forecasting methods carried out so far with LSTM and GBM. The author of reference [1] proved that SVR outperformed the other models such as Convolutional Neural network-LSTM, for stock market forecasting. The works in [2] define that GBM fits well in the data which can evaluate the CPE level effectively. The study [4] examined and compared the models with GBM and LSTM for predicting ether prices. The author of reference [10] proposed a hybrid algorithm with XGB and K-means in predicting similar days of the time series data where the XGB performed well in feature selection. In shortcomings and among many algorithms, the GBM framework performs best [9]. The power load consumption case study LightGBM performs well compared to GRU [10]. As a limitation, there is no capable framework for mining the energy production data [9] The author references [11] proposed an XGBoost model to forecast the stock market time series which is used to frame the decision. The authors of [12] developed a hybrid model of GBM and genetic programming to detect electricity theft. For that, they have used the electricity consumption data of 4000 families.

On the whole, GBM is a successful method to predict the forecasting data. despite many GBM applications in different fields being successful, prediction in electricity energy production and consumption from this is very low. Therefore, we try to analyze the efficiency of the GBM framework on electricity production data in this paper by comparing it with LSTM.

3 MATERIAL AND METHODS:

We selected a time series dataset from the Kaggle public database [14]. This data is about electricity production which involves two columns as attributes. One is the ‘date’ of the production and two has the ‘production count’. This data set has 397 instances which are collected from the time 1985 – 2018 and the production quantiles range from 55.3 – 129.

3.1 LSTM

A branch of machine learning models is an artificial neural network that is used to build a model based on the principles of a human neural network in which recurrent neural network (RNN) is a type. It is regarded as the information’s flow of direction. We know that ANN is capable of classification, data processing, regression analysis, time series prediction, and fitness and modelling. Whereas the bi-directional RNN has the ability to the tasks like speech recognition or handwriting recognition. Long short-term memory (LSTM) is an RNN network designed to pact the gradient problem in RNN. It solves the time gap between input and feedback by having multiple gate structures that propagate the error even in a long sequence or deep network

[5]. LSTM is different in the structure of repeating modules without only a single layer [6].

3.2. XGBoost

The gradient boosting method is a well-suited method for predicting problems (regression and classification), it can handle dissimilar feature issues in prediction and can stabilize the scalability and robustness[15,16]. When computing it has less complexity by having multiple discriminates. Among gradient boosting, the important applied algorithm is Extreme Gradient Boosting which can help to predict in many cases like industry, forecasting, and internet applications where its performance in regression and classification is high. It can deal with the parameters of complex modules severely to sidestep overfitting. Also, it can perform on missing values and initiate cross-validation inevitably [7].

3.3 LightGBM

LightGBM is a framework of gradient boosting that is based on a decision tree and ranking. Contrast to other algorithms. It can find appropriate splits for data samples based on one-side sampling. By its leaf-wise growth pattern, it can minimize the loss to predict better [8]. It has two techniques which are Gradient-based side sampling and exclusive future sampling. Both describe the characteristics of the LightGBM algorithm which increase the proficiency of the model. It will fasten the training process by converting continuous attributes to discrete ones. Also, it can perform well even in large data in a short time.

3.4 Proposed Investigation

After the data selection and pre-processing, we use the following specific approach to conduct mining on the data to predict future electricity production development. The proposed methodology will perform well on the test data only when it is bagged with training data which helps to validate and evaluate our model that may assist to improve the analysis of power generation based on time series. In this study, we utilized RMSE values derived from the data's train-test ratio to assess the effectiveness of methods such as LSTM, XGBoost, and LightGBM for power generation forecasting using time series data. Reference [13] demonstrates that by adjusting the LightGBM objective parameter to 'multiclass,' algorithms outperform in network anomaly detection, providing superior predictive accuracy.

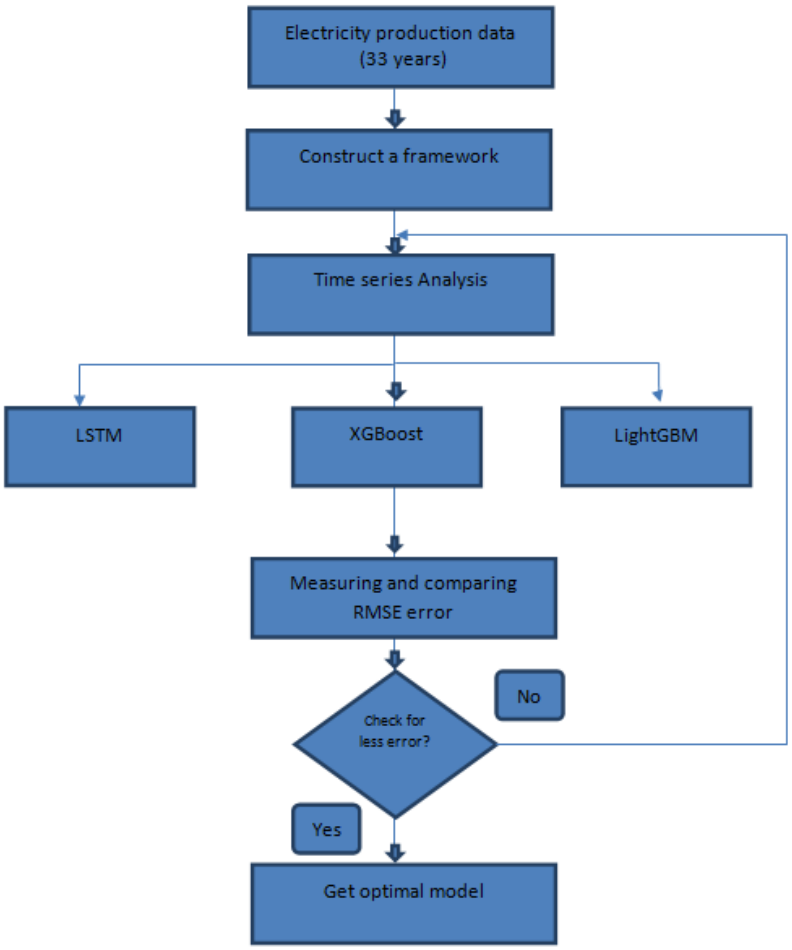

Figure 1: Proposed Framework for Gradient Boosting and LSTM for Power Distribution

4 RESULTS AND DISCUSSION

For estimating the power production dataset, we have performed a cross-validation train-test ratio using Stratified. We have used some current network frameworks and boosting methods to measure the prediction performance of the data. The models were run mostly with default parameters.

Table 1 presents the comprehensive assessment and RMSE values of each approach across various train-test split ratios, revealing the distribution of RMSE errors in power generation forecasting.

Table 1: RMSE Error Values of Methods

S. No	Train-Test Split Ratio	LSTM RMSE Values	XGBoost RMSE Values	LightGBM RMSE Values
1	90:10	8.333782408	3.67246741	3.473369121
2	80:20	9.131604594	3.438439473	3.484333131
3	70:30	9.078793884	3.950814258	3.87642361
4	60:40	10.29599931	4.246434853	4.301240513

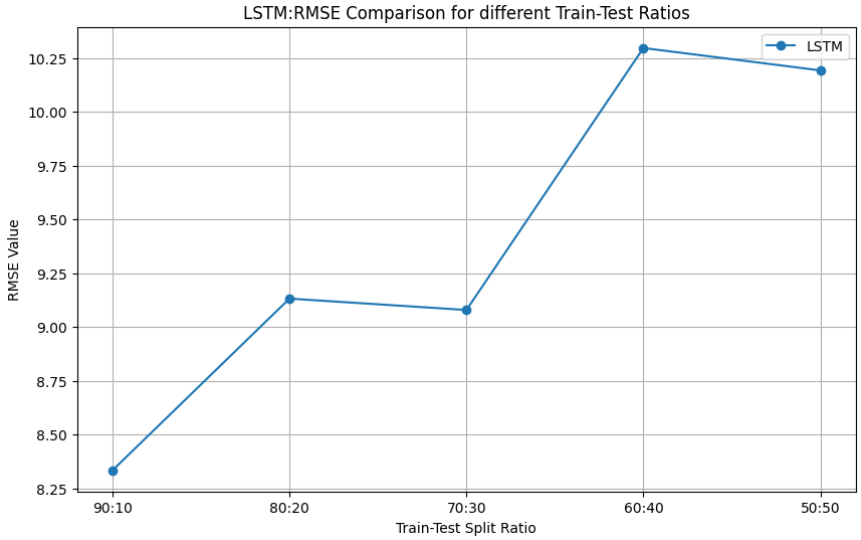

Figure 1. LSTM and RMSE Comparison for Different Train-Test Rations

The lower plot represents the 90:10 ratio, the upper end shows the 60:40 ratio and the error value of 70:30 is comparably low among the best-split ratios 70:30 and 80:20. There is a high variance with a 90:10 split. In this case, the data have more instances than 100, we ignore that ratio. The other ratio is reflected in high significance with others as large errors. An information of comparison to match the error scattering is pictured in Figure 1.

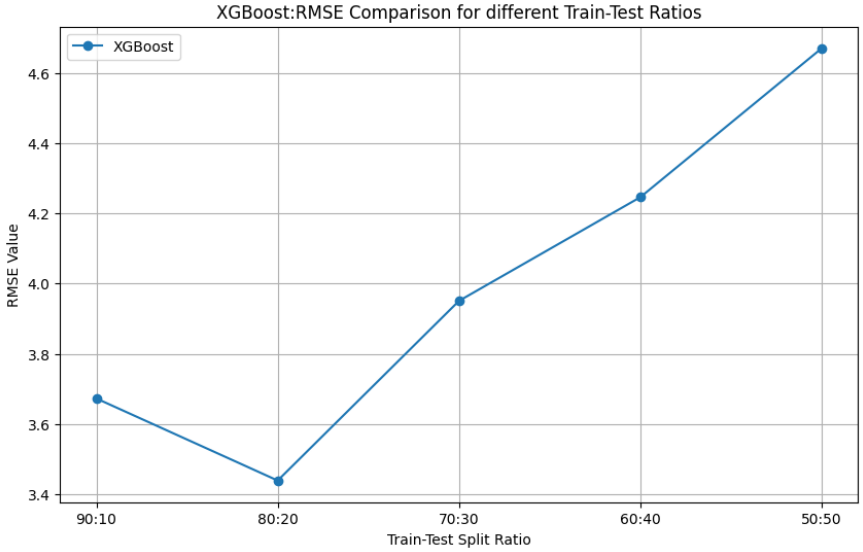

Figure 2. XGBoost, RMSE Comparison for different Train-Test Rations

The lower plot represents the 80:20 ratio, the upper end shows the 50:50 ratio, and the error value of 80:20 which is the best-split ratio is comparably low among all ratios. The other ratio is reflected in high significance with others. A summary of the XGBoost error's distribution is shown in Figure 2.

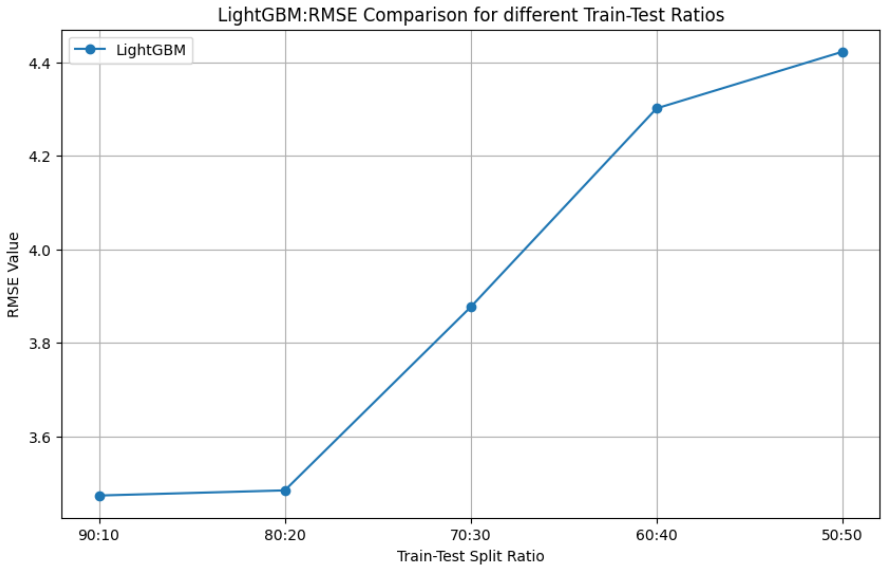

Figure 3. LightGBM: RMSE Comparison for different Train-Test Rations:

The lower plot represents the 90:10 ratio, the upper end shows the 50:50 ratio, and the error value of 80:20 which is the best-split ratio is comparably low among all ratios. There is a high variance with a 90:10 split. In this case, the data have more instances than 100, we ignore that ratio. The other ratio is reflected in high significance with others. A summary of the XGBoost error's distribution is pictured in Figure 3.

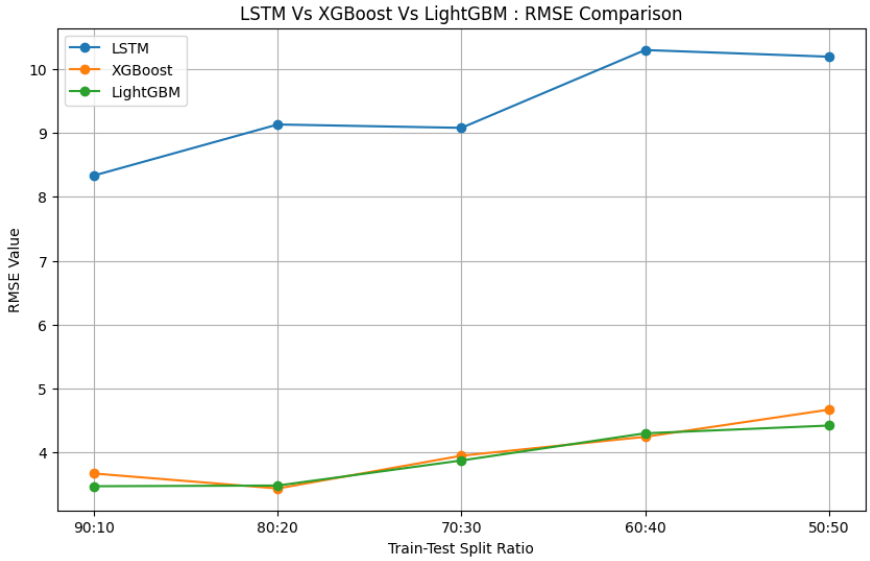

Figure 4: Comparing_Performances_by_LSTM_with_Xtra_and _Light_Boosting_ML _ Ensemble

The above Figure 4 shows the comparison between the methods which explains XGBoost method carries out better than other methods in the time series forecast of power production. Compared with LSTM, boosting methods (XGBoost, LightGBM) claim fewer features that may be used to predict power production in the next period. The XGBoost accomplishes better than the other models due to the data size is small. The root mean square error value of XGBoost of 80:20 split ratio is 3.43 which is comparatively low. Therefore, the XGBoost is the best prediction model for LSTM and LightGBM in power generation prediction.

5 CONCLUSION

In this study, three machine learning methods of forecasting (LSTM, KGBoost, LightGBM) are tested and compared in the prediction of the average power production of 33-year time series data. The results obtained for the gradient boosting tree-based algorithm, XGBoost outperformed the other models with less RMSE value of 3.43. Yet, LightGBM produces the best performance, and XGBoost achieved a

lower error value in the best training-test split ratio than LightGBM. To determine the results precisely, several methods should be tested and further research investigated to improve the performance of Boosting methods for other imminent forecasting analyses.

FUNDING: NA

DATA AVAILABILITY: https://www.kaggle.com/datasets/shenba/time-series-atasets?select=Electric_Production.

CONFLICTS OF INTEREST: The authors declare no conflict of interest.

REFERENCES

1. Oukhouya, Hassan, and Khalid El Himdi. 2023. "Comparing Machine Learning Methods—SVR, XGBoost, LSTM, and MLP— For Forecasting the Moroccan Stock Market" Computer Sciences & Mathematics Forum 7, no. 1: 39. https://doi.org/10.3390/IOCMA2023-14409.
2. D. Tarwidi, S. R. Pudjaprasetya, D. Adytia, and M. Apri, “An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach,” MethodsX, vol. 10, p. 102119, 2023, doi: 10.1016/j.mex.2023.102119.
3. I. Paliari, A. Karanikola and S. Kotsiantis, "A comparison of the optimized LSTM, XGBOOST and ARIMA in Time Series forecasting," 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA), Chania Crete, Greece, 2021, pp. 1-7, doi: 10.1109/IISA52424.2021.9555520.
4. K. Sathiyapriya, S. Vankadara, K. S. Babu and M. Muralidharan, "Performance Comparison of LSTM and XGBOOST for Ether Price Prediction from Spam Filtered Tweets," 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), Coimbatore, India, 2023, pp. 650-655, doi: 10.1109/ICISCoIS56541.2023.10100425.
5. F. A. Gers, J. Schmidhuber, and F. Cummins. “Learning to forget: continual prediction with LSTM”. In: 1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470). Vol. 2. Sept. 1999, 850–855 vol.2. doi: 10.1049/cp:19991218.
6. Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-term Memory”. In: Neural computation 9 (Dec. 1997), pp. 1735–80. doi: 10.1162/neco.1997.9.8.1735.
7. Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System.” In: KDD. Ed. by Balaji Krishnapuram et al. ACM, 2016, pp. 785–794
8. Guolin Ke et al. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: Advances in Neural Information Processing Systems 30. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 3146–3154

9. Xie Bingbing, Zhu Chenliang, Zhao Liang, Zhang Jun,A gradient boosting machine-based framework for electricity energy knowledge discovery Frontiers in Environmental Science,10,2022
10. Aguilar Madrid E, Antonio N. Short-Term Electricity Load Forecasting with Machine Learning. *Information*. 2021; 12(2):50. https://doi.org/10.3390/info12020050.
11. Dezhkam, A.; Manzuri, M.T. Forecasting stock market for an efficient portfolio by combining XGBoost and Hilbert–Huang transform. Eng. Appl. Artif. Intell. 2023, 118, 105626.
12. Razavi, R., Gharipopur, A., Fleury, M., and Akpan, I. J. (2019). A practical feature engineering framework for electricity theft detection in smart grids. Appl. Energy 238, 481–494. doi:10.1016/j.apenergy.2019.01.076.
13. Islam, Md. Khairul & Hridi, Prithula& Hossain, Md Shohrab& Narman, Husnu. (2020). Network Anomaly Detection Using LightGBM: A Gradient Boosting Classifier. 1-7. 10.1109/ITNAC50341.2020.9315049.
14. https://www.kaggle.com/datasets/shenba/time-series-atasets?select=Electric_Production.
15. Lu, W.; Li, J.; Li, Y.; Sun, A.; Wang, J. A CNN-LSTM-based model to forecast stock prices. Complexity 2020, 2020, 6622927.
16. Dezhkam, A.; Manzuri, M.T. Forecasting stock market for an efficient portfolio by combining XGBoost and Hilbert–Huang transform. Eng. Appl. Artif. Intell. 2023, 118, 105626.