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

Filip Geppert

Thesis to obtain the Master of Science Degree in

Energy Engineering and Management

Supervisors: Prof. Carlos Augusto Santos Silva

Dr. Henrique Ramalho Monteiro Latourrette Pombeiro

Examination Committee

Chairperson: Prof. Duarte de Mesquita e Sousa

Supervisor: Prof. Carlos Augusto Santos Silva

Member of the Committee: Prof. Susana Margarida da Silva Vieira

June 2018

Acknowledgments

Author would like to thank the following parties for the participation in thesis work:

- Watt IS Company – for sharing datasets and providing technical support,
- InnoEnergy Master School – for the whole Master School program without which this thesis wouldn’t be created.

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.

Resumo

O objetivo desta tese é criar um modelo para predição do consumo de eletricidade utilizando algoritmos de machine learning e análise dos padrões de atividade. Para dois contextos diferentes: predição para a próxima hora e predição para o próximo dia. Os dados dos casos de estudo foram fornecidos pela empresa Watt-IS e incluíam apenas o consumo de energia para duas clínicas de análises de sangue e um restaurante. A análise exploratória dos dados revelou que existem padrões específicos de atividade, como variações entre o dia e a noite e variações sazonais. Estes padrões de atividade foram extraídos de cada um dos casos de estudo, como por exemplo informação sobre feriados ou dias da semana. Desta forma os dados dos casos de estudo foram enriquecidos com esta informação para a aplicação dos métodos machine learning. O caso do restaurante foi ainda enriquecido com dados meteorológicos. Em geral, a utilização desta informação adicional sobre os padrões de atividade melhora os modelos.

Diferentes métodos de machine learning foram desenvolvidos e testados, em particular: random regression trees, gradient boosting regression and neural networks with LSTM, para prever o consumo de eletricidade na próxima hora e no próximo dia. A comparação dos modelos demonstrou que o método de gradient boosting regression teve os melhores resultados entre todos os modelos. A implementação destes modelos de machine learning pode ter várias aplicações comerciais para ter a otimização da operação da rede elétrica, como a implementação de modelos de gestão de procura.

Palavras-chave: machine learning, previsão de demanda de eletricidade, redes neurais, análise estatística.

Table of contents

Acknowledgments ii

Abstract iii

1. Introduction 1

1.1. Problem definition 1

1.2. Research goal 2

1.3. Research questions 3

1.4. Structure of the document 3

2. Literature review 4

3. Methodology 9

3.1. Exploratory data analysis 9

3.1.1. Description of the analyzed case studies 9

3.1.2. Data transformation procedure 9

3.2. Machine learning algorithms 10

3.2.1. Gradient Boosting Regression Trees 10

3.2.2. Random Forest Regression Trees 11

3.2.3. Long short-term memory (LSTM) recurrent neural network 11

3.3. Metrics used in the validation of the machine learning algorithm 12

3.3.1. Mean squared error (MSE) 12

3.3.2. Mean absolute error (MAE) 13

3.3.3 Coefficient of determination (R^2) 13

3.3.3. Pearson correlation coefficient 13

4. Results & Discussion 14

4.1. Exploratory analysis 14

4.1.1. Blood analysis clinic A dataset 14

4.1.2. Blood analysis clinic B Clinics dataset 17

4.1.3. Restaurant dataset 21

4.2. Random forest regression 24

4.2.1. Blood analysis clinic A Dataset 24

4.2.2. Blood analysis clinic B Dataset 24

4.2.3. Restaurant Dataset 25
4.3. Gradient boosting regression 26
4.3.1. Blood analysis clinic A Dataset 26
4.3.2. Blood analysis clinic B Dataset 26
4.3.3. Restaurant Dataset 27
4.4. RNN with LSTM cell 27
4.4.1. Blood analysis clinic B Dataset 28
4.5. Different resolution of predicted values (one hour ahead, one day ahead) 30
4.5.1. Dataset description 30
4.5.2. Random Forest Regression – Results. 31
4.5.3. Gradient Boosting Regression - Results. 32
4.6. Discussion 34
4.6.1. What are the cyclic patterns that can be observed form the data produced by different types of electricity consumers? 34
4.6.2. What are the strategies and methods that can improve the forecasting ability of the model? 34
4.6.3. Does the same model perform equally on two different datasets? 35
5. Business application of the solution 36
5.1. How the solution can be implemented 36
5.2. What are the possible markets that could benefit from those algorithms 37
5.3. Potential benefits for the user 38
6. Conclusions 39
7. References 40

List of tables

Table 1 - Literature Review 6
Table 2 - macro values calculated on Blood analysis clinic A dataset. 15
Table 3 – macro value calculated on Blood analysis clinic B dataset. 18
Table 4 - macro values calculated on Restaurant dataset. 22
Table 5 - grid search parameters 24
Table 6 – random forest regression. Hyperparameters chosen for the model. 24
Table 7 - Random Forest Regression. Results - Blood analysis clinic A. 24
Table 8 - Random Forest Regression. Results – Blood analysis clinic B.. 25
Table 9 - Random Forest Regression. Results - Restaurant.. 25
Table 10 - grid search parameters for gradient boosting regression. 26
Table 11 - Gradient boosting regression. Hyperparameters chosen for the model 26
Table 12 - Gradient Boosting Regression. Results - Blood analysis clinic A. 26
Table 13 - Gradient Boosting Regression. Results – Blood analysis clinic B. 26
Table 14 – Gradient Boosting Regression. Results – Restaurant. 27
Table 15 – grid search parameters defined for RNN with LSTM cell. 27
Table 16 - Neural Net with LSTM cell - used hyperparameters. 27
Table 17 – RNN with LSTM cell. Blood analysis clinic B - Results. 28
Table 18 - one day ahead prediction - model hyperparameters. 31
Table 19 - Random Forest Regression - Results. 31
Table 20 - Gradient Boosting Regression - hyperparameters. 32
Table 21 - Gradient Boosting Regression - Results. 32
Table 22 - comparison of results for different datasets 35
Table 23 - possible solution implementation areas. 37

List of figures:

Figure 1 - LSTM neural net schema. Source: [17]... 12
Figure 2 - Blood analysis clinic A - dataset distribution based on crucial parameters. 14
Figure 3 - Blood analysis clinic A - Pearson correlation matrix. 15
Figure 4 - average power distribution in each hour during day calculated on Blood analysis clinic A dataset. ... 16
Figure 5 - average power usage in each day of week calculated on Blood analysis clinic A dataset. 16
Figure 6 - average power usage in each month of a year calculated on Blood analysis clinic A dataset. 17
Figure 7 – Blood analysis clinic B - dataset distribution based on crucial parameters. 17
Figure 8 - correlation matrix - Blood analysis clinic B dataset 18
Figure 9 - average power usage in each hour of a day calculated on the Blood analysis clinic B dataset. 19
Figure 10 - average power usage in each month calculated on Blood analysis clinic B dataset. 19
Figure 11 - average power usage in each day of a week calculated on Blood analysis clinic B dataset. 20
Figure 12 – Restaurant - dataset distribution based on crucial parameters. 21

Figure 13 - Restaurant - Pearson correlation matrix. 21
Figure 14 - average power usage in each hour of a day calculated on Restaurant dataset. 22
Figure 15 - average power usage in each day of a week calculated on Restaurant dataset. 23
Figure 16 - average power usage in each month calculated on Restaurant dataset. 23
Figure 17 - Random Forest Regression Results - Blood analysis clinic B dataset 25
Figure 18 - Gradient Boosting Regression Results - Blood analysis clinic B Dataset 27
Figure 19 - loss function distribution over training process. 28
Figure 20 - LSTM net. test set results. 29
Figure 21 - Blood analysis clinic B - one day data groupings and their distribution. 30
Figure 22 - Blood analysis clinic B dataset - 1 day ahead prediction - Pearson correlation matrix. 31
Figure 23 - Random Forest Regression – 1 day ahead prediction 32
Figure 24 - Gradient Boosting Regression - 1 day ahead prediction. 33

1. Introduction

1.1. Problem definition

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. Ultimately, forecasting can significantly contribute to the goals set by EU and help to reduce CO_2 emission to the atmosphere.

1.2. Research goal

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 will be explored, and the model will be trained and tested with the use of historical data coming from different type of buildings.

The requirements for the model are the following:

- Model improvement achieved by data exploratory analysis – based on the available data, crucial information is extracted and is given as an input to the predictive model;
- Electricity prediction – determination of the model performance measured by mean absolute error, mean squared error and coefficient of determination;
- Response time – based on the model latency, the ability of the model to be used on real-time data is determined. Additionally, the computational cost is presented to estimate potential business usability.

1.3. Research questions

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.

1.4. Structure of the document

In the Introduction section, where the problem and research goals and questions have been defined, the document will follow with the Literature review section, where various works in the field of electricity load forecasting are investigated. These references lead to the choice of three algorithms to model energy demand in our experimental cases: random forest regression, gradient boosting regression and neural network with long-short-term-memory cell.

In the Methodology section, the thesis methodology is presented. The selected algorithms are described in detail from the perspective of the mechanisms and logic behind their formulation. Additionally, the procedure for exploratory input dataset analysis is proposed. Finally, we present and describe in detail the validation metrics for assessing if the proposed algorithms work well on our dataset.

In the Results & Discussion section, the results of model application are presented. Different models were tested with the use of the three datasets. Results are presented in both numerical and graphical form. Different performance metrics were calculated on the prediction results with the goal to determine the applicability of the algorithms on real-time data. The results obtained in this section allow us to respond to the research questions of this thesis.

In the Business application of the solution section, potential business applications for the proposed methods are presented. The section describes the ways of possible implementation as well as markets where this application of the methods could generate positive value for people.

In the Conclusion section, this thesis is summarized, and key conclusions are made based on the work that has been carried out.

2. Literature review

This section presents an analysis of different scientific papers that refer to the thesis goals. The scope of analysis in the papers often covers only partially the scope of the thesis. For that reason, a careful investigation of those can help significantly to improve further implementation of models and contribute to better results than those achieved in this thesis. The prediction target in case of all papers was set to 24h ahead. The most important conclusions from the papers are:

- With the decrease of the system size, the complexity of the problem increases. It is more difficult to develop models that will predict electricity consumption for a small building than the model for the whole power system.
- Various machine learning algorithms can improve forecasting ability and contribute to the monetary saving for the end customer even on the scale of medium size buildings. On the other hand, those algorithms need to be carefully selected as different buildings have different electricity consumptions.
- The problem of electricity forecasting in small and medium buildings is not yet solved. Various algorithms are proposed, but the achieved results show that there is a space for improvement.
- Based on the papers that were studied, the best forecasting ability was achieved with the use of regression tree, support vector regression and neural net models.

Table 1 gathers all the literature that was analyzed in the area similar to the scope of thesis.

Name	**Author**	**Year**	**Inputs for the model**	**Model**	**Target**	**Dataset**
Electricity forecasting on the individual household level enhanced based on activity patterns [1]	Krzysztof Gajowniczek, Tomasz Zabkowski	2017	Time, Load, Temperature, Humidity	3 ML Techniques were compared with standard measurement models.	2 step analyses: 1. Determination of the behavioral patterns, 2. Use of ML to determine 24 load values.	WikiEnergy Dataset.
Two-Stage Electricity Demand Modeling Using Machine Learning Algorithms [5]	Krzysztof Gajowniczek, Tomasz Zabkowski	2017	Power Curve, Time, Day, Schedule	1step: Classification of peak values 2nd step: Comparison of different machine learning	One Day ahead load prediction for the electricity system	Historical dataset from around 10 years.

				models: RF, Neural Net, SVM		
Short-term residential load forecasting: Impact of calendar effects and forecast granularity [6]	Peter Lusisa, Kaveh Rajab Khalilpoura, Lachlan Andrewa, Ariel Liebmana	2017	Load data	Comparison of different machine learning models	Check the influence of the parameters on the final prediction	3-year household load from Australia.
DR-Advisor: A data-driven demand response recommender system [7]	Madhur Behl, Francesco Smarra, Rahul Mangharam	2016	Temperature, Humidity, Wind, Time, Day, Schedule, Chilled Water Temp, Zone Temp, Power	Family of Regression Trees, mbCRT algorithm	Real-time Estimate of Power, The best control action, Optimization of the controllable parameters	University of Pensylvania buildings (different types)
Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques [8]	Sergio Jurado, Angela Nebot, Fransisco Mugica, Narcís Avellana	2015	Power Curve, Time, Day, Schedule	Comparison of different models: FIR, Random Forests, Neural Net	Hourly One day ahead load prediction	Restaurant, Library, Office at UPC in Catalonia

A review and analysis of regression and machine learning models on commercial building electricity load forecasting [4]	B. Yildiz, J.I. Bilbao, A.B. Sproul	2017	Temperature, Time, Day, Schedule, Power	Comparison of different machine learning models: Regression Tree, Neural Net, SVR	Hourly One day ahead load prediction	Different University buildings
A short-term power load forecasting model based on the generalized regression neural network with decreasing step fruit fly optimization algorithm [9]	Rui Hua, Shiping Wena, Zhigang Zenga, Tingwen Huangc	2017	Power Curve, Time, Day, Schedule	Neural Net with the stress on the optimization algorithm	One Day ahead load prediction for the electricity system	
Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-Metering Devices [10]	Yangyang Fua, Zhengwei Lia, Hao Zhangb, Peng Xua	2015				Office building.

Table 1 - Literature Review

Subsequent sections contain detailed description of five papers, four focusing on the electricity forecast for small or medium buildings and the last focusing on the electricity forecast for the whole power system. Those five papers were defined as ones that have the biggest impact on the thesis and can provide a lot of valuable information.

In the [1] authors focus on the development of a model that will forecast the electricity consumption for 24 ahead at the individual household level. The authors have used a very large dataset that covered the consumptions from particular devices, as well as aggregated one. Different forecasting algorithms were tested, and input data was improved by the household activity patterns. The authors have achieved mean absolute percentage errors (MAPE) equal to 41.7% and 23.64% in the case of the best performing model for two different datasets. Additionally, clustering and sequence recognition algorithms have proved to be good tools for identification of patterns in behavior. General trends can be caught, especially taking into account the hours when devices were turned on and off. Moreover, the forecasting ability can be improved when behavioral data is added. The authors showed that addition of such data as new input features reduced MAPE by 8.5% as compared to the model without those features. Finally, the authors used support vector regression, neural nets, random forest regression. Those algorithms yielded better results as compared to the contemporary models. Neural nets proved to be the best tool for forecasting.

In [6] the authors investigate how calendar effects, granularity of forecasting and size of the training data affect the performance of the forecasting model. The goal of the model is to predict electricity consumption for 24 hours ahead. The authors have tested several machine learning techniques for prediction: regression trees, neural networks and support vector regressions. The dataset that was used covered historical load profiles, weather data and influence of calendar effects such as holidays. Authors discovered that one year of historical data used as a training set for machine learning model is sufficient to develop a well-performing prediction model. The increase of the dataset size had marginal effect on the improvement in results. Additionally, calendar effects and load forecasting granularity improve the performance and predictive ability of the 24h ahead model. Calendar effects were presented as an input in the form of binary variables. Moreover, calendar effects have bigger effect on the forecasting ability when historical consumption data is not available. Finally, regression tree model and support vector regression had the best performance from all tested models.

In [4] the authors present the review of different load forecasting models. Predictive abilities of models are compared with the use of University of New South Wales campus data. Different models were tested to predict daily peak in the electricity demand. The authors confirmed that peak forecasting turned out to be more difficult task than prediction of electricity demand for a 24h horizon. Moreover, forecasting electricity consumption of a single building was acknowledged to be more difficult task than forecasting the overall load of University campus. Finally, the selection of the appropriate model is very time consuming and iterative process, yet the effort can bring significant improvements in the overall forecasting ability.

In [7] the authors developed a demand response model for large commercial buildings was developed with the use of regression tress. The data-driven control system developed by the authors had outperformed a regular system and allowed to save € 45k. The tool developed by authors – DR-Advisor - had achieved prediction accuracy greater than 93% tested on 8 buildings at the Penn's campus. The authors integrated different machine learning algorithms in one system what improved the general forecasting ability of the system. Moreover, the authors showed that use of real time machine learning based algorithms combined with elastic electricity tariff systems can generate significant potential savings for the customer. Finally, regression tree models had proved to be to best performing machine learning algorithm out of all that were studied.

In [5] the authors developed a machine learning model for the national power system for the 24h ahead electricity demand prediction. Dataset used in the exercise consisted of hourly electricity load in the Polish Power System between 1 January 2008 and 31 December 2015. The dataset was improved with the weather data that contained: temperature, humidity, sunrise time and sunset time. Moreover, authors investigated a two-stage approach: firstly, predicting the demand peak with the use of classification models and secondly by predicting the demand value. Different algorithms are tested both for classification and regression problem. Ultimately, the influence of various input features was investigated. The authors showed that it is possible to accurately predict peak load up to 24h ahead and this task can be done with high precision. The model developed by authors had reached the accuracy equal to 91%. Additionally, load forecasting models had reached very low errors, the lowest equal to 2.7% in case of artificial natural nets. Moreover, addition of the information from peak classification model and weather data as an input to the load forecasting model had improved forecasting capability. Finally, machine learning algorithms had performed better than the benchmarking model used for comparison. The best performing algorithm for classification problem was a support vector machine (SVM) algorithm, whereas for regression problem artificial neural networks (ANN) turned out to have the best forecasting ability.

Based on the literate review that was performed, three machine learning methods were chosen to test in the thesis. Those methods are the following:

- Regression trees – as those were the best performing models in the [7], [6] and [1].
- Neural nets – as those were the best performing models in the [5] and [1].

3. Methodology

In order to appropriately create and validate the models that were created for the purpose of this thesis, this methodology section describes all applied procedures in detail, presenting the development of all algorithms used in this thesis work as well as the validation metrics that were used on the models. The algorithms that were chosen are the ones that performed the best in the papers analyzed in the literature review combined with the best performing contemporary machine learning models.

All the analysis was carried out with the use of Python as a programming language. The following libraries were used in various scripts:

- Pandas [11] – open source library providing high performance data analysis tools.
- Matplotlib [12]– open source library for data visualization.
- Sckitlearn [13]– open source machine learning library.
- Keras [14] – open source neural networks library.

3.1. Exploratory data analysis

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.

3.1.1. Description of the analyzed case studies

Three case studies were analyzed:

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

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

3.1.2. Data transformation procedure

Firstly, datasets were cleaned, and missing values (representing less than 2% of the total dataset) were replaced with the arithmetic means from the month where the missing value was located. Once the datasets were organized and ready to work with, the set of operations presented below was implemented.

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

1. Addition of day, month, year, day of the week and hour information. This information was derived from the timestamp included in the input dataset.
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 the Restaurant dataset, information about meteorological data: temperature, pressure and wind speed were added. Meteorological data was downloaded from [15] 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.

3.2. Machine learning algorithms

The following machine learning algorithms have been used based on the conclusions made in Literature review section: Gradient Boosting Regression Trees, Random Forest Regression Trees and Long short-term memory (LSTM) recurrent neural network. Additionally, goals set in Research goal section were considered during algorithm selection process. The details of each method are given in the following sub-sections. Hyperparameters chosen for the model were defined through the exhaustive grid search which means that from the set of different values defined for one parameter, all combinations of different parameters are tested. The best combination is characterized with the lowest error. The values for parameters with the lowest error are used in the model.

3.2.1. Gradient Boosting Regression Trees

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.

The algorithm involves three main parts [16]:

1. Definition of the loss function that is optimized during models training process,
    a. Loss function must be differentiable,
    b. In regression problems, mean squared error is used.
2. Creation of weak learner to make predictions,
    a. Regression decision tree is used as a weak learner,
    b. Depth of decision tree is set to be equal to 5 to avoid overfitting problem.
3. Addition of series of weak learners to minimize loss function.

a. Gradient descent procedure is used to minimize loss when adding trees – after calculating error or loss in the function, coefficients in a regression equation are updated to minimize that error.
b. After coefficients are updated for each tree, all outputs are added together with an effort to improve the final output of the model.

3.2.2. Random Forest Regression Trees

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. The disadvantage of the algorithm is that some single trees can be trained on data that is very specific and does not give a good perspective on the overall structure of the dataset. For that reason, it is important to assure that enough time is spent to carefully adjust number of trees in the forest and hyperparameters of those such as: depth of a single tree and criteria for creating a split in the tree.

The algorithm involves four main parts [16]:

1. From the whole dataset, a random sample of data is taken.
2. A single decision tree is trained on the random sample with the use of the most optimal split to predict the target value.
3. Operations 1 and 2 are repeated n times, where n stands for number of trees that are predefined.
4. Once the test data is given, each tree comes with the prediction and prediction that was forecasted by the majority of single trees is taken as a model one.

3.2.3. Long short-term memory (LSTM) recurrent neural network

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. The crucial element of the LSTM cell is a cell state that is represented by the horizontal line on the top of the diagram presented in Figure 1. The LSTM adds or removes an information to the cell state with the use of the activation functions that can be either sigmoid or tanh one.

The algorithm (at state t) involves the following steps [17]:

1. Output weights from the previous cell (represented by t-1 in the Figure 1) together with the cell state and new input are delivered as input parameters to the cell (represented by t in the Figure 1).
2. Based on the new information and forgetting factor that are both controlled by different activation function, the cell state is updated.
3. Based on the new cell state, a prediction value is generated.

4. Cell state and weights are passed as an input to the next cell (represented by t+1 in the Figure 1).
5. A series of cells forms a LSTM network that can either output the prediction value at every step like it is presented in the Figure 1 (many to many network) or output a single value at the end of the network (many to one network).

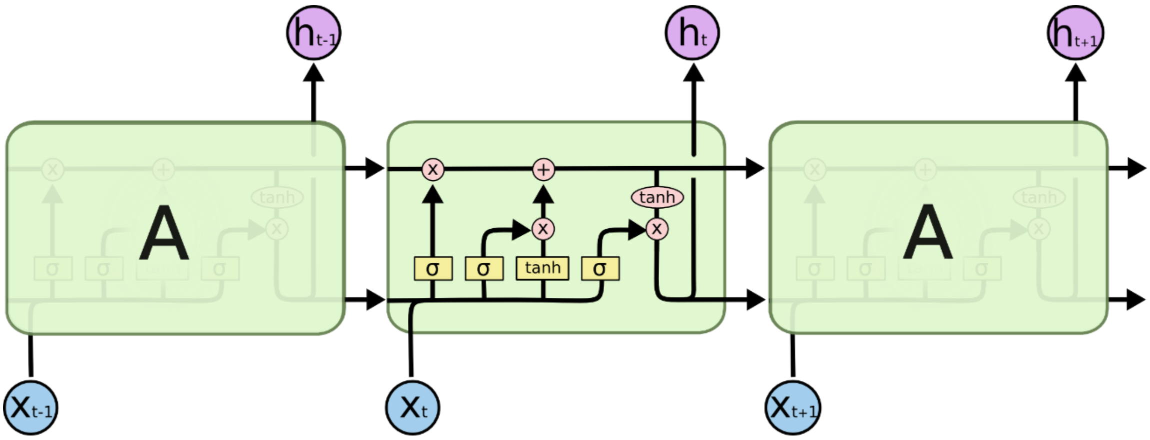

Figure 1 - LSTM neural net schema. Source: [17]

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

The following metrics were used in the thesis as either validation metric that shows what is a performance of the algorithm, or as loss metric that is used as an optimization function that is differentiated during the learning period.

3.3.1. Mean squared error (MSE)

The MSE is calculated using the formula:

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

where:

n – number of observations, i – instance of observation,

Di – true value for observation, Xi – predicted value for observation.

3.3.2. Mean absolute error (MAE)

The MAE is calculated using the formula:

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

where:

n – number of observations, i – instance of observation,

Di – true value for observation, Xi – predicted value for observation.

3.3.3 Coefficient of determination (R^2)

R^2 is calculated using the formula:

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

where:

y_i – true value for observation, f_i –predicted value for observation.

$\bar{y}$ – mean of the observed data.

3.3.3. Pearson correlation coefficient

It is a measure of a linear correlation between two variables in a sample. It varies from 1 to -1 where:

- 1 means a positive correlation,
- 0 means no linear correlation,
- -1 means a negative correlation.

Pearson correlation coefficient is calculated using the formula [18]:

$$\rho_{X,Y} = \frac{covariance(X,Y)}{\sigma_X \sigma_Y} \qquad \text{IV}$$

where:

X – variable X in sample, Y – variable Y in sample.

σ – standard deviation.

4. Results & Discussion

4.1. Exploratory analysis

Three datasets were examined with the use of the use of the methodology described in Data transformation procedure section. The subsequent sections contain results of the analysis. Finally, an attempt is made to conclude common features that can be used as an input for the machine learning model and improve its general forecasting ability.

4.1.1. Blood analysis clinic A dataset

Figure 2 shows the behavior of the data that is included in the dataset. Target value presented on the vertical axis is a value of electricity consumption for 1 hour ahead prediction.

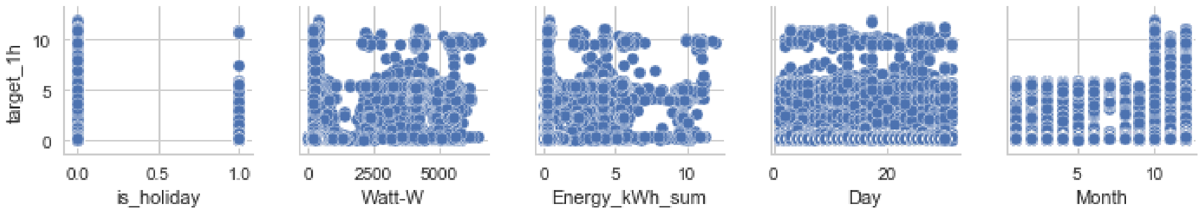

Figure 2 - Blood analysis clinic A - dataset distribution based on crucial parameters.

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 both electricity consumption and power, and a target value which has a Pearson coefficient equal to 0.56 and 0.61. This correlation means that with the increase of electricity consumption the value of electricity prediction is increasing as well. Other values that have Pearson correlation coefficient close to 0 mean that there is small influence of those parameters on the target value. The reason for this may be caused by the fact that in this case one-hour ahead value is predicted and if the prediction window was wider, there is a chance that those parameters would increase their influence.

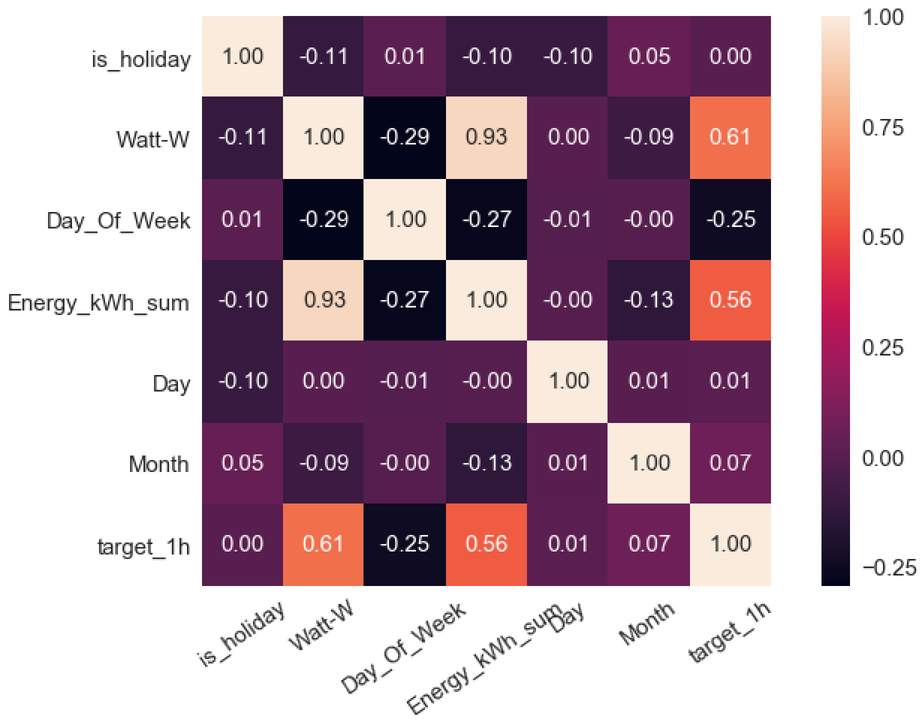

Figure 3 - Blood analysis clinic A - Pearson correlation matrix.

Table 2 presents the summary of the macro values calculated for the Blood analysis clinic A dataset. One can notice that standard deviation is higher than average, which means that dataset is very diverse and contains a lot of different points that have significantly different values. Moreover, the maximum value that is around seven times higher than average one, which is another indicator of how diverse the dataset is.

Table 2 - macro values calculated on Blood analysis clinic A dataset.

Parameter	Value
Average	1450 W
Standard deviation	2068 W
Min	0 W
Max	8417 W

Figure 4 shows the average power usage taking into account only the hour of a day. One can notice that electricity consumption varies significantly between daytime and nighttime. Two periods of electricity consumption can be distinguished: nighttime when the average consumption is equal to 200W (probably it accounts for base consumption of the building) and daytime when average consumption is equal to 2700 W.

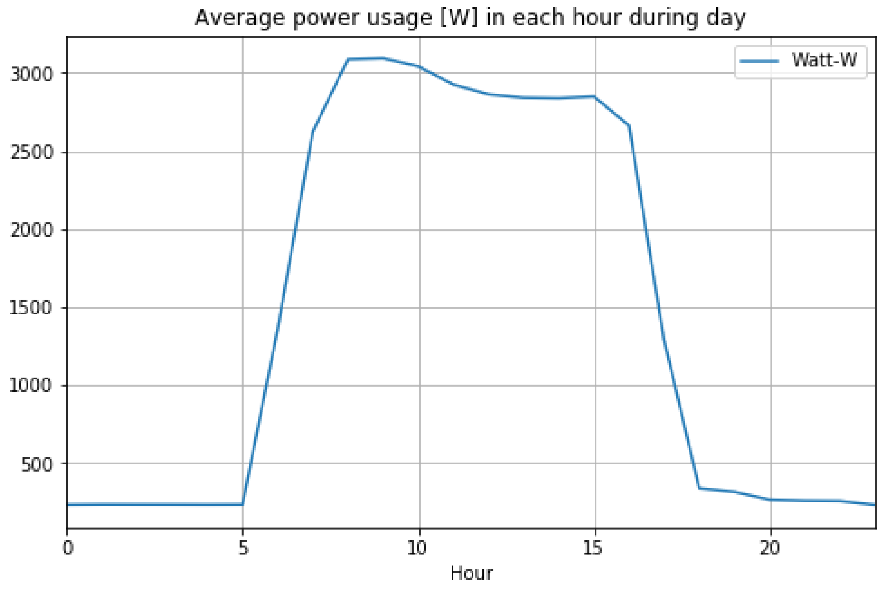

Figure 4 - average power distribution in each hour during day calculated on Blood analysis clinic A dataset.

Figure 5 shows an average power usage in each day of the week. There is a huge difference between workdays and weekend in average power consumption. The lowest power usage takes on average place on Sunday, whereas the highest on takes place on Wednesday.

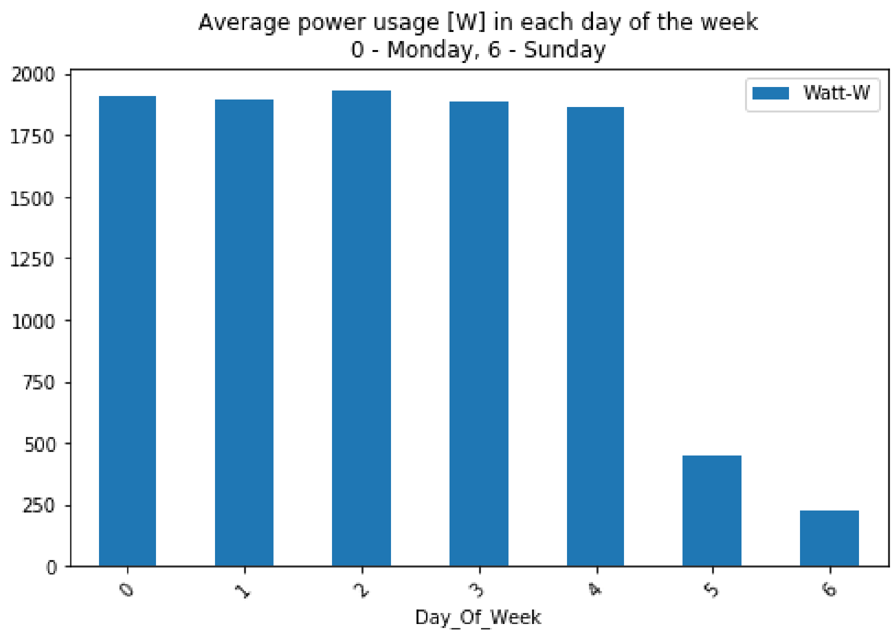

Figure 5 - average power usage in each day of week calculated on Blood analysis clinic A dataset.

Figure 6 presents average power usage in each month over a year. From the figure it is hard to grasp a clear pattern that may bring a valuable information. The highest average power usage takes place in June and the lowest in August (holidays).

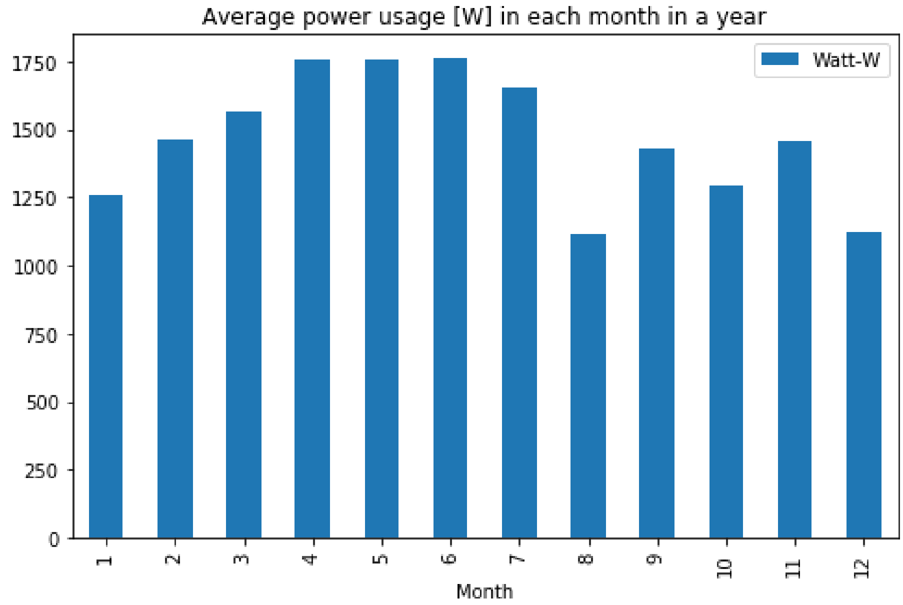

Figure 6 - average power usage in each month of a year calculated on Blood analysis clinic A dataset.

4.1.2. Blood analysis clinic B Clinics dataset

Figure 7 shows the behavior of the data that is included in the dataset. Target value presented on the vertical axis is a value of electricity consumption for 1 hour ahead prediction.

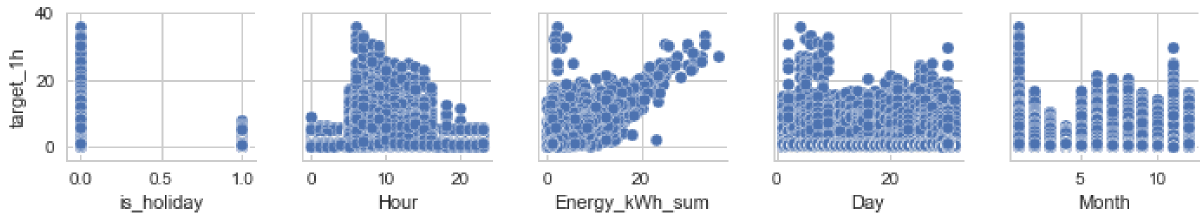

Figure 7 – Blood analysis clinic B - dataset distribution based on crucial parameters.

Figure 8 shows the correlation matrix that was calculated based on the Pearson correlation coefficient. Once can notice that there is a clear correlation between electricity consumption and a target value which has a Pearson coefficient equal to 0.83. This correlation means that with the increase of electricity consumption the value of electricity prediction is increasing as well. Similarly, as in Figure 3 other parameters have very small influence on the target value, probably due to small prediction window.

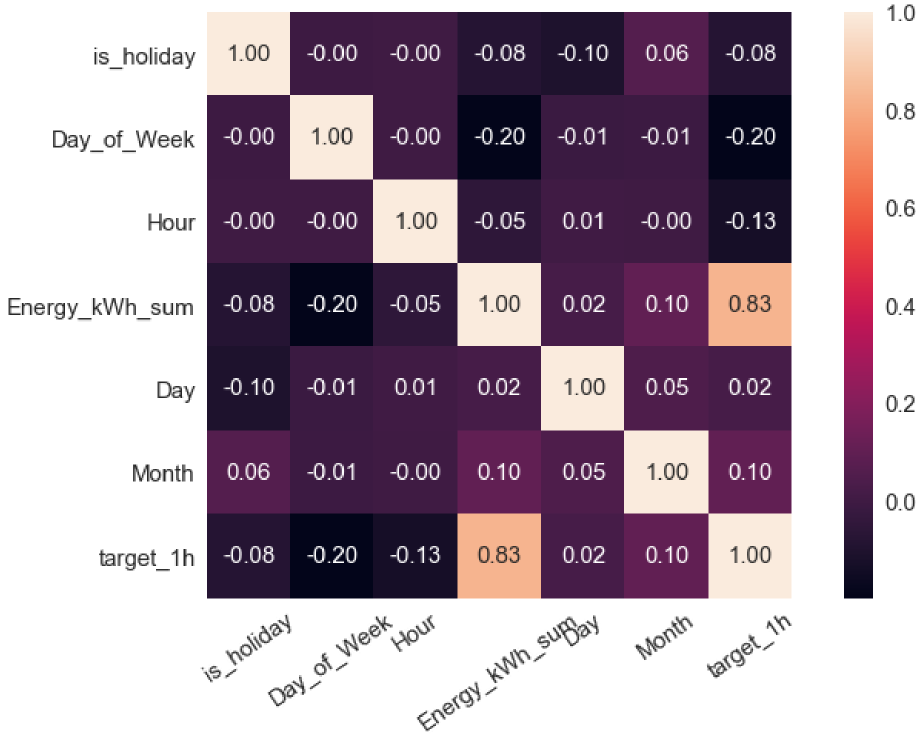

Figure 8 - correlation matrix - Blood analysis clinic B dataset

Table 3 presents the summary of the macro values calculated for the Blood analysis clinic B dataset. One can notice that standard deviation is higher than average, which means that dataset is very diverse and contains a lot of different points that have significantly different values. Moreover, the maximum value equals to 28 kW, which is around eight times higher than average one shows how diverse the dataset is.

Table 3 – macro value calculated on Blood analysis clinic B dataset.

Parameter	Value
Average	3086 W
Standard deviation	4016 W
Min	0 W
Max	28505 W

Figure 9 shows the average power usage taking into account the hour of a day. One can notice that electricity consumption varies significantly between daytime and nighttime. Three periods of electricity consumption can be distinguished: nighttime when the average consumption is equal to 1000W (probably it accounts for base consumption of the building), morning high electricity consumption from 6am to 11am and afternoon medium level electricity consumption from 11am to 5pm.

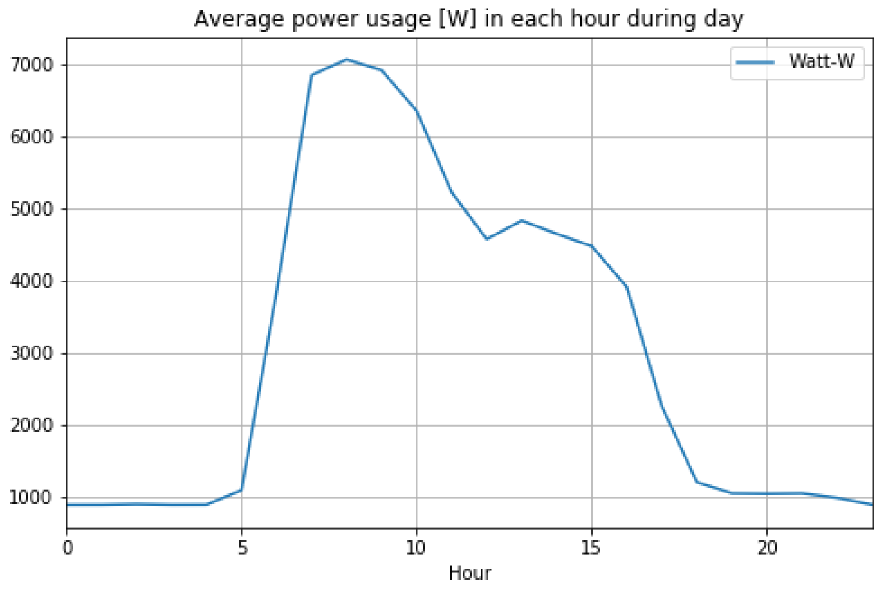

Figure 9 - average power usage in each hour of a day calculated on the Blood analysis clinic B dataset.

Figure 10 presents average power usage in each month over a year. From the figure it is hard to grasp a clear pattern that may bring a valuable information. The highest average power usage takes place in June and the lowest in February.

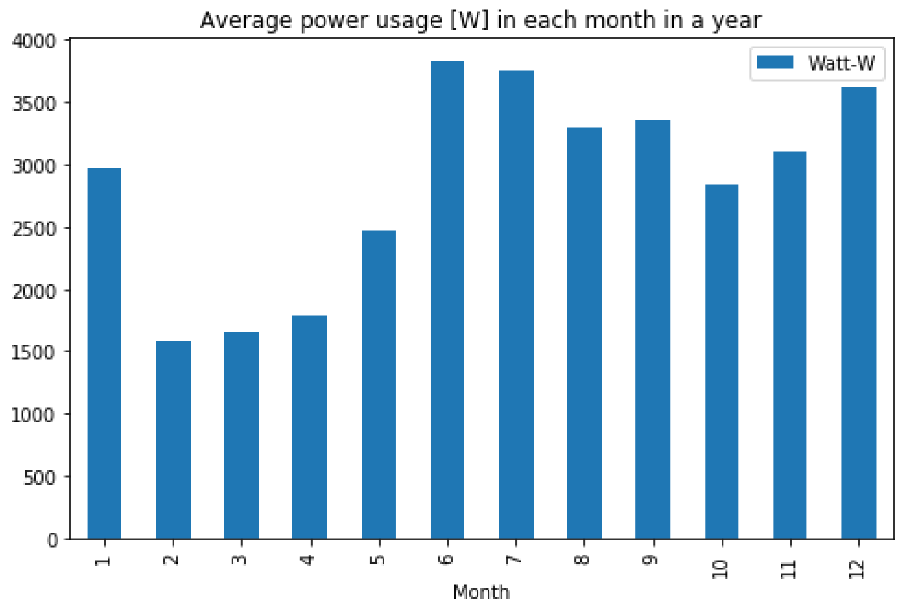

Figure 10 - average power usage in each month calculated on Blood analysis clinic B dataset.

Figure 11 shows an average power usage in each day of the week. There is a huge difference between workdays and each day of a weekend in average power consumption. The lowest power usage takes place on Sunday, whereas the highest on takes place on Monday.

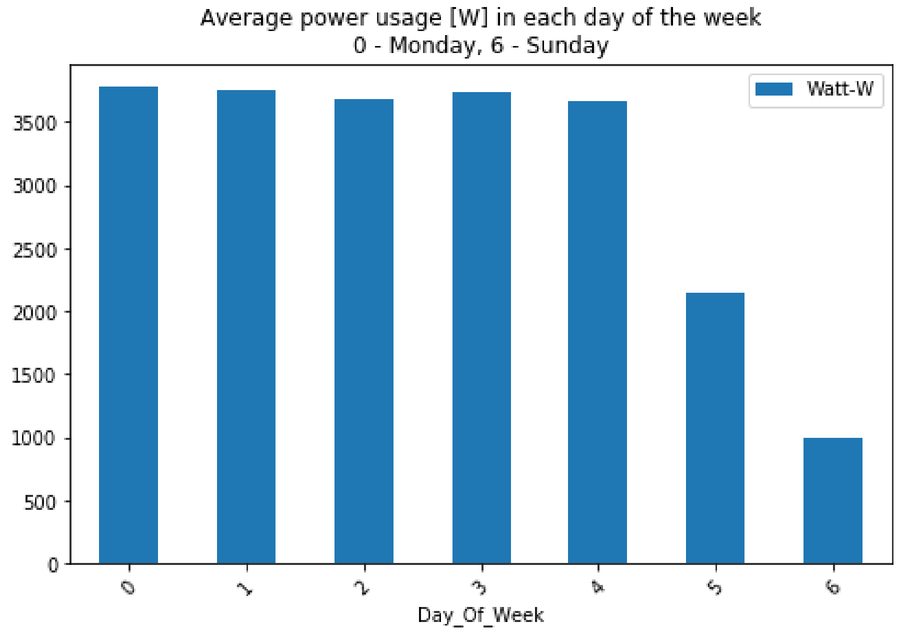

Figure 11 - average power usage in each day of a week calculated on Blood analysis clinic B dataset.

4.1.3. Restaurant dataset

Figure 12 shows the behavior of the data that is included in the dataset. Target value presented on the vertical axis is a value of electricity consumption for 1 hour ahead prediction.

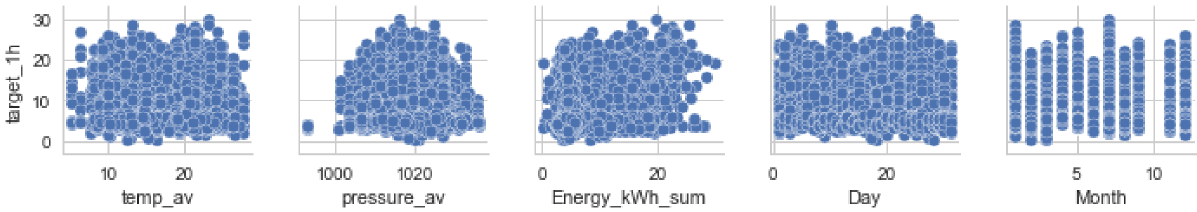

Figure 12 – Restaurant - dataset distribution based on crucial parameters.

Figure 13 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. This correlation means that with the increase of electricity consumption the value of electricity prediction is increasing as well. Similarly, as in Figure 3 and Figure 8 other values have small influence on the target value, probably due to small prediction window.

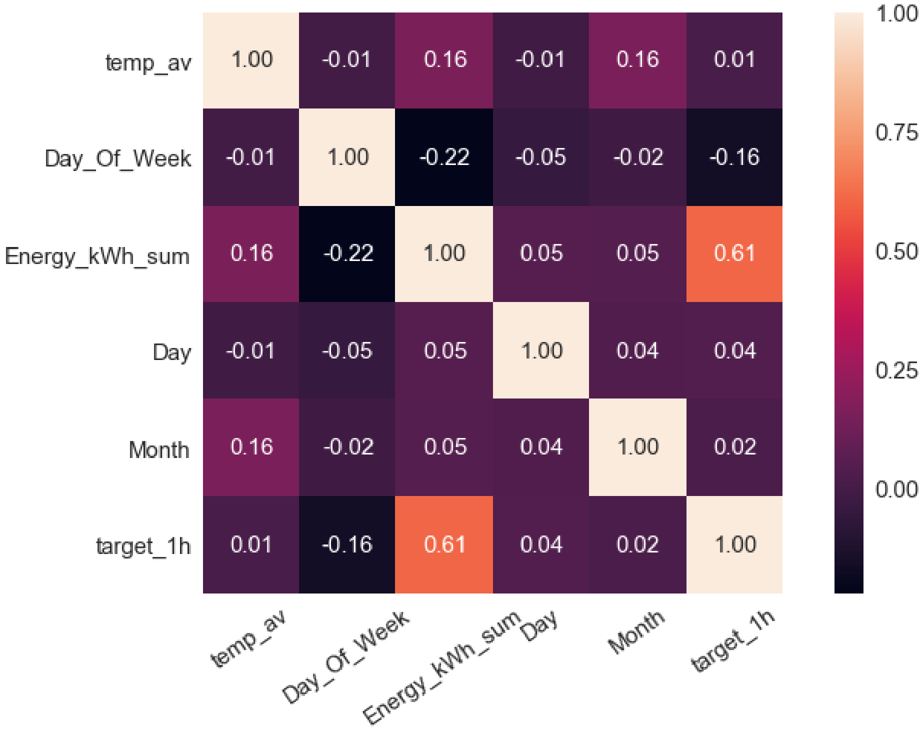

Figure 13 - Restaurant - Pearson correlation matrix.

Table 4 presents the summary of the macro values calculated for the Restaurant dataset. One can notice that maximum value equal to 38 kW is around 5 times higher than average one. This shows how diverse the dataset is.

Table 4 - macro values calculated on Restaurant dataset.

Parameter	Value
Average	7602 W
Standard deviation	5681 W
Min	107 W
Max	38040 W

Figure 14 shows the average power usage taking into account the hour of a day. One can notice that electricity consumption varies significantly between daytime and nighttime. Two periods of electricity consumption can be distinguished: nighttime when the average consumption is equal to 5000 W (probably it accounts for base consumption of the building) and daytime with a clear peak consumption at 1pm equal to 16 000 W.

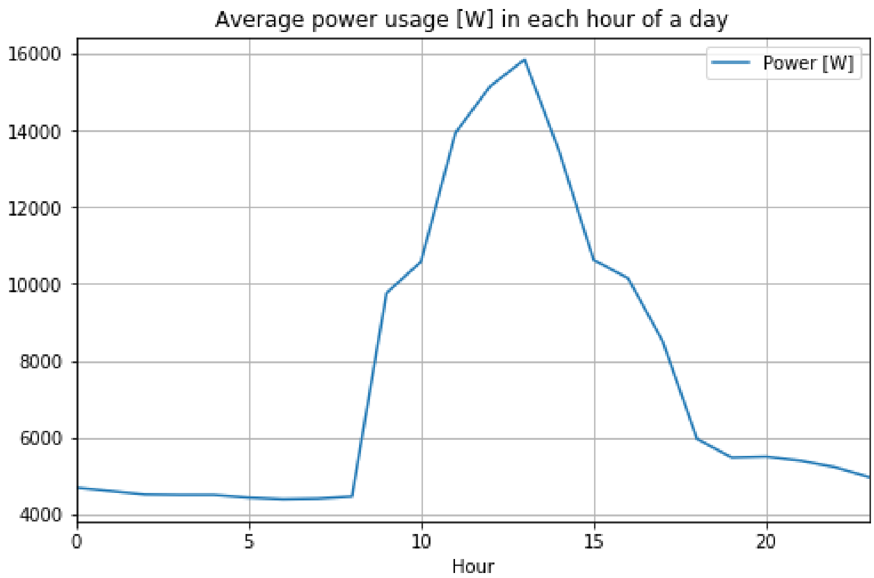

Figure 14 - average power usage in each hour of a day calculated on Restaurant dataset.

Figure 15 shows an average power usage in each day of the week. There is a huge difference between workdays and weekend days in average power consumption. The lowest power usage takes place on Sunday, whereas the highest on takes place on Thursday.

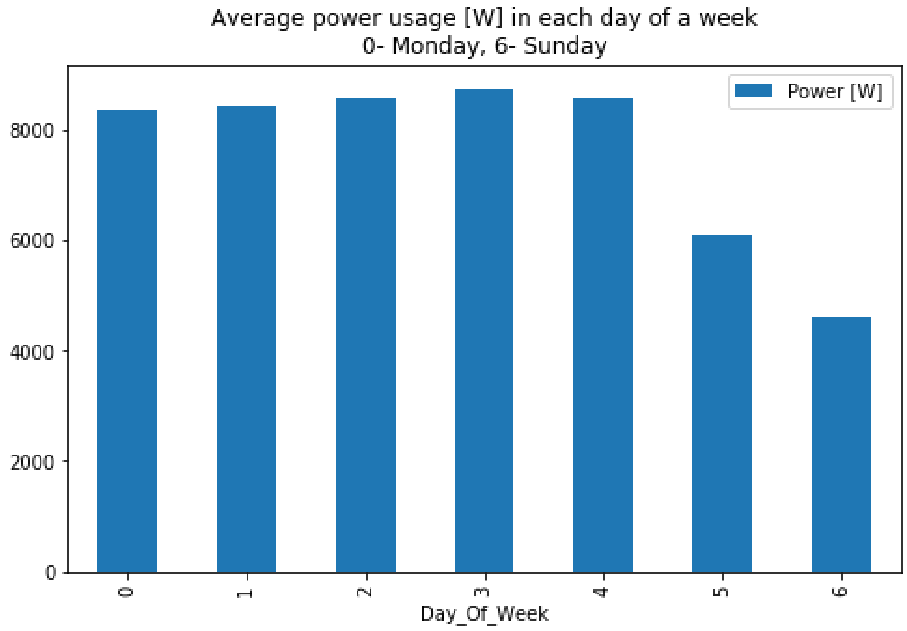

Figure 15 - average power usage in each day of a week calculated on Restaurant dataset.

Figure 16 presents average power usage in each month over a year. From the figure it can be concluded that cooling season takes place from May to August when average power is greater than 8000 W. The highest average power usage takes place in November and the lowest in April.

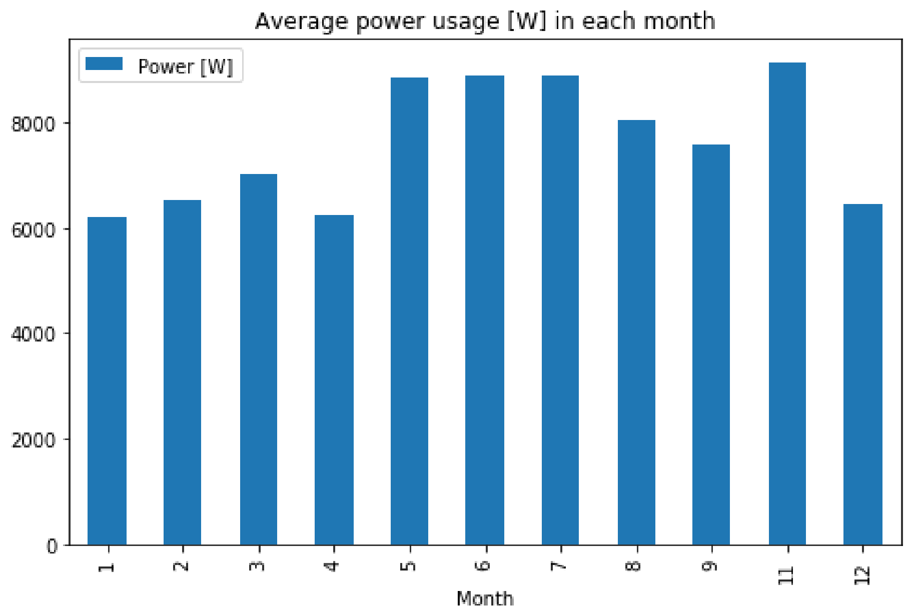

Figure 16 - average power usage in each month calculated on Restaurant dataset.

Conclusions made from exploratory analysis

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.

4.2. Random forest regression

Table 6 shows the hyperparameters that were defined for the model and determined empirically through the exhaustive grid search from the parameters defined in Table 5. The goal of the grid search was to maximize coefficient of determination on the test set.

Model parameter	Value
Max depth	2,4,6,8
Number of estimators (tree in series)	500, 1000, 1500

Table 5 - grid search parameters

Model parameter	Value
Max depth	8
Number of estimators (trees in series)	1000

Table 6 – random forest regression. Hyperparameters chosen for the model.

4.2.1. Blood analysis clinic A Dataset

Based on the input variables described in Data transformation procedure section the following results that are presented in Table 7 were obtained.

	Training Set	Validation Set	Test Set
R^2 Score	0.822	0.802	0.798
Mean Absolute Error (MAE) [kWh]	0.524	0.523	0.567

Table 7 - Random Forest Regression. Results - Blood analysis clinic A.

4.2.2. Blood analysis clinic B Dataset

Based on the input variables described in Data transformation procedure section the following results were obtained. Table 8 shows the results of the model forecast with the use of defined performance metrics.

	Training Set	Validation Set	Test Set
R^2 Score	0.9	0.86	0.82

Mean Absolute Error (MAE) [kWh]	0.57	0.65	0.6

Table 8 - Random Forest Regression. Results – Blood analysis clinic B.

Graphical presentation of model results on test dataset is presented in Figure 17.

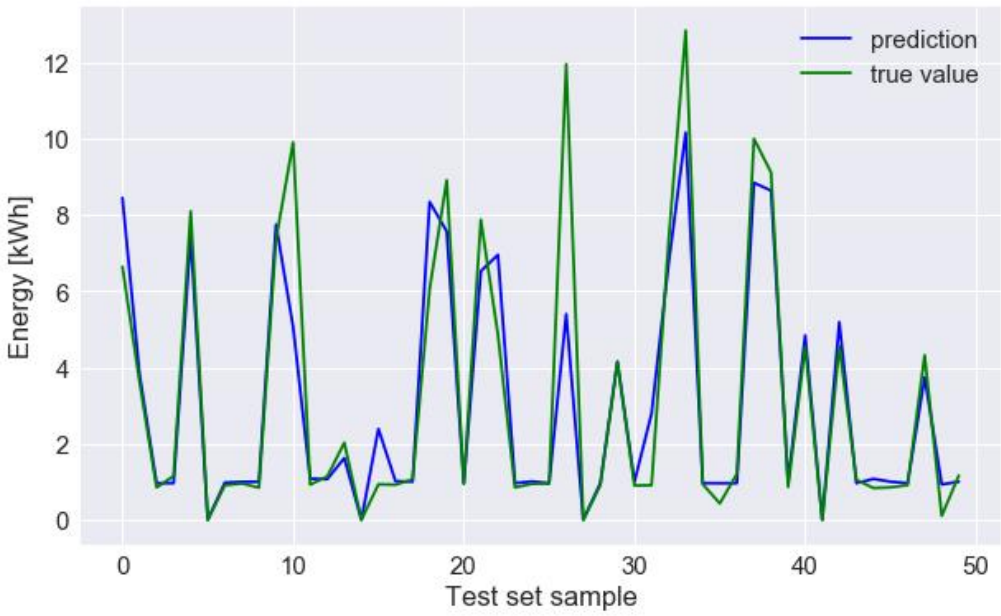

Figure 17 - Random Forest Regression Results - Blood analysis clinic B dataset

4.2.3. Restaurant Dataset

Based on the input variables described in Data transformation procedure section the following results were obtained. Table 9 shows the results of the model forecast with the use of defined performance metrics.

	Training Set	Validation Set	Test Set
R^2 Score	0.86	0.76	0.77
Mean Absolute Error (MAE) [kWh]	1.22	1.55	1.57

Table 9 - Random Forest Regression. Results - Restaurant.

4.3. Gradient boosting regression

Table 11 shows the hyperparameters that were defined for the model and determined empirically as a result of the exhaustive grid search. Grid search was performed with the use of the parameters showed in Table 10. The goal was to maximize coefficient of determination on the test set.

Model Parameter	Value
Learning rate	0.5, 0.05, 0.005
Number of estimators (trees in series)	1000, 2000, 3000,
Max depth	2,4,6,8

Table 10 - grid search parameters for gradient boosting regression.

Model parameter	Value
Learning rate	0.005
Number of estimators (trees in series)	3000
Max depth	6

Table 11 - Gradient boosting regression. Hyperparameters chosen for the model

4.3.1. Blood analysis clinic A Dataset

Based on the input variables described in Data transformation procedure section the following results were obtained. Table 12 shows the results of the model forecast with the use of defined performance metrics.

	Training Set	Validation Set	Test Set
R^2 Score	0.95	0.908	0.905
Mean Absolute Error (MAE) [kWh]	0.277	0.352	0.374

Table 12 - Gradient Boosting Regression. Results - Blood analysis clinic A.

4.3.2. Blood analysis clinic B Dataset

Based on the input variables described in Data transformation procedure section the following results were obtained. Table 13 shows the results of the model forecast with the use of defined performance metrics.

	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 13 - Gradient Boosting Regression. Results – Blood analysis clinic B.

Graphical presentation of the model results on test dataset is presented in Figure 18.

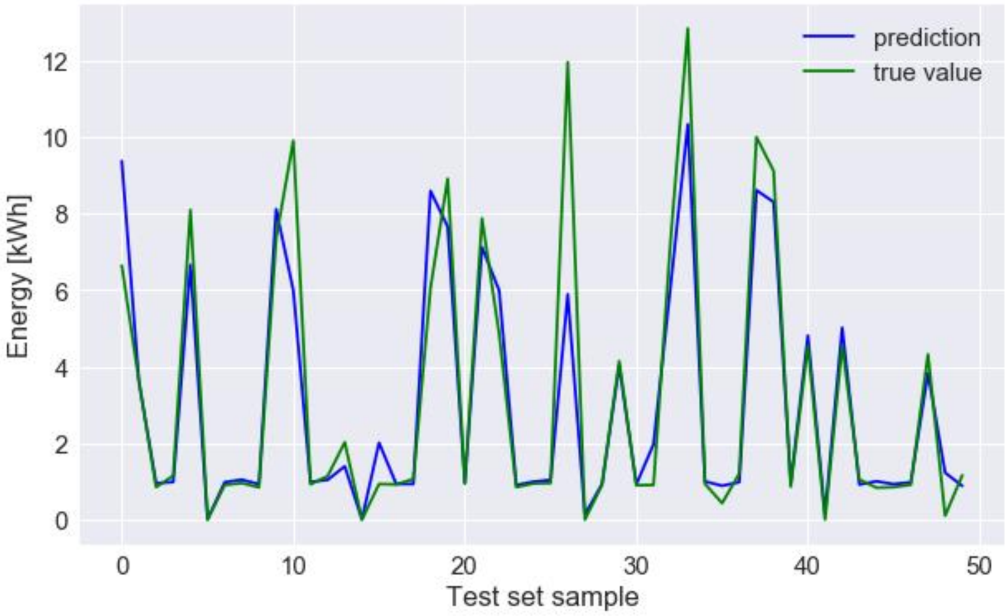

Figure 18 - Gradient Boosting Regression Results - Blood analysis clinic B Dataset

4.3.3. Restaurant Dataset

	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 14 – Gradient Boosting Regression. Results – Restaurant.

4.4. RNN with LSTM cell

Table 16 shows the hyperparameters used in the neural net model that are the result of the exhaustive grid search carried out on the parameters defined in Table 15. The hyperparameters presented are the effect of the iterative process where the goal was to maximize coefficient of determination.

Model training parameters	Value
Number of epochs	500, 1000, 1500
Loss function	512, 1024, 2048

Table 15 – grid search parameters defined for RNN with LSTM cell.

Model training parameters	Value
Number of epochs	1000
Batch size	1024
Loss function	Root squared error

Table 16 - Neural Net with LSTM cell - used hyperparameters.

4.4.1. Blood analysis clinic B Dataset

Due to long computation time and limited computational power resources, only Blood analysis clinic B dataset was analyzed. The results are presented in Table 17.

	Training Set	Validation Set	Test Set
R^2 Score	0.68	0.56	0.62
Mean Absolute Error (MAE) [kWh]	1.23	1.37	1.22

Table 17 – RNN with LSTM cell. Blood analysis clinic B - Results.

The distribution of loss function obtained during the training process is presented in Figure 19. From the figure it can be concluded that at the end of a training process model is reaching a plateau phase, which means that a training limit has been met. To improve the performance, one can increase the number of epochs or increase the network size (impossible in this project due to computational power limitations).

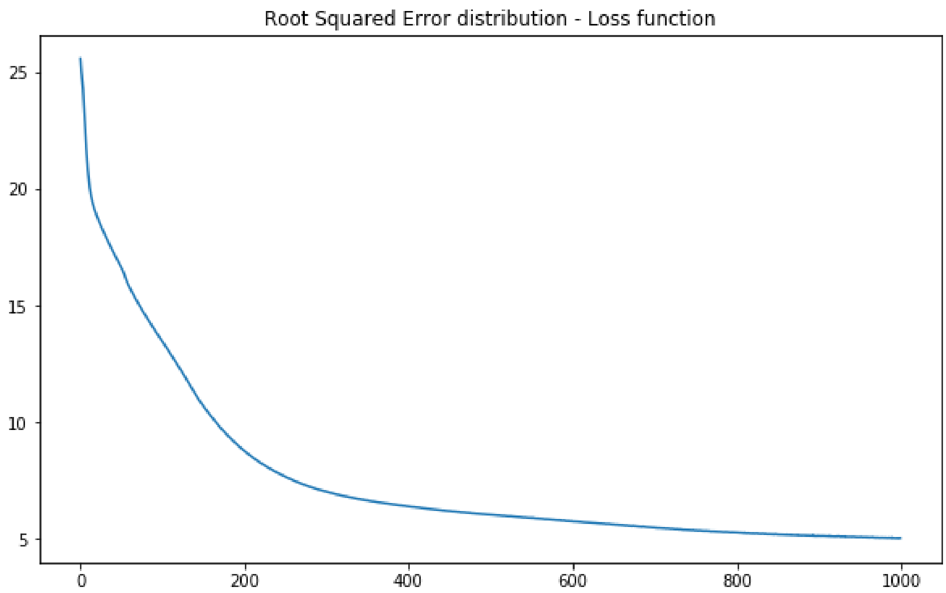

Figure 19 - loss function distribution over training process.

The model prediction is presented in Figure 20. One can notice that prediction values are covering the real ones, what is the goal of the forecasting model. The outliers' problem which are the values above 10 kWh is very apparent. The model finds it difficult to train a pattern when those values occur. On the other hand, those values take place very rarely.

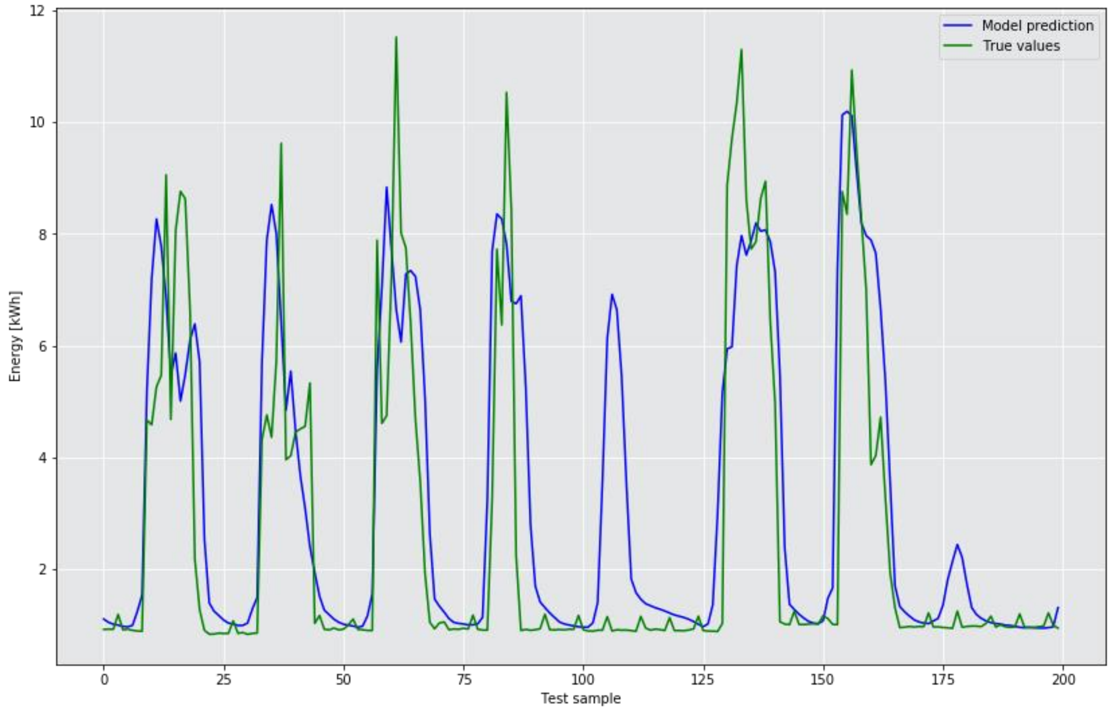

Figure 20 - LSTM net. test set results.

It is worth to emphasize that neural nets perform very well when dataset size is large, so that the training process can be carried out precisely and the loss function can be optimized on the various data.

4.5. Different resolution of predicted values (one hour ahead, one day ahead)

In the previous part, the prediction target was set to one hour ahead predictions (their application in real life is discussed in the Business Application of the Solution section). However, it is worth to consider longer prediction periods due to potential business application described in Business application of the solution section. In this section the one-day ahead prediction will be discussed with the use of Radom Regression Trees and Gradient Boosting Regression Trees. All the calculations performed in this section were carried out with the use of the Blood analysis clinic B dataset that was randomly selected from available datasets.

4.5.1. Dataset description

The dataset size (data needed to be grouped in one-day single readings) was compressed to 669 readings (669 days of recorded period) because one-day readings were needed to perform the analysis. Figure 21 presents the distribution of data based on different parameters, referred to the target value which was determined as electricity consumption in next 24 hours (1 day ahead).

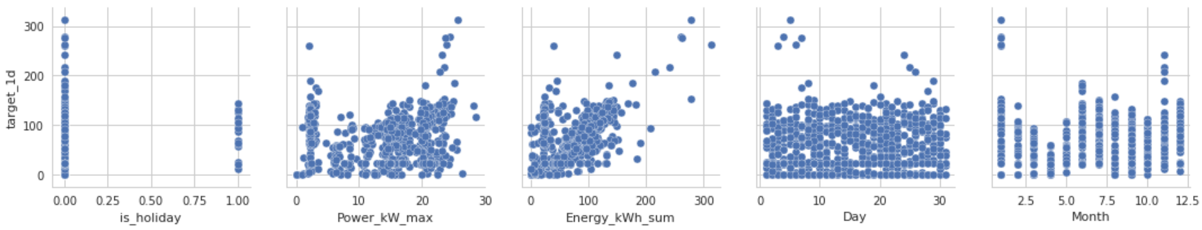

Figure 21 - Blood analysis clinic B - one day data groupings and their distribution.

Figure 22 presents the Pearson correlation matrix between various parameters that were taken into account. One can notice that there is a positive correlation between maximum power during day with the targeted value and between energy consumption during the day and one-day-ahead prediction.

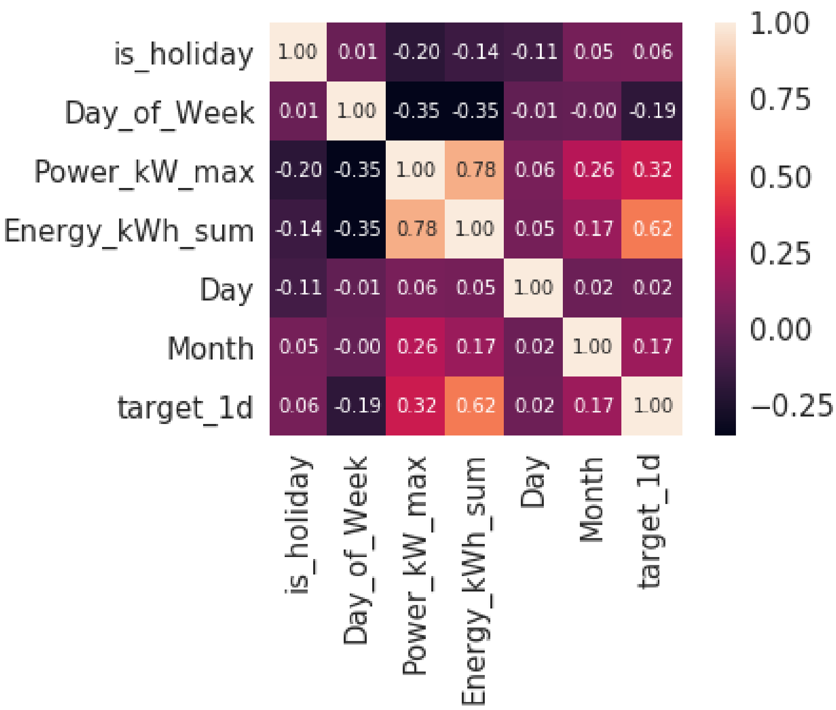

Figure 22 - Blood analysis clinic B dataset - 1 day ahead prediction - Pearson correlation matrix.

4.5.2. Random Forest Regression – Results.

Table 18 shows the hyperparameters that were defined for the model and determined empirically with the goal to maximize coefficient of determination on the test set.

Model parameter	Value
Max depth	5
Number of estimators (trees in series)	3000

Table 18 - one day ahead prediction - model hyperparameters.

Results of the random forest regression model are presented in Table 19.

	Training Set	Validation Set	Test Set
R^2 Score	0.85	0.47	0.68
Mean Absolute Error (MAE) [kWh]	65	104	95

Table 19 - Random Forest Regression - Results.

Graphical presentation of model results on test dataset is presented in Figure 23.

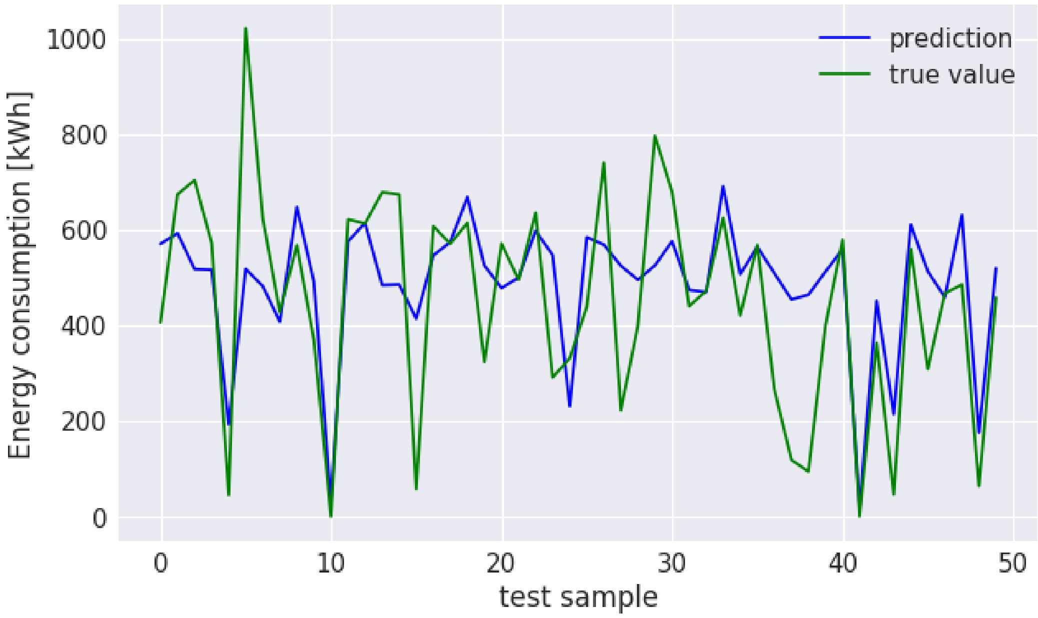

Figure 23 - Random Forest Regression – 1 day ahead prediction

4.5.3. Gradient Boosting Regression - Results.

Table 20 presents hyperparameters used in the model.

Model parameter	Value
Learning rate	0.1
Number of estimators (trees in series)	100000
Max depth	3

Table 20 - Gradient Boosting Regression - hyperparameters.

Table 21 presents results calculated by model.

	Training Set	Validation Set	Test Set
R^2 Score	0.99	0.54	0.52
Mean Absolute Error (MAE) [kWh]	0.001	116	115

Table 21 - Gradient Boosting Regression - Results.

Graphical presentation of model results on test dataset is presented in Figure 24.

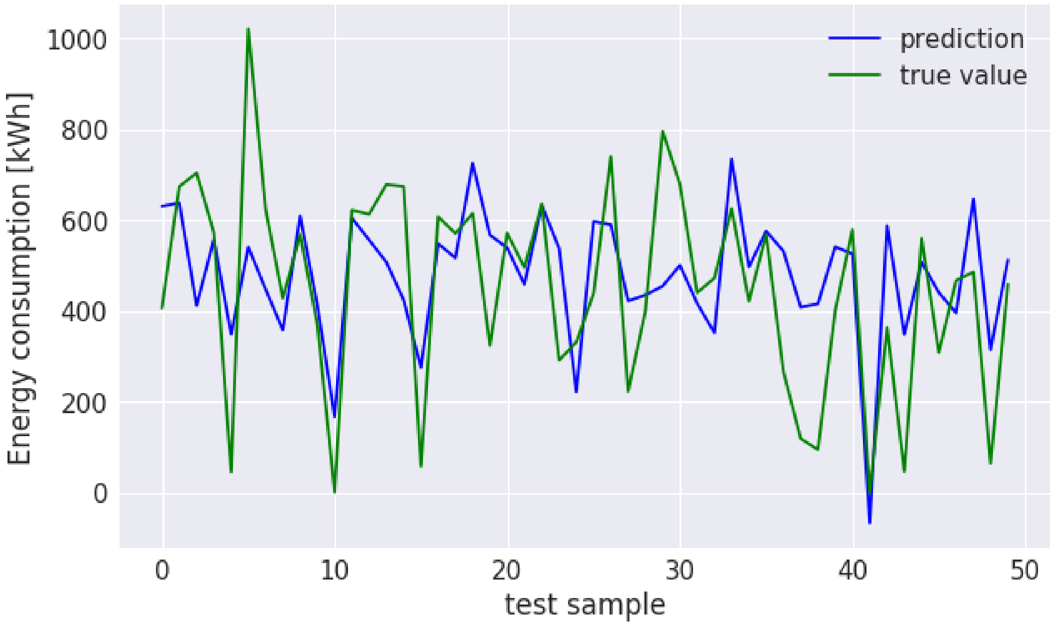

Figure 24 - Gradient Boosting Regression - 1 day ahead prediction.

One can conclude that due to the small size of the dataset, the overfitting problem is very apparent. For that reason, it is difficult to adjust performance of the model on the validation and test set.

4.6. Discussion

This section answers research questions stated at the beginning of the thesis in section Research questions. Based on the results presented in Exploratory analysis, Random forest regression section, Gradient boosting regression section and RNN with LSTM cell section, the following questions are answered.

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

As presented in Figure 2, Figure 3, Figure 7, Figure 8, Figure 12 and Figure 13 and the results of the exploratory analysis, there are certain patterns that occur in each analyzed dataset. These patterns are:

- 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.

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

The 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, the 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 ahead 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, the learning ability was limited.

To sum up, it can be concluded that machine learning brings 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.

4.6.3. 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 22 - comparison of results for different datasets

Results provided in Table 22 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.

5. Business application of the solution

Apart from the technical applicability there must be a market need in order to implement the solution and offer it to customers. As the big data technologies became to be implemented by electricity retailers, the use of machine learning in those technologies is predicted to increase in coming years. An estimate of savings from a 1% reduction in the mean absolute percentage error (MAPE) for a utility with 1GW peak load is according to [19] equal to $300,000 per year from short-term load forecasting. Short-term forecasting can vary from minutes up to few days ahead [20]. Solution can be offered both as a business to business solution where model is sold to an utility company and as a business to customer where intelligent controller is sold to customer.

There are many areas of implementation for such systems and different markets that can take advantage of those implementations. Implementation areas and different markets are discussed in the following sections.

5.1. How the solution can be implemented

The implementation of the solution is defined as a product that contains particular technology that is either the main added value of product or one of many values that are offered within a product. Table 23 presents potential areas of implementation of the models that were presented in Thesis. Applications of two forecasting targets (one hour ahead and one day ahead) are presented separately, as their use however similar can be applied in slightly different way. Two main areas of the business applications are: demand-response model for the on-grid installations and smart metering control systems for off-grid installations.

One-hour-ahead prediction	One-day-ahead prediction
• Demand-response models implemented by electricity retailers [21] (on-grid installations)– information from model can be used as an input in the demand response model that will control particular devices such as washing machine or dishwasher to minimize the cost of electricity paid by customer. • Smart controllers installed individually by consumers on-site [22] (off grid installations) – local power generation systems such as PV panels with accumulation systems can take advantage of the input from prediction model in order to appropriately steer the amount of electricity that is accumulated.	• Demand-response models implemented by electricity retailers [21] (on-grid installations)– information from model can be used as an input in the demand response model that will control particular devices such as electric vehicle charging station or electric heating or air-conditioning systems to minimize the cost of electricity paid by customer. • Smart meters installed individually by consumers on-site [22] (off-grid installations) – local power generation systems such as PV panels with accumulation systems combined with local demand-response steering devices can take advantage of the input from prediction model in order to appropriately steer the

	amount of electricity that is accumulated, consumed and produced during each day.

Table 23 - possible solution implementation areas.

5.2. 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 [23] 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 [23]. In [24], authors show how New York City power grid can benefit from implementation of machine learning based systems. In [25] 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, using machine learning to 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 – these systems are very often composed by a renewable electricity generator, an accumulation system and a controller that could be enabled with machine learning algorithms that would be responsible for electricity flows among consumption, production and energy accumulation.

5.3. Potential benefits for the user

There are several benefits that can be an effect of machine learning based forecasting systems. The ultimate goal of forecasting is to offer added value to customers and there are two main areas where this can be achieved. 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.

6. Conclusions

The goal of the thesis was to show applicability of various machine learning algorithms in small and medium buildings. Three different datasets 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 dataset:

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

For Clinic B dataset:

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

For Restaurant dataset:

- 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.
- 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.

7. References

[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] T. E. Demand, M. Using, and M. L. Algorithms, "Two-Stage Electricity Demand Modeling Using Machine Learning Algorithms," *Energies*, vol. 10, no. 10, p. 1547, 2017.

[6] P. Lusis, K. R. Khalilpour, L. Andrew, and A. Liebman, "Short-term residential load forecasting: Impact of calendar effects and forecast granularity," *Appl. Energy*, vol. 205, no. June, pp. 654–669, 2017.

[7] M. Behl, F. Smarra, and R. Mangharam, "DR-Advisor: A data-driven demand response recommender system," *Appl. Energy*, vol. 170, pp. 30–46, 2016.

[8] S. Jurado, À. Nebot, F. Mugica, and N. Avellana, "Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques," *Energy*, vol. 86, pp. 276–291, 2015.

[9] R. Hu, S. Wen, Z. Zeng, and T. Huang, "A short-term power load forecasting model based on the generalized regression neural network with decreasing step fruit fly optimization algorithm," *Neurocomputing*, vol. 221, no. September 2016, pp. 24–31, 2017.

[10] Y. Fu, Z. Li, H. Zhang, and P. Xu, "Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-metering Devices," *Procedia Eng.*, vol. 121, pp. 1016–1022, 2015.

[11] W. McKinney, "Data Structures for Statistical Computing in Python," *Proc. 9th Python Sci. Conf.*, vol. 1697900, no. Scipy, pp. 51–56, 2010.

[12] J. D. Hunter, "Matplotlib: Python plotting — Matplotlib 2.2.2 documentation." 2007.

[13] V. Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, P. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, A. and and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, and E. Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, "Scikit-learn: Machine Learning in

Python.” 2011.

[14] F. and others Chollet, “Keras.” 2015.

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

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

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

[18] A. Müller and S. Guido, *Introduction to Machine Learning with Python*. 2016.

[19] T. Hong, “Crystal Ball Lessons in Predictive Analytics,” *EnergyBiz Mag.*, vol. Spring, pp. 35–37, 2015.

[20] “Energy Forecasting: Very Short, Short, Medium and Long Term Load Forecasting,” *blog.drhongtao.com*.

[21] D. Zhou, M. Balandat, and C. Tomlin, “Using Machine Learning with Observational Data.”

[22] A. Bielecki, M. Bielecka, and S. Ernst, “Proposal of an Intelligent, Predictive Fuzzy Controller for Off-Grid Devices,” *IFAC-PapersOnLine*, vol. 49, no. 25, pp. 523–528, 2016.

[23] 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].

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

[25] 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].