International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 11, pp 261-266, November-2023

https://doi.org/10.47001/IRJIET/2023.711036

Assessing Machine Learning Models for Power Plant Generation Prediction

[1]Chamoda De Silva, [2]Vihan Jayawardana, [3]Pasindu Gunasekara, [4]Tharani Medawatta, [5]Anuradha Jayakody [6]Shashika Lokuliyana

[1]Dept. of Information Technology, Sri Lanka Institute of Information Technology, Galle, Sri Lanka

[2,4,5,6]Dept. of Computer Systems Engineering, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka

[3]Dept. of Information Technology, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka

Authors E-mail: [1]chamodadesilva6@gmail.com, [2]vihan1968@gmail.com, [3]pasindukalanagunasekara@gmail.com, [4]tharani.c.medawatte@gmail.com, [5]anuradha.j@sliit.lk, [6]shashika.l@sliit.lk

***Abstract* - The difficulty of maintaining a balanced energy supply and demand has emerged as a key issue, creating a number of issues around the globe as a result of the rapidly increasing global population and industrial growth. This situation emphasizes the value of performing additional study on power generation and demand forecasts using machine learning approaches, particularly in Sri Lanka. In this analysis, we concentrate on projecting how much electricity two particular power plants, Laxapana and Mahaweli, will produce in the future. Our method involves parameterizing the net electricity generating data for each power plant as well as the hydro inflow data from multiple sub-power stations. We use a number of machine learning methods, such as Lasso Regression, Random Forest, and XGBoost. To further improve our models, we use methods like feature engineering and hyperparameter tuning with GridSearchCV and RandomizedSearchCV. In order to further enhance the predictive performance, we use model stacking.**

Keywords: Generation, Power Plants, Machine Learning, Accuracy, Electricity.

I. INTRODUCTION

A global energy supply, which has a huge number of sources and ways to generate, transport and distribute power in various countries and regions, is very intricate and also an ever changing system. A high management of energy with reduced energy loss is required in order to ensure the sustainability of energy. Therefore, the future of power generation and power grids should provide extraordinary levels of resilience in energy management. Electricity generation forecasting is crucial in the growth and operation of power systems. It gives a chance for the energy suppliers to measure the usage of electricity and plan for demands that may occur in the future. It also gives the opportunity to power distributors to manage and measure up to production of future electricity with demand. Therefore, Electricity generation forecasting has lately obtained substantial attention; sometimes, this is a demanding subject [1]. Sri Lanka's power supply is mainly sourced from thermal power plants, with a considerable contribution from hydroelectric power plants and a little amount from wind and solar power plants. Currently, the installed capacity of the country's power plants is around 4000MW, and the peak demand for electricity is around 2,500MW. The country's three main electricity sectors can be classified as industry, transport and household, and last as commercial sectors. As it is impossible to store energy in massive amounts, it is crucial to plan energy facilities and the production of energy to meet the country's demand. Sri Lanka's peak demand for electricity has been growing at an annual average rate of 3.4% during the past 20 years. The country's daily electricity demand profile has three distinguishable periods: night peak, day peak, and off-peak. Though the night peak has the highest electricity demand, the daytime demand is believed to increase in years to come [2]. Therefore, to match this increasing demand, it is crucial to keep track of future electricity generation and future electricity demand. Therefore, this paper addresses this issue by applying a methodology based on machine learning algorithms such as Lasso Regression, Random Forest, and XGBoost for energy generation forecasting of main two power plants of the country. The main contribution of this work is the development and implementation of a forecasting methodology with good performance and a good accuracy level to contribute to one of the great challenges of the developing world.

II. RELATED WORKS

Deep learning techniques were utilized by Ayvaz et al. to forecast electricity consumption. They used a time series dataset that included hourly electricity consumption over a year to forecast daily consumption. First off, all Prerequisite preparations for the analytical step have been finished. The analysis of time series techniques LSTM, GRU, and ARIMA

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 11, pp 261-266, November-2023

https://doi.org/10.47001/IRJIET/2023.711036

were evaluated on the data set. Accurate estimation has been achieved by hyperparameter adjustment. They looked for the factors that gave the greatest results for each modeling type and then compared those parameters to the outcomes. The LSTM model produced the most accurate forecast with a 13% rate compared to the other model's test results. Compared to the LSTM, the GRU technique provided more accurate predictions. The ARIMA approach was discovered. [3].

Tokgoz et al. used time series forecasting techniques based¨ on recurrent neural networks (RNN), long short-term memory (LSTM), and gated repetitive units (GRU) for Turkey's electricity consumption forecasting in their thesis study. Compared to earlier ARIMA and artificial neural network investigations, their experiments yielded a better average absolute error percentage [4].

In their thesis study, Nespoli et al. looked at the precision of load forecasting one day ahead, dependent on the data type employed in the training of the LSTM. The years 2017 and 2018 were used to evaluate a genuine Italian industrial energy load study, with data being monitored every 15 minutes [5].

Le et al. suggest a Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (Bi-LSTM)-based Electrical Energy usage Forecasting model. The Combination they used outperformed cutting-edge methods in terms of a variety of evaluation measures for electrical energy usage forecasting on a variety of variations of the real-time, short-term, medium-term, and long-term of one single household electrical power consumption data set, according to the experimental results [6].

Bedi and colleagues used Chandigarh's system to implement their thesis. The LSTM algorithm was used to forecast electricity demand depending on seasonal, daily, and interval data. To enhance the accuracy of the predictions, the notion of active window-based learning was incorporated into this work[7].

Because of the nonlinear and non-stationary nature of the time series, Zheng et al. investigated LSTM-based Recurrent Neural Networks (RNN) to solve the short-term electric charge forecasting problems. By utilizing the lengthy forecasting capacity of the LSTM-based RNN, they successfully forecasted complex electric charge time series. [8].

Abbasi et al. transformed load data from the Australian Energy Market Operator into time series to forecast a smart grid's future load. They predicted the electrical demand for one individual time delay by extracting features from the data using the XGBoost method. It has been noted that XGBoost works remarkably well regarding the effective use of memory and processing time. [9]

III. ALGORITHMS

This section briefly explains the machine learning algorithms used.

- Random Forest - It is a synthetic learning method for classification, regression, and other applications that works by training decision trees, also known as random decision forests. These trees are growing at a parallel rate within the algorithm, so they don't interact with each other as they develop.
- XGBoost - Extreme Gradient Boosting is a scalable endto-end tree boosting system. A gradient tree approach is applied most frequently to this area. More specifically, the gradient boost is a supervised learning process in which a few weak and simple model predictions are combined to predict our target variable accurately.
- Lasso Regression - The Least Absolute Shrinkage and Selection Operator are referred to as LASSO. This Lasso Regression model is commonly known as L1 regularization, a famous method for estimating the associations between variables and making predictions. This method is preferred over regression techniques to get a more accurate forecast outcome.

IV. METHODS TO IMPROVE ACCURACY

Below methods are used to improve the accuracy of the above machine learning models.

A) Stacking Ensemble Methods

It is primarily intended to apply, optimise and combine the best algorithms. These include methods that increase efficiency and flexibility by using the common complementarity of underlying models. Basic modeling and meta-modeling training stages are generally part of the Stacking Group Training. Consideration shall be given to regression and average models in this part of the study. We are thus using ensembles, models that combine a mixture of the two components. They are able to improve accuracy and reduce variability.

In this instance, this stacking ensemble method is used in one place. For once, it presents an ensemble framework consisting of three base models: Random Tree, XGBoost, and lasso Regressor. It is found that for the specific ranges of the target parameter, all Machine Learnings are better than each other. This observation led to create a dynamic weighted average ensemble model that in detailed comparison, was

ISSN (online): 2581-3048
Volume 7, Issue 11, pp 261-266, November-2023
https://doi.org/10.47001/IRJIET/2023.711036

more accurate than every other standalone utilized model. This is clearly visualized in the below image 1.

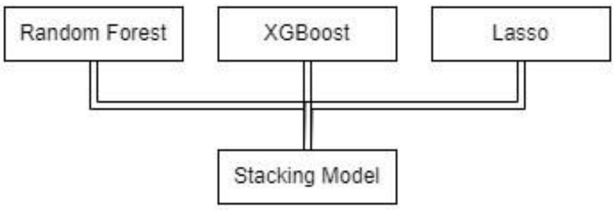

Figure 1: Stacking Ensemble Method Used

B) Feature Engineering

The feature engineering process is to build a forecast model from machine learning or statistical modelling and use domain knowledge to choose and modify the best possible variables that are derived from real data. Various engineering techniques, such as the creation of lag features, scaling or handling outliers, are used. In this aforementioned models standardization has been used as a feature engineering technique to increase accuracy. To scale the features of the dataset to a mean of 0 and a standard deviation of 1, standardisation is a common feature engineering technique used in machine learning. This will facilitate the use of automated algorithms to learn what features are connected, and it can make a number of models much more efficient, in particular those that respond to the great scale of input functions as support vector machines or k-nearest neighbours.

C) Hyperparameter Tuning

When training machine learning models, each dataset and model requires a different set of hyperparameters, which is a type of variable. The only way to determine them is to run some tests, choose a set of hyperparameters, and run them in the model. This is called hyperparameter tuning. There may be a lot of hyperparameters in models and finding the best combination of parameters is an issue that can be considered to be searching. GridSearchCV and RandomizedSearchCV are the two most effective ways of tuning hyperparameters.

• GridSearchCV

For each hyperparameter value, this method evaluates the artificial neural network model. It calls this approach GridSearchCV because it will search for the best set of hyperparameters based on a grid of hyperparameter values. Here, this method is used in the decision tree model and will get the best set of hyperparameters for the decision tree and then be used to make predictions with the optimized model. This improves the model's prediction accuracy by fine-tuning its configuration.

• RandomizedSearchCV

The disadvantage of GridSearchCV is overcome by this method, because it only accepts a fixed number of hyperparameters. To locate the most suitable set of hyperparameters, it will travel in a random way across the mesh. There is a reduction in unnecessary calculations with this approach.

V. EVALUATION METRICS

The results of these algorithms shall be assessed using a set of metrics to compare with the techniques described below. Next, are the math formulas and their brief descriptions.

A) R-squared (R2)

The variance of the error is compared to that of the model data by the coefficient of determination. In other words, this is the proportion of variance described as "explained" by a Predictive Model and so compared to subsequent error based measures, higher its value will be better suited. The (R2) value can be calculated below (1).

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

$$= 1 - \frac{\sum (y_i - f_i)^2}{\sum (y_i - \bar{y})^2} \qquad (1)$$

In the context of the coefficient of determination (R^2), SS_{res}, this is the sum of squares of residual errors. In contrast, SS_{tot} corresponds to the total sum of squares, which is proportional to the variance of the data. This is where y_i corresponds to the actual power load value, y is the mean of the actual values, and f_i belongs to the estimated energy load value.

B) Mean absolute error (MAE)

It is relatively simple to calculate the MAE, as it only entails adding an absolute number of errors which respond to differences between actual and projected values, then dividing that total error by a few observations. For all errors, MAE considers them to be the same weights as any other statistical method. The equation for this is given below (2).

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{n} |y_i - f_1| \qquad (2)$$

If the number of data points is N, the actual value will be shown in y_i as well as f_i that corresponds to the expected values.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 11, pp 261-266, November-2023

https://doi.org/10.47001/IRJIET/2023.711036

C) Mean squared error (MSE)

The MSE indicates the fit by calculating the squared difference between the i-th observation and the corresponding model prediction and then finding the mean of all errors. Squared gives greater value for bigger differences, apart from removing all negative signs. It is of course important to note that the lower the MSE, the more favourable the forecast. The equation for this is given below (3).

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{n}(y_i - f_1)^2 \qquad (3)$$

Where the number of data points is N, then y_i corresponds to the actual values and f_i represents the expected values.

D) Root mean squared error (RMSE)

RMSE is the square root of the mean squared difference between the true and the expected values, the square root of the MSE. Although both measures have a similar formula, RMSE is more commonly applied because it's proportional to the variable in its same units. The RMSE applies more weight to larger errors because the impact of the error on the sum is proportional to its square, not its magnitude, according to its mathematical definition.

$$\sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - f_i)^2} \qquad (4)$$

Where the number of data points is N, then y_i corresponds to the actual values and f_i represents the expected values.

VI. RESEARCH DESIGN

This study aims to solve the problem of electricity generation forecasting of two main power plants to maximize the operations of power plants. Our approach is to generate ML techniques using and comparing multiple algorithms. Tuning these three algorithms and implementing an ensembles approach for predictions is our goal at the end. Python programming languages and libraries such as Pandaplot, Numpy, Matlib, Sklearnetc were used for the implementation. The proposed methodology consists of two phases: data collection and design, algorithmic tuning as well as model prediction.

A) Data Collection and engineering

Monthly data was collected to train the machine learning models from the year 2000 till the year 2019, giving the time to identify the needed variations and changes. This data was collected from hard copies such as books and reports, which were later added to an Excel file and taken as the feeding data set for the ML models.

The two data sets have eight parameters and nine parameters for the predictions of laxapana complex and mahaweli complex respectively: month and date, hydro inflow data to the relevant sub power stations, and total generation in gigawatt hours. As this data was collected from the yearly and generation reports from years back, no values were missing.

B) Algorithmic tuning and model prediction

This sub-section includes the detailed architectures of the three models used for electricity generation prediction in this problem. Here both the models use the same algorithms and method. At first the basic models were trained with the methods to increase the accuracy. First a feature engineering method is used, namely Standardization. Randomized Search is used to find the best hyperparameters for the Random Forest model. Then GridSearch is used in lasso regression and XGBoost.

The model's configurations, parameters used, and each and every model's implementation are depicted in Table 1.

Table I: Configurations of the models

Model	Configuration
Random Forest	n estimators: [100, 200, 400] max depth: [5, 10, 15] min samples split: [2, 5, 10] min samples leaf: [1, 2, 4] criterion: 'absolute error' max features: 'auto'
XGBoost	n estimators: [100, 200, 400] max depth: [5, 10, 15] learning rate: [0.01, 0.1, 0.2]
Lasso Regression	alpha: 1.0

Then the evaluation and comparison of three regressions models Random Forest, XGBoost and Lasso Regression were done in this section of the research paper. In order to determine whether each model performs in a given category and range of target values, it was aimed at evaluating its performance on the test dataset. To that end, two thresholds have been set out in the first definition which serves as the criteria for classifying target values into 3 different categories: "High," "Medium" and "Low." The special characteristics of the data distribution were also taken into account in these thresholds. Then, for each of these categories based on the defined thresholds, creation of a a categorization function to automatically assign all target values.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 11, pp 261-266, November-2023

https://doi.org/10.47001/IRJIET/2023.711036

Subsequently, on the test data for each of the three models Random Forest, XGBoost and Lasso Regression forecasts were generated. Then the calculation of key performance metrics, including the R-squared (R2) score, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) for each of the models were done.

The results of this analysis (Table 2 and Table 3) were able to give us a better understanding of how the two models worked in separate subsets of these data, which is characterised by various target values. This understanding has given a complete insight into the strengths and weaknesses of each model within different scenarios in order to help choose which model is best suited for particular data subsets, as well as shed light on possible areas where modelling can be improved.

Table II: Performance metrics for different models by category for mahaweli plant

Cat.	Model	R2	MSE	RMSE
High Gen.	Random Forest	0.962	62.155	7.884
	XGBoost	0.944	92.650	9.625
	Lasso	1.000	0.00012	0.01078
Low Gen.	Random Forest	-4.129	40.860	6.392
	XGBoost	-0.583	12.611	3.551
	Lasso	0.99998	0.00018	0.01328
Med. Gen.	Random Forest	0.573	37.995	6.164
	XGBoost	0.925	6.646	2.578
	Lasso	1.000	0.00017	0.01290

Table III: Performance metrics for different models by category for laxapana plant

Category	Model	R2	MSE	RMSE
High Gen.	Random Forest	NaN	365.65	19.12
	XGBoost	NaN	8.60	2.93
	Lasso	NaN	0.00	0.05
Low Gen.	Random Forest	0.89	2.95	1.72
	XGBoost	0.95	1.28	1.13
	Lasso	1.00	0.00	0.01
Medium Gen.	Random Forest	0.23	31.60	5.62
	XGBoost	0.70	12.27	3.50
	Lasso	1.00	0.00	0.02

After going through this it has able to determine which performed better in the specific range of the target parameter. The result of this observation has been the creation of a dynamic Weighted Average Ensemble Model which is more accurate in detailed comparison against all existing independently used models.

VII. RESULTS

This section presents the results of the electricity generation forecast. After training the models with the above algorithms the decision to go ahead with an ensemble model was made. In this section of the research paper, a method for calculating model weights based on performance metrics and subsequently using these weights to create ensemble predictions was made. The goal is to leverage the strengths of multiple regression models (Random Forest, XGBoost, and Lasso) and combine their predictions in a way that optimally contributes to the overall predictive accuracy. This aggregation technique is intended to improve the reliability of forecasts by giving different weightings to each model in order to reflect its performance on a variety of subsets of data.

The ensembles methodology gives a systematic approach for exploiting the various strengths of different regression models, and targeting their effects to relevant subsets of data, resulting in an improved forecast accuracy on a wide range of target values. To respond to variability and complexity of realworld data, combined models by weighted ensembles forecasts are a complex strategy that could increase the accuracy and reliability of predictions in actual applications. This eradicates the overfitting issue of the lasso regression and increases the accuracy of the entire model.

The evaluation metrics of these models are shown in the Table 3 below.

Table IV: Ensemble model evaluation results for mahaweli plant

Metric	Value
R-squared (R2)	0.9999996509530101
Mean Squared Error (MSE)	0.0007655029847366551
Mean Absolute Error (MAE)	0.019046828852416525
Root Mean Squared Error (RMSE)	0.027667724603527757

Table V: Ensemble model evaluation results for laxapana plant

Metric	Value
R-squared (R2)	0.9999836974557256
Mean Squared Error (MSE)	0.002178434313857788
Mean Absolute Error (MAE)	0.032749090438404065
Root Mean Squared Error (RMSE)	0.046673700451729645

VIII. CONCLUSION

This study concentrates on the analysis of time series data in the context of forecasting energy generation for two power facilities. It analyses factors with separate timestamps to improve the precision of power demand predictions. It also includes metrics like R2, MAE, MSE, and RMSE to evaluate each model's predictive performance. According to the

ISSN (online): 2581-3048
Volume 7, Issue 11, pp 261-266, November-2023
https://doi.org/10.47001/IRJIET/2023.711036

study's findings, each model performs best in particular target parameter ranges. This finding inspires the creation of a dynamic weighted average ensemble model, which, upon close examination, outperforms all previously employed individual models. Our study's findings highlight the value of regression models for predicting future power generation and production. We can considerably increase prediction accuracy by using strategies like model stacking and hyperparameter adjustment. These findings enhance well-informed energy sector planning and decision-making by adding insightful information to power forecasting.

ACKNOWLEDGMENT

We would like to express our special gratitude to Sri Lanka Institute of Information Technology.

REFERENCES

[1] Mystakidis, Aristeidis, et al. "Energy Generation Forecasting: Elevating Performance with Machine and Deep Learning." Computing, vol. 105, no. 8, 25 Feb. 2023, pp. 1623–1645, https://doi.org/10.1007/s00607-02301164-y. Accessed 28 Aug. 2023.

[2] Ceylon Electricity Board Long Term Expansion Plan 2022-2041.

[3] Ayvaz, S., & Arslan, O. (2020, October). Forecasting electricity consumption using deep learning methods with hyperparameter tuning. In 2020, the 28th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[4] Tokgoz, A., Unal, G. (2018). A RNN based time series approach for¨ forecasting Turkish electricity load. In 2018, the 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[5] Nespoli, A., Ogliari, E., Pretto, S., Gavazzeni, M., Vigani, S., Paccanelli, F. (2021). Electrical Load Forecast using LSTM: The Impact of Data Quality. Forecasting, 3(1), 91-101.

[6] Le, T., Vo, M. T., Vo, B., Hwang, E., Rho, S., & Baik, S. W. (2019). Improving electric energy consumption prediction using CNN and BiLSTM. Applied Sciences, 9(20), 4237.

[7] Bedi, J., Toshniwal, D. (2019). Deep learning framework to forecast Electricity demand. Applied energy, 238, 1312-1326.

[8] Zheng, J., Xu, C., Zhang, Z., Li, X. (2017). Electric load forecasting in smart grids using long-short-term-memory-based recurrent neural Network. In 2017, the 51st Annual Conference on Information Sciences and Systems (CISS) (pp. 1-6). IEEE.

[9] Abbasi, R. A., Javaid, N., Ghuman, M. N. J., Khan, Z. A., & U Rehman, S. (2019). Short-term load forecasting using XGBoost. In Workshops of the International Conference on Advanced Information Networking and Applications (pp. 1120-1131). Springer, Cham.

Citation of this Article:

Chamoda De Silva, Vihan Jayawardana, Pasindu Gunasekara, Tharani Medawatta, Anuradha Jayakody, Shashika Lokuliyana, "Assessing Machine Learning Models for Power Plant Generation Prediction" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 11, pp 261-266, November 2023. Article DOI https://doi.org/10.47001/IRJIET/2023.711036

****