DOI: 10.xx/ijdsml.2023.xx

LONG-TERM FORECASTING OF ELECTRICAL LOAD USING MACHINE LEARNING ALGORITHMS - A CASE STUDY

Uttam S. Satpute, Suresh D. Mane and S.S. Deshpande
Department of Electrical Engineering, Dr. D.Y. Patil Pratishtan's College of Engineering, India

Abstract

Long-term Electrical Load Forecasting (ELF) is essential for infrastructure planning and the proper functioning of substations. ELF reduces the overall planning uncertainty added by the intermittent production of renewable energy sources. It helps to minimize the hydrothermal electricity production costs in a power grid. Although there is some research in the field and even several research applications, there is a continual need to improve forecasts. The use of Machine Learning Algorithms (MLAs) for prediction purposes is increasing in recent times. The paper presents the results of electrical load forecasting using various MLAs and their comparative analysis. The electrical load data for training the models are obtained from the 110/33/11kV substation of Haliyal, District: Uttara Kannada, Karnataka, India. Other features such as temperature and salary are included for enhancing the prediction. The MLAs are implemented in Python using Scikit-Learn. The performance of various MLAs is measured in terms of Root Mean Square Error (RMSE). The model validation is done using cross-validation. The comparative analysis shows that the Decision Tree Algorithm gives better results for the prediction of electrical load as compared to others. It is further concluded that MLAs prove to be an effective tool for substation planning, expansion, and proper functioning.

Keywords:

Machine Learning, Forecasting, Planning, Functioning

1. INTRODUCTION

The electric power system is operated continuously. It requires real-time coordination among the power plants and substations (primary and secondary) to operate securely and reliably. Before the real-time operation, it is necessary to consider the renewable energy production behavior, the power plants, and grid maintenance, and operate the hydrothermal resources, so the electricity production meets a projected demand. This real-time balance between energy generation and load should be sustained to ensure the secure and reliable operation of the grid [1].

The time scope of power system operational planning can be categorized into three frames: short-term, mid-term, and long-term [2]. The short-term timeframe ranges from 1 day to 1 week, focusing more on the power system's operational and security aspects. The mid-term timeframe ranges from several weeks to several months, focusing on managing production resources and avoiding energy deficits. Consequently, the long-term timeframe ranges from years to decades, intending to define the installation of new power plants or changes to the transmission system. At the outset, the paper focuses on long-term forecasting of electric load on substations intended for infrastructure planning and for secured operation.

During Covid and post-Covid period, the electricity consumption patterns have changed a lot. And in fact, electricity consumption is a continuously evolving process. New machine learning algorithms are emerging, encouraging the examination to update the forecasting methods with the most efficient approach [3]-[7]. The paper aims to investigate the efficient machine learning algorithm for long-term forecasting of electrical load on substation. The models will be evaluated with the electrical load data for obtained from the 110/33/11kV substation of Haliyal, District: Uttara Kannada, Karnataka, India.

2. METHODOLOGY

2.1 DATA

The electrical load data is obtained from the 110/33/11kV substation Haliyal, District: Uttara Kannada, Karnataka, India. The substation consists of two 110kV incoming lines and two 33kV and ten 11kV outgoing lines supplying electrical load to Haliyal city and surrounding villages. The data is obtained over 5 years since 2017. Figure 1 shows the load on the substation year-wise.

It can be seen that the pattern of load over the year is almost similar having a high load during summer (March-August) and a low load during rain (August-September). The intended rise in load is found to be declined in 2020 due to corona.

The electrical load is tail heavy on the right-hand side about its mean as shown in the histogram (Fig.2). But most of the MLAs predict better for bell-type shaped data. So the data need to be transformed to normal distribution. The Table.1 describes the data in terms of the count, mean, standard deviation, minimum, maximum, and 25%, 50%, and 75% percentile.

In the data which has been recorded in the substation register, it is observed that some readings were missing and some readings were entered wrong (outliers). So initially data cleaning and transformation has been performed to prepare the data for machine learning algorithms.

2.2 DATA PREPARATION

The missing values in the data are replaced by the mean value. It is done year-wise, i.e. the missing values in the particular year (say 2017) are replaced by the mean value of that year's data only. Whereas the outliers are handled by adopting a different strategy. Normally the outliers are due to errors while entering the data into the record register. The readings greater than two times the power transmission capacity of the feeder are considered outliers and they are replaced by mean value as discussed above for missing values. The data so obtained is then transformed to the normal distribution (scaled in the range 0 to 1) using Eq. (1).

$$X_{normal}=(X-\mu)/\sigma \qquad (1)$$

where: X is the data value, μ is the mean value, and σ is the standard deviation. The normalization is also done year-wise. Data preparation has been implemented in Python using the "StandardScaler" function of the Scikit Learn library.

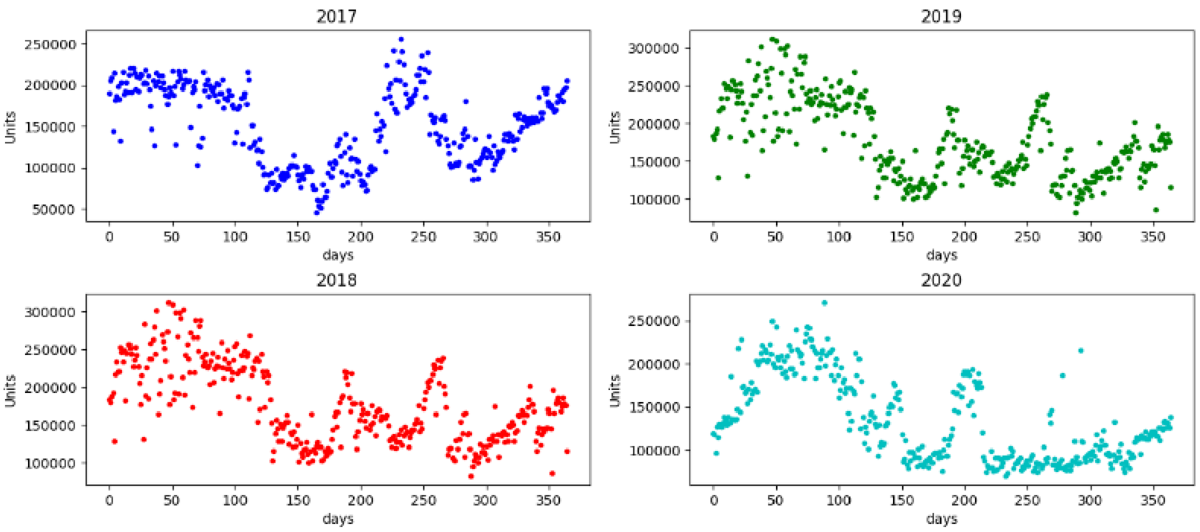

Fig.1. Electric Load on the substation – year wise

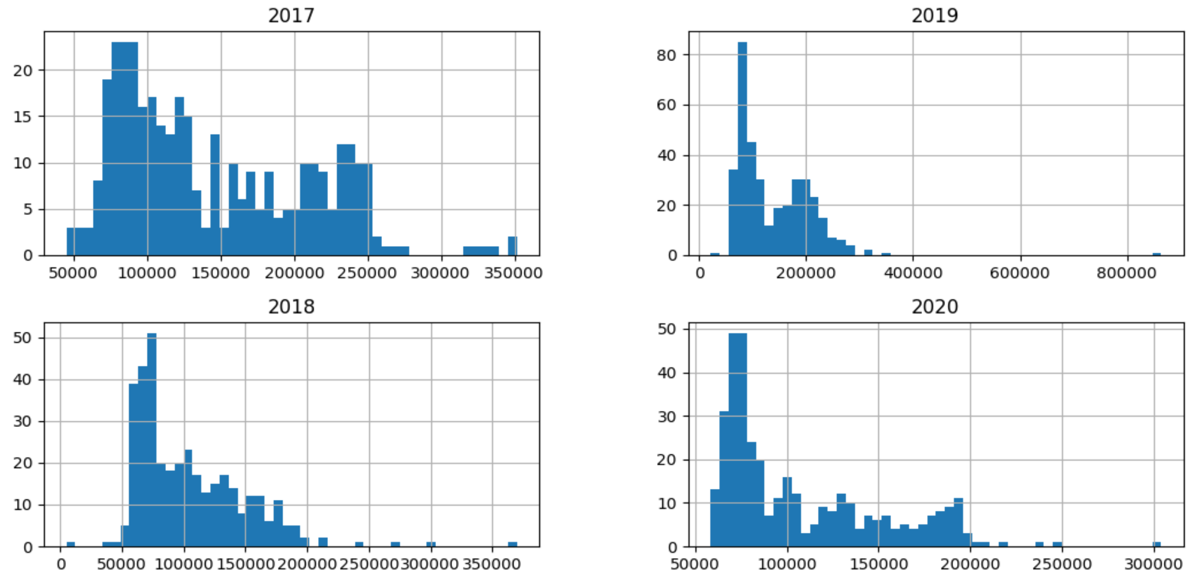

Fig.2. Histogram of electrical load year wise

Table.1. Description of electrical load year wise

	2017	2018	2019	2020
Count	365	365	365	365
Mean	144946.6521	138578.5	104777.6	107387.6
Std	64382.86184	72859.12	43970.15	43887.57
Min.	44948	21560	5000	58500
25%	90500	81400	70800	73200
50%	124900	113200	94000	86400
75%	202580	187900	131300	134800
Max.	351700	863900	369800	304000

2.3 MACHINE LEARNING ALGORITHMS

The present problem of electrical load prediction is a category of supervised learning because the training set that we feed to the algorithm includes the desired solutions i.e. electrical load on the substation; further, the problem is categorized as batch learning because here the entire set of available data (electrical load of previous years) is utilized for training the algorithms; further, the problem is categorized as model-based learning, because here the models are developed using various MLAs for electrical load prediction.

The following MLAs are implemented for load prediction:

- Linear Regression
- Polynomial Regression
- Stochastic Gradient Descend (SGD)
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest

The theoretical background of MLAs has not been presented here, as it is available in the literature and importance has been given to their implementations. The electrical load data from 2017 to 2020 is used for training and data from 2021 is used for testing the models.

2.3.1 Linear Regression:

The load is modeled using a linear equation of the form:

$$y = Ax + C \tag{2}$$

where, y is the load, x is the feature, A is the coefficient, and C is the intercept.

It is implemented using the "Linear Regression" function of the Scikit-Learn library [8]. All the parameters are set to default values. The coefficient and intercept of the linear model are obtained as follows: A: -29157.93 and C:150314.30.

The Root Mean Square Error (RMSE) for the testing data is obtained as follows: RMSE:46395.19

The Fig.3 shows the training set data for four years from 2017 to 2020 and the linear model obtained.

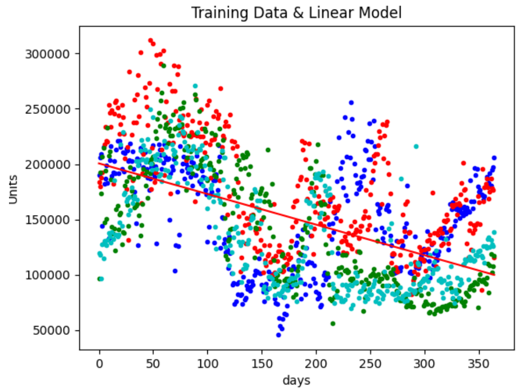

Fig.3 Linear Model and the training data

The Stochastic Gradient Descent (SGD) algorithm is also implemented using "SGDRegression" function, and similar results are obtained as Linear Regression.

2.3.2 Polynomial Regression:

The load is modeled using a quadratic equation of the form:

$$y = Ax^2 + Bx + C \tag{3}$$

where: y is the load, x is the feature, A and B are the coefficient, and C is the intercept.

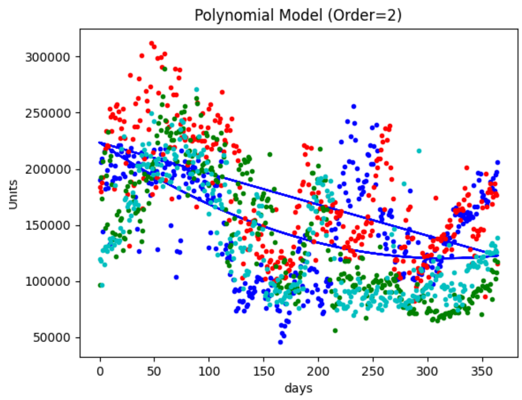

Fig.4. Polynomial model (Quadratic) and training the data

It is implemented using the "PolynomialRegression" function of Scikit-Learn library [8]. The coefficient and intercept of the linear model are obtained as follows: A: -29157.93; B: 11484.92 and C:138829.37

The Fig.4 shows the training set and the quadratic model so obtained. The RMSE determined based on testing data is as follows: RMSE=47355.06

Further, it observed that as the order of the polynomial is increased the prediction gets improved as shown in Fig.5, which shows, how the RMSE goes on reducing as the order is increased. For the present electrical load, the 7th-order polynomial is found to be best suited for prediction.

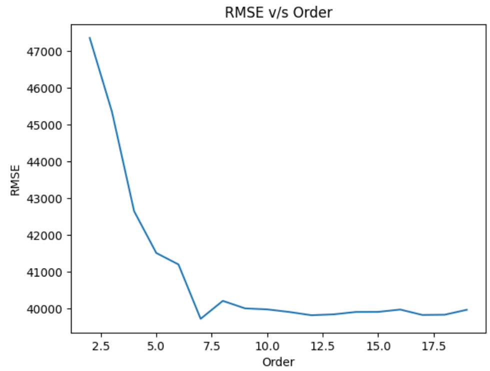

Fig.5. RMSE v/s order of the polynomial

2.3.3 Support Vector Machine:

Linear and nonlinear SVMs are implemented for regression tasks using the "Linear SVR" and "SVR" functions of Scikit-Learn [8]. The hyper-parameter margin value is set to 1.5. Fig.6 shows the Linear SVM model and the training set data. The RMSE is obtained as follows: RMSE=128600

The Fig.7 shows the Nonlinear SVM model and the training set data. The RMSE is obtained as follows: RMSE=53101.15

2.3.4 Decision Tree:

The Decision Tree algorithm has been implemented using "DecisionTreeRegressor" function of Scikit Learn [8]. The model so obtained is shown in Fig.8. The RMSE is obtained as follows: RMSE=41376.156.

2.3.5 Random Forest:

The Random Forest algorithm has been implemented using "RandomForestRegressor" function of Scikit Learn [8]. The model so obtained is shown in Fig.9. The RMSE is obtained as follows: RMSE=41505.70

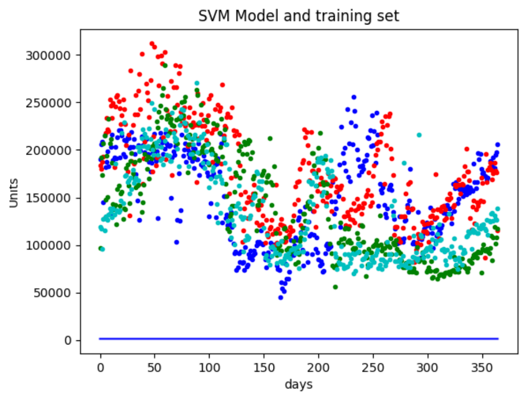

Fig.6. Linear SVM model and training the data

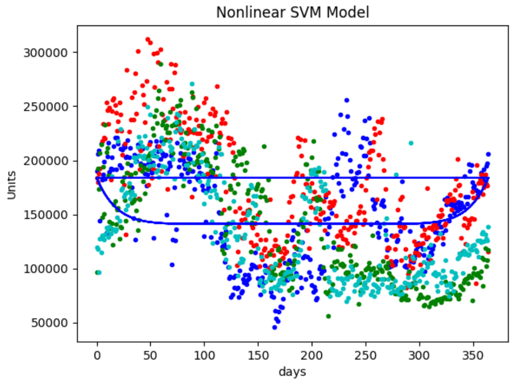

Fig.7. Nonlinear SVM model and training the data

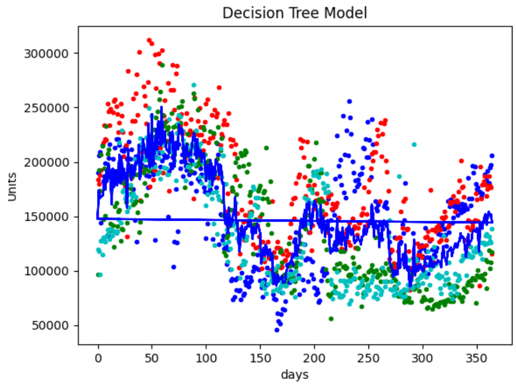

Fig.8. Decision Tree model and training the data

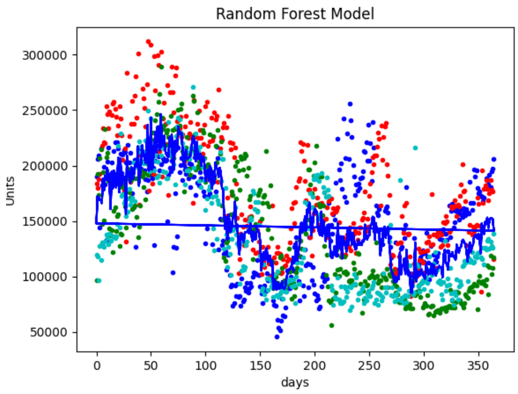

Fig.9. Random Forest model and training the data

2.3.6 Artifical Neural Network (ANN):

The Artificial Neural Network has been implemented using "KERAS API" function of Tensorflow [8]. The ANN model has been prepared with seven hidden layer, each consisting of 50 neurons. The model has been trained with "RMSPROP" optimizer and with "Mean Squared Error" as loss function. Each of the optimization algorithm is iterated for 20 iterations. The RMSE on testing data is obtained as follows: RMSE=47295

2.4 VALIDATION OF MODELS

All the models developed above are validated using the cross-validation concept. The total training data has been divided into 10 batches. It is done using "Cross_Value_Score" function of Scikit Learn [8]. Table.2 and Fig.10 gives the result of RMSE and validation results for each of the algorithms.

Following are the various inferences that can be drawn from the results:

- Decision Tree and Random Forest algorithms are proven to be best suited for load prediction.
- High order polynomial model above 7 also gives a better result.
- All the models fit the training data well but perform poorly on testing data, which implies that the models are over-fitted and not able to generalize the data.
- The above problem can be overcome by including other features such as population, climatic conditions, lifestyle, salary, operating conditions, etc.

2.5 INCLUSION OF OTHER FEATURES

The prediction can be improved by including other features such as climatic conditions, population, salary, etc. Fig.11 shows the temperature, dew, humidity, wind speed, population data, and salary data of the Haliyal region during 2017-2020.

Climatic parameters are obtained from the Belaum Weather Station which is 80km away from Haliyal [9]. Population data and salary data are obtained from [10] [11] respectively. The correlation between various features is obtained in Table.3. It can be seen that temperature and salary have positive relation with the electrical load on the substation whereas other features have negative relation.

All the MLAs as discussed are again implemented with new additional features. The Table.4 and Fig.12 give the consolidated results. It can be seen that almost all the results are improved except a few. The decision tree gives zero RMSE on training data, it means that the algorithm prediction of the load is accurate on training data, but it is performing poorly on testing data. Random forest on other hand is ranked second.

Based on testing data, Artificial Neural Network is performing better. Altogether it is concluded that the inclusion of features has improved the predictions on training data but performed poorly on testing data. It implies that the models could not able to generalize the data and need to either reduce the features or else go for a better model (algorithm).

Table.2. Comparative analysis of Machine Learning Algorithms

Algorithm	Order	RMSE based on Training Data	RMSE based on Testing Data	Cross Validation	
				Mean	Standard Deviation
Linear Regression (LR)		44581	46395	45945	7257
Stochastic Gradient Descent (SGD)		44581	46395	45945	7257
Polynomial Regression (PR)	2	43381	47355	46123	6832
	4	41075	42648	44748	7568
	6	39146	41200	44315	5226
	8	37890	40211	43751	5539
	10	37890	39979	43733	5900
Linear SVM (LSVM)		158098	128600	155308	30657
Nonlinear SVM (NSVM)		52493	51910	55789	11533
Decision Tree (DT)		35357	41376	46609	7078
Random Forest (RF)		35412	41396	46600	7088
Artificial Neural Network (ANN)	Hidden layers=7	42504	47295	46459	6950

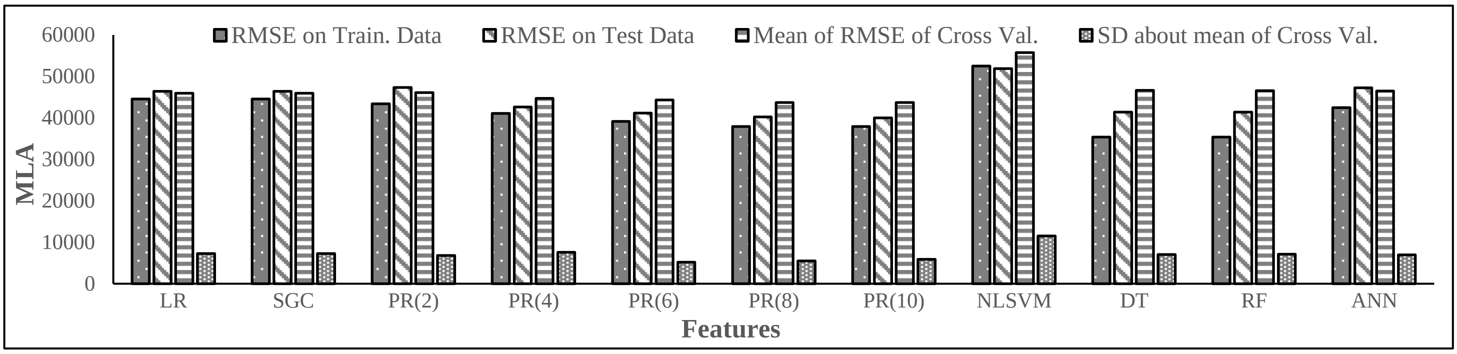

Fig.10. Comparison of Results of MLAs

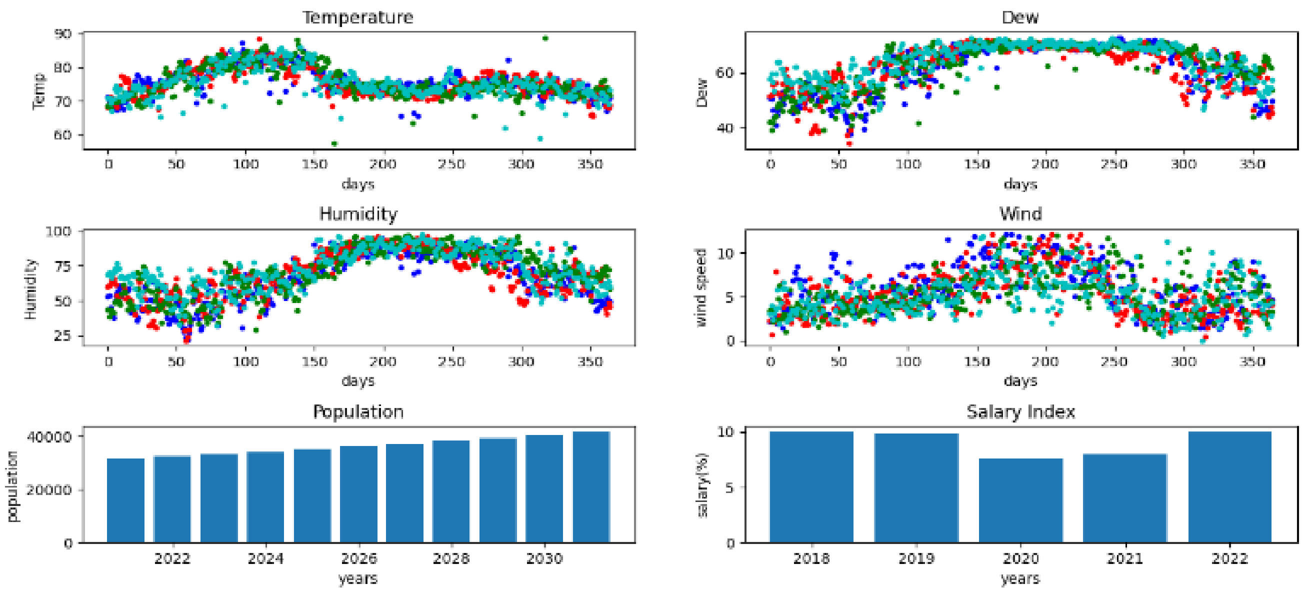

Fig.11. Climatic parameters, population and salary etc.

Table.3. Correlation between various features

Feature	Correlation
Electrical load units	1.0000
Temperature	0.2725
Dew	-0.4898
Humidity	-0.5702
Wind Speed	-0.2409
Population	-0.3653
Salary	0.2378

Table.4. Comparative analysis of Machine Learning Algorithms including all the features

Algorithm	RMSE based on Training Data	RMSE based on Testing Data	Cross Validation	
			Mean	Standard Deviation
Linear Regression (LR)	39870	44189	43150	9911
Stochastic Gradient Descent (SGD)	39880	44547	43030	9920
Polynomial Regression (PR) (Order=2)	36254	41997	48649	9957
Linear SVM (LSVM)	158098	128600	155646	30664
Nonlinear SVM (NSVM)	51139	48123	54809	9974
Decision Tree (DT)	0	54099	40891	10691
Random Forest (RF)	7893	45543	37653	10147
Artificial Neural Network (ANN)	40540.41	36358.64	34320	8540

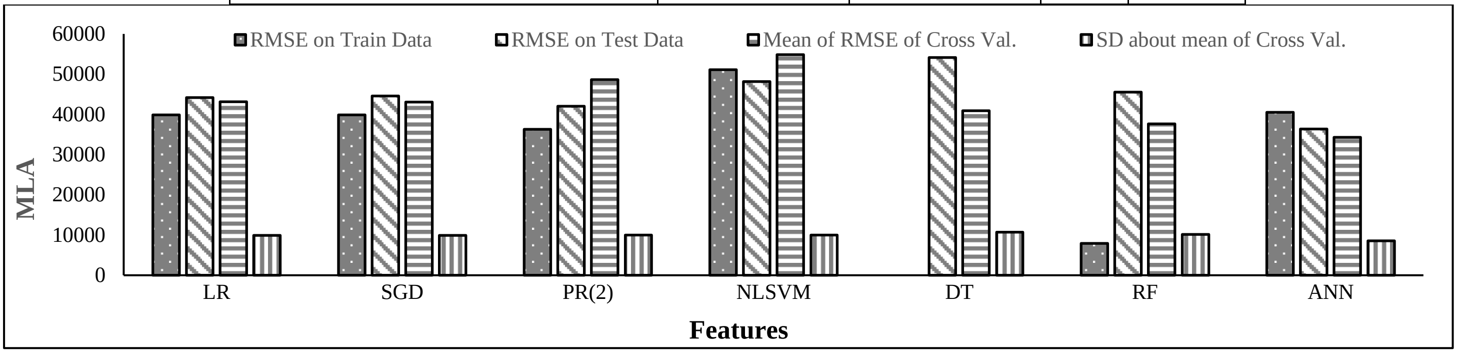

Fig.12. Comparison of MLAs with all features

3. CONCLUSIONS AND FUTURE SCOPE

The paper has presented the long-term forecasting of electrical load on substation using Machine Learning Algorithms. The substation load forecasting is critical for infrastructure planning, secured and reliable operation of the substation. All the MLAs are implemented in Python using SCIKIT Learn with climatic parameters, population and salary as additional features. The results show that prediction gets improved with additional features on training set but performs poor on testing data. So it is concluded that, some complex model is required for prediction or some irrelevant features need to be omitted from the database after deliberate feature engineering. Neural network or Deep Neural Network may be used for prediction and if time-scope is reduced to week level, the results can be still be improved. The paper has attempted to explore the use of MLAs for load forecasting and their evaluation but the same can be extended to other forecasting applications as well.

REFERENCES

[1] A.J. Wood, B. Wollenberg and G. Sheble, "*Power Generation, Operation, and Control*", John Wiley and Sons, 2013.

[2] S. Hossein and S.S. Mohammad, "*Electric Power System Planning*"; Springer, 2011.

[3] E. Becirovic and M. Cosovic, "Machine Learning Techniques for Short-Term Load Forecasting", *Proceedings of International Symposium on Environmental Friendly Energies and Applications*, pp. 1-4, 2016.

[4] H.O. Sarmiento and W. Villa, "Load Forecasting with Neural Networks for Antioquia-Choco Region", *Proceedings of IEEE/PES Conference on Transmission and Distribution*, pp. 1-7, 2008.

[5] O. Adeoye and C. Spataru, “Modelling and Forecasting Hourly Electricity Demand in West African Countries”, *Applied Energy*, Vol. 242, pp. 311-333, 2019.

[6] S. Dutta, F.H. Choo and H.B. Puttgen, “Load and Renewable Energy Forecasting for a Microgrid using Persistence Technique”, *Energy Procedia*, Vol. 143, pp. 617-622, 2017.

[7] N. Paterakis, B. Stappers and W. Van Alst, “Deep Learning Versus Traditional Machine Learning Methods for Aggregated Energy Demand Prediction”, *Proceedings of International Conference on Innovative Smart Grid Technologies*, pp. 1-6, 2017.

[8] Aurilien Geron, “Hands On Machine Learning with Scikit-Learn, Keras and TensorFlow”, Oreilly Publisher, 2019.

[9] Weather Underground, “Belgaum Karnataka, India, Weather History”, Available at https://www.wunderground.com/history/monthly/in/belgaum/VABM/date/2019-2 , Accessed on 2023.

[10] Haliyal Town Panchayat City Population Census 2011-2023, Available at https://www.census2011.co.in, Accessed on 2023.

[11] India Records Highest Salary Increase, Available at https://www.livemint.com, Accessed on 2023.