On Linear Regression and other advanced algorithms for electrical load forecast using weather and time data

K R Khadiev and L I Safina
Kazan Federal University, Kazan, Russia
E-mail: kamilhadi@gmail.com

Abstract. Energy field plays an important role in commercial world and makes a vital part of humanity. The paper considers the prediction problem concerning spent electrical energy in the Republic of Tatarstan, Russia. In fact, the task of electrical load prediction was set for each hour of the specified period. We solve the problem using Machine learning methods. Four models are considered. These are Linear Regression, Decision Tree, Random Forest and Gradient Tree Boosting. Linear Regression method shows good results, the average relative error is 3.98%. Decision Tree and Random Forest show the worst result, the average relative error is 10.44%. Gradient Tree Boosting show the best result, the average relative error made 2.17%. At the same time, Linear Regression model is much faster than other ones and more useful in industry. In the paper we show that using several techniques we can improve results for Linear Regression, such that it will be close to another advanced algorithms. The average relative error that is less than 5% is considered as a high enough result. The solution of the problem with a small error allows us to prevent the accidents related to electric overload. We assume that load depends on date, time and temperature. Then input variable models were obtained from these data. It is common approach to use that parameters as input data but let us note that the results obtained by other researchers are not suitable for this area, since each area has individual climatic, geological and social features.

1. Introduction
Today energy world is really big commercial marketplace. We should use energy optimally including electric energy.

Cost saving, different necessary value energy at different time, prediction accident on electric networks. This is reason, why companies wish to predict electrical load [1–3].

We solve this problem for Republic Tatarstan in Russia using machine learning approach. Such method is popular in USA and Europe. But decision for USA or Europe cannot be used directly because problem of an electrical load prediction is individual for each country. Each country has own weather, geography, social and economic situation. These options have influence for required energy. Note that, even in Russia results for one regions are not applicable for other region. The reason is a big area of the country and different weather and economic conditions.

Researchers use different machine learning models for the problem [4]. The most frequent models are the artificial neural networks [5]. But other models are shown well in use. The solutions of load forecasting problem for the city of Ulyanovsk, for example, were obtained by the recovery of the regression equations. The researcher obtained rather high results. The same method was used by the researchers from the United States. Serbian scientists, for example, used fuzzy logic methods. We use
following algorithms and models: Gradient Tree Boosting [6] Random Forest [7], Decision Tree [8, 9] and Linear Regression [10].

The best result is received on method Gradient Tree Boosting (the average relative error was 2.17%). At the same time, the running time of this method is big. We decided to improve results on Linear Regression. After adding and changing our input data model, we get the average relative error 3.98% for Linear Regressions. The worst value is 10.44%. It was obtained using the methods Decision Tree and Random Forest.

Main steps of the solution are Preprocessing and applying model. In Section 2 we discuss models. Section 3 contains results for Gradient Tree Boosting, Decision Tree and Random Forest models. We apply additional methods of preprocessing for models for getting better result in Section 4. Section 5 contains conclusion and analysis of the results.

2. Models
There are different prediction methods for solving the problem. One of them is developing physical model. However, machine learning methods are cheaper and effective. Machine learning allows to reveal complex patterns due to the large volume of collected statistical data. You can learn more about machine learning models in [11–13]. Machine learning is very useful tool that used in different areas, for example [14].

Let us introduce basic definitions. Let us consider a problem of machine learning. The input \( X = \{x_1, \ldots, x_N\} \) is a set of \( N \) features. Feature is the result of some object characteristic measurement. \( Y \) is the correct solution of the problem for the input \( X \). \( Y' \) is the solution that is returned by a model (algorithm) on the input \( X \).

The algorithm uses a set of input and output objects for learning. This set is called a training set.

2.1. Linear Regression
Linear regression is the method based on mathematical statistics approach. We assume that a target function \( F \) is linear depends on input features. We consider all input attributes as real numbers and construct function in the following form: \( F(x) = a_0 + \sum_{i=1}^{N} a_i x_i. \) Here \( a_i \) is a coefficient, \( x_i \) is a parameter (feature), \( N \) is a number of parameters.

A main problem for linear regression method is a finding the coefficients \( a_i \) that reflect actual results most significantly. We obtain \( a_i \) by minimization of sum for value concerning error squared (squared difference between the actual value and the received built function). The method is based on Euclidean distance minimization between the vector of actual results and the vector, obtained by regression recovery.

2.2. Decision Tree and Random Forest
Decision Tree is a set of rules and conditions for the input features and operation outcomes that provide some results. The determination of rules is tree constructing process that is invoked on a training set. The leaves are the output results. During the solution of the regression recovery problem the tree leaves are real numbers. Random Forest is a model based on Decision Tree and uses bagging method and probability mechanism.

2.3. Gradient boosting
The Gradient Tree Boosting is the model which uses a decision tree and boosting method. The model is the sum of decision trees. New tree components are added to the sum by the greedy empirical risk minimization. The lost function is based on the obtaining of the least absolute deviation. The combination of several machine learning methods allows to improve results considerably. Boosting is the procedure of an algorithm composition consistent development for machine learning, where each following algorithm seeks strives to compensate the shortcomings of the previous algorithm compositions. The gradient boosting is sufficiently stable to retraining, thus a large number of component trees tend to influence the results positively.
3. Results for Gradient Tree Boosting, Decision Tree and Random Forest Models

Let us discuss a base result that is given by applying the models of machine learning. We will improve the results in the next section.

3.1. Input Data

Time and weather conditions are considered as input parameters. Electrical load values for the current hour (the shares of electricity) are output data.

Grid load data are taken from the website of OJSC “Tatenergosbyt” that is the company which is a guaranteeing supplier at the territory of the Republic of Tatarstan [15]. The data represent electricity volume share from every hour of the monthly energy consumption by a guaranteeing supplier from the wholesale and retail markets for each month within the period of 2010-2014 (since January 2010 till December 2014).

Climatic conditions in the Republic are taken from the site of the weather schedule [16]. Weather data are presented on the website with 3 hours step for the period from January 1, 2010 to December 31, 2014.

The following time and climatic parameters are taken as the input features:

- day of a month (1-31);
- month number (1-12);
- year (2010-2014);
- day of a week (1-7);
- hour (0-23);
- working day/day off (1 or 0, As day offs we consider Saturday, Sunday, public holidays);
- temperature.

3.2. Preprocessing

We use One-hot-encoding method for achieving the independence of influence on the result for each day of the week. The method replaces the variable $x$ which is equal to a week day $i$, by 7 variables $x_k$ according to the following rule: $x_k = 1$, if $x = k$; and $x_k = 0$, otherwise. For example, if $x = 3$ then $(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = (0, 0, 1, 0, 0, 0, 0)$.

The One-hot-encoding algorithm is applies to all categorical attributes.

Finally, we get following input variables:

- $x_1 = $ a day of a month (range of values: 1 - 31);
- $x_2 = $ a number of a month (range of values: 1 - 12);
- $x_3 = $ a year (range of values may be presented by any natural number);
- $x_4 = $ the first day of a week (Is it Monday today? Yes is 1, No is 0);
- $x_5 = $ the second day of a week;
- $x_6 = $ the third day of a week;
- $x_7 = $ the fourth day of a week;
- $x_8 = $ the fifth day of a week;
- $x_9 = $ the sixth day of a week;
- $x_{10} = $ the seventh day of a week;
- $x_{11} = $ the value of the current hour (value range: 0 - 23);
- $x_{12} = $ a working day or day off (Is it Saturday, Sunday or public holiday today? Yes is 1, No is 0);
- $x_{13} = $ a temperature (an any real number);
- $x_{14} - x_{25} = $ months;

3.3. Preprocessing

We evaluate the models by the average relative deviation of the obtained result from the correct value. Maximum and minimum values of a relative deviation are analyzed for each model. The average relative deviation is calculated by the following formula: 

\[ \sum_{i=1}^{M} \frac{|Y_i - Y'_i|}{M}, \]

where \( Y \) is the correct result, \( Y' \) is the predicted result, \( M \) is a number of elements in a test set.

The percentage ratio of tests is also evaluated where a relative error is more than 5% using formula:
\[ \frac{k}{M} \]

where \( k \) is a number of tests, at which the relative error \( R_i = \frac{|Y_i - Y'_i|}{Y_i} \) is more than 5%.

### 3.4. Applying models

Results of models work are presented in Table 1. The gradient boosting method provides the best result: the average relative error is 2.17%. The best result is showed by the system using 500 of such trees, the depth of which is equal to 10 (results are worse at lesser depth). Researchers have not use this model to solve the problem of electricity load forecasting before.

Linear regression shows a relatively good result. This result is also important, because it is one of most popular methods for industry. The reason is high computational speed. Decision tree provides the worst values.

| Method              | Mid  | Min    | Max    |
|---------------------|------|--------|--------|
| Linear Regression   | 7.91%| 0.024% | 37.25% |
| Decision Tree       | 10.44%| 0.00096%| 37.81% |
| Gradient Boosting   | 2.17%| 0.00044%| 18.42% |

Note, that in case of GTB model we have less than 7.5% of tests that have error more than 5%. It is normal, because this situation corresponds to Force Majeure.

The obtained predictions may positively influence the performance of supplying companies and customers. The results allow to prevent most emergencies, to purchase the required amount of electric energy and minimize costs.

#### 4. Improving Results for Linear Regression

Recall, that Linear Regression is fastest method for forecasting. But it is sensitive to preprocessing. So it is typical situation when preprocessing gives benefits. We try to improve results for the model.

##### 4.1. Preprocessing

We add season and holidays. We take day of week and season as category variables. After that we apply one-hot-encoding method to these variables. These changes give us small improvement. The error becomes 7.42%.

After that new characteristic is added, it is peak hours. The peak hours are time when people use more electrical energy. This time is morning and evening. And we have experience with different value peak hours. We get improvement and we decide to apply the one-hot-encoding method for value of current hour. These steps give us profit. Error became 4.07%.

We add another one characteristic. It is value of temperature raised to the 4 degree Celsius. We try to consider a temperature from 2 to 8 degrees. The best result is obtained at 4 degree. The average relative error becomes 3.98%.

So to solve the problem we use the following characteristics:

- \( x_1 \) = a day of month number (integer number from 1 to 31),
- \( x_2 - x_5 \) = a season by using method one-hot-encoding (0 or 1),
- \( x_6 - x_{17} \) = a number of month by using one-hot-encoding (0 or 1),
- \( x_{18} \) = a number of year (integer number, current year),
- \( x_{19} - x_{42} \) = the current hour by using one-hot-encoding (0 or 1),
- \( x_{43} \) = a value of temperature (real number),
- \( x_{44} \) = a value of temperature raised to the 4 degree (non-negative real number),
• $x_{45} - x_{51}$ = a number of week by using one-hot-encoding method (0 or 1),
• $x_{52}$ = public holidays in Russia (1 or 0).
After these procedures we get following results:

| Method                    | Mid    |
|---------------------------|--------|
| Linear Regression         | 3.98%  |
| Decision Tree             | 10.44% |
| Gradient Boosting         | 2.17%  |

**Table 2.** The relative errors.

5. Conclusion

The prediction of electrical load by Linear Regression gave good results. Information technology helps companies-supplies electricity to get fast, correct enough and useful method to predict required value.

Linear Regression is faster than other considered methods of machine learning. Choosing the right features as input variables helped us to improve the result for linear regression. As possible additional variables we can add economic characteristics. We hope that it will help to get better results.

Acknowledgments

The research was funded by the subsidy allocated to Kazan Federal University for the state assignment in the sphere of scientific activities, project no. 1.13556.2019/13.1.

References

[1] Domanov V I and Bilalova A I 2014 Analysis of forecasting the energy consumption with various data bases
[2] Hong T et al. 2010 Modeling and forecasting hourly electric load by multiple linear regression with interactions *Power and Energy Society General Meeting IEEE* 1–8
[3] Ruzic S, Vuckovic A and Nikolic N 2003 Weather sensitive method for short term load forecasting in Electric Power Utility of Serbia *IEEE Transactions on Power Systems* Vol 18 1581–1586
[4] Lu C N, Wu H T and Vemuri S 1993 Neural network based short term load forecasting *Power Systems, IEEE Transactions on*. 336–342
[5] Dayhoff J E and De Leo J M 2001 Artificial neural networks *Cancer* S8 1615–1635
[6] Friedman J H 2002 Stochastic gradient boosting *Computational statistics & data analysis* 38 (4) 367–378
[7] Liaw A and Wiener M 2002 Classification and regression by random Forest *R news* 2 (3) 18–22
[8] Breiman L, Friedman J, Olshen R and Stone C 1984 Classification and Regression Trees *Wadsworth, Belmont, CA*
[9] Quinlan J R 1986 Induction of Decision Trees
[10] Montgomery D C, Peck E A and Vining G G 2015 Introduction to linear regression analysis *John Wiley & Sons*
[11] Draper N and Smith H 1998 Applied Regression Analysis
[12] Dietterich T G 2000 Ensemble methods in machine learning *International workshop on multiple classifier systems. Springer, Berlin, Heidelberg*
[13] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, and Vanderplas J 2011 Scikit-learn: Machine learning in Python *Journal of machine learning research* 12(Oct) 2825–2830
[14] Khadiev K, Zulfat M and Sidikov M 2016 Collaborative filtering approach in adaptive learning *International Journal of Pharmacy and Technology*
[15] Website: Tatenergosbit https://tatenergosbyt.ru/ (accessed 30.07.2019)
[16] Website: Weather url= https://rp5.ru (accessed 30.07.2019)