Using machine learning methods to forecast the number of power outages at substations

Chulpan Minnegalieva*, and *Alina* Gainullina

Institute of Computational Mathematics and Information Technologies, Kazan Federal University, 18, Kremlyovskaya street, Kazan, 420008, Russia

Abstract. Forecasting in the energy sector is of great importance for suppliers and for consumers. Optimum power consumption depends on many factors. Due to natural or any other external conditions, accidents are possible. In order to minimize emergency consequences, it is necessary to be prepared for possible outages in advance in order to reduce the time for their elimination and decision-making. This article considers the problem of forecasting power outages at substations. The enterprise provided a summary table of outages at substations due to natural disasters on specific days. To solve the problem, a machine learning method was chosen – binary classification. Five different algorithms were considered. The models were tested on data from the first half of 2022. The most effective algorithm for 20% of the test sample was the binary classification algorithm using generalized additive models (GAM). This algorithm is also one of the best with a sample of 50%. A model has been prepared for further use in predicting the probability of outages at the enterprise. The model can be used in other organizations; for this, it is first necessary to train the model on the data of the corresponding region.

1 Introduction

At present, the reliability of enterprise systems and financial results depend on the accuracy of forecasts. Solving forecasting problems in the energy industry is important not only for the industry, but also for consumers. This issue is constantly being studied by industry specialists, economists and researchers. Forecasting is based on the use of modern technology, statistical data on electricity consumption and climatic conditions. Both electricity suppliers and consumers use the results of these analyses in their work for the greatest benefit. The problems of forecasting, economical consumption of electricity are also important for the environment [1].

Optimal power consumption depends on many factors, such as the serviceability of the equipment, the number of consumers, and the number of industrial enterprises in a given area [2]. Various methods are being explored to use new approaches to more accurately forecast consumption. Artificial neural networks are used for long-term planning of power distribution. For example, a spatial-temporal load forecasting method for recognizing and

* Corresponding author: mchulpan@gmail.com

predicting patterns of development using historical dynamics is proposed [3]. Anomaly detection in power consumption data is investigated [4].

As has been said, climatic conditions also affect electricity consumption. Modern technology is also widely used in weather forecasting. Deep neural networks are used to predict the local temperature for the next 24 hours. [5]. Forecasting of solar irradiation (as an alternative energy source) can be performed using ensemble methods [6].

Accidents are possible due to natural or any other external conditions. This problem of emergency stops is also relevant in various sectors of the economy. Different approaches are applied. Methods for improving avalanche forecasting using machine learning algorithms are considered [7]. Data on power outages and duration of outages in a number of countries were collected and analyzed, and explicit expressions for the probability and for the duration of power outages were confirmed [8].

The article proposes the use of machine learning methods to forecast the number of power outages at a substation. In order to minimize emergency consequences, it is necessary to be prepared for possible outages in advance in order to reduce the time for their elimination and decision-making. This forecast will be relevant for local branches of enterprises, since large-scale systems reveal forecasts for the scope of the organization as a whole, and a separate service will allow you to specify the forecast, which will increase its accuracy. Moreover, the weather, geological, and social conditions of one area or city can differ significantly even from those of neighboring areas.

2 Methods for solving the problem of forecasting

The problem of forecasting power outages at substations of the branch is considered. The input data for forecasting were archival weather data in the regions of the Republic of Tatarstan for 2019-2021 and a summary table of outages at substations of one branch due to natural disasters. Among the weather data were air temperature, wind speed, whether there was snow, rain and thunderstorms. Historical outage data for each day was obtained in the form of Excel spreadsheets. The resulting tables include the necessary information about the substation where the outage occurred, the date, the reason for the outage, and some weather data.

It was decided to consider machine learning methods for probabilistic assessment of the possibility of power outages, select methods, train the model, evaluate the results and offer the enterprise the most optimal algorithm for forecasting possible outages.

Machine learning methods are widely used in solving consumption forecasting problems, in solving problems of transport, meteorology, and medicine [5]. To solve this problem, we chose a machine-learning method: binary classification. Classification is a supervised machine learning task that categorizes data instances into several categories, in our case there will be two categories. The method of training the model by binary classification can be implemented by different algorithms, therefore, in order to obtain the maximum accuracy of the forecast, we need to find the algorithm with the best estimate.

In the binary classification the output variable can take only two values. That is, the question of whether an object belongs to one of two classes is decided. In our case, states 1 and 0 will be used (whether or not an emergency outage is possible).

Let's consider approaches for solving this problem.

FastTree is an efficient implementation of the gradient boosting algorithm. Gradient boosting is a machine learning method used, in particular, in regression and classification problems. It builds each regression tree step by step, uses a standard loss function to measure the errors in each step and correct them in the next. For the binary classification problem, the output data are converted to probability using one or the other calibration option.

Binary logistic regression using the stochastic dual coordinate ascent method (SdcaLogisticRegression). SdcaLogisticRegression uses empirical risk minimization (i.e., ERM) to generate an optimization problem created from the collected data. Empirical risk is measured by applying a loss function to model predictions from collected data points.

Field-aware machine factorization model trained using the stochastic gradient method (FieldAwareFactorizationMachine). Unlike other binary classifiers, which can support only one feature column, multiple feature columns can be used in this case. Each column is considered as a container of some functions, and such a container is called a field. All feature columns must be floating point vectors, but their sizes may vary. The motivation for separating features into different fields is to model features from different distributions independently.

A linear logistic regression model trained using the L-BFGS (LbfgsLogisticRegression) method is used with a large number of features. The implemented optimization technique is based on the limited memory of the Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method.

Binary classification using generalized additive models (GAM). Here, the linear response variable depends linearly on the unknown smooth functions of some predictor variables, and interest is focused on inferences about these smooth functions. GAM were originally developed to combine the properties of generalized linear models with additive models.

3 Results

The implementation of the solution included two stages: preprocessing and model training using the C# programming language and ML.NET. Preprocessing included the processing of archive data received from the enterprise, the identification of features suitable for training, the preparation of training and test samples. When working with machine learning, interactions are reduced to a data-model relationship. The predictive result of the model depends entirely on the settings passed by the programmer and on the quality of the data set on which the training takes place. Therefore, in the process of data preprocessing, they are brought into line with the requirements determined by the task.

In order for forecasting to give a result with maximum accuracy, it is necessary to determine the factors that affect the probable emergency outage. The following parameters were highlighted: name of the substation; month in which the outage occurred; air temperature; wind speed; snow; rain; thunderstorm. Since the first and second attributes are textual, the categorical data values were converted to numbers using the OneHotEncoding method, which assigns different numeric key values to different values in each of the columns.

All of the models discussed above were trained on two years of outage data. The models were tested on data from the first half of 2022. Models were analyzed with test samples of 20% and 50% of the training data.

The most effective algorithm for 20% of the test sample was the binary classification algorithm using generalized additive models (GAM). This algorithm is also one of the best with a sample of 50%. The linear logistic regression model trained with L-BFGS (LbfgsLogisticRegression) and FastTree showed relatively good results. The field-aware machine factorization model trained using the stochastic gradient method (FieldAwareFactorizationMachine) gave the worst results (Tables 1, 2). In the tables, the Accuracy measure shows the number of correctly classified class labels, AUC (Area Under Curve) is the area under the graph showing the relationship of correctly classified positive class objects to falsely positively classified negative class objects, and F1-Score is the harmonic mean of Precision and Recall.

Table 1. Indicators of model accuracy estimates with a test sample of 20%.

	Accuracy	**AUC**	**F1-Score**
FastTree	0.92	0.92	0.91
Binary logistic regression using the stochastic dual coordinate ascent method (SdcaLogisticRegression)	0.91	0.94	0.91
Field-aware machine factorization model trained using the stochastic gradient method (FieldAwareFactorizationMachine)	0.46	0.62	0.19
A linear logistic regression model trained using the L-BFGS method (LbfgsLogisticRegression)	0.92	0.94	0.91
Binary classification using generalized additive models (GAM)	0.92	0.98	0.91

Table 2. Indicators of model accuracy estimates with a test sample of 50%.

	Accuracy	**AUC**	**F1-Score**
FastTree	0.94	0.96	0.94
Binary logistic regression using the stochastic dual coordinate ascent method (SdcaLogisticRegression)	0.95	0.96	0.95
Field-aware machine factorization model trained using the stochastic gradient method (FieldAwareFactorizationMachine)	0.63	0.68	0.56
A linear logistic regression model trained using the L-BFGS method (LbfgsLogisticRegression)	0.94	0.97	0.93
Binary classification using generalized additive models (GAM)	0.95	0.97	0.94

Thus, binary classification using generalized additive models showed a fairly good result. Using the trained model, we predicted outage on new records and compared the results with real values (Table 3, outage was (1) or not (0)).

Table 3. Results of the experiment.

	Predicted value	**Actual value**	**Outage probability**
Substation_Name="Kamskoe ust'e", Date_Outage="January", Temperature=-5, Wind_speed=14, Snow=1, Rain=0, Thunder=0	1	1	0.89
Substation_Name="Sviyazhsk", Date_Outage="January", Temperature=-16,	0	0	0.03

Wind _speed=4, Snow=0, Rain=0, Thunder=0			
Substation_Name="Apastovo", Date_Outage="May", Temperature=10, Wind_speed=13, Snow=0, Rain=1, Thunder=0	1	1	0.91
Substation_Name="Karatalga", Date_Outage="June", Temperature=20, Wind_speed=1, Snow=0, Rain=0, Thunder=0	0	0	0.06
Substation_Name="Verhnij Uslon", Date_Outage="June", Temperature=14, Wind_speed=17, Snow=0, Rain=1, Thunder=0,	1	1	0.97

4 Conclusion

Thus, the paper proposes a solution for forecasting power outages due to natural factors. The signs for compiling training and test samples are determined, an analysis is carried out to select the appropriate algorithm for training the model. It was found that binary classification using generalized additive models gives the best result. The model was trained for further use in predicting the probability of outages at the enterprise. The model can be used in other organizations by pre-training the model on the data of the corresponding region.

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program ("PRIORITY-2030").

References

1. P. Schnaars, The Electricity Journal **35**, 107074 (2022)
2. M. Parejo Guzmán, B. Navarrete Rubia, P. Mora Peris et al, Electr Eng **104**, 1681-1696 (2022)
3. S. Zambrano-Asanza, R. E. Morales, Joel A. Montalvan, John F. Franco, JEPE **148**, 108906 (2023)
4. C. Chahla, H. Snoussi, L. Merghem et al, Energy Efficiency **13**, 1633-1651 (2020)
5. D. Kreuzer, M. Munz, S. Schlüter, MLWA **2**, 100007 (2020)
6. N. Rahimi, S. Park, W. Choi, et al, J. Electr. Eng. Technol **18,** 719-733 (2023)
7. R. Fromm, C. Schönberger, MLWA **10**, 100405 (2022)
8. Romney B. Duffey, Dependability **20(3)**, 3-14 (2020)