Ensemble Machine Learning Modelling for Medium to Long Term Energy Consumption Forecasting

Bhawna Dhupia[a], M. Usha Rani[b]

[a] Research Scholar, [b] Professor, Dept. of Computer Science,
[a, b] SPMVV, Tirupati, India
[a] bhawnasgn@gmail.com, [b] musha_rohan@yahoo.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract: Electricity demand forecasting is an important research area that is gaining popularity among researchers these days. Forecasting energy consumption is a critical task as it affects the overall functioning of the power industry. There are many types of research already done in the field of short-term energy forecasting, but there is a scarcity of researches in medium to long-term energy forecasting. This paper focuses on medium to long-term energy forecasting using machine learning and the ensemble approach. The machine learning methods include Linear Regression (LR), Random Forest (RF), Least Absolute Shrinkage and Selection Operator (LASSO), and Gradient Boosting Regressor (GBR). For the comparative analysis the performance metric selected are R2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE)..

Keywords: Ensemble Model, Machine Learning, Energy Consumption Forecasting, Comparative Analysis

1. Introduction

The prediction of energy consumption can give great help to the decision-makers of the energy sector. There is extensive research work done in the field of household energy consumption prediction. But there is a lack of research in the field of the industrial sector for power consumption analysis and prediction. This research paper focuses on the industrial sector. This research can provide a valuable guideline for the energy management of industrial companies thatare major energy consumers [1]. Since the consumption of energy is based on various factors, it is important to study and analyzeall the factors, before developing an effective model for power consumption prediction.The data generated by the smart grid is very huge. To process the data we utilize, big data analytics technique efficiently. In recent researches, Machine Learning (ML) based algorithms have become very popular for providing predictions. With the help of these models, we can exploit the pattern of energy usage of each customer and can offer customized solutions. Various popular machine learning models are Linear Regression (LR), Random Forest (RF), Gradient Boosting (GBR), and LassoRegression are widely used in various applications for energy prediction [2][3][4]. Nowadays, Deep Learning (DL) is also contributing a lot to energy forecasting research.

In some of the research, DL algorithms performed better in some of the parameters as compared to ML models. Due to the ability to model non-linearity[4] of the data Artificial Neural Network (ANN) was used by many researchers. The structure of ANN helps to improve the accuracy of the model for load forecasting for Short Term Load Forecasting (STLF) [5]. Another researcher [6] proposed a novel energy forecasting methodology using two algorithms namely a standard Long Short-Term Memory (LSTM) and an LSTM with sequence architecture, which gave promising results. Rahman et al. [7] present dual RNNs for medium to long-term electric load prediction for residential buildings. In some researches,the hybrid model also gave a better result as compare to the individual model. The hybrid model can be a combination of more than one model to get an improved result on desired parameters. Zheng et al. [8] developed a model by combining LSTM, Xgboost for feature selection, and K-means for clustering to get an improved result. This approach substantially improved the prediction by LSTM.

The research focuses on short to medium-term forecasting. Medium-term forecasting lasts from months to a year, based on the historical energy consumption data. In this paper, we propose an ensemble model for demand-side load forecasting over short and medium-term forecasting. The performance of the forecasting model is based on the most popular evaluation metrics such as R^2, mean absolute error (MAE) and root mean square error (RMSE) The contribution of the paper is as follows:

- Comparison and findings on ML and DL models resultin model evaluation parameters.
- Proposal of an ensemble model (ML) to improve the accuracy of the prediction. The model is designed to predict the accuracy of energy consumption for medium to long term prediction. The proposed model has shown improvement on all the parameters of evaluation metrics namely, R^2, MAE, and RMSE.

The rest of the paper is organized as follows. Section II introduces the data source, and the pre-processingdetails of data used. In Section III, the details about the model and different phases of the research are

discussed. The main contribution of the paper, that is the proposed ensemble method is also included in this section. Section IV provides the result and discussion on the implementation phase of the existing ML and DL models and proposed ensemble model. Section V will conclude the chapter and discuss the future scope of the research.

2. Data Introduction

Historical Electricity Data: The data was collected from Himachal Pradesh State Electricity Board for the industrial sector situated in Kala Amb, in India. The duration of data collected is from Aug'2018 to July'2020. The data collected from the smart meter device based on the daily consumption of energy by the 65 companies. The reading taken by the smart meter is after every 15 minutes interval. As the research is based on big data analytics, so the amount of data taken for analyses was approximately 4.5 million data samples. There were 6 attributes of the dataset such as Meter_number, Company_name, Account_no, Capture_time, KW_imp, and KV_Imp. The research is based on the prediction of energy consumption for medium to long-term, so we dropped columns of less importance while doing the feature selectionprocess. We visualize the consumption on monthly basis to see the effect of the season on consumption, but it does not show any considerable change. So, we just considered the historical data of energy consumption for further research processing.

Clustering:The dataset we considered is quite big. So, to reduce the size of the data for research purposes, the clustering method is often used in big data analytics [9]. we clustered the whole databased on energy consumption, to evaluate the models on all range of energy consumptions. This paper is a continuous work on pre-published paper titles "Smart Profile: Smart Meter Data-Oriented Customer profiling" [10]. The Elbow method is applied to identify, the adequate number of clusters to categorize the data. The data is divided into four clusters according to the result of the Elbow method. Two companies are chosen randomly from each cluster to justify the implementation of ML and DL models.

Model Evaluation Indexes: Generally, performance metrics for dataset evaluation are mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) [11]. MSE is the average of the squared difference between the target values and the predicted by the regression model. Another widely used metric is RMSE considered for regression analysis. It is the square root of the target value and the value predicted by the model. It penalizes the larger error terms and tends to become increasingly larger than MAE for outliers. MAE is the absolute difference between the target value and the predicted value. All these metrics are explained as the equation for better understanding:

$$MSE = \frac{1}{n}\sum(y_i - \hat{y_i})^2 \quad (1)$$

$$RMSE = \sqrt{\sum_{i=0}^{n} \frac{(y_i - \hat{y_i})^2}{N}} \quad (2)$$

$$MAE = \sqrt{\frac{\sum_{i=0}^{n} |y_i - \hat{y_i}|}{n}} \quad (3)$$

where $\hat{y_i}$ is the predicted and y_i is actual energy consumption.

3. Proposed Model

The research is focused on exploiting ML and DL models to design a solution for the prediction of energy consumption demand with improved accuracy and reliability as compared to the previous researches done. The ensemble model proved to give satisfactory results in many prediction types of research. With the ensemble model, we mean to combine more than one model to process the data samples. The proposed model is divided into four phases, namely Data Acquisition, Data Pre-processing, Prediction, and Performance Evaluation.

Phase-1 In this phase, real-time consumption data from the industrial sector is collected for 2 years starting from Aug'2018 to July'2020. The frequency of data recorded is 15 minutes gap.

Phase-2 This phase deals with the pre-processing of the data before implementation. While collecting the data, it may have some, missing values, outliers, etc. which can give the wrong prediction. In this phase basically, the cleaning process of the data is done.

Phase-3 The clusterization is applied on the dataset, to categorize the customerbased on their power consumption pattern. It will help to implement the model on different ranges of power consumption categories.

Phase-4 The implementation of ML and DL models are done in phase-3. After comparing the results from several models, a hybrid model is proposed. This proposed model increases the accuracy of the prediction with a minimal error rate.

Phase-5 This phase discussed the results incurred in phase-4 based on the evaluation metrics such as R^2, MAE, and RMSE.

The model proposed in the research is based on an ensemble approach. Three base ML models and one meta-model are stacked to improve the prediction accuracy with a very low error rate. The model selected are Linear Regression, Random Forest, and LASSO as a base and Gradient Boosting Regressor (GBR) as the meta-model. The RF and LR contributed towards the accuracy of the model in form of R^2 and LASSO help to decrease the error rate. The metric to evaluate the model is R^2, mean absolute error (MAE) and root mean square error (RMSE).F*igure -2* shows the detailed structure of the model proposed.

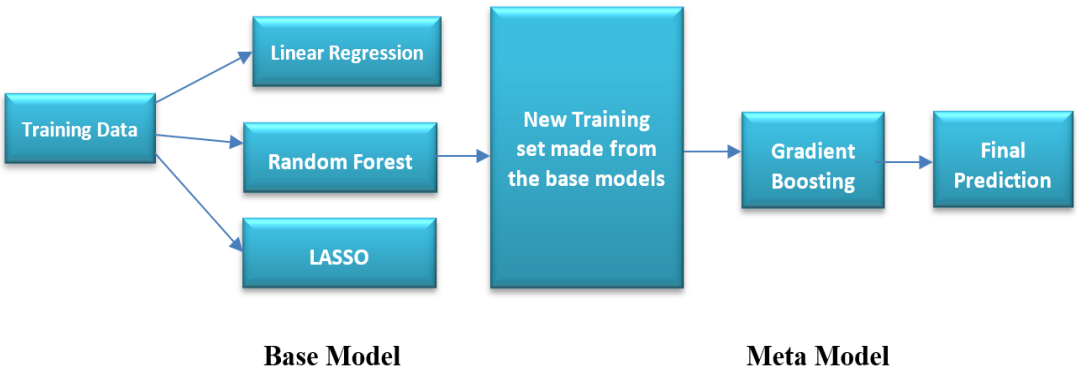

Figure-2: Proposed Model

4. Results & Discussion

After this selection process, the data for 2 years are taken for research. The data collected from HPSEB, India for power consumption from the industrial sector through a smart meter. The total sample of data was 4.5 million. The frequency of the record is after every 15 minutes. The data is implemented on LR, RF, and LASSO individually and then implemented in the proposed ensemble model using Python 3.2 using Keras and Tensor flow in the backend. The base model used are LR, RF, LASSO, and the meta-model is Gradient Boosting (GB). Before proposing the ML ensemble model, the data was tested with the most popular DL model such as RNN and LSTM also. The result incurred using ML models were far better than DL models. All the machine learning models evaluated on R^2, MAE, and RMSE. R^2 gives the accuracy of the model after comparing the actual and predicted value by the model implemented. The following table gives a descriptive detail of ML,DL, and ensemble model based on R^2, MAE, and RMSE.

Modal Evaluation Comparison			
MODEL	R^2	RMSE	MAE
RF	95.37%	0.0058	0.0011
LR	93.35%	0.0067	0.0012
RNN	94.33%	0.0365	0.0068
LSTM	93.23%	0.0400	0.0016
RF+LR+LASSO+GB	**96.78%**	**0.0054**	**0.0010**

Table -1 Evaluation Metric Results for 4 ML and 1 Ensemble Model

Initially, the data samples are tested with all ML and DL models individually. The ratio considered for training and testing data was 80:20. Table-1 describes the performance of the individual models. The performance of the DL models namely RNN and LSTM did not perform better on all the parameters. Although the performance of the RNN model was better in R^2 after RF it performed badly on RMSE and MAE parameters. Moreover, both the models took an exceptionally long time for execution also. So, considering the performance of ML models and less time for execution, an ensemble model is proposed on ML models. RF and LR performed better on all the parameters. LASSO is a type of linear regression, which contributes towards the improvement of the accuracy and GBR helps to reduce the error rate especially in the ensemble model. So, the stacking technique of ensemble model is applied with Python along with

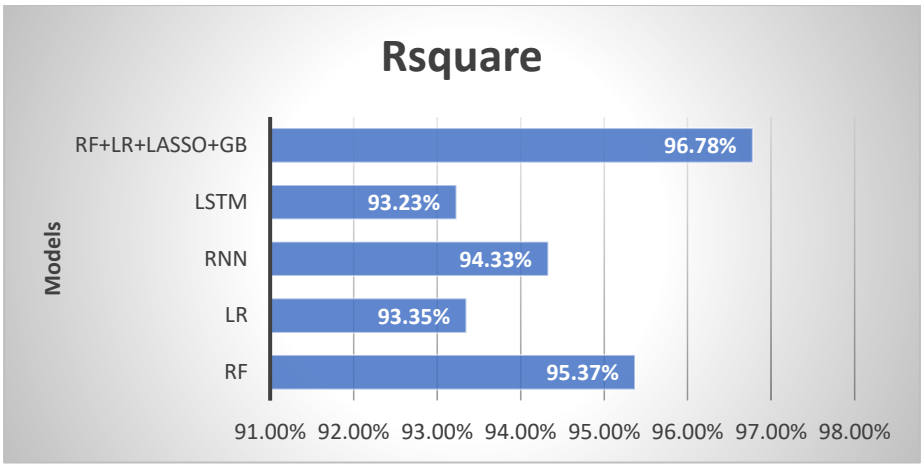

Figure-3: The Comparison of Model Accuracy (R^2)

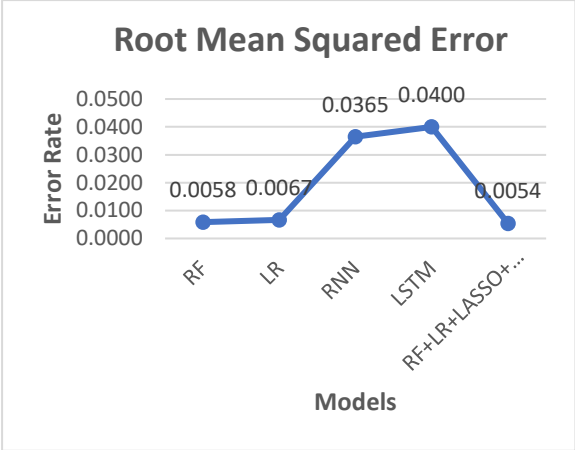

Figure-4(a)

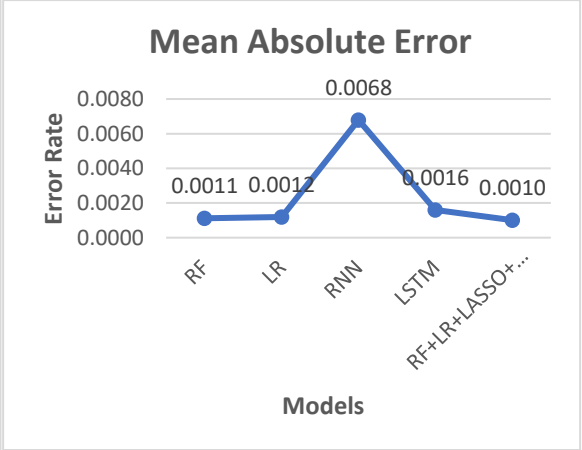

Figure-4(b)

The comparisons of RMSE 4(a) and MAE 4(b) for models

Figure-3 and figure-4(a)(b) presents the comparison of all the four models along with proposed model performance on R2, MAE, and RMSE. It can be visualized that the proposed ensemblemodel gave considerably improved results on all the parameters.

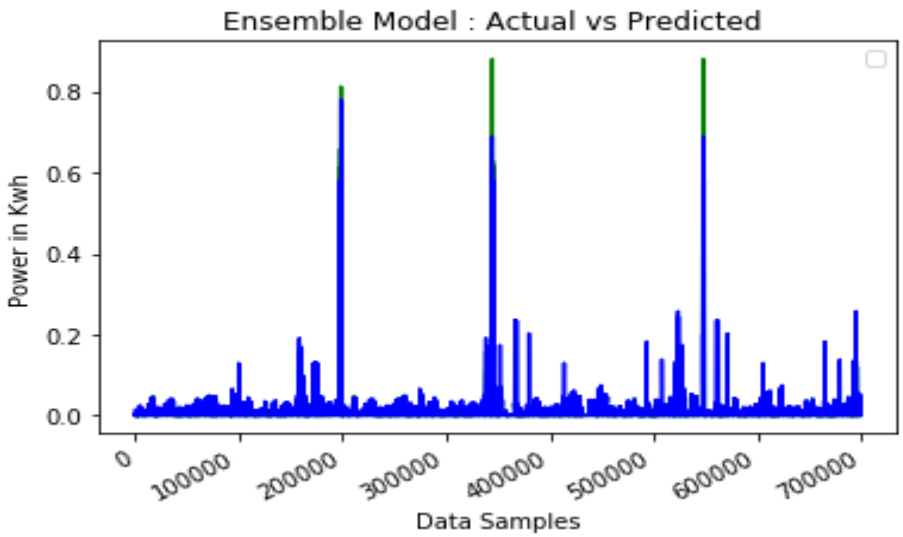

Figure-5: Proposed Ensemble Model Result Actual vs Predicted

Figure-5 elaborates the graphical representation of actual vs predicted result. The ensemble model performed very well with an accuracy percentage of 96.78% and a minimal error rate of 0.0054 (RMSE) and 0.0010 (MAE).

5. Conclusion

This paper proposed an ensemble machine learning model to predict the consumption of energy for medium to long term forecasting. The model considered making ensemble models are LR, RF, LASSO, and, GB. LR, RF, LASSO is taken as base models using the estimator function of python, and GB is taken as the final estimator. R^2 is considered to evaluate the accuracy of the model along with the error metrics MAE ad RMSE. The result of the comparison proves that the proposed ensemble model gave the best results on all the evaluation parameters. The dataset is also implemented on the most popular DL models namely, RNN and LSTM. After comparing all the parameters, machine learning models were demonstrated to give the best results.

References

1. Z. Zhang, X. Wang, and Y. Ji.,"The power load forecasting of SVR based on hadoop." In 2018 37th Chinese Control Conference (CCC), pages 4484–4488, July 2018.
2. Zhou, Xin, Zhezhuang Xu, Xiaotong Yu, and Yuxiong Xia., "The Analysis and Prediction of Power Consumption in Industry Based on Machine Learning." In 2019 Chinese Automation Congress (CAC), pp. 5738-5742. IEEE, 2019.
3. He, Yaoyao, Yang Qin, Shuo Wang, Xu Wang, and Chao Wang., "Electricity consumption probability density forecasting method based on LASSO-Quantile Regression Neural Network." Applied energy 233 (2019): 565-575.
4. Y. Cheng, L. Jin, and K. Hou.,"Short-term power load forecasting based on improved online elm-k. In 2018 International Conference on Control, Automation and Information Sciences (ICCAIS), pages 128–132, Oct 2018.
5. Hippert, H.S.; Pedreira, C.E.; Souza, R.C.,"Neural networks for short-term load forecasting: A review and evaluation." IEEE Trans. Power Syst. 2001, 16, 44–55. [CrossRef]
6. Marino, D.L.; Amarasinghe, K.; Manic, M.,"Building energy load forecasting using Deep Neural Networks." In Proceedings of the IECON 42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, Italy, 23–26 October 2016; pp. 7046–7051.
7. Rahman, A.; Srikumar, V.; Smith, A.D.,"Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks." Appl. Energy 2018, 212, 372–385. [CrossRef]
8. Zheng, H.; Yuan, J.; Chen, L.,"Short-Term Load Forecasting Using EMD-LSTM Neural Networks with a Xgboost Algorithm for Feature Importance Evaluation." Energies 2017, 10, 1168. [CrossRef]
9. Chen, Haiwen, Shouxiang Wang, and Yingjie Tian. "A new approach for power-saving analysis in consumer side based on big data mining." In 2018 IEEE Power & Energy Society General Meeting (PESGM), pp. 1-5. IEEE, 2018.
10. Dhupia B.; M Usharani. "Smart Profile: Smart Meter Data Oriented Customer Profiling", TEST Engineering & Management 83, May-Jun 2020, 5690-5695
11. Yildiz, Baran, Jose I. Bilbao, and Alistair B. Sproul. "A review and analysis of regression and machine learning models on commercial building electricity load forecasting." Renewable and Sustainable Energy Reviews 73 (2017): 1104-1122