Electricity load forecasting with the use of machine learning and activity patterns

Filip Geppert

filip.geppert@gmail.com

Instituto Superior Técnico, Universidade de Lisboa, Lisboa

June 2018

Abstract: The goal of thesis is to create a model for electricity load forecasting with the use of machine learning and activity patterns. Two target values were defined: one-hour-ahead prediction and one-day-ahead. Thesis was carried out thanks to the courtesy of Watt IS company that shared all necessary data to create and validate proposed models. Three datasets containing electricity consumption were tested: two for blood clinics and one for restaurant. The behavioral analysis revealed that there are activity patterns such as daytime and nighttime variation in consumption or seasonal variation. Those patterns can be extracted by analyzing separately a particular dataset. With the use of information about holidays and applied statistical methods, datasets were enriched with additional information that were used as an input in the machine learning model. Moreover, Restaurant dataset was enriched with meteorological data. It was determined that thanks to the additional information, performance of the model increases.

Various machine learning models were implemented and tested, namely: random regression trees, gradient boosting regression and neural nets with LSTM cell, to predict one-hour-ahead electricity consumption on one-day-ahead target values. A performance comparison shows that gradient boosting regression resulted in the best coefficient of determination among all tested models. For the three different datasets, coefficient of determination in case of gradient boosting regression was equal to 0.91, 0.85 and 0.85. Implementation of machine learning models can have various business applications that optimize cost of grid utilization, such as implementation of demand-response models or off-grid intelligent controllers.

Key-words: machine learning, electricity demand forecasting, neural nets, statistical analysis.

I. Introduction

Electricity consumption forecasting has recently become an area of interest and therefore research and development among companies related to electricity (Distribution System Operators DSO, retailers, energy service companies, etc.) [1]. Over last years, there is an increasing interest in knowing how different buildings consume electricity with the lowest level of granularity. Moreover, the advancement in metering technologies and their decreasing cost has enabled a lot of consumers not only to start thinking of smart meter use, but also to optimize their electricity consumption based on the information that these meters collect. Apart from consumer awareness seen as one of benefits, electricity retailers started offering different electricity

consumption tariffs that encourage customers to optimize their usage. [2]

Dynamic tariffs, based on demand-response concept, are already being implemented and planned to be implemented in some markets [2], where consumers are offered to buy electricity per particular hour with the price adjusted to optimize the whole generation and transmission system. Moreover, the transition to the distributed energy generation based on intermittent (and hard-to-predict) energy sources such as photovoltaic panels or wind turbines combined with electricity storage systems has created a significant challenge in the market concerning the grid management [2]. In fact, the main European utilities have lost over $1 trillion in their market share because, with the unpredictability of the decentralized production, overproduction was occurring, forcing the grid to sell electricity at low prices or even paying to other consume that excess of electricity [3]. Therefore, the market has been forced to start thinking of the precise metering and forecasting systems in order to avoid the occurrence of such problems.

For example, photovoltaic panels that are installed on the roof require new operational strategies to optimize the production and guarantee that production is either covered by the demand or stored [2]. Delivering robust and reliable metering technology, enhanced by the forecasting part, has become one of the highest priority for many smart metering manufacturers.

Electricity forecasting enables to predict how the building is going to consume electricity over the short period of time in the future, starting from one hour ahead to even one week. Knowledge about future consumption can be very useful and would allow to anticipate a series of issues for both on-grid consumers, who are connected to the external grid, and off-grid ones who only use electricity produced by their own resources. However, based on the literature review, consumption forecasting on the low voltage level is more difficult than the forecasting on the high voltage one [4]. Households and SMEs change their electricity consumption very quickly when different devices are turned on and off. The load from a single consumer is much less predictable than one from a whole group of consumers, where averaging values very often yields accurate predictions and forecasts.

Electricity demand, especially at the small consumer scale, depends on different factors. Examples of such factors are: weather conditions, time of a day and socio-economic constrains. Weather conditions can influence greatly the consumption of the building especially during either heating or cooling seasons, when a lot of buildings use heating, ventilation and air conditioning (HVAC) systems. Time of a day is another indicator of high and low electricity consumption is and can give a lot of valuable information. The most difficult influencing parameter is the behavior of the consumer that cannot be generalized as every consumer uses the electricity in a different way. On the other hand, careful study of consumer's electricity behavior can give a lot of useful information that can improve significantly the forecasting ability.

To sum up, appropriate electricity forecasting is key to assure the optimal energy generation and consumption. Forecasting can not only improve the usage of the intermittent generation sources, but also optimize the utilization cost without losing usage comfort and improve the awareness of end users.

II. Research goal and questions

The research goal of the thesis is to develop a model that will accurately predict electricity consumption with different future time horizons. To do that, different machine learning algorithms were explored, and the model was trained and tested with the use of historical data coming from different type of buildings.

To answer this thesis goal, a set of research questions will be addressed:

- What are the cyclic patterns that can be observed form the data produced by different types of electricity consumers?

For this thesis, thanks to the courtesy of Watt-IS Company, different electricity consumption datasets were examined: clinic and restaurant datasets. The thesis will try to present the conclusions that can be made based on the analysis of different power curves and ultimately determine the difference between different electricity consumers.

- What are the strategies and methods that can improve the forecasting ability of the model?

This thesis will examine different machine learning methods and check their performance on different datasets with the ultimate goal to determine the best performing algorithm and the one that can be used in the industry to forecast electricity consumption.

- Does the same model perform equally on two different datasets?

This thesis compares different datasets on the same model and determines if the same model can be applied to two different sources of data, namely restaurants and clinics.

III. Methodology

In order to understand the behavior of data, certain set of actions were applied to the dataset with the goal to prepare it as a valuable input for the machine learning model. Ultimately, from the two features in the initial dataset (timestamp and power usage), the input for the machine learning model was extended and contained 36 features that are the effect of applied statistical operations. Additionally, input was enriched with external information such as meteorological data.

i. Description of the case studies:

Three case studies were analyzed:

1. Blood analysis clinic A.
2. Blood analysis clinic B.
3. Restaurant.

All three case studies included datasets consisted of two columns: timestamp – with the precise information when the reading was collected and power usage in Watts.

ii. Data transformation procedure

The following steps were implemented to explore the datasets and extract valuable information:

1. Addition of day, day of the week, month, year and hour information.
2. Calculation of macro values on the analyzed dataset: average, min, max and standard deviation.
3. Data aggregation by hourly time interval. Calculation of average power usage within this interval.
4. Calculation of energy usage based on the average power value.

5. Summary of data based on hourly electricity consumption.
6. Calculation of average electricity consumption in a particular day of the week.
7. Calculation of average electricity consumption per each month in the year.
8. For Restaurant dataset, information about meteorological data: temperature, pressure and wind speed were added. Meteorological data was downloaded from [5] and contains information gathered at Lisbon Airport station.
9. Addition of information regarding public holidays in Portugal over analyzed time.
10. Calculation of rolling averages, minimum and maximum values for the past values: 3, 6, 12 and 24 hours. Values were calculated considering power usage, temperature, wind speed and pressure.
11. Calculation of rolling sum values in the past for values: 3, 6, 12 and 24 hours. Values were calculated considering electricity consumption.
12. Selection of analyzed period of data: 1st January 2015 – 31st December 2017.

The result of the following procedure is a vector containing 36 features that was used as an input in the machine learning models.

Secondly, input dataset was investigated with the goal to determine patterns in electricity consumption. The following procedure was implemented:

1. Distribution of the dataset was plotted with regards to the targeted prediction value.
2. Correlation matrix based on Pearson Correlation coefficient was created.
3. Statistical metrics such as mean, standard deviation, maximum and minimum value were calculated on the dataset.
4. Dataset was grouped and plotted based on average usage within each hour during day.
5. Dataset was grouped and plotted based on day of week.
6. Dataset was grouped and plotted based on month in year.

iii. Machine learning algorithms

The following machine learning algorithms have been used:

- Gradient Boosting Regression Trees [6] - The main idea of the gradient boosting is to compute a sequence of very simple prediction trees, where each one is constructed based on the error of the previous one and tries to improve it.
- Random Forest Regression Trees [6] - Random Forest is a supervised learning algorithm from the group of decision trees, that based on small subsets of data, creates a whole group of trees by merging them to obtain an accurate final prediction. If a sufficient number of single trees is provided, algorithm is prevented from overfitting the original problem.
- Long short-term memory (LSTM) recurrent neural network [7] - LSTM are a special kind of Recurrent Neural Nets. They found to be very efficient with sequential data as LSTM nets can deal with vanishing gradient problems. In other words, the LSTM cell can take advantage of the

information that was taken place in the past and is not a current input to the model as an opposite to the regular neural net that takes only current input as an input information. LSTM cells have a chain structure, which means that the output weights from one cell is an input of the next one together with new input data.

iv. Metrics used in the validation of the machine learning algorithm

- Mean squared error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Di - Xi)^2 \quad \text{(I)}$$

n – number of observations, i – instance of observation, Di – true value for observation, Xi – predicted value for observation.

- Mean absolute error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Di - Xi| \quad \text{(II)}$$

n – number of observations, i – instance of observation, Di – true value for observation, Xi – predicted value for observation.

- Coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum(y_i - f_i)^2}{\sum(y_i - \bar{y})^2} \quad \text{(III)}$$

y_i – true value for observation, f_i – predicted value for observation, $\bar{y}$ – mean of the observed data.

IV. Results

Based on the exploratory analysis carried out on three different datasets, the following conclusions can be made:

1. Each building even when belonging to the same group (clinics) has different power curve distribution and should be analyzed separately in order to extract valuable information.
2. There is a clear difference in the nighttime and daytime power usage in all three analyzed datasets.
3. There is a clear difference in power usage between workdays (from Monday to Friday) and weekend days (Saturday and Sunday).
4. Presence of a seasonal cooling that takes place in July and August is a conclusion that can be made from the monthly power usage aggregation.

Figure 1 shows the correlation matrix that was calculated based on the Pearson correlation coefficient. One can notice that there is a clear correlation between electricity consumption and a target value which has a Pearson coefficient equal to 0.83.

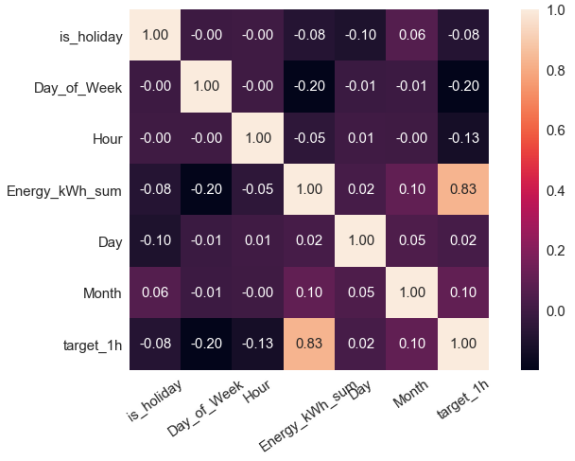

Figure 1 - Pearson correlation matrix for Clinic B case study.

Figure 2 shows the prediction of gradient boosting regression model trained on the Clinic B dataset.

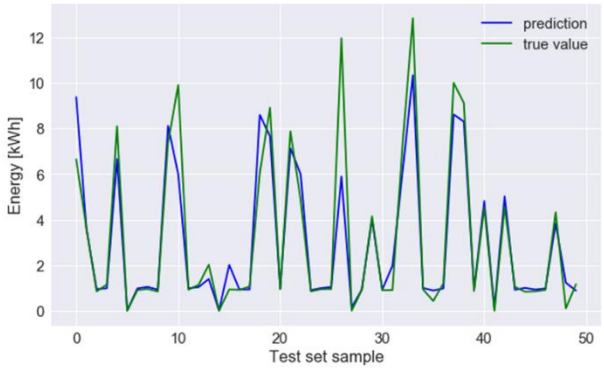

Figure 2 - Gradient Boosting Regression – Prediction for Clinic B Dataset

Table 1 shows numerical results of performance metrics used to evaluate predictive ability of the model for the Clynic B case study. Coefficient of determination greater than 0.85 is a proof of a good performance ability of the model.

	Training Set	Validation Set	Test Set
R^2 Score	0.97	0.88	0.85
Mean Absolute Error (MAE) [kWh]	0.35	0.61	0.67

Table 1 - Gradient Boosting Regression - Results Clinic B case study

Similarly, the procedure was carried out for the restaurant case study. Figure 3 shows the correlation matrix that was calculated based on the Pearson correlation coefficient. One can notice that there is a clear correlation between electricity consumption and a target value which has a Pearson coefficient equal to 0.61.

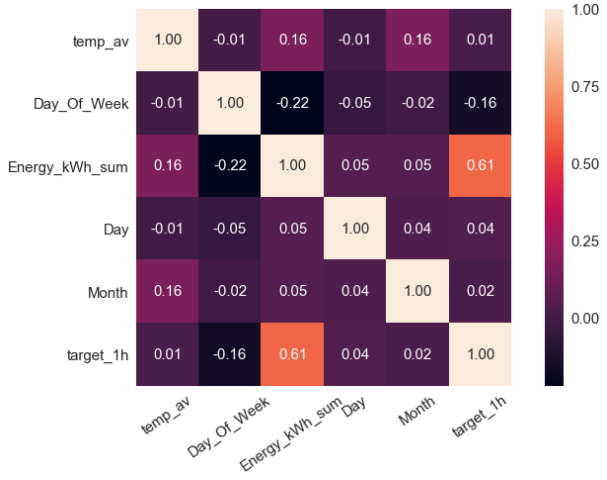

Figure 3 - Pearson correlation matrix for Restaurant case study.

Table 2 shows numerical results of performance metrics used to evaluate predictive ability of the model for the Restaurant case study. Coefficient of determination greater than 0.85 is a proof of a good performance ability of the model.

	Training Set	Validation Set	Test Set
R^2 Score	0.97	0.86	0.85
Mean Absolute Error (MAE) [kWh]	0.56	1.15	1.22

Table 2 – Gradient Boosting Regression – Results Restaurant Case Study

V. Discussion

i. What are the cyclic patterns that can be observed form the data produced by different types of electricity consumers?

- Decreased electricity consumption during weekend time in both clinics and restaurant dataset.
- Nighttime and daytime different electricity consumptions as well as seasonal ones.
- High correlation of electricity consumption with both one-hour-ahead and one-day ahead prediction values.

ii. What are the strategies and methods that can improve the forecasting ability of the model?

Addition of the meteorological data and information about holidays gives the model more information and based on the improved input, the forecasting ability of model increases. Moreover, with the use of rolling sums and averages based on the historical data, the information about past consumption is added as an input. Similarly, calculation of maximum and minimum values in a given historical period adds the quality to the model input and improve its performance on the test datasets.

Results from different models and with the use of different machine learning algorithms (Random Forest Regression, Gradient Boosting Regression and LSTM Neural Net) proved that good forecasting ability can be achieved on all three analyzed datasets with one-hour ahead prediction target. Results achieved on one-day ahed prediction dataset showed worse performance as compared with one-hour ahead prediction, but it is important to emphasize that the dataset that was used was smaller in size. Thus, learning ability was limited.

To sum up, it can be concluded that machine learning brings a quality to the forecasting of electricity consumption in both one-hour-ahead forecasting period and one-day- ahead one if input datasets are sufficiently large.

Does the same model perform equally on two different datasets?

	Blood Clinics A – test – R^2	Blood Clinics B – test – R^2	Restaurant – test – R^2
Random Forest Regression	0.798	0.82	0.77
Gradient Boosting Regression	0.905	0.85	0.85
Neural Net with LSTM cell	-	0.62	-

Table 3 - comparison of results for different datasets

Results provided in Table 3 show that the same model performs differently on two different datasets. With the high variance in the input data, the training part is crucial to achieve appropriate results and predictions on the dataset. For that reason, it can be stated that in order to achieve good forecasting ability, the training period cannot be omitted. It is a crucial part of the process. However, in every analyzed model, the biggest influence on the forecasting ability was made by value of electricity consumption sum in the past. Additionally,

the maximum value of power in the past period influenced the forecasting ability.

iii. What are the possible markets that could benefit from those algorithms?

Apart from the business application, there must be a market that is willing to adapt the solutions that are designed based on machine learning algorithms. For example, thanks to machine learning Swiss grid can successfully balance supply and demand of electricity [8] that flows from different countries. Before installation of machine learning based system, the losses from inaccurate load production were equal to 48.18 M Euro per year [8]. In [9], authors show how New York City power grid can benefit from implementation of machine learning based systems. In [10] author describes different applications of machine learning for the utility companies. There are several markets for which the need for adaptation was characterized and that are currently seeking for the innovative solutions in these areas. Those are:

Pay-as-you-go electricity market – market which is based on the model where customer pays for the amount of electricity that is used with no additional payments. This market is created by people that would like to pay for the exact amount of electricity that is used with no additional curtailed payments. Very often low price of electricity is a driver and these customers would like to decrease cost of their electricity bill. For that reason, a tool that would optimize the cost of electricity and maintain the usage comfort, is desired.

Electricity production clusters – defined as group of off-grid electricity producers that would like to take advantage of many small generation sources and decrease the risk of electricity shortage. Machine learning algorithms can help to predict what will be consumption and production from each source. Once the prediction is made, certain actions can be taken to appropriately manage energy flows among cluster uses.

Retail on-grid electricity market – constant optimization in electricity generation and distribution is a priority for large producers and retailers. For that reason, use of machine learning that would produce information about future system demand could decrease the losses that are the effect of a mismatch between demand and supply of electricity in the distribution grid.

Off-grid systems – apart from off-grid solutions that very often are consisted of: renewable electricity generator and accumulation system, the controller could be enabled with machine learning algorithm that would be responsible for electricity flows among consumption, production and energy accumulation.

There are several benefits that can be an effect of machine learning based forecasting systems. Those are:
Reduction of electricity cost – forecasting can decrease the cost of electricity usage with the same comfort being maintained. Stable energy usage is the most optimal and the cheapest one and it can be achieved by shifting peak consumptions to periods when consumption is low.
Increased usage of renewables – some national power grids face the problem of electricity over production when renewables produce electricity, but there is no demand for those in particular moment. If one had information about demand earlier, certain actions could be taken to prevent the situation when electricity that is not used.

VI. Conclusions

The goal of the thesis was to show applicability of various machine learning algorithms in small and medium buildings. Three different case studies were investigated. Two coming from clinic buildings and one coming from a restaurant building. The datasets were provided with the courtesy of Watt-IS Company and consisted of power curves for two-years with one-minute readings. The exploratory analysis of the dataset showed that there are certain patterns that can be noticed through the analysis of the power curve.

For Clinic A case study:

- Nighttime and daytime periods of consumption,
- Weekday and weekend periods of consumption.

For Clinic B case study:

- Nighttime, low and high daytime periods of consumption,
- Weekday and weekend periods of consumption.

For Restaurant case study:

- Nighttime and daytime periods of consumption,
- Weekday and weekend periods of consumption,
- Cooling season period of consumption.

It can be concluded that patterns may vary significantly depending on the dataset, so to determine the consumption patterns each instance must be investigated carefully.

With the use of information about holidays and statistical methods, Clinic A and B as well as Restaurant datasets were enriched with additional information that were used as an input in the machine learning model. Moreover, Restaurant dataset was enriched with meteorological data. It was determined that thanks to the additional information, performance of the model increases. With the use of Random Forest Regression Trees, Gradient Boosting Regression Trees and Neural Nets with LSTM cell as machine learning algorithms, the models for forecasting energy consumption for one-hour-ahead and one-day-ahead values were build. The coefficient of determination for the three different datasets was on average equal to 0.81 (with the maximum theoretical value equal to 1). The most effective algorithm was Gradient Boosting Regression Forest with the coefficient of determination for three different datasets equal to 0.91, 0.85 and 0.85 and mean absolute error equal to 0.374 kWh, 0.65 kWh and 1.22 kWh respectively for clinic A, clinic B and Restaurant datasets.

Various business applications of the forecasting solution designed with machine learning algorithms were defined. Solution can be implemented in many areas such as: grid load optimization or maximization of electricity generation and a lot of different users could take advantage from the use of forecasting models. Solution can be implemented for example in demand-response models implemented by electricity retailers or smart controllers installed by off-grid consumers on-site. The ultimate goals that a more effective forecasting solution yields are: the decrease of the electricity usage costs and fostering the penetration of renewable energy sources in the grid.

VII. Bibliography

[1] K. Gajowniczek and T. Zabkowski, “Electricity forecasting on the individual household level enhanced based on activity patterns,” *PLoS One*, vol. 12, no. 4, pp. 1–26, 2017.

[2] Y. Oualmakran, J. M. Espeche, M. Sisinni, T. Messervey, and Z. Lennard, “Residential Electricity Tariffs in Europe : Current Situation , Evolution and Impact on Residential Flexibility Markets †,” no. November 2016, pp. 1–5, 2017.

[3] The Economist, “How to lose half a trillion euros - European utilities.” [Online]. Available: https://www.economist.com/news/briefing/21587782-europes-electricity-providers-face-existential-threat-how-lose-half-trillion-euros. [Accessed: 04-May-2018].

[4] B. Yildiz, J. I. Bilbao, and A. B. Sproul, “A review and analysis of regression and machine learning models on commercial building electricity load forecasting,” *Renew. Sustain. Energy Rev.*, vol. 73, no. December 2016, pp. 1104–1122, 2017.

[5] tempo.pt, “Histórico do tempo para Lisboa - tempo.pt.” [Online]. Available: https://www.tempo.pt/lisboa-sactual.htm. [Accessed: 25-Apr-2018].

[6] S. Raschka, *Python Machine Learning Unlock*, vol. 22, no. 2. 2015.

[7] C. Olah, “Understanding LSTM Networks -- colah’s blog,” 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/. [Accessed: 25-Apr-2018].

[8] Siemens, “Power Transmission: Forecasting Electricity Demand - Energy & Efficiency - Pictures of the Future - Innovation - Home - Siemens Global Website.” [Online]. Available: https://www.siemens.com/innovation/en/home/pictures-of-the-future/energy-and-efficiency/power-transmission-forecasting-electricity-demand.html. [Accessed: 13-May-2018].

[9] C. Rudin *et al.*, “Machine Learning for the New York City Power Grid,” vol. 34, no. 2, pp. 328–345, 2012.

[10] Ben Packer, “7 reasons why utilities should be using machine learning | Oracle Utilities Blog.” [Online]. Available: https://blogs.oracle.com/utilities/utilities-machine-learning. [Accessed: 13-May-2018].