Electrical Load Forecasting Using Hybrid of Extreme Gradient Boosting and Light Gradient Boosting Machine

Eric Nziyumva, Rong Hu, Chih-Yu Hsu, and Jovial Niyogisubizo

Abstract Ensemble learning methods have been used to improve performance accuracy through bias-variance trade-off techniques. However, there is still room to improve. This paper proposes an ensemble model to forecast the electrical load behavior based on a hybrid of Extreme Gradient Boosting (XGBoost) and Light gradient boosting machine (LGBM). Extreme gradient boosting (XGBoost), a Light gradient boosting machine (LGBM) and a hybrid of XGBoost and LGBM models are trained, evaluated, and compared. The experiments show that the proposed model outperforms other methods by reducing more than 1% in mean absolute percentage error (MAPE), root mean squared percentage error (RMSPE), and mean absolute error (MAE). The dataset from the Pennsylvania-New Jersey-Maryland interconnection power grid was used to validate the evolutionary capability of the proposed method and the finding of optimal accuracy of the model.

Keywords Electrical load forecasting · Ensemble learning · Extreme gradient boosting machine (XGBoost) · Light gradient boosting machine (LGBM)

1 Introduction

Energy power prediction is very important in our daily life. It is the first approach to the best power system management and plays a great significance for all-electric power-related activities. Moreover, this prediction not only presents its strongness to the reliable grid operation but also for safe electricity planning, modern transportation, communication and has a great positive impact on national security. Thus, accurate electric load prediction is essential for power systems since accurate prediction leads to the economic development of any country through substantial savings

E. Nziyumva · J. Niyogisubizo
Fujian Key Lab for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou 350118, China

R. Hu (✉) · C.-Y. Hsu
Fujian Provincial Key Laboratory of Big Data Mining, Fujian University of Technology, Fuzhou, China
e-mail: hurong@fjut.edu.cn

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

J. Yao et al. (eds.), *The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)*, Lecture Notes in Electrical Engineering 813,
https://doi.org/10.1007/978-981-16-6963-7_95

in operating and maintenance costs [1]. However, achieving the desired accuracy is difficult due to the various factors influencing the electric load behavior include human social activities, country policies, climate change, and economic development [2].

Previously, electric load forecasting (ELF) had been almost entirely limited to traditional statistical methods. The classical researches proposed include the adaptive time-series auto-regressive moving average (ARMA) model presented a good performance of reducing the error compared to the other models. Due to its simplicity and effectiveness, ARMA was popular and extensively used in ELF researches however it is limited to only being used for stationary time-series data. Cheng-Ming Lee and Chia-Nan Ko [3] proposed a new hybrid algorithm based on auto-regressive integrated moving average (ARIMA) which has the advantage of introducing non-stationary time-series data. Compared to other models used in their work, the simulation results indicated also the highest forecasting performance. Unfortunately, the random noise which disturbs the whole process, ARMA and ARIMA models use only time and load as input data which implies the ARMAX and ARIMAX to be discovered for introducing the exogenous variables. Then, Indian researchers Shilpa G N and Dr. G S Sheshadri had used the ARIMAX model, an extension of ARIMA with an exogenous variable for ELF [4]. However, the classical models are limited due to only focus on the relationship between the dependent and independent variables. In addition, their forecasting accuracy is not good enough therefore the modern models were introduced for making the most possible accurate predictions.

With modern science progress, load prediction technologies have been considerably developed. Lately, the introduction of machine learning (ML) theories in the electrical power engineering field became more and more popular which implies great efficiency for improving the performance of forecasting models [2]. The widely used ML models include multiple linear regression (MLR) applied for ELF and gave successful results. Even if it is easy for results interpretation, the regression models may lead to erroneous and misleading results due to the wrong assumptions [5]. Salkuti, S. R. proposed an ANN-based hybrid for predicting short-term electrical load demand in which a better result was found. Besides, to avoid different drawbacks of ANNs such as falling into the trap of local minima during the parameter optimization process, Salkuti, S. R. used a hybrid approach of combining ANN, wavelet transforms (WTs), and evolutionary-based differential evolution algorithm [6]. On the other hand, Mohamed proposed a full wavelet packet transform and neural network-based ensemble method. The simulation results show that the proposed approach reduces MAPE by 20% in comparison with the traditional neural network method [7]. Later, Chengdong developed a wavelet transforms-based model in which the proposed method combines the fuzzy inference system and the periodicity knowledge to generate accurate forecasting results [8]. Due to its double major advantages of being used in optimization and prediction fields, the genetic algorithm (GA) has been well-suited with nonlinear systems and it conducts a particular optimization based on the natural selection of the optimal solutions found from a wide range of forecasting model candidates' population [9]. Then, expert systems-based models had been increasingly developed for handling prediction issues.

Although the machine learning models had been widely used in various forecasting issues include ELF, their models' performance present various gaps of erroneous results due to variance, bias, and noise. This affects the ELF which leads to significant losses due to maintenance costs, unsafe power system operation, and all power planning-related activities [10]. Moreover, inaccurate load forecasting has great negative impacts on energy generating capacity scheduling which leads to inadequate operating.

In this paper, the proposed approach aims to upgrade the forecasting performance of machine learning algorithms. This is achieved by reducing the error between actual and predicted values through the bias-variance trade-off. The main principle behind this work is the combination of ensemble learning. At first, the extreme gradient boosting (XGBoost), light gradient boosting machine (LGBM), Adaptive boosting (AdaBoost), and random forest are firstly compared according to their accuracy and training time. Then, the hybrid of XGBoost-LGBM has been done to enhance the performance accuracy. The innovations of the proposed approach are such as the combination of two models and performance improvement compared to the remaining models used in this paper which leads to significant loss reductions.

The rest of this paper is organized as follows: Sect. 2 describes the methods used in this paper. Section 3 evaluates, discusses, and compares the performance results of models. Section 4 concludes the paper.

2 Methodology

The methodology used in this paper is graphically represented through Fig. 1.

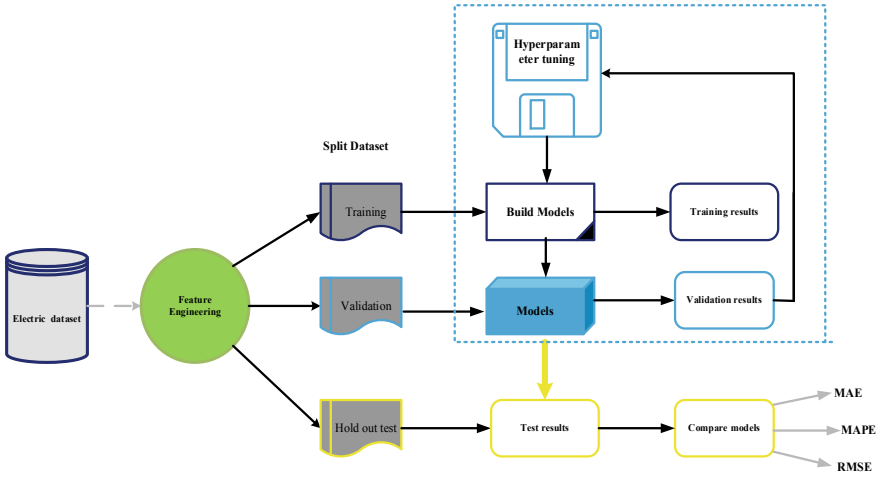

Fig. 1 Overall design of the study

According to their performance accuracy and training time, the boosting-based models include XGBoost, LGBM, Adaboost, and random forest, a bagging-based ensemble learning model, are compared with a hybrid of XGBoost-LGBM, the proposed model. The hyperparameter tuning of the models has been also computed. This section explains the research methodology process. Finally, the performance comparison of the models is done.

2.1 Overall Research Design

Figure 1 represents the development of the overall research design proposed to conduct the study for enhancing electric load forecasting. After identifying the inaccurate forecasting problems, the helpful steps of overcoming them have been proposed as follows: First, the electric dataset was gained. Then, feature engineering was conducted. Third, the ensemble machine learning-based techniques (Random Forest, XGBoost, LightGBM, Adaboost, and XGBoost-LGBM) were trained, evaluated, and compared. The mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square percentage error (RMSPE) were analyzed.

2.2 Description of Dataset

In this paper, the dataset used comes from Pennsylvania-New Jersey-Maryland (PJM) website [11]. The extracted information holds data of 87,600 samples with a sampling frequency of 1 h. The database covers a range of times starting from July, 1st 2008, and the hourly load demand recorded from twelve electrical networks is expressed in megawatts (MW). Various important techniques have been applied to this dataset for improving the outcome of the models. The first technique is dataset filtering done for selecting the sub dataset to be used for viewing and analysis. The second concerns dataset training which has the role of training the algorithm so that it can predict accurately. The third technique emphasis the evaluation of the dataset which is considered as the performance comparison tasks done with the help of a validation dataset to find the smallest error network. A model may overfit during the validation procedure therefore the performance evaluation might be done to the testing dataset. Cross-validation as the technique of evaluating the models for limited data by resampling had been used. The attributes after data filtering include the day of the week, holidays, season, the hour of the day, month, and energy consumption. The target variable is the consumption of electrical energy expressed in megawatts (MW).

2.3 Feature Engineering

Feature engineering expressed as the way of extracting variables from raw data plays an important role in determining the key variables that will be useful to win the upcoming applications and then to classify them as either low and high according to their impacts. Since ML framework development is a process that begins by carefully defining the requirements [12], the iterative process starts which involves building and testing various models over a dataset. To explore and analyze the information, the data gained from the dataset should be pre-processed and transformed. In the end, the relationship between independent features such as the day of the week, holidays, season, the hour of the day, month, year, and target variable (consumption of electrical energy) is seen.

The next stage is model building to try and evaluate different models. Here, the data is organized into three different split sets. With the help of the training and validation sets, there is an optimization possibility of model parameters using cross-validation procedures then hyperparameter tuning. The third set namely the "hold-out test" is used for final testing and model comparison.

2.4 Ensemble-Based Machine Learning Models

With the fast improvement of machine learning, various techniques to increase accuracy have been proposed. All the previous works have been done to reduce the errors. Here, we present a brief description of ensemble-based machine learning techniques used in this paper.

Random forest: The random forest algorithm has been firstly invented by Tin Kam Ho using attribute bootstrap aggregating (bagging) in 1995 [13]. The functionality of random forest consists of three main parts. Firstly, the samples are selected through the bagging techniques which are used for extracting the N (number) times training datasets from original data. Then, the prediction from each tree-based learner is found. Finally, the result is found through the combination methods such as averaging or voting. Random forests algorithm works very well compared to the decision tree as it corrects the overfitting but its accuracy is lower than that of the gradient boosted trees [14]**.**

Adaptive Boosting (AdaBoost): Adaptive boosting is the first boosting-based algorithm developed by the joint of Freund and Schapire [15]. The class of boosting algorithm takes its description as the machine learning approach to increase the forecasting performance level based on the combination of various weak learners and inaccurate rules. As the first practical boosting algorithm, adaptive boosting is widely used and studied and then applied in numerous fields. Its advantages were seen in regression and classification issues handling.

Extreme Gradient Boosting (XGBoost): Through the research project conducted by Tianqi Chen, the XGBoost that works under the principle of boosting gradient

tree had officially come out on March 27, 2014 [16]. This model could be used for handling regression or classification issues. Mathematically, the gain leads to a regularized boosting techniques is defined by [17]:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\beta} + \frac{G_R^2}{H_R+\beta} - \frac{(G_R+G_L)^2}{H_R+H_L+\beta}\right] - \alpha \tag{1}$$

where the first term is the score of the left child, the second the score of the right child and the third the score if we do not split. β and α are ridge and lasso regularization coefficients respectively.

Light Gradient Boosting Machine (LGBM): The Gradient-Based One-Side Sampling (GOSS) decreases the computation costs since it uses a subset of smaller instances rather than using all instances. Exclusive Feature Bundling (EFB) converts exclusive features into less dense features. The combination of both GOSS and EFB techniques produces LGBM [18]. Compared to the traditional gradient boosting concept, the LGBM has an accelerated training process and higher performance accuracy. Furthermore, this kind of gradient boosting can deal with the large dataset since it supports GPU and parallel learning [10].

Hybrid of XGBoost-LGBM: The hybrid of XGBoost and LGBM, boosted-based ML models, in which the individual's components are sequentially coupled for building a powerful meta-learner. The idea behind the proposed model is to find the approach of reducing the contribution of both bias and variance to error since the errors and predictions in any ML models are adversely influenced by bias, variance, and noise [19]. A high bias and variance bring about the underfitting and overfitting of the training data respectively while noise is considered as an irreducible error caused by improper cleaning of data. The proposed approach is found through grouping the individual models in sequential. The numerical simulations with cross-validation (CV) and hyperparameters tuning verify the power of the new meta-learner algorithm by giving the best results compared to the single model as shown in Table 1. The expected prediction error of a regression model using squared-error loss is expressed as:

Table 1 Performance evaluation of models: LGBM, XGBoost and Random forest

	LGBM			XGBoost			Random forest		
CV	MAPE	RMSPE	MAE	MAPE	RMSPE	MAE	MAPE	RMSPE	MAE
K1	1.88	2.12	199.66	1.94	2.04	200.1	2.16	2.58	205.86
K2	1.94	2.24	188.64	1.98	2.56	199.72	2.12	2.84	205.38
K3	1.78	1.94	189.30	1.84	2.52	199.04	2.14	2.92	200.72
K4	1.74	1.96	195.78	1.90	2.48	198.72	1.96	2.70	200.36
K5	1.72	1.78	193.72	1.74	2.46	197.84	1.90	2.66	220.04
K6	1.46	1.68	188.10	1.64	2.24	197.82	1.88	2.54	200.02
Mean	1.74	1.94	192.52	1.84	2.38	198.87	2.02	2.70	205.38

$$\begin{aligned}\Psi\left(y, \hat{F}(x_i)\right) &= E\left[\left(y - \hat{F}(x_i)\right)^2 | x = x_i\right] \\ &= \sigma_\varepsilon^2 + \left[E\left(\hat{F}(x_i)\right) - F(x_i)\right]^2 - E\left[\hat{F}(x_i) - E\left(\hat{F}(x_i)\right)\right]^2 \\ &= \sigma_\varepsilon^2 + bias^2\left(\hat{F}(x_i)\right) + variance\left(\hat{F}(x_i)\right) \end{aligned} \quad (2)$$

where the first term represents an irreducible error, the second term is the contribution of squared bias to error while the last term is the contribution of variance to error. F and $\hat{F}$ represent the actual and predicted values respectively while E represents expected values then $x = x_i$ represents point value.

2.5 Evaluation Criteria

To evaluate and compare the performance of ML models, several performance metrics are used. The mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean squared percentage error (RMSPE) have been utilized. Mathematically, they are defined:

$$\text{MAE} = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{N}} \left|\text{Y}_\text{i} - \hat{\text{Y}}_\text{i}\right| \quad (3)$$

$$\text{MAPE} = \frac{100}{\text{N}} * \sum_{\text{i}=1}^{\text{N}} \left|\frac{\text{Y}_\text{i} - \hat{\text{Y}}_\text{i}}{\text{Y}_\text{i}}\right| \quad (4)$$

$$\text{RMSPE} = 100 * \sqrt{\frac{\sum_{\text{i}=1}^{\text{N}} \left(\text{Y}_\text{i} - \hat{\text{Y}}_\text{i}\right)^2}{\text{N Y}_\text{i}}} \quad (5)$$

Y_i and $\hat{Y}_i$ are the actual and predicted values respectively while N is the observation number. The more the values of MAE, MAPE, and RMSPE, the worse prediction accuracy.

3 Results and Discussions

The performance evaluation of four ensemble learning-based methods was judged and compared with the proposed model. In this section, the results are presented and discussed.

Tabel 2 Performance evaluation of models: AdaBoost and XGBoost-LGBM

	AdaBoost			XGBoost-LGBM		
CV	MAPE	RMSPE	MAE	MAPE	RMSPE	MAE
K1	2.14	2.44	202.24	1.78	1.92	199.06
K2	2.10	2.78	203.76	1.70	1.30	196.04
K3	2.06	2.76	200.32	1.62	1.44	188.62
K4	1.92	2.64	200.04	1.50	1.28	184.68
K5	1.86	2.56	199.7	1.46	1.08	182.04
K6	1.82	2.38	199.06	1.24	0.90	178.52
Mean	1.98	2.58	200.84	1.54	1.32	188.16

3.1 Comparison

The ML performance metrics are often used to compare the models for selecting the best suitable for ELF. According to the literature, different models have been proposed and compared based on the performance evaluation results with the execution time. Since the existing models have not yet satisfied the desired forecasting quality, the research is still undergoing. Here, the hybrid of XGBoost-LGBM and four single models have been trained, tested, and then compared. For making an accurate performance evaluation, the simulation process is repeated six times. The technique used here is popularly known as k-fold cross-validation (CV) which prevents the model to overfit the new data [20]. Then, hyperparameter tuning has been also applied.

The six-fold cross-validation results for each model are resumed in Tables 1 and 2. The performance results before applying CV techniques are worse than those obtained after using it. The errors were higher which indicates significant losses due to the inaccurate electric load prediction.

According to the mean absolute percentage error (MAPE) which is considered as a loss function to define the error termed by the model evaluation. The mean results for the k-folds cross-validation reveal that the proposed hybrid successfully limits the forecasting error to 1.54% compared to the other remaining single models.

Figure 2 represents the comparison of the MAPE and MAE obtained from various models used in this paper. The proposed model, a hybrid of XGBoost and LGBM, shows the highest accuracy compared to the other remaining models.

In addition, the root mean squared percentage error (RMSPE) of the single models is greater than that of the proposed model. Based on the experiment results, the hybrid of XGBoost-LGBM reduces the error to 1.32%. Afterwards, the mean absolute error (MAE) expressed as the measure of how far the predictions were from the actual output shows that the proposed model has the least values of error. Based on the aforementioned discussions and performance metrics principle (the smaller the values of MAPE, RMSPE and MAE, the better the model performance) therefore it is reasonable and proves that the proposed model outperforms the other models

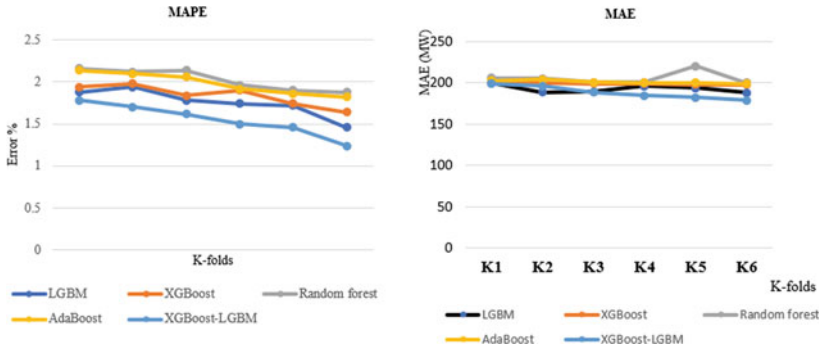

Fig. 2 Comparison of the MAPE and MAE obtained

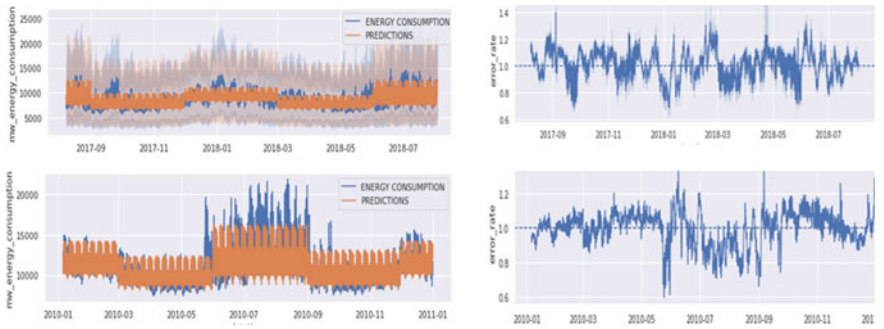

Fig. 3 Energy consumption versus Estimation and Error rate versus datetime

tested in this paper. However, the proposed model takes a longer time to be trained compared to single models.

Figure 3.shows ELF with two months ahead. The left graphs represent the energy consumption and prediction versus datetime while the right graphs show the error rate versus datetime. For the first-row graphs, there are fewer losses as the actual energy consumption is almost equal to the predicted electrical energy therefore the utilization of generated energy is maximized. While for the second-row graphs, the significant error values are greatly remarkable which indicates a significant difference between predicted and actual values. Consequently, the first row-graphs show an accurate model which leads to the main goal of ELF.

4 Conclusion

The overall development of any country is based on proper load forecasting. The inaccurate forecasting affects negatively the planning and may lead to various significant

losses. Hence, the performance accuracy should be improved. This paper proposes the techniques of enhancing accuracy for electric load forecasting. The advantages of both XGBoost and LGBM are combined for achieving the target. The proposed model, a hybrid of XGBoost and LGBM, shows the highest accuracy compared to the other remaining models tested in this paper. According to the mean absolute percentage error (MAPE), the mean results for the k-folds cross-validation reveal that the proposed hybrid successfully limits the forecasting error to 1.54% compared to the other remaining single models. In addition, the root mean squared percentage error (RMSPE) of the single models are greater than that of the proposed model which is not good. On the other hand, the mean absolute error (MAE) shows that the proposed model has the least values of prediction error which leads to attractive results. Moreover, the accurate forecasting obtained from the proposed approach leads to the reduction of the significant losses since the utilization of power-generated energy is maximized. Furthermore, the contributions of this proposed approach include safe power system operation, proper planning of transmission and distribution facilities, proper financing (future expenditure and earnings), substantial savings in operating and maintenance costs, and then safe power planning related activities while its innovations are such as the combination of two models and performance improvement compared to the other models tested in this experiment. Based on the aforementioned discussions and performance metrics principle (the smaller the values of MAPE, RMSPE, and MAE, the better the model performance) therefore it is reasonable and proves that the proposed model outperforms the other models tested in this paper.

References

1. Jung, S.-M., Park, S., Jung, S.-W., Hwang, E.: Monthly electric load forecasting using transfer learning for smart cities. Sustainability **12**(16), 6364 (2020)
2. Wang, R., Wang, J., Xu, Y.: A novel combined model based on hybrid optimization algorithm for electrical load forecasting. Appl. Soft Comput. 82, 105548 (2019)
3. Lee, C.-M., & Ko, C.-N.: Short-term load forecasting using lifting scheme and ARIMA models. Expert Syst. Appl. 5902–5911 (2011)
4. Shilpa, G.N., Sheshadri, G.S.: ARIMAX model for short-term electrical load forecasting. Int. J. Recent Technol. Eng. (IJRTE) **8**(4) (2019)
5. Divina, F., Gilson, A., Goméz-Vela, F., García Torres, M., Torres, J.: stacking ensemble learning for short-term electricity consumption forecasting. Energies **11**(4), 949 (2018)
6. Salkuti, S. R.: Short-term electrical load forecasting using hybrid ANN–DE and wavelet transforms approach. Electr Eng **100**(4), 2755–2763 (2018)
7. El-Hendawia, M., Wang, Z.: An ensemble method of full wavelet packet transform and neural network for short term electrical load forecasting. Electr. Power Syst. Res. 182, 106265 (2020)
8. Li, C., Tang, M., Zhang, G., Wang, R., Tian, C.: A hybrid short-term building electrical load forecasting. Int. J. Fuzzy Syst. **22**(1), 156–171 (2020)
9. Al-Douri, Y.K., Al-Chalabi, H., Lundberg, J.: Time Series forecasting using genetic algorithm. In: The Twelfth International Conference on Advanced Engineering Computing and Applications in Sciences (2018)
10. Quinto, B.: Next-Generation Machine Learning with Spark: Covers XGBoost, LightGBM, Spark NLP, Distributed Deep Learning with Keras, and More (2020)

11. PJM.PJM load forecast, [Online]. Available https://dataminer2.pjm.com/list. Accessed 11 Nov 2020
12. Al Mamun, A., Sohel, M., Mohammad,N., Sunny, M.S.H., Dipta, D.R., Hossain, E.: A comprehensive review of the load forecasting techniques using single and hybrid predictive models. IEEE Access 8, 134911–134939 (2020)
13. Ho, T.K.: Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition 1, 278–282 (1995)
14. Piryonesi, S. M., El-Diraby, T. E.: Role of data analytics in infrastructure asset management: overcoming data size and quality problems. J. Transp. Eng. Part B: Pavements **146**(2), 04020022 (2020)
15. Schapire, R.E.: Explaining AdaBoost. Empirical Infer. 37–52 (2013)
16. Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, San Francisco (2016)
17. Omar, K.B.A.: Xgboost and LGBM for Porto Seguros Kaggle Challenge: A Comparison (2018)
18. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: a highly efficient gradient boosting decision tree. In: Proceedings of the Advances in Neural Information Processing Systems 30, 3146–3154 (2017)
19. Géron, A.: Hands-on machine learning with Scikit-Learn, Keras, Canada: O'Reilly Media, Inc., Sebastopol 1492032611 (2019)
20. Lin, Y., Luo, H., Wang, D., Guo, H., Zhu, K.: An ensemble model based on machine learning methods and data preprocessing for short-term electric load forecasting. Energies **10**(8), 1186 (2017)