Optimizing Short Term Load Forecast: A study on Machine Learning Model Accuracy and Predictor Selection

Pande Popovski[1], Goran Veljanovski[1], Mitko Kostov[1] and Metodija Atanasovski[1]

Abstract – **This paper focuses on the importance of choosing the proper predictors when training a Machine Learning model for Short Term Load Forecasting, as well as to demonstrate the usefulness of Machine Learning in the field of power load forecasting. For the goals of the study, a correlation analysis was performed in order to observe the impact of some factors on the changes of power consumption. In addition, a number of models were created using machine learning where combinations of predictors were used based on their correlation to power load. The performance of these models was evaluated and the results are shown in this paper.**

Keywords – **Short Term Load Forecast, Machine Learning, Regression analysis, Correlation analysis, Decision Trees, Support Vector Machines.**

I. Introduction

Load forecasting represents a vital part of the process of planning and exploitation of the power system. Each of the different types of load forecast which we are familiar with, more specifically Short, Medium and Long Term Load Forecasting (abbreviated as STLF, MTLF and LTLF), allow us to effectively assume how power system load might change over a short and relatively longer period of time. Having this information available greatly helps with maintaining the balance between power generation and consumption, as well as planning the infrastructure of the power system itself.

With this in mind, the usage of new and improved methods for power load forecasting should always be a priority, as it would greatly benefit the process of planning and exploitation.

Machine learning (ML) is a versatile tool which has proven its usefulness across many fields, including its usage in power load forecasting. ([1],[2],[3],[4]).

This paper focuses on the importance of choosing the proper predictors when training a ML model for STLF, as well as to demonstrate the usefulness of ML in the field of power load forecasting. The study uses data for the power system of the Republic of North Macedonia.

For the goals of the study, a correlation analysis is performed on a number of factors in order to observe their impact on the changes of power load. After this, two types of machine learning algorithms are used to train a number of models where combinations of predictors are implemented based on their correlation to power load. The performance of these models is evaluated and the results are shown in this paper.

The paper is organized as follows. Section II briefly covers factors that have the highest influence on the accuracy of a load forecast. Section III covers the correlation analysis performed for the goals of the study. Section IV covers the ML model analysis, and Section V concludes the paper.

II. Factors Influencing Forecast Accuracy

There are a number of factors that have a noticeable influence on the accuracy of a load forecast. This is due to the fact that these factors have an influence over the behavior of consumers, and with this on the changes of power system load itself. By effectively combining them as input variables, we could make improvements in the accuracy of our load forecasts.

The following classification can be made [5]: Time factor, Economic factor, Weather factor and Customer factor.

A. Time factor

Regardless of what time span is being observed (day, week, month etc.), a trend can be noticed in the way system power load changes. Typical load curves can be constructed for these time periods, which to a certain degree can give a correct assumption on how power load would change in the future.

The behavior of consumers throughout a given time period explains this. An example would be how we spend our time throughout the day: part of our day is spent working, which would increase power consumption in the commercial sector. In the evenings, we spend time at home which would increase power consumption in the domestic sector.

B. Economic factor

Circumstances concerning the economy can also have a drastic influence over the changes in power consumption. Most often, attention is given to economic factors when conducting a LTLF. However they have noticeable importance and influence in regard to the remaining two types of load forecast as well. A few examples of economic factors would be the following [5]:

- *Industrial development*: development in the industry of a certain region cause an increase in power consumption;
- *Population increase*: an increase in population would also causes an increase in power consumption;

[1]Pande Popovski, Goran Veljanovski, Mitko Kostov and Metodija Atanasovski are with the Faculty of Technical Sciences-Bitola, St. Kliment Ohridski University, Republic of N. Macedonia, E-mail: pande.popovski@uklo.edu.mk

978-1-6654-8500-5/22/$31.00

- *Price of electricity*: a drop in electricity price is usually followed with an increase in power consumption, and vice versa.
- *Time of use*: time of use pricing has an effect on the duration and the time of occurrence of peak load.

C. Weather factor

Weather conditions are closely correlated with power system load. Changes in weather have a direct influence over the behavior of consumers. Factors that are often taken into consideration when conducting a forecast are air temperature and humidity, solar radiation and wind speed.

- *Air temperature*: changes in the temperature can directly affect power consumption. During the winter period the use of heating appliances increases. Similarly for the summer period, the use of cooling appliances increases. A detailed analysis on the correlation between air temperature and system power load is shown in ([6],[7]);
- *Air humidity*: air humidity to a certain extent has influence over changes in power consumption. Generally higher humidity levels are followed by an increase in power load and vice versa;
- *Wind speed*: windy days in general are also followed by an increase in power consumption;
- *Solar radiation*: also very closely correlated with power load. Solar radiation has an effect on the changes of air temperature. Cloudy days are characterized with lower air temperatures, thus the increase in power consumption.

The correlation between the aforementioned factors and system power load is studied in [8]. The study concludes that air temperature and solar radiation have a strong correlation to power load. The study shows a change in this correlation when separate seasons of the year are observed.

D. Consumer factor

Electricity producers serve a broad range of consumers which can be categorized into domestic, commercial and industrial load. Things such as the number of load units, their category and their size can be considered consumer factors. A typical load curve for each type of consumer can be constructed based on these mentioned factors [5].

III. Correlation Analysis

Before the ML models were trained, a correlation analysis was performed using a number of different variables. The correlation between these variables and system power load was observed. This in practice would be of help in the process of choosing predictors for the ML models.

The data used for the correlation analysis consists of the hourly values for the following variables, given for the years 2015-2018 (35064 observations in total):

- *Hour of day (HR)*: values 0-23;
- *Day of year (DY)*: values 1-365(366 for leap years);
- *Day of week (DW)*: values 1-7 (Monday to Sunday);
- *Month of year (YM)*: values 1-12 (January to December);
- *Season of year (YS)*: values 1-4 (Spring to Winter);
- *Holiday (HD)*: values 0 or 1 (No/Yes);
- *Weekend (WE)*: values 0 or 1 (No/Yes);
- *Air temperature (T)*: expressed in ^{o}C;
- *Air humidity (H)*: expressed in %;
- *Wind speed (W)*: expressed in m/s;
- *Cloud cover (C)*: expressed in %;
- *Power load previous day, same hour (PL)*: express. in MW;
- *Power load (current hour)*: expressed in MW.

Data for the weather conditions was acquired from the Internet [9], and it originates from a weather station located in the city of Skopje. The data for the power load is for the power system of the Republic of North Macedonia. The last variable on the previous list is the dependent variable, whereas all the rest are the independent variables.

Two measures were used to test the correlation. One is the Pearson's correlation coefficient (PCC), denoted by R, which can have values from -1 to 1. PCC indicates the strength of a linear correlation between two variables, and its value can be interpreted in the following way:

- $R = \pm 1$ - a perfect positive or negative linear correlation;
- $0,0 < |R| < 0,09$ - very weak linear correlation;
- $0,1 < |R| < 0,29$ - weak linear correlation;
- $0,3 < |R| < 0,49$ - a moderate linear correlation;
- $0,5 < |R| < 5$ - a strong linear correlation;
- $R = 0$ - no linear correlation.

The second measure used is the Distance correlation coefficient (DCC). This measure is used as an indicator for the strength of a correlation that is not necessarily linear. It can have values from 0 to 1, and can be interpreted the same way as PCC. When DCC equals 0, it means that there is no correlation whatsoever between the two variables.

The correlation between the variables and the system power load was analyzed using two approaches: the first time the entirety of the data was used to compute the coefficients, and the second time separate seasons of the year were observed. The results are shown in Tables 1 and 2.

The more notable values for the coefficients in Tables 1 and 2 are bolded. The results show that the following variables: system load for previous day (same hour), air temperature, season of the year, hour of day, day of year and month of year might be suitable as predictors in a ML model because of their relatively high correlation to the dependent variable.

It is important to consider that correlation does not equal causation: only by experimenting with these variables we can confirm that they truly are suitable for use as predictors.

978-1-6654-8500-5/22/$31.00 ©2022 IEEE

TABLE I
CORRELATION ANALYSIS RESULTS (ENTIRE PERIOD 2015-2018)

Measure	HR	DY	DW	YM	YS	HD	WE	T	H	W	C	PL
R	**0,45**	-0,08	0,02	-0,09	**0,48**	-0,02	0,00	**-0,51**	0,05	0,07	0,19	**0,96**
DCC	**0,45**	**0,34**	0,03	**0,34**	**0,50**	0,02	0,02	**0,57**	0,20	0,09	0,19	**0,95**

TABLE II
CORRELATION ANALYSIS RESULTS (SEASONS)

Measure	Season	HR	DY	DW	YM	HD	WE	T	H	W	C	PL
R	Spring	**0,49**	**-0,51**	0,04	**-0,51**	-0,12	0,01	**-0,33**	-0,12	0,12	0,28	**0,94**
	Summer	**0,67**	0,12	-0,01	0,11	-0,01	-0,04	**0,58**	**-0,57**	0,26	0,01	**0,93**
	Autumn	**0,60**	**0,57**	0,01	**0,53**	-0,07	-0,01	-0,26	-0,03	0,12	0,10	**0,95**
	Winter	**0,58**	0,12	0,05	0,11	-0,01	0,01	-0,23	-0,20	0,05	-0,02	**0,91**
DCC	Spring	**0,48**	**0,49**	0,04	**0,49**	0,12	0,03	**0,38**	0,23	0,14	0,27	**0,93**
	Summer	**0,72**	0,14	0,04	0,12	0,02	0,04	**0,55**	**0,55**	0,26	0,11	**0,92**
	Autumn	**0,61**	**0,55**	0,03	**0,50**	0,07	0,02	**0,35**	0,21	0,12	0,10	**0,94**
	Winter	**0,59**	0,15	0,06	0,15	0,02	0,04	0,22	0,22	0,05	0,04	**0,89**

IV. MACHINE LEARNING MODEL ANALYSIS

The data used for the training of the ML models is the same one used for the correlation analysis (years 2015-2018, 35064 observations). The models trained using this dataset are then tested using a separate set of data, which consists of the hourly values for the same variables mentioned before, except for the year 2019 (8760 observations).

The study is conducted in the following way: a number of training iterations are performed, such that in the first iteration only one predictor is used. In the following iterations additional predictors were added and the change in model accuracy is monitored through each step. If a variable had a positive effect on model accuracy it is kept for the following iterations, and if not then it is removed. The process is also known as the stepwise training method.

The correlation results from the previous analysis helped in choosing the most suitable predictors.

Two ML algorithms were used to train the models: Decision Trees (DT) and Support Vector Machines (SVM). The performance of the trained models is tested using data from the year 2019 as previously mentioned. Hourly forecasts for system power load are conducted, and the forecasted values are compared to the real ones. The following measures are used to evaluate model performance:

- *Mean Absolute Error – MAE* (measured in MW);
- *Mean Squared Error – MSE* (measured in MW^2);
- *Root Mean Squared Error – RMSE* (measured in MW);
- *Coefficient of Determination – R^2* (unitless);

Of the previous measures, RMSE can be considered the most practical when evaluating the performance of the model.

In total, fifteen training iterations were performed. The model which showed the best results used the following variables as predictors: Hour of day, Day of year, Day of week, Holiday, Air temperature and Power load from the previous day. The test results from this model are shown in Table 3, where a comparison is made with results obtained in a previous study [3]. The indicators in the results refer to the error the model has made while forecasting the entire year of 2019. Results that follow are shown for the model trained with SVM, as this model showed more accurate forecasts.

TABLE III
FORECAST RESULTS FOR YEAR 2019

	This study (SVM)	Forecast results from [3]		
		Bagged trees	Gaussian SVM	Exponential GPR
MAE [MW]	30,04	43,63	41,44	40,83
MSE [MW^2]	1758,13	3447,72	3064,59	3000,08
RMSE [MW]	41,93	58,71	55,36	54,77
R^2	0,96	0,94	0,95	0,95

The average system load for the year 2019 was 842,47 MW, whereas the forecasted value is 838,31MW.

The minimum load for the power system occurred on October 21, 03:00 AM (352 MW). The model forecasted a minimum power load on October 22, 03:00 AM (432,82 MW).

Maximum power load occurred on January 9, 03:00 PM (1466 MW). The model forecasted a maximum power load on January 13, 03:00PM (1411,2 MW).

The minimum summer power load occurred on July 11, 03:00 AM (430 MW). The model forecasted a minimum summer load on July 12 (453,73 MW).

Forecasts for a number of daily diagrams were also created using the model and compared to the real daily diagrams for those days. Results from a few daily diagram forecasts are listed. The RMSE calculated for these daily diagrams is from the 24 hours of that day.

- *January 9*: RMSE = 36,95 MW;
- *February 1*: RMSE = 10,19 MW (day with smallest error);
- *July 11*: RMSE = 62,61 MW;
- *August 13*: RMSE = 14,31 MW;

978-1-6654-8500-5/22/$31.00 ©2022 IEEE

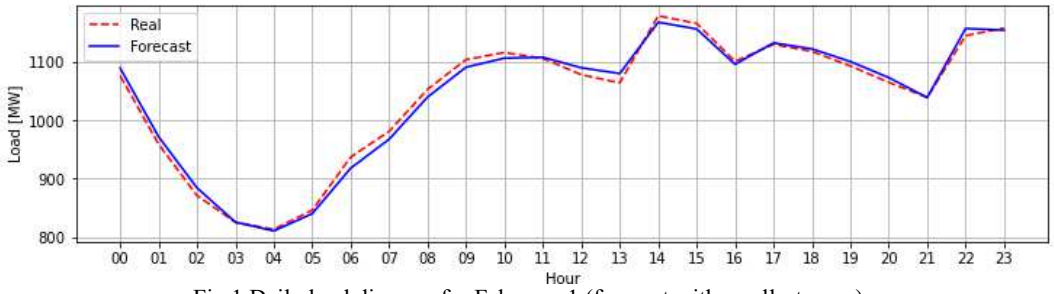

Fig.1 Daily load diagram for February 1 (forecast with smallest error)

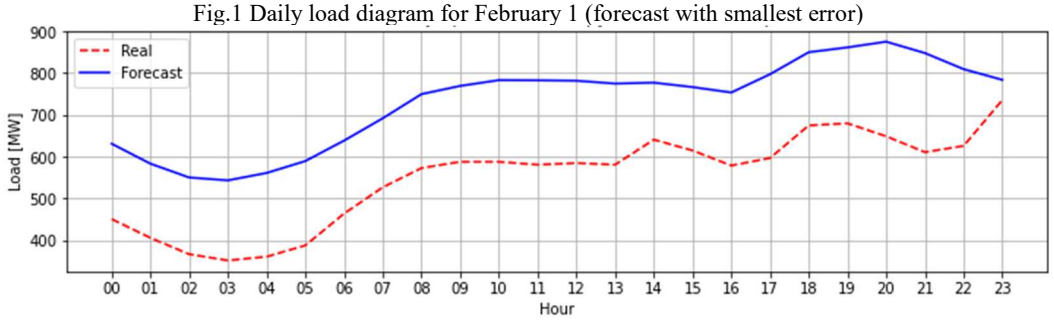

Fig.2 Daily load diagram for October 21 (forecast with largest error)

- *October 21*: RMSE = 183,93 MW (day with largest error);
- *December 1*: RMSE = 50,46 MW;

Figure 1 shows the daily diagram, both forecasted and real for February 1. This is the day for which the model made the smallest forecast error. Figure 2 shows results for October 21, the day for which the model made the largest forecast error.

Interesting to note, the system power load for October 21 during the years 2015-2018 (training dataset) hasn't dropped below 500 MW, whereas for the year 2019 the minimum power load for that day was 352 MW. This leads to an assumption that the large error for this day is owed to this.

V. Conclusion

A correlation analysis was performed on a number of factors in order to observe their impact on the changes of power consumption. After this, a number of ML models were trained where combinations of predictors were used based on their correlation to power load. These models were then used to conduct hourly load forecasts.

Judging by the performance of the trained models, the results are quite satisfactory. The models forecast peak and valley loads, as well as the whole daily load curves with a relatively high accuracy. There are exceptions, like the example shown with the hourly forecast for October 21.

The weather data used in the study is from the city of Skopje, whereas the power load data is for the whole power system. Better results might be achieved if forecast was conducted on separate parts of the power system located in specific regions, where weather data for that region in particular would be used. Additional variables, such as solar radiation should be used in future studies.

Acknowledgement

This research is supported by the EU H2020 project TRINITY (GA no. 863874) This paper reflects only the author's views and neither the Agency nor the Commission are responsible for any use that may be made of the information contained therein.

References

[1] M. Kostov, M. Atanasovski, G. Janevska, and B. Arapinovski, "Power System Load Forecasting by using Sinuses Approximation and Wavelet Transform", ICEST - Ohrid, North Macedonia, 27-29 June 2019.

[2] M. Atanasovski, M. Kostov, B. Arapinoski, M. Spirovski: "K-Nearest Neighbour Regresion for Forecasting Electricity Demand", *Proceedings of papers 55th ICEST 2020*, Nis, Republic of Serbia, 2020.

[3] P. Popovski, M. Kostov, **M. Atanasovski,** G. Veljanovski,: " Power Load Forecast for North Macedonia Using Machine Learning", *Proceedings of papers 55th ICEST 2020*, Nis, Republic of Serbia, 2020

[4] G. Veljanovski, M. Atanasovski, M. Kostov, P. Popovski: "Application of Neural Networks for Short Term Load Forecasting in Power System of North Macedonia", *Proceedings of papers 55th ICEST 2020*, Nis, Republic of Serbia, 2020

[5] N. Phuangpornpitak, W. Prommee: „A Study of Load Demand Forecasting Models in Electric Power System Operation and Planning",GMSARN International Journal 10, 2016.

[6] M. Atanasovski, M. Kostov, B. Arapinoski, I. Andreevski: „Correlation between Power System Load and Air Temperature in Republic of Macedonia ", ICEST, 2018, Sozopol Bulgaria.

[7] M. Kostov, M. Atanasovski, B. Arapinoski, G. Janevska: "Comparison of the Sine and Polynomial Approximation of the Dependence of the Power System Load on the Air Temperature", Cigre North Macedonia, 2019.

[8] L. Hernández, C. Baladrón, J. M. Aguiar, L. Calavia, B. Carro, A. S.-Esguevillas, D. J. Cook, D. Chinarro, J. Gómez: „A Study of the Relationship between Weather Variables and Electric Power Demand inside a Smart Grid/Smart World Framework", Sensors Journal, 2012.

[9] https://openweathermap.org/