proceedings

MDPI

Proceedings

Short-Term Electric Demand Forecasting for the Residential Sector: Lessons Learned from the RESPOND H2020 Project †

Meritxell Gómez-Omella [1,2,*], Iker Esnaola-Gonzalez [1] and Susana Ferreiro [1]

[1] TEKNIKER, Basque Research and Technology Alliance (BRTA), C/Iñaki Goenaga, 5, 20600 Eibar, Spain; iker.esnaola@tekniker.es (I.E.-G.); susana.ferreiro@tekniker.es (S.F.)

[2] Faculty of Informatics, University on the Basque Country (UPV/EHU), Paseo Manuel Lardizabal, 1, 20018 Donostia-San Sebastian, Spain

* Correspondence: meritxell.gomez@tekniker.es

† Presented at the Sustainable Places 2020, Online, 28–30 October 2020.

Published: 6 January 2021

Abstract: RESPOND proposes an Artificial Intelligent (AI) system to assist residential consumers that would like to make use of Demand Response (DR) and incorporate it into their energy management systems. The proposed system considers the forecast energy consumption based on the data acquired. This work compares the results obtained by different forecasting methods using the Root Mean Square Error (RMSE) as a measure of the forecast performance. The ARIMA, Linear Regression (LR), Support Vector Regression (SVR) and k-Nearest Neighbors (KNN) models are tested, and it is concluded that the results achieved with the KNN obtain a better fit. In addition to obtaining the lowest RMSE, KNN is the algorithm that spends less time in obtaining the forecasts.

Keywords: time series forecasting; k-nearest neighbor; electric demand; RESPOND project

1. Introduction

Buildings use more than 35% of global energy use but a significant amount can be saved if they are properly operated. Apart from the large energy consumption of buildings, peak energy demand certainly attracts lots of attention because of its negative impact on energy grid capital, operational cost and environmental pollution to name a few. This impact is a direct consequence of the carbon-intense generation plants that grid operators deploy to satisfy energy demand during these peak periods [1]. Demand Response (DR) activities including load shifting or peak shaving have a huge potential to match energy demand with energy supply side, thus avoiding these undesirable peaks [2]. The implementation of DR activities is especially promising in the residential sector, where the full capabilities of the DR are yet to be unlocked.

This is where the RESPOND H2020 project (https://project-respond.eu) originates, aiming to bring DR programs to neighborhoods across Europe. More specifically, RESPOND aims to deploy an interoperable energy automation, monitoring and control solution to deliver DR programs at a dwelling, building and district level to neighborhoods across Europe. To do so, RESPOND proposes an Artificial Intelligent (AI) system to assist residential consumers that would like to make use of DR and incorporate it into their energy management systems [3]. The proposed system considers the forecast energy consumption and production based on the data acquired by the deployed IoT equipment and looks for modifications that would mitigate potential instabilities in the energy supply network by applying optimal energy use and load shifting. This article focuses on the energy consumption forecasting part of the AI system, where the effectiveness of different models of time series and machine learning has been evaluated.

The rest of the article is structured as follows. The data available to make the study is introduced in Section 2. Then, the theoretical principles of the forecasting methods are explained in Section 3. In Section 4, the results obtained are summarized. Finally, conclusions of the lesson learned are written in Section 5.

2. Data Availability

The available electrical consumption data (measured in Wh) is obtained from different houses located in three different places: Madrid (Spain), Aarhus (Denmark) and the Aran Islands (Ireland).

Some indicators are calculated to assess the quality of the data available. The explanation of the data quality metrics used is out of the scope of this paper. There are some indicators that take values between 0 or 1 depending on the quality of the data in some respects. A value of 0 is considered the worst possible quality and a value of 1 represents the perfect quality for that indicator. In general, the quality metrics obtain values close to 1 except the Timeliness indicator. When a sensor fails for whatever reason, it stops sending data, including the time value. Due to failures in sensors, waiting times occur in the time variable. This fact is reflected in the low values of Timeliness indicators.

Even though all the data collection processes began on the same date, on 1 January 2019 it can be observed that not all the pilot sites have the same data availability. The analysis was performed on 6 March 2020. It is decided to eliminate from the study the houses with more than 30% of missing data. Following this criterion, data are available from 10 houses in Madrid, 11 in Aarhus and 8 in the Aran Islands.

3. Methodology

Traditionally, the energy demand forecasting has been addressed via data-driven algorithms due to their high performance. Therefore, RESPOND's Energy Demand Forecasting Service has targeted these algorithms with views to having the best performance possible.

In what follows, a short description of some well-known forecasting methods are presented. Auto Regressive Integrated Moving Average (ARIMA) are models designed to make forecasting in time series and past values of the response variables are used to estimate the future values. Linear Regression (LR), Support Vector Regression (SVR) and k-Nearest Neighbors (KNN) are machine learning algorithms that use explanatory variables to estimate the values of the response variable. First, we decide that the explanatory input variables in the predictive models were extracted from the time variable. On the one hand, this agreement provides simplicity to the models and allows the results to be explained. On the other hand, continuity in time allows the imputation of the missing data in case of sensor failure. These variables, also called features, were *day*, *sinMonth*, *cosMonth*, *sinHour*, *cosHour*, *Season*, *weekday* and *workingDay*. Before creating models, we identified some outlier values. These are values that excessively exceed the typical values for electrical consumption. After observing the behavior of the consumption data for the different houses, we concluded that a common pattern would lack precision. Finally, we decided to remove values greater than 3000 Wh. These values are considered meaningless and possibly caused by a failure in the data collection method.

The Root Mean Square Error (RMSE) is used to measure the standard deviation of the prediction errors. Therefore, the forecasts with the lowest RMSE are considered the best fit.

The calculations have been executed with the statistical software R. The *caret* and *forecast* packages have implemented the necessary functions to carry out the predictions, perform cross validation and automatically search for the optimal values of the necessary parameters in each algorithm.

3.1. ARIMA

In the process of finding the best predictive model, we started with Autoregressive Integrated Moving Average *ARIMA(p,d,q)* models. Those models are fitted to time series data to predict future points where data show evidence of non-stationarity. Time series can be transformed into stationary by differentiation d times. Once the series is stationary, we used the classic explanatory methods to

choose the orders p and q based on the comparation of *Akaike Information Criterion* (AIC) and *Bayesian Information Criterion* (BIC).

3.2. Linear Regression

Linear regression (LR) attempts to model the relationship between two variables by fitting a linear equation to observed data [4]. A linear regression line has an equation of the form $Y = \beta_0 + \beta_1 X$, where X is the explanatory variable and Y is the dependent variable. Multiple Linear Regression (MLR) uses more than one explanatory variable to fit the response variable.

3.3. Support Vector Regression (SVR)

SVR uses the same principles as Support Vector Machine (SVM) but it used in a regression method, so we can use SVR for working with continuous values. SVR allows the definition of the width of the band around the error in our model and the discovery of an appropriate hyperplane to fit the data. The objective function of SVR is to minimize the coefficients, not the squared error. The error term is instead handled in the constraints, where we set the absolute error less than or equal to a specified margin, called the maximum error, ϵ (epsilon) (https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2) [5].

3.4. K-Nearest Neighbors

K-nearest neighbors algorithm (KNN) is a supervised machine learning algorithm that can be used to solve regression models. First, the distance between the explanatory variables of the point x and the rest of observations should be calculated. Therefore, each point has a distance value associated. The k nearest neighbors to the point x are the observations with the lower distance. These k observations are used to compute the value of the predictor Y. The value of Y in some point is the average of the dependent variable of its k nearest neighbors [6].

4. Results and Discussion

This section summarizes the results obtained from the forecasts using the different methods mentioned above. The RMSE shown is the mean of the errors obtained in the 29 available houses.

4.1. ARIMA

Due to high amount of data, the search for the optimal p and q was neither simple nor satisfactory. It has multiple functions for the treatment of time series in R. Specifically, we used a method that finds the best Seasonal Autoregressive Integrated Moving Average (SARIMA) model. The idea is that SARIMA models are $ARIMA(p, d, q)$ models whose residues are ARIMA(P, D, Q). The RMSE value achieved using ARIMA model was 1540.07 Wh whereas the RMSE was 1294.81 Wh using SARIMA model.

4.2. Linear Regression

Linear relationship between dependent variables and independent variable was found. Although the RMSE was significantly lower than in the results obtained with ARIMA, the Coefficient of determination (R^2) was less than 0.3 in all fitted models. The best RMSE value was 540.22 Wh and it was obtained when train the model with all the time related features. Furthermore, the predictions seem to repeat the same pattern every day.

4.3. Support Vector Regression (SVR)

Electricity consumption took negative values in some cases using SVR and this makes no sense. Electric consumption can never be less than 0 Wh. Although RMSE obtained was lower than the

previous, the method was rejected because this problem could not be controlled. Furthermore, SVR is computationally expensive and takes too much time and resources to forecast.

4.4. K-Nearest Neighbors

The best fit with the KNN algorithm was obtained using all the mentioned exploratory variables. In all cases the RMSE takes values between 156 and 439 Wh. Furthermore, the k value indicating the number of neighbors to be used was tested from 1 to 10. All forecasts were obtained using less than 5 neighbors.

5. Conclusions

Data quality is important to ensure that forecasting models work properly and take advantage of the useful information provided by historical data. The low RMSE obtained with the KNN shows that it is an optimal algorithm with which to make the forecast of electricity demand. Depending on the data availability, performance of forecasters varies. Therefore, predictive models are periodically re-trained as they are expected to improve their performance as a bigger historical data size is available.

Author Contributions: Data Curation, Formal Analysis and Conceptualization, M.G.-O.; Investigation and Writing M.G.-O. and I.E.-G., Supervision S.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by RESPOND (integrated demand REsponse Solution towards energy POsitive NeighbourhooDs) project grant number 768619.

Acknowledgments: This work is partly supported by the RESPOND (integrated demand REsponse Solution towards energy POsitive NeighbourhooDs) project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 768619.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Collins, L.D.; Middleton, R.H. Distributed demand peak reduction with non-cooperative players and minimal communication. *IEEE Trans. Smart Grid* **2017**, *10*, 153–162.
2. Albadi, M.H.; El-Saadany, E.F. Demand response in electricity markets: An overview. In Proceedings of the 2007 IEEE Power Engineering Society General Meeting, Tampa, FL, USA, 24–28 June 2007; pp. 1–5.
3. Esnaola-Gonzalez, I.; Diez, F.J.; Pujic, D.; Jelic, M.; Tomasevic, N. An artificial intelligent system for demand response in neighbourhoods. In Proceedings of the Workshop on Artificial Intelligence in Power and Energy Systems (AIPES 2020), Santiago de Compostela, Spain, 4 September 2020. doi:10.13140/RG.2.2.30279.32163.
4. Aalen, O.O. A linear regression model for the analysis of life times. *Stat. Med.* **1989**, doi:10.1002/sim.4780080803.
5. Drucker, H.; Surges, C.J.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. *Adv. Neural Inf. Process. Syst.* **1996**, *9*, 155–161.
6. Keller, J.M.; Gray, M.R. A Fuzzy K-Nearest Neighbor Algorithm. *IEEE Trans. Syst. Man Cybern.* **1985**, doi:10.1109/TSMC.1985.6313426.