LINEAR RANDOM FOREST TO PREDICT ENERGY CONSUMPTION

Gianpaolo Zammarchi [1]

[1] Department of Economics and Business Science, University of Cagliari, Via Sant'Ignazio da Laconi, 17, 09123, Cagliari (Italy). (e-mail: gp.zammarchi@unica.it)

ABSTRACT: Forecasting electricity consumption is a relevant task to ensure that the supply of energy fed into the grid always equals the demand. In this study we compare the performance of random forest and linear random forest in the prediction of daily electricity consumption in Italy. We show that both implementations reach a good performance in this task, with the best results obtained by linear random forest in a model including different features such as lags, difference variables and day - month variables.

KEYWORDS: linear random forest, time series, energy consumption

1 Introduction

Due to the rapid increase in world population and the global economic growth, the energy consumption is expected to increase in most countries. In particular, electricity is one of the main energy sources for homes, offices, factories and many other public and private places. A relevant problem is to ensure that the supply of energy fed into the grid always equals the demand or, in other words, to guarantee the equilibrium between the production of electricity and the consumption. For this reason, different companies and researchers have developed methods to forecast electricity daily consumption (Zhang *et al.*, 2021). In this study we assessed the performance of two different implementations of random forest in the prediction of energy consumption in Italy and compared their results with the effective consumption and with the prediction of Terna (the company that manages the Italian national transmission system).

The rest of the paper is organized as follows: first, we give an overview of the problem, the data collection and the methodology in Section 2. Then, we present the results in Section 3, and a brief summary and the future developments in Section 4.

2 Methods

In this section we will describe how data were collected, the features engineering process, and how these features were used to build the model used for predictions.

2.1 Data collection

The data related to the forecasts made by Terna's model, together with the actual consumption detected by the company (in Megawatt, MW), are published daily in the form of PDF files (Terna S.p.A., 2023). The files were downloaded and read in R (R Core Team, 2023). The data set included day-by-day hourly consumption values and forecasts for all days ranging from August 1, 2022 to March 31, 2023. Subsequently, these values were aggregated as follows: we computed $\upsilon_i = \{v_1, v_2, ..., v_j\}$ with $j = 1, ..., 24$, and a vector $\mathbf{V} = \{\upsilon_1, \upsilon_2, ..., \upsilon_i\}$ with $i = 1, ..., 243$ in order to obtain a vector with daily values obtained as the sum of individual hourly values.

2.2 Random forest

Random forest is a popular machine learning technique based on the combined use of decision trees, bootstrap, and ensemble methods (Breiman, 2001). It incorporates the output of several decision trees to produce a single evaluation. In this study we used the classical random forest implementation as well as a recently developed linear random forest variation based on the implementation of a ridge regression in the leaves (Künzel *et al.*, 2022). In this variation the returned value is computed using a linear aggregation function: $\hat{\mu}(x_{new}) := x^t_{new}(X^t_S X_S + \lambda I)^{-1} X^t_S Y_S$, where X_{new} is a new observation, S is a leaf, Y is the response variable, $\mathbf{X}$ the design matrix for the training set, and λ is a regularization parameter. The optimal splitting point is defined with a greedy strategy and the stopping criteria is based on an R^2 improvement threshold (Künzel *et al.*, 2022).

2.3 Feature engineering

We created lagged and difference variables to be used as predictors for the random forest. We therefore defined k as the number of lags that can be created starting from the response variable, the daily consumption of electricity. Difference variables were also created as in the following equation: $y_i - y_{i-t}$ where y_i is the energy consumpion during the day i. During the model evaluation phase, various configurations were tested, using a number of lags $k \in \{k_1, ..., k_m\}$ with $m = 30$ and t equals to 7 and to 14. In addition, two variables relative to the day of the week (Monday-Sunday) and the month (from August 2022 to March 2023) have been included.

2.4 Models evaluation

The daily consumption values up to the end of February were used as the training set to predict daily consumption in March (test set). The predictions have been evaluated using two widely used metrics: root-mean-square error (RMSE) and mean absolute percentage error (MAPE). Terna's prediction has also been included as a benchmark to further compare the magnitude of the errors. Errors are computed using a moving window scheme.

3 Results

In this section we will present the results obtained using the two different implementations of random forest and compare these results with the effective consumption and with Terna's prediction. Figure 1 shows a comparison of the RMSE for both implementations of random forest compared with Terna's prediction. While both implementations of random forest showed a good performance in the prediction of the daily consumption of energy, Terna's model showed a lower error (RMSE: 8,796; MAPE: 1.05%). Linear random forest and classical random forest obtained an RMSE ranging from 11,853 to 16,886, and from 12,355 to 16,548, respectively, based on the different models we tested. As shown in Table 1, the best results were obtained by linear random forest in the configuration including 15 lags and the two difference variables. This configuration proved to be the best also for the classical implementation of random forest.

4 Conclusions

To conclude, we showed that random forest can provide accurate predictions even when used with time series. The two implementations of random forest used to forecast the energy consumption provides similar results and this might be due to, among other things, the specific properties of the time series used for the evaluation. As a future development we plan to further investigate the role of lags, differentiation and size of the training set.

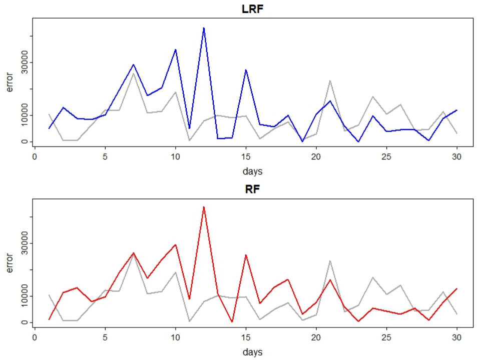

Figure 1. *Error of LRF (blue), RF (red) and Terna (grey) in the prediction of the electricity consumption. Abbreviations: LRF, linear random forest; RF, random forest*

Table 1. *Error of LRF and RF in the prediction of the electricity consumption*

Lags	Differences	RMSE LRF	RMSE RF	MAPE LRF	MAPE RF
5	-	14,740	14,400	1.78%	1.71%
15	-	15,241	15,643	1.80%	1.84%
30	-	16,886	16,548	1.99%	1.97%
5	2	15,585	13,076	1.88%	1.57%
15	2	**11,853**	**12,355**	**1.42%**	**1.48%**
30	2	12,763	13,236	1.54%	1.59%

In bold the best result (smallest error) for both models. Abbreviations: LRF, linear random forest; RF, random forest; RMSE, root-mean-square error; MAPE, mean absolute percentage error

References

Breiman, Leo. 2001. Random forests. *Machine learning*, **45**, 5–32.

Künzel, Sören R, Saarinen, Theo F, Liu, Edward W, & Sekhon, Jasjeet S. 2022. Linear aggregation in tree-based estimators. *Journal of Computational and Graphical Statistics*, **31**(3), 917–934.

R Core Team. 2023. R: A Language and Environment for Statistical Computing.

Terna S.p.A. 2023. https://www.terna.it, Last access: April, 5 2023.

Zhang, Liang, Wen, Jin, Li, Yanfei, Chen, Jianli, Ye, Yunyang, Fu, Yangyang, & Livingood, William. 2021. A review of machine learning in building load prediction. *Applied Energy*, **285**, 116452.