Machine Learning Algorithm Comparison for Short-Term Electricity Consumption

Emre Çakmakyurdu
Department of Data Informatics
Middle East Techinal University
Ankara, Turkey
emre.cakmakyurdu@metu.edu.tr

***Abstract*— Electricity prediction research is widely investigated since electricity consumption and production is vital for modern human life. Electricity needs to be utilized efficiently and effectively. Production of electricity needs to supply the consumption demand; thus, electricity consumption prediction plays a significant role in electricity production. This paper uses machine learning methods and statistical analysis to address two critical aspects of electricity consumption prediction. Firstly, it examines the impact of temperature on power consumption in different zones using ANOVA. Weather data is an important aspect in predicting electricity. Secondly, this paper evaluates the performance of various machine learning models, including linear regression, feedforward neural networks, random forest, support vector regressor in predicting electricity consumption in Tetouan City over varying timeframes, such as every 10 minutes. Linear regression used as a baseline predictor. To increase performance of the models, parameters are optimized using Grid Search method.**

Keywords*— *Electricity consumption prediction, linear regression, random forests, forecasting, machine learning, ANOVA

I. Introduction

Electricity is crucial in our modern daily life. It is widely used from lighting our houses to industry. A lot of money have been spending on electric production and there is an ongoing effort to maintain this production to be effective. To achieve this effectiveness, production and consumption of the electricity should be balanced. Abdulwahed Salam et. al. [1] constructed an equation that describes the relationship between produced electricity and consumption.

$$Production - (consumption + ELT) = v$$

ELT represents the energy lost in the production. Ideally, v should be zero which means the electrical consumption and production is equal and balanced. If it is a large positive number, production exceeds the consumption and there is an unwanted electric production which will be lost. On the other hand, if v is a large negative number, there are not enough electricity is produced.

Electrical consumption prediction is classified under three different categories.

- Short-term electrical consumption forecast: From an hour until a week.
- Medium-term electrical consumption forecast: From a week until a year.
- Long-term electrical consumption: More than a year.

Electricity cannot be stored for long-term so, investigating and predicting the short-term behavior can be helpful to electric power companies to plan their next hour and 10-minute electricity production.

Machine learning is a part of artificial intelligence where computers learn from data and experience using algortihms and statistical tests. Machine learning algortihms can be classified as two types: supervised, where the computer learns from the data and produce outputs and unsupervised, where computer finds patterns on its own. Statistical tests are used for determining the likelihood that observed pattern or relationship is occurred by chance or random.

In this paper, the effect of temperature on power consumption of each zone will be investigated using ANOVA test. Also, the best machine learning model on the short-term (10 minutes) electrical consumption on Tetouan City database will be investigated.

II. Literature Review

1) This article compares various machine learning algorithms for predicting energy consumption in Tetouan City. It focuses on predicting power usage every 10 minutees and every hour. It uses feedforward neural networks with backpropogation, decision trees, random forest and support vector machines. It aims to find efficient machine learning model fro efficient planning and operations. [1]

2) This article emphasizes the machine learning model “Prophet” developed by Facebook for electrical consumption forecasting. It emphasize the effectiveness of the model especially in the short-term consumption prediction. This article compares the traditional machine learning algortihms with “Prophet” and highlight its advantages over traditional models. Also, it emphasize that “Prophet” can handle seasonality which is an important factor in electrical consumption forecasting. [2]

3) This article compares Artificial Neural Networks with multiple regression analysis using data obtained from 2006 to 2016. The satudy finds that ANN models with backpropogation algortihms outperform MRA models. This study uses root mean squared error and mean absolute percentage error as metrics. It concludes that Artifical Neural Networks best perform when data has complex patterns and has noise in the data. [3]

4) This article is an overview of the research articles published from 2016 to 2019 in the field of energy prediction. It reviews 31 articles and seperates them based on the dataset features, machine learning algorithm used and evaluation metrics. Also, it focuses on how short-term energy forecasting is affected with different methodologies and metrics. [4]

III. OVERVIEW OF MACHINE LEARNING ALGORITHMS

A. Linear Regression

Linear regression is one of the simplest machine learning algorithm used in statistical uses. It is based on the linear relation between the dependent and independent variables. [1] It is very similar to this linear equation.

$$y = b + b_1X_1 + b_2X_2 + ... + b_nX_n \quad (1)$$

$$y = b_0 + \Sigma\, b_iX_i \quad (2)$$

In eqution (1) and (2), y is the target output and x is the independent variable and summation of independent variables with their biases gives the dependent variable. Error is calculated using Root Mean Squared Error (RMSE). It is generally used as baseline model for comparision.

B. Feed Forward Neural Networks

Feed forward neural networks 3 or more layers. An input l layer, one or more hidden layers and an output layer. Each neuron in one layer conncects to the next layers' neurons, however there are no back loops. It uses weights and biases to investigate the relations in the data, applies activation functions like tanh, RELU, sigmoid for nonlinear relations prensent in the data. Hidden layeres do not expose to the input or output, they are the computational backbone of the algorithm. Output layer produces the output. Feed forward neural networks consists of two phases.

- Feedforward phase: Input data flows thorugh the network. At each layer, a non-linearity is added to the inputs by calculating the weighted sum of the inputs and applying activation function. This phase ends after output layer generates a prediction.
- Backpropogation phase: Once the prediction is made, error, difference between the prediction and the actual result, is calculated and propogated through the network. Weights are updated by considering the error and tries to minimize the error function.

The output of the neuron is calculated using (3) where w_i are the weights, x_i are the inputs, b is the bias and f is the activation function.

$$y = f(\Sigma(w_ix_i) + b) \quad (3)$$

The calculate the error function, Mean Squared Error (MSE) or Root Mean Squared Error is used (RMSE) where x_i is the actual result and y_i is the predicted result. MSE i calculated using (4). RMSE is the squared root of (4).

$$y = \frac{1}{n} \Sigma^n_{i=1}(y_i - x_i)^2 \quad (4)$$

C. Random Forest

Random forest is an advanced ensemble method which combines multiple decision trees to increase the accuracy more than a single decision tree achieves. More than one regression tree is constructed and each tree contributes to the final prediction. Final prediction is the average of the all individual tree predictions. At each split, trees are splitted randomly which increases the diversity of the model and prevent overfitting. This process also involves out of bag samples to estimate accuracy without a seperate validation set. A random forest regresiion predictor is defined as [1]:

$$f^N_{rf}(x) = \frac{1}{N}\sum_{n=1}^{N} T_{tree}(x) \quad (5)$$

Where T_{tree} is referred as decision tree and x is a p-dimensional vector of inputs.

IV. DATASET

Data used in this study is the time series data of power consumption in Teuotan City which is collected for every 10 minutes between 2017-01-01 00.00.00 and 2017-12-31 23.50.00. There is approximately a year long data is available. It contains measurements taken from sensors such as temperature, humidity, wind speed, general diffuse flows and diffuse flows.

Dataset contains no missing values.To investigate dataset, state of the art libraries like Pandas, matplotlib, seaborn and statsmodel api is used. Missing values checked using Pandas "isna()" method. Then, dataset is checked whether it contains any disguised missing data by using Q-Q plots. Disguised missing data is not explicitly represented as such, but entered some values which can be interpreted as correct data values. This causes the model performance to degrade.

Correlation between the features are also investigated using Pandas "corr()" method. Understanding correlation between features is an important step to increase model performance and understand the data by feature reduction and giving proper weights to some features. Table 1. summarize the correlation between features.

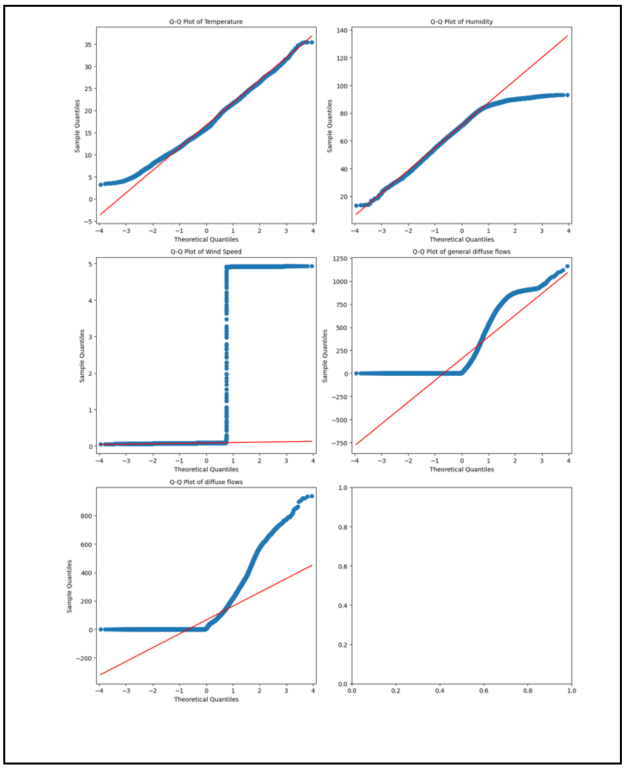

Fig. 1. Q-Q Plots of the dataset

Index	Temperature	Humidity	Wind Speed	General diffuse flows	Diffuse Flows	Zone 1 Power Consumption	Zone 2 Power Consumption	Zone 3 Power Consumption
Temperature	1	-0.46	0.47	0.46	0.19	0.44	0.38	0.49
Humidity	-0.46	1	-0.13	-0.47	-0.25	-0.28	-0.29	-0.23
Wind Speed	0.47	-0.13	1	0.13	-0.001	0.16	0.14	0.27
General diffuse flows	0.46	-0.46	0.13	1	0.56	0.18	0.15	0.06
Diffuse Flows	0.19	-0.25	-0.001	0.56	1	0.08	0.04	-0.04
Zone 1 Power Consumption	0.44	-0.28	0.16	0.18	0.08	1	0.83	0.75
Zone 2 Power Consumption	0.38	-0.3	0.14	0.15	0.04	0.83	1	0.57
Zone 3 Power Consumption	0.49	-0.23	0.27	0.06	-0.04	0.75	0.57	1

Table 1. Correlation matrix between features

Dataset contains 7 'float64' columns an done 'object' column. Object column needs to be converted to 'datetime' to become an input to the model. This process is done using Pandas. Table 2. Demonstrates the object type of the dataset after conversion.

Table 2. Data types of the features

Index	Temperature
DateTime	Datetime64 (ns)
Temperature	float64
Humidity	float64
Wind Speed	float64
General diffuse flows	float64
Diffuse Flows	float64
Zone 1 Power Consumption	float64
Zone 2 Power Consumption	float64
Zone 3 Power Consumption	float64

V. METHODOLOGY

A. Machine Learning Models

In this paper, 4 different model is analyzed. For each zone three different model is generated. However, feed forward neural network model is constructed as one model. Before moving on to analysis part, feature importance, selection and generation is constructed. For feature selection a random forest model is constructed to obtain the importance of each feature. Random forest model is selected since it is easy to construct, and training time is significantly lower. Based on the results of random forest, the most important feature is 'Temperature' followed by 'Humidity'. However, results indicate that each feature contributes to the model performance significantly. The least contributing feature is 'diffuse flows' which has an importance of almost 10% which is quite high. So, all the features are selected based on the analysis. For feature elimination, PCA is used. Based on the results of PCA, first three principal components explain most of the variance in the data.

Preprocessing is applied to the dataset for the best performance. Box-Cox transformations and Standard Scaler is applied to data. Box-Cox transformations transforms non-normal data into a normal distribution. Standard scaler centers and scales the features. Box-Cox method is effective as it can handle non-normal data and since the dataset contains non-normality it is appropriate to apply here. All the features are scaled centered with mean of zero and standard deviation of 1.

Hyperparameters are effective to achieve the best performance of the machine learning model that is used. In this research, randomized search is used to select the best hyperparameters which boost the model performance. The algorithms mentioned is implemented using scikit-learn, Keras and PyTorch.

Three different error metrics are used in the evaluation of the algorithms. Mean Squared Error (MSE), Mean Absolute Error (MAE) and R2 score. Each have different effect on the evaluation. R2 score measures the proportion of the variance in the dependent variables that is predictable from dependent variables. The higher the R2 score, the higher the accuracy. It is easy to interpret however it can be misleading if dataset contains outliers. MAE calculates the average of the absolute errors in the dataset and all differences have equal weight. It is less sensitive to outliers however it does not penalize the errors. MSE calculates the average of the squared errors. It penalizes large errors however it is more sensitive to outliers.

Linear Regression algorithm does not have any hyperparameters to tune. It is used as a baseline model for comparison.

The most important parameters to tune in Random Forest Regressor is maximum depth of the tree, number of trees in the forest, minimum number of samples required to split an internal node and minimum number of samples required to be at the leaf node. Consecutively, tried numbers are [10, 20, 30, None], [100], [2, 5, 10] and [1, 2, 4, 6, 7, 8]. Based on the results of the randomized search, for Zone 1 parameters are chosen as [30, 100, 2, 1] consecutively. For Zone 2, parameters are chosen as [None, 100, 5, 1] and for Zone 3 parameters are chosen as [20, 100, 2, 1] consecutively.

The most important parameters to tune in Support Vector Regressor is regularization parameter, kernel coefficient and kernel type. Consecutively, tried numbers are [1, 10, 100, 1000], [0.1, 0.001, 0.0001] and ['rbf']. Based on the results of the randomized search, best parameters are for Zone 1, [1000, 0.01, rbf], for Zone 2, [1000, 0.01, rbf] and for Zone 3 [100, 0.01, rbf] consecutively.

For feed forward neural network, two layered model is constructed with PyTorch. The hyperparameters are selected using manual hyperparameter tuning. Different batch size, optimizer and activation function are experimented. Batch size parameters are 10,20. Optimizer parameters are 'Adam' and 'SGD', and activation functions are 'relu' and 'tanh'. Learning rate parameter is not tried for manual hyperparameter tuning for speed purposes and it is set as 0.001. Batch size is the number of training samples used to perform one update of the model weights. Small batch sizes lead to faster convergence however, large batch sizes can lead more accurate results of the model. Optimizer tries to reduce the loss by updating the weights and learning rate. Stochastic Gradient Descent (SGD) has good generalization capability whereas Adam is an adaptive learning rate optimizer have faster convergence. Activation functions decide whether a specific neuron should be activated or not by looking the weighted sum. In this study, RELU and tanh is used. Based on the analysis, best parameters are chosen as (10, 'adam', 'tanh').

Influence of each feature to decision making is investigated using shap library. In Figure 2, influence of each feature on decision making can be observed for linear regression In Figure 3, random forest model can be observed.

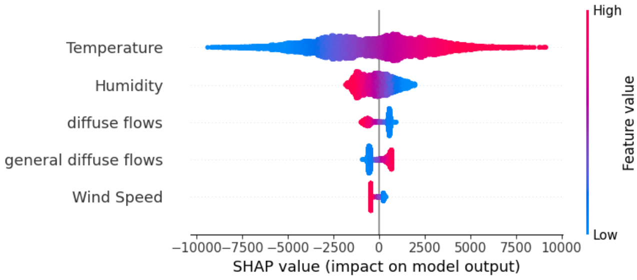

Fig. 2. Output of SHAP for linear regression model

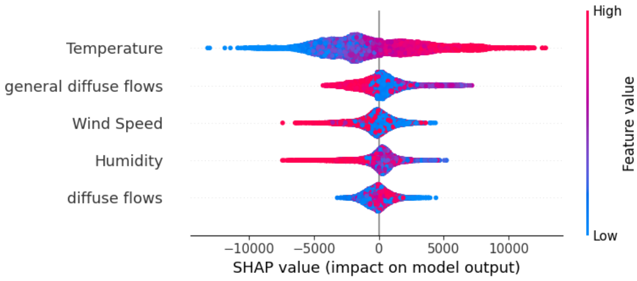

Fig. 3. Output of SHAP for random forest model

Effect of temperature in decision making is significant.

B. *Statisctical Test*

The null hypothesis states that there is no significant difference in power consumption in different temperature zones. Significance level is chosen as 0.05.

$H_o =$ there is no significant difference in power consumption in different temperature zones.

Alternative hypothesis states that there is significant difference consumption across different temperature zones.

$H_1 =$ there is no significant difference in power consumption in different temperature zones.

For this purpose, ANOVA test is used. Temperature is categorized into 4 different categories indicating low, normal, high and very high temperatures. Then, normality of the variables are observed since ANOVA test is gives accurate results when distribution is normal. For this purpose, Shapiro-Wilk test is used and Q-Q plots is used for observation of normality. After that, log transformation applied to dataset to normalize the features. Then, ANOVA test is constructed.

VI. Results

A. *Machine Learning Models*

Table 3 summarize the results for linear regression, random forest, support vector regressor and feed forward neural network models. Linear regression is used as baseline model. Experimental results states that, best model for this task is random forest model. Applied performance criterations are MAE, MSE and R2 score.

B. *Statistical Test*

Output of ANOVA test results states that there are significant differences in power consumption across different temperature zones and it also states that temperature is highly effective in power consumption. Finding can be found in Table 4.

Table 4. Output of ANOVA test

	ANOVA RESULTS		
	ZONE 1	***ZONE 2***	***ZONE 3***
p-value	0	0	0
F-statistic	3533	2492	4830

VII. Conclusion

This research is focused on finding the best machine learning model for electricity consumption and the effect of temperature on power consumption. It is concluded as support vector regressor is the most effective model for predicting electricity consumption and temperature has a significant effect on power consumption. The potential future work in this topic includes applying a statistical test which is more robust to non-normality of data, developing deep learning methods to further establishes the hidden relationships in the data and extending the dataset by adding weather related and socio-economic parameters.

Table 3. MAE, MSE and R2 scores of ML model

Table Head	**ZONE 1**			**ZONE 2**			**ZONE 3**		
	MSE	***MAE***	***R2***	***MSE***	***MAE***	***R2***	***MSE***	***MAE***	***R2***
LR	4077382	5221	0.2	22557736	3825	0.17	31852791	4484	0.284
RF	23456283	3424	0.54	12139593	2461	0.555	13021458	2461	0.7
SVR	41843040	4897	0.186	22186070	3645	0.187	32400445	4349	0.27
FFNN	284413760	14357	-5.5	284413760	14357	-5.5	284413760	14357	-5.5

References

[1] A. Salam and A. E. Hibaoui, "Comparison of Machine Learning Algorithms for the Power Consumption Prediction : - Case Study of Tetouan city –," 2018 6th International Renewable and Sustainable Energy Conference (IRSEC), Rabat, Morocco, 2018, pp. 1-5, doi: 10.1109/IRSEC.2018.8703007.

[2] S. Desai, T. Dalal, S. Kadam and S. Mishra, "Electrical Load Forecasting using Machine Learning," 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), Puducherry, India, 2021, pp. 1-6, doi: 10.1109/ICSCAN53069.2021.9526444.

[3] S. L. Karunathilake and H. R. K. Nagahamulla, "Artificial neural networks for daily electricity demand prediction of Sri Lanka," 2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2017, pp. 1-6, doi: 10.1109/ICTER.2017.8257823.

[4] N. S. Md Salleh, A. Suliman and B. N. Jørgensen, "A Systematic Literature Review of Machine Learning Methods for Short-term Electricity Forecasting," 2020 8th International Conference on Information Technology and Multimedia (ICIMU), Selangor, Malaysia, 2020, pp. 409-414, doi: 10.1109/ICIMU49871.2020.924