Topias Henrik Jokiniemi
&
Talal Saleh

Report on Energy forecasting by regressors

Vaasa 2020

Contents

1 Introduction 4
1.1 Gradient boosting regressor 5
1.2 Linear regressor 5
1.3 Multi-layer perceptron regressor 5
2 Methodology 6
2.1 Datasets 6
2.2 Feature extraction 6
2.2.1 Feature importance 7
2.3 Code snippets 8
2.4 Comparison of regressors 9
2.5 Results obtained 12
3 Possible improvements 14
4 Conclusion 15
Bibliography 16

Figures

Figure 1**:** 2018 weekday mean and 2018 hourly mean

Figure 2: Feature importance

Figure 3: Model performances

Figure 4: January 2019 energy prediction

Figure 5: January 2020 energy consumption

Tables

Table 1: Accuracy scores of models

1. Introduction

Increase in human population and technological advancement has created an increase in demand of electricity. Electricity demand and supply curve greatly fluctuates which also have an impact on the stability of power systems. Integration of distributed energy resources (DERs), complex loads and latest demand response technologies on demand side such as electric vehicles and IoT has further aggravated power system dynamics (Zhao, J.,Gómez-Expósito, A., Netto, M., Mili, L., Abur, A., Terzija, V., ... & Huang, Z. 2019). Balancing supply and demand will become more challenging in the future especially when considered along with the increase in renewable energy resources (Boßmann, T., & Staffell, I. 2015).

To ensure power system stability and reliability they are planned to be operated and controlled to deal with the power system dynamics (Zhao, J.,Gómez-Expósito, A., Netto, M., Mili, L., Abur, A., Terzija, V., ... & Huang, Z. 2019). Control of the power system depends upon the load demand. Energy from various generators may be fed into the system or removed from the energy network based on the electricity demand. To manage load, system reliability and stability energy forecasting can play a great part which could help in pre-planning the system operation which will help in balancing energy supply and demand.

In this project work, a study is conducted to achieve a model to forecast the energy load in Finland. In Finland the weather varies a lot between seasons. It was assumed that temperature would be one of the most important factors for energy forecasts as in colder weather the houses need to be heated. Finnish meteorological institute (Ilmatieteenlaitos) provides great weather data in all cities. The data contains air temperature, pressures, wind speeds, humidity, snow depth and even cloud amount. All of this data is tracked hourly throughout the whole year. Eventually it was found while analysing the electricity load data, that electricity load varies based on hour and weekday. This information was further used to improve the models.

1.1.Gradient boosting regressor

Gradient boosting regressor use decision trees to solve the problem. It works in a way that it modifies the sample by setting sample to negative gradient and keeps the distribution constant (Duffy, N., & Helmbold, D. 2002). It uses gradient descent to minimize the loss. The regressor calculates the difference between the prediction and known value. The difference remaining is known as residual.

1.2.Linear regressor

Linear regression, as the name suggests, forms a linear relationship between the input and the output where an independent variable tries to predict a dependent variable. It is a type of regression which is mostly used as predictive modeling (Kumar, P., Ambekar, S., Kumar, M., & Roy, S. 2020).

1.3.Multi-layer perceptron regressor

Multi-layer perceptron network is a class of feedforward artificial neural networks (ANN) which are mostly used for machine learning and data mining. Feedforward neural networks (FF NN) are also the first ever devised neural networks (Schmidhuber 2015). MLP is a supervised learning method which requires a certain set of datasets which will be used to predict future outcomes or classes. MLPR consists of three layers namely input layer, hidden layer and the output layer. Similar to a human brain, or a neural network, the layers consist of neurons. These neurons are arranged in each layer and are connected only with the neurons of the former layer which all together builds up a perceptron; no neurons in the same layers are interconnected (Choubin, B., Khalighi-Sigaroodi, S., Malekian, A., & Kişi, Ö. 2016). The number of neurons in the input layer depends upon the data fed into it; the hidden layer can be one or more whereas the output layer consists of neurons having the result produced.

2. Methodology

2.1. Datasets

The objective of this study was to forecast the electricity consumption of Finland based on weather data. Weather data was provided by Finnish meteorological institute (Ilmatieteenlaitos, *https://en.ilmatieteenlaitos.fi/download-observations*). The weather data used was hourly data of year 2019 and 2020. Hourly energy consumption datasets of year 2018-2020 were obtained from ENTSO-E, (The European Network of Transmission System Operators for Electricity, *transparency.entsoe.eu*). ENTSO-E has hourly load and their own daily prediction in megawatts (MW).

2.2. Feature extraction

Weather data consisted of Cloud amount, Pressure (msl), Precipitation amount, Relative humidity, Precipitation intensity, Snow depth, Air temperature, Dew-point temperature, Horizontal visibility, Wind direction, Gust speed, Wind speed.

While analysing the load data it was found that average load varies based on hours and weekdays. During nights there are smaller loads and weekends also have smaller loads than other days. Sundays seemed to have 500MW less load on average than workdays. As can be read from the chart below.

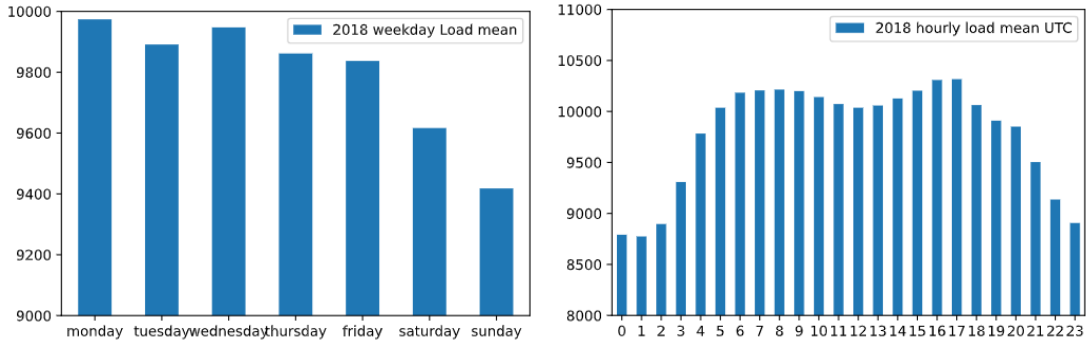

Figure 1: 2018 weekday mean and 2018 hourly mean

Hourly averages, weekday averages and the weather data were applied into the same

dataframe (X axis). To predict load based on this data the models were trained based on actual load (y axis). The result would be a regressor model that could be used to predict load if it was given a weather forecast data. Feature importance of all the features were calculated and only most important features (Snowdepth, air-tepmerature, dew-point temp., month, actual energy, average hour & average weekly energy) were used in training the regressor model.

2.2.1. Feature importance

0	Cloud amount	1/8	0.01
1	Pressure (msl)	hPa	0.00
2	Relative humidity	%	0.01
3	Precipitation intensity	mm/h	0.00
4	Snow depth	cm	0.18
5	Air temperature	degC	0.16
6	Dew-point temperature	degC	0.07
7	Horizontal visibility	m	0.00
8	Wind direction	deg	0.00
9	Gust speed	m/s	0.00
10	Windspeed	m/s	0.00
11	Actual load of previous year (at the same day & hour)	MW	0.43
12	month		0.05
13	day		0.00
14	hour		0.06
15	weekday		0.02

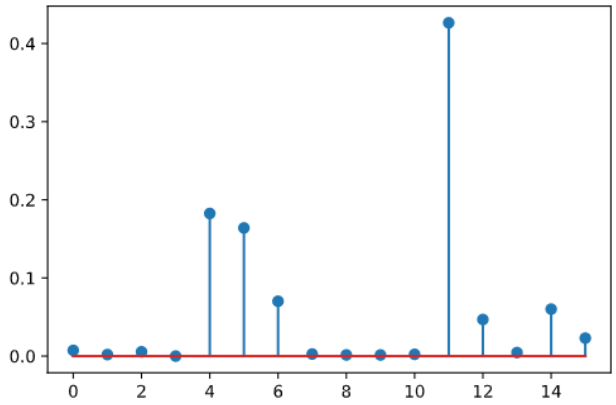

Figure 2: Feature importance

2.3.Code snippets

In the preprocessing unimportant columns were removed and empty values were filled with mean values. All of the datasets were downloaded as a CSV file type which was read using pandas.read_csv() function.

```
wea_data20=pd.read_csv("weather_helsinki_2020.csv")
weather20=wea_data20.T.iloc[5:].T
weather20 = weather20.fillna(weather20.mean())
load20=pd.read_csv("load 2020.csv")
load20=load20.iloc[:8760,2]
load20=load20.str.replace('-','') #Replacing '-' in data with empty
value
load20=pd.to_numeric(load20,errors='coerce') #converting object to
integer
load20=load20.fillna(load20.mean()).rename('energy 2020') #fillna
values with mean
```

The mean values of energy weekly and hourly were mapped on weekday and hour columns.

```
#mapping daily average of year 2018
feature['weekday']=feature['weekday'].map({'Monday':monday,
                    'Tuesday':tuesday,'Wednesday':wednesday,
                   'Thursday':thursday,'Friday':friday,
                   'Saturday': saturday,'Sunday':sunday})
```

```
#mapping hourly average of year 2018
feature['hour'] = feature['hour'].map({0: hourss[0],
                                       1:hourss[0],2:hourss[0],
                                      3:hourss[3],4:hourss[4],
                                      5:hourss[5],6:hourss[6]...})
```

2.4.Comparison of regressors

Three regression models namely gradient boosting, linear and MLP regressors are trained with the features extracted for year 2019. For the MLP regressor three hidden layers with different numbers of neurons were used to train the MLPR model. To compare each model performance seaborn library was used to plot the results.

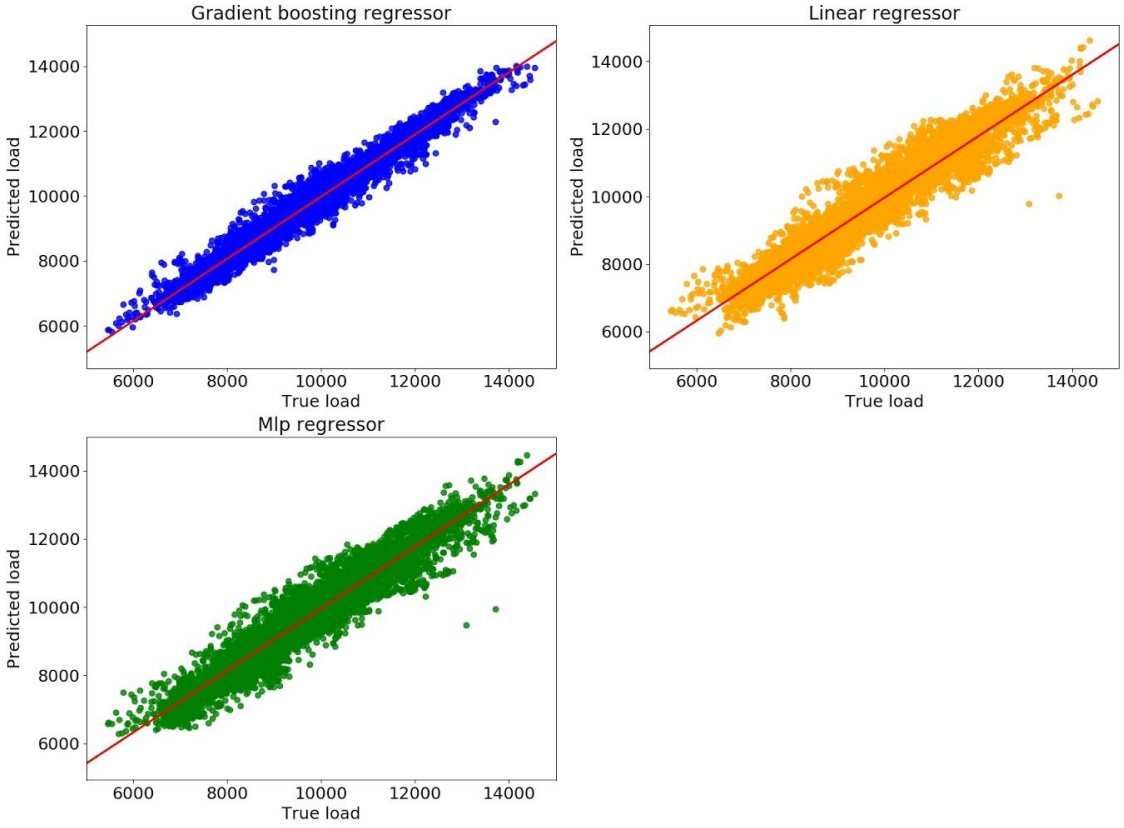

Figure 3: Model performances

To further evaluate models, accuracies of each model was found by calculating cross validation score, training and test score. Following table below shows the accuracy obtained by each model.

Table 1: Accuracy scores of models

Models	Cross validation score	Training score	Test score
Gradient boosting	0.958542	0.964054	0.959817
Linear regressor	0.906316	0.906715	0.912245
MLPR	0.899452	0.891872	0.896538

It was found that gradient boosting regressor outperformed linear and MLP regressors. To visualize the predicted energy against the actual energy of year 2019, plotting of each model was made using matplotlib library to predict energy of january 2019.

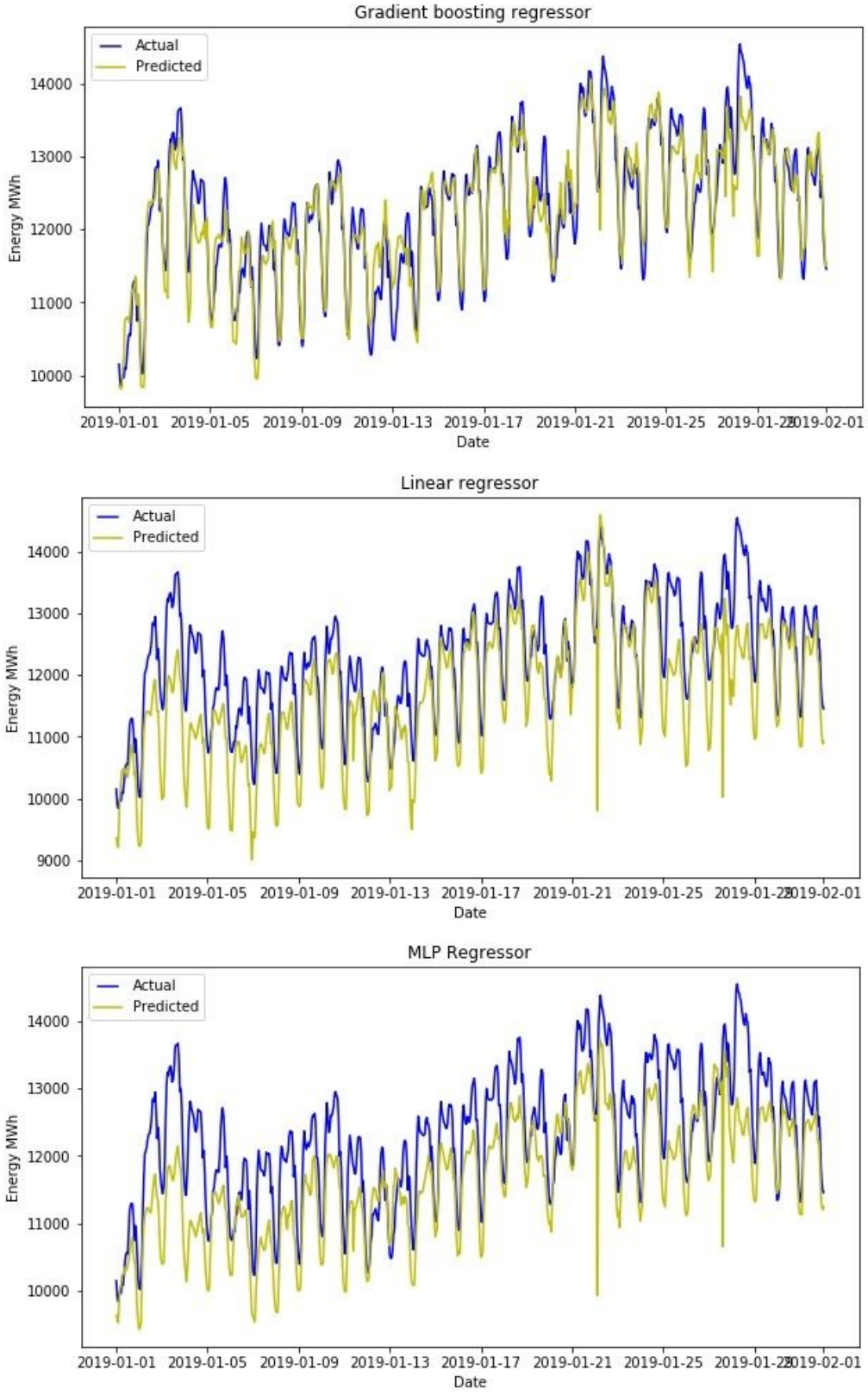

Figure 4: January 2019 energy prediction

2.5.Results obtained

All the three models were further required to predict energy consumption of year 2020. Separate feature extraction were carried out which included important features of weather data 2020 (i.e. Snow depth, air temperature and dew point temperature), months of year 2020, average daily and hourly consumption of year 2019 and energy consumed in year 2019. To predict energy consumption of year 2020, previously trained models were used. All the models can predict hourly energy consumption of the whole year, month, week or day.

Although the accuracy of the gradient boosting regressor was higher than the other two models, it was not able to accurately predict the energy consumption of year 2020. It was however more likely predicting the energy during peak hours and was not able to accurately follow the pattern of load. The second model, MLPR model was found to predict the energy consumption better than the gradient boosting model. However, during peak hours or off-peak hours it was not able to form a pattern similar to the actual energy consumption. Linear regression model was better able to predict energy consumption as compared to the other two models. However, during low energy consumption hours the model sometimes predicted less energy consumption as compared to the actual.

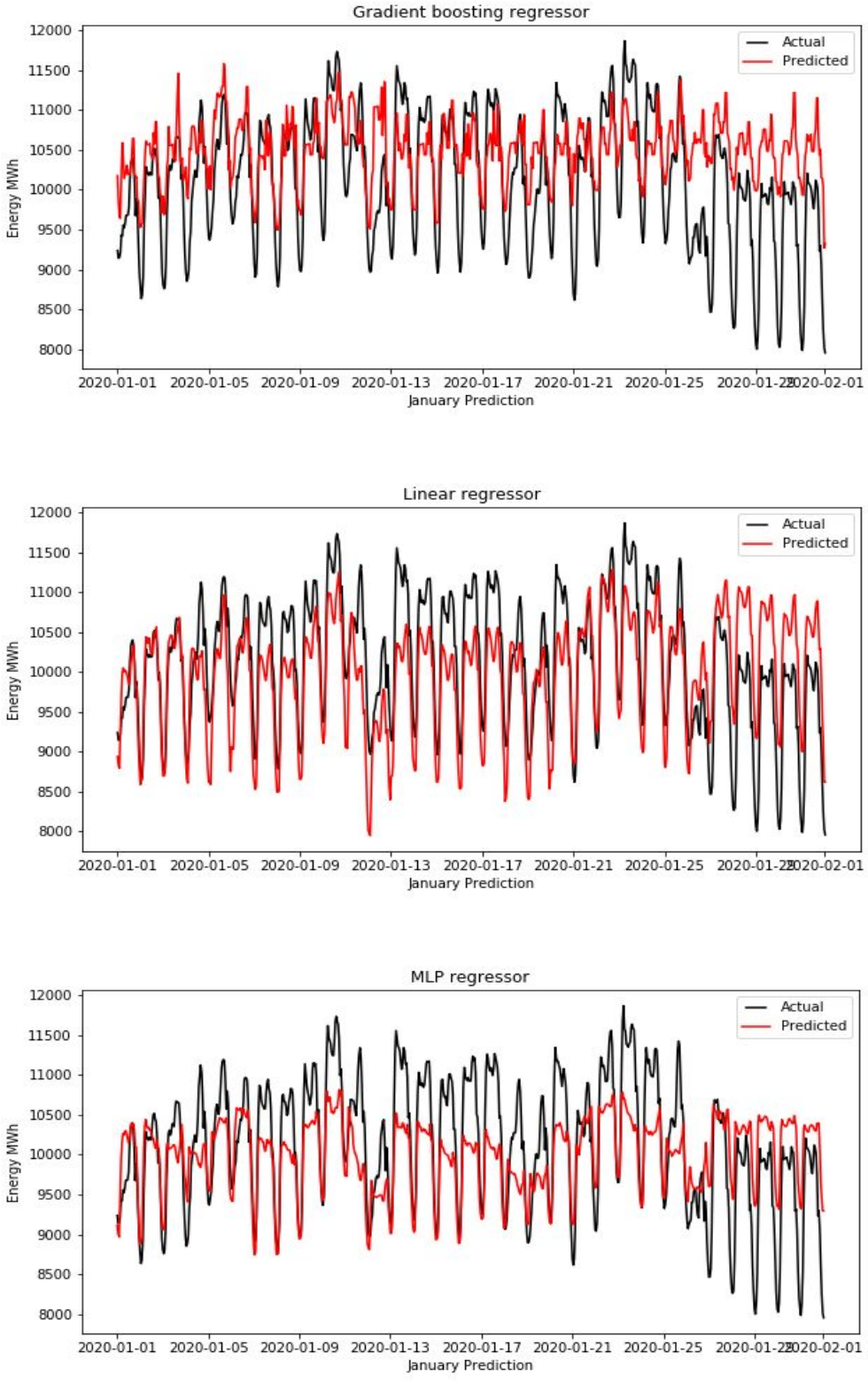

Figure 5: January 2020 energy consumption

3.Possible improvements

Improvements could be made especially to the weather data collection. Only weather data from Helsinki was used but since Helsinki is not in the middle of Finland and it doesn't represent the Finnish weather average the best. It might be possible to use weather from the largest cities and average all the weathers by giving coefficients based on population.

The model could also be improved by optimizing parameters, MLP regressor layers could be optimized or other activation functions could be tested to improve accuracy. Different regression methods could also be considered and also other algorithms such as evolutionary computation.

4.Conclusion

The study suggests that weather can have a certain effect on the energy consumption in an indirect manner. Low weather temperature or more snow can result in an increase in industrial and household load when heaters are turned on. Frequent visits to recreational places in winters such as saunas which operate by consuming more electricity can also be one of the reasons which causes peak in electricity demand.

It was assumed in this study that a relationship exists between weather and energy which could help in accurately forecasting energy. The weather data and other features were used to train three regression models that were able to predict energy consumption on an hourly basis. Further it was found that the linear regression model was better in forecasting energy for the next year (i.e. 2020) than gradient boosting and MLP regression models.

Bibliography

Zhao, J.,Gómez-Expósito, A., Netto, M., Mili, L., Abur, A., Terzija, V., . & Huang, Z. (2019). Power system dynamic state estimation: Motivations, definitions, methodologies, and future work. IEEE Transactions on Power Systems, 34(4), 3188-3198. *https://ieeexplore.ieee.org/abstract/document/8624411*

Boßmann, T., & Staffell, I. (2015). The shape of future electricity demand: Exploring load curves in 2050s Germany and Britain. Energy, 90, 1317-1333. *https://www.sciencedirect.com/science/article/abs/pii/S0360544215008385*

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117. *https://www.sciencedirect.com/science/article/abs/pii/S0893608014002135*

Choubin, B., Khalighi-Sigaroodi, S., Malekian, A., & Kişi, Ö. (2016). Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals. Hydrological Sciences Journal, 61(6), 1001-1009. *https://www.tandfonline.com/doi/full/10.1080/02626667.2014.966721*

Kumar, P., Ambekar, S., Kumar, M., & Roy, S. (2020). Analytical Statistics Techniques of Classification and Regression in Machine Learning. In Data Mining-Methods, Applications and Systems. IntechOpen. *https://www.intechopen.com/online-first/analytical-statistics-techniques-of-classification-and-regression-in-machine-learning*

Duffy, N., & Helmbold, D. (2002). Boosting methods for regression. Machine Learning, 47(2-3), 153-200. *https://link.springer.com/article/10.1023/A:1013685603443*

Finnish meteorological institute *https://en.ilmatieteenlaitos.fi/download-observations*

The European Network of Transmission System Operators for Electricity (ENTSO-E) *https://www.entsoe.eu/about/inside-entsoe/objectives/*