RESEARCH PAPERS
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA
SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA

2019 Volume 27, Number 45

ADDING ADDITIONAL FEATURES TO IMPROVE TIME SERIES PREDICTION

Dmitrii BORKIN[1], Martin NÉMETH[1], German MICHAĽČONOK[1], Olga MEZENTSEVA[2]

[1]SLOVAK UNIVERSITY OF TECHNOLOGY IN BRATISLAVA
FACULTY OF MATERIALS SCIENCE AND TECHNOLOGY IN TRNAVA
INSTITUTE OF APPLIED INFORMATICS, AUTOMATION AND MECHATRONICS
ULICA JÁNA BOTTU Č. 2781/25, 917 24 TRNAVA, SLOVAK REPUBLIC
e- mail: martin.nemeth@stuba.sk, dmitrii.borkin@stuba.sk, german.michalconok@stuba.sk

[2]THE INSTITUTE OF INFORMATION TECHNOLOGIES AND TELECOMMUNICATIONS OF NCFU,
355017 STAVROPOL, ST. PUSHKIN 1, RUSSIAN FEDERATION
e-mail: omezentceva@ncfu.ru

Received 19 August 2019, accepted 08 October 2019, published 29 November 2019

Abstract

This paper aims at the time-series data analysis. We propose the possibility of adding additional features to the existing time series data set, to improve the prediction performance of the prediction model. The main goal of our research was to find a proper method for building a prediction model for the time-series data, using also machine learning methods. In this phase of research, we aim at the data analysis and proposal of the ways to add additional features to our dataset. In this paper, we aim at adding derived parameters from one of the original features. We also propose incorporating LAG's into the dataset as new features, to enhance the prediction performance on the time series based data.

Key words

Time series, Prediction, Machine learning, Statistics

INTRODUCTION

Time series data can be found in various areas, like industry, economy, healthcare and so on. They are describing important processes and have huge knowledge potential. By analyzing these data we can obtain valuable information about these processes, and we can also build a prediction models to predict future values of the time series data. Such prediction models can improve the overall control, minimizing failures in the system and even maximizing the effectivity of the investigated process [2, 3]. Many times it is not enough to use only available original time series data and it is needed to search for additional features to improve the prediction performance. In this paper we aim at proposing two possible ways of adding new

features to the original dataset. Proposed methods are applicable in cases, where there are no external data to be added to the data set and it is needed to derive new features from the original data set [3, 4, 5].

DESCRIPTION OF THE TIME SERIES DATA

In this research, we are working with time series data from a thermal plant. The original dataset consists of two parameters. First is the date time feature and second parameter is the thermal power output. Whole dataset consists records from three years. The measurement frequency was set 10 minutes, which means 144 measurements / records per day.

The dataset can be seen in Figure 1, with time on the horizontal axis and thermal power output on the vertical one. At first sight it is clear that there is obvious seasoning in these data.

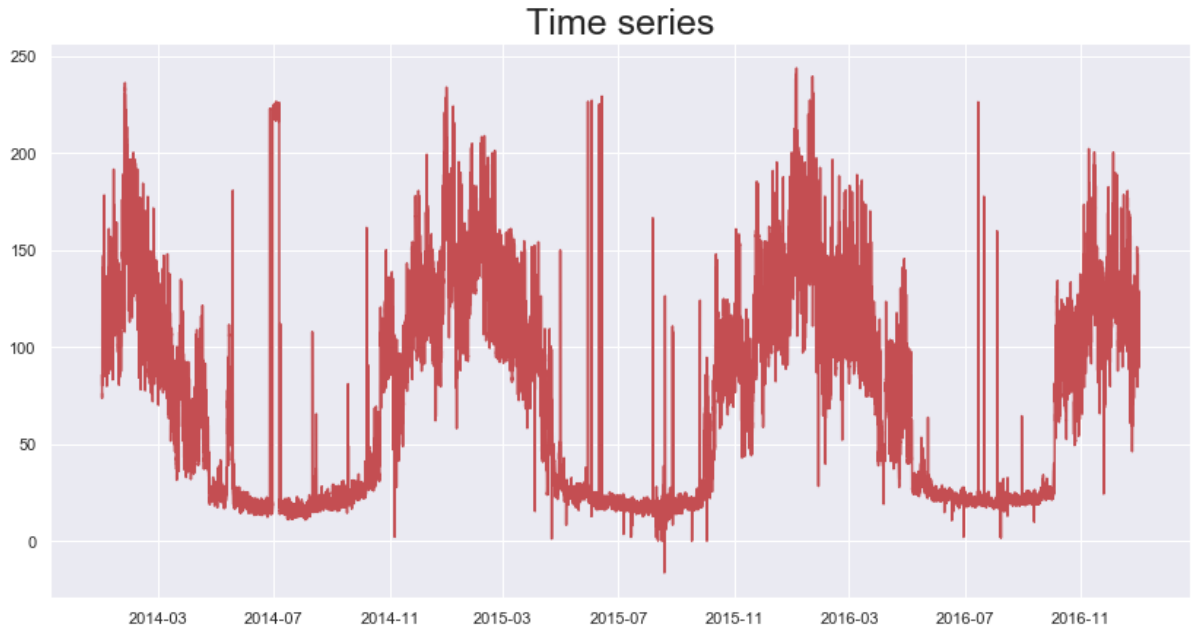

Figure 1 *The data set - obvious seasoning in the data*

Figure 2 shows the distribution of our target variable. In our case, the target variable is the thermal power output. These figures clearly indicate that the data can be described by the bimodal distribution.

This distribution is used with the data with two or more peaks. Bimodal distribution in general can indicate different groups in the data set, and also the fact that the data is of a sinusoidal character.

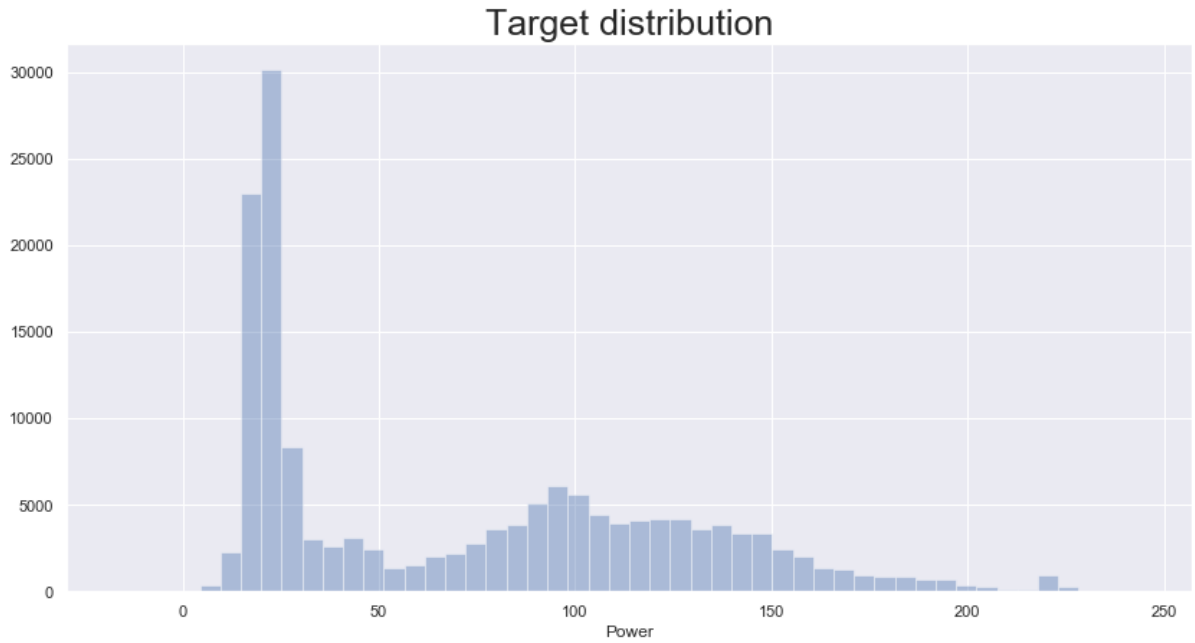

Figure 2 *Distribution of our target variable*

Adding new features to the data set

In this part of our paper, we would like to present adding additional features to our time series data set. One way of adding these new features to the time series data set is by using lags. When working with the time series data, we assume that we know all past values of the target variable.

By adding lags, we simply add new columns of data to the original dataset. These columns are filled with past values of the target variable. For example if we know that the information (value of our target) should not significantly change for 10 records, our new feature *lag01* will start at the 11-th record of the target, but with the first value of our target [1, 3, 5].

This concept of adding lags to our data set should improve the performance of the prediction model. The amount of added lags depends on the performance result of the prediction model.

After adding lags to our dataset, we also added other additional features. We derived these features from the existing original feature of the dataset (date-time). From this feature, we were able to derive the parameters such as *year, hour, day of the week, and weekend*. All additional features including lags were added by using Python libraries.

The accuracy of the model will be evaluated using MAPE (mean absolute percentage error) [2, 3, 4]:

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i} \tag{1}$$

Testing new features with a simple linear regression model

After adding new features to our dataset, we tested how much they would improve prediction performance. For testing purposes, we decided to use a simple linear regression model. We tested the additional parameters in three scenarios.

In the first test case, we ran the regression model twice: first, only with the original dataset, and then, with lags, without using the original target variable. After running these testing scenarios, we compared the performance of each case, and we also computed the mean absolute percentage error, which, in this case, was equal to 10.05%. Figure 3 shows a plot of the two cases; it is clear that the model was sufficiently accurate.

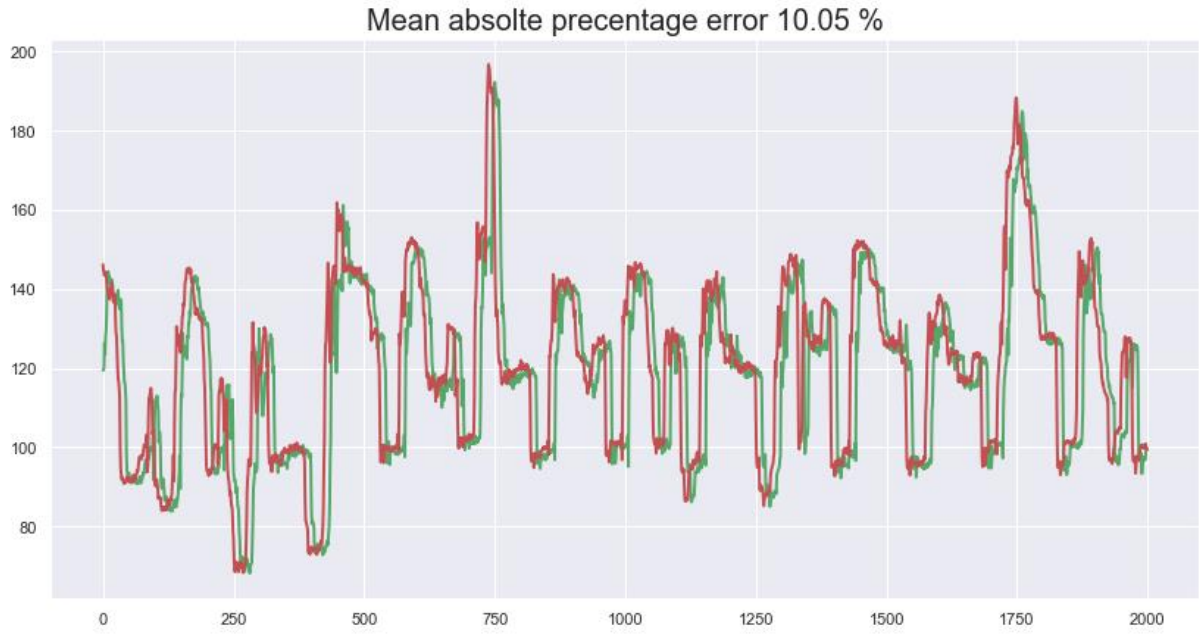

Figure 3 *Plot of two cases; t is clear that the model was sufficiently accurate (red line – original data, green line – prediction)*

For the second test scenario, we used the derived parameters instead of using lags. Figure 4 shows the difference between running a regression model with the original target variable and with only the derived parameters. The mean absolute percentage error is much higher than in the first scenario, where lags were used. In this case, MAPE is equal to 18.15%.

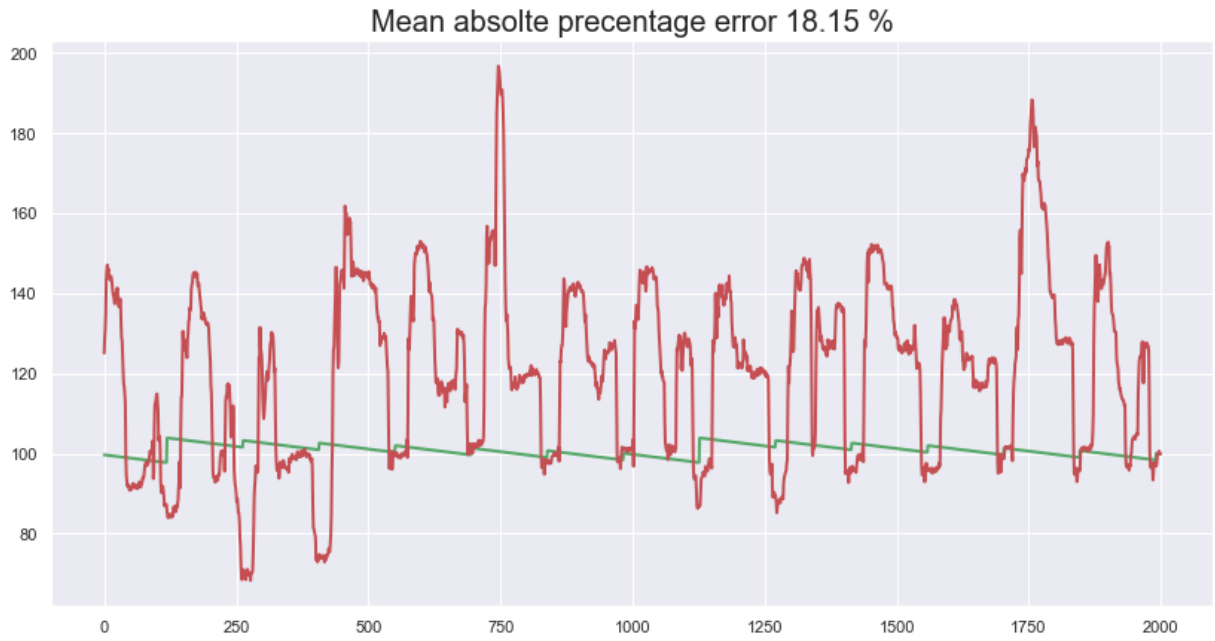

Figure 4 *The difference between running a regression model with the original target variable and with only the derived parameters (red line – original data, green line – prediction)*

In the third test scenario, we decided to run the regression model with both lags and new derived parameters. Figure 5 shows also the plotted results and differences between using the original target variable and the group of additional parameters. In this case, we can see that the mean absolute percentage error is equal to 9.56%, which is less than in the first scenario, where only lags were used. This result makes it clear, that the combination of lags and derived parameters can improve the prediction performance of the linear model.

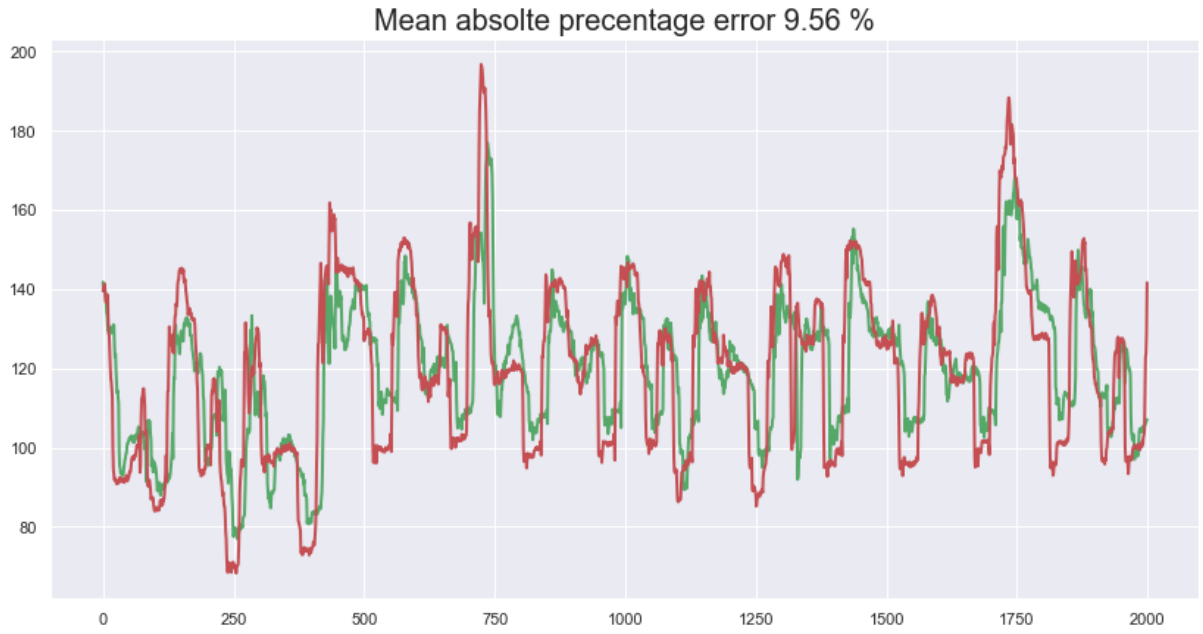

Figure 5 *Results and differences between using the original target variable and the group of additional parameters (red line – original data, green line – prediction)*

ATTAINED RESULTS

Figure 6 illustrates the importance plot of the additional parameters in our dataset. As we can see, the combination of lags and derived parameters can give us better results in respect of the mean absolute percentage error. However, Figure 6 shows that the most important parameter is the lag_12, followed by the parameter of year. In this plot, we can also see the lags which have only little to no effect on the regression performance. Based on this Figure, we can remove these lags from our dataset, which can lead to a smaller data file, thus reducing the computing requirements to run a prediction model.

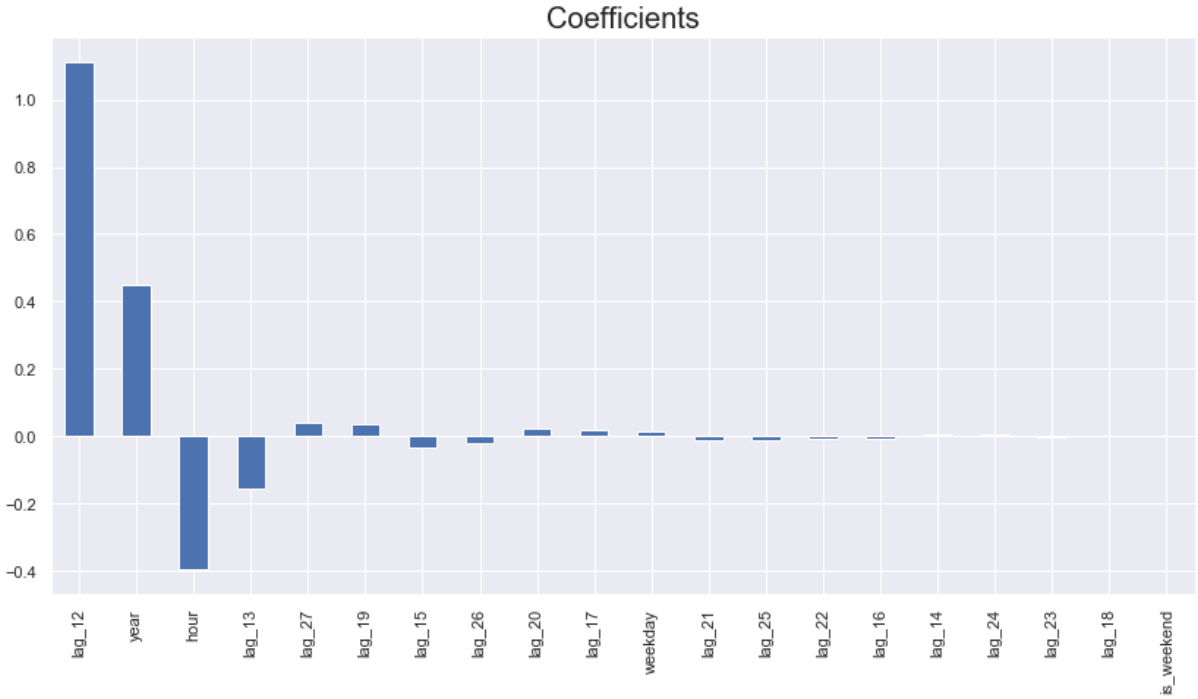

Figure 6 *The importance plot of the additional parameters in our dataset*

CONCLUSION

In this paper, we were dealing with adding new features to the dataset to enhance the performance of the prediction model. In our research, we dealt with the time series data, which is a special type of data that should be treated with certain methods. Time series data can be used to predict future behaviour of the examined process based on the previous values and trends of the target variable. In our research, we dealt with the time series data from a thermal plant; the target variable in this case was the thermal power output measured in megawatts. The original dataset contained only the date time and the target variable. In this paper, we presented how to add new features to the data set, and we also tested the impact of these new features on the prediction performance. For testing purposes, we decided to use a simple linear regression model. In conclusion we can say that the use of additional features can improve the prediction performance of the prediction model. In our future work, we will aim at building a machine learning based prediction model to predict the future values of the target variable.

Acknowledgement

This publication has been written thanks to support of the Operational Program Research and Innovation for the project: Research of advanced methods of intelligent information processing; ITMS code: 313011T570 co-financed by the European Regional Development Fund.

References

[1] BLUM, Avrim L., LANGLEY, Pat. 1997. Selection of relevant features and examples in machine learning. Artificial intelligence, 97.1-2: 245-271.
[2] LANGLEY, Pat. 1994. Selection of relevant features in machine learning. In: *Proceedings of the AAAI Fall symposium on relevance*, pp. 1-5.
[3] BONTEMPI, Gianluca, TAIEB, Souhaib Ben, LE BORGNE, Yann-Aël. 2012. Machine learning strategies for time series forecasting. In: *European business intelligence summer school*. Springer, Berlin, Heidelberg, pp. 62-77.
[4] MITCHELL, Tom M. 1999. Machine learning and data mining. *Communications of the ACM*, 42.11.
[5] FAWCETT, Tom; PROVOST, Foster J. 1996. Combining Data Mining and Machine Learning for Effective User Profiling. In: *KDD*, pp. 8-13.
[6] MITCHELL, Tom M. 1997. Does machine learning really work? *AI magazine*, 18.3: 11-11.
[7] KOSTELICH, Eric J., SCHREIBER, Thomas. 1993. Noise reduction in chaotic time-series data: A survey of common methods. *Physical Review E*, 48.3: 1752.
[8] BAR-JOSEPH, Ziv. 2004. Analyzing time series gene expression data. *Bioinformatics*, 20.16: 2493-2503.