**UDK 004.8**
**IRSTI 27.43.17**
https://doi.org/10.55452/1998-6688-2023-20-2-103-114

**Knaytov Ye. N.*[1], Akzhalova A. Zh.[1], Sadok Ben Yahia[2]**
[1]Kazakh-British Technical University, 050000, Almaty, Kazakhstan
[2]Tallinn University of Technology, Tallinn, Estonia,
*E-mail: y_knayatov@kbtu.kz

# TIME SERIES-BASED APPROACHES FOR IMPROVING WIND POWER GENERATION FORECAST ACCURACY

**Abstracts**. This study provides a detailed analysis and prediction of power generation at wind farms in Germany using Lasso, LightGBM, and CatBoost machine learning models. Feature Engineering was used on the data, which allowed the extraction of more detailed data, which was used to improve the quality of the models. Through Extensive Data Analysis (EDA), the authors identify and develop lagged and moving features from the energy production time series, under the assumption that accurate predictions can significantly improve the stability of energy systems, especially in the context of increasing dependence on renewable energy sources. The performance of each model is evaluated based on the Mean Absolute Error(MAE), Mean Squared Error(MSE), and Root Mean Squared Error(RMSE) metrics, with CatBoost exhibiting the highest accuracy. In conclude, pointing to opportunities for further research aimed at optimizing these models and adapting them to other regions, emphasizing the comprehensive and long-term potential of this study in the context of energy field.

**Key words:** wind energy, forecasting, time series, Lasso, LightGBM, CatBoost.

**Қнаятов Е.Н.*[1], Акжалова А.Ж.[1], Садок Бен Яхия[2]**
[1] Қазақстан-Британ техникалық университеті, 050000, Алматы қ., Қазақстан
[2]Таллин технологиялық университеті, Таллин қ., Эстония
*E-mail: y_knayatov@kbtu.kz

# ЖЕЛ ЭНЕРГИЯСЫНЫҢ ӨНДІРІСІН БОЛЖАУДАҒЫ ДӘЛДІКТІ ЖАҚСАРТУ ҮШІН УАҚЫТ ҚАТАРЛАРЫ НЕГІЗІНДЕГІ ТӘСІЛДЕР

**Аңдатпа.** Бұл зерттеуде біз Lasso, LightGBM және CatBoost машиналық оқыту моделдерін пайдалана отырып, Германиядағы жел электр станцияларында электр энергиясын өндіруді егжей-тегжейлі талдау мен болжауды ұсындық. Деректерді өңдеу үшін модельдердің сапасын жақсарту мақсатында пайдаланылған ендік және уақыттық ақпарат арқылы егжей-тегжейлі деректерді алуға мүмкіндік беретін Feature Engineering әдісі қолданып, жаңа дерекпен толтырылды. Жетілдірілген деректерді талдау (Extensive data Analysis, EDA) арқылы біз дәл болжамдардың сапасын нақтырақ энергетикалық жүйелердің тұрақтылығын, әсіресе жаңартылатын энергия көздеріне тәуелділіктің артуы жағдайында айтарлықтай жақсарта алатындығына сүйене отырып, энергия өндірудің уақыт сериясынан кешігу және жылжымалы белгілерді анықтадық және моделдердің сапа көрсеткішін жақсарттық. Әрбір моделдің өнімділігі орташа абсолютті қате (MAE), орташа квадраттық қате (MSE) және түбір асты орташа квадраттық қате (RMSE) статистикалық көрсеткіштері негізінде бағаланады. Осы модельдердің ішінде CatBoost барлық көрсеткіштер бойынша ең жоғары дәлдікті көрсетеді. Қорытындыда осы модельдерді оңтайландыруға және оларды басқа аймақтарға бейімдеуге бағытталған әрі қарайғы зерттеулердің мүмкіндіктері көрсетіледі, энергетикалық сала контекстінде осы зерттеудің кешенді және ұзақ мерзімді әлеуеті атап өтіледі.

**Тірек сөздер:** жел энергиясы, болжау, уақыт қатарлары, соңғы, LightGBM, CatBoost.

**Кнаятов Е.Н.*[1], Акжалова А.Ж.[1], Садок Бен Яхия[2]**
[1]Казахстанско-Британский технический университет, 050000, г. Алматы, Казахстан
[2]Таллинский университет технологий, г. Таллин, Эстония
*E-mail: y_knayatov@kbtu.kz

# ПОДХОДЫ НА ОСНОВЕ ВРЕМЕННЫХ РЯДОВ ДЛЯ ПОВЫШЕНИЯ ТОЧНОСТИ ПРОГНОЗА ВЕТРОЭНЕРГЕТИКИ

**Аннотация.** В данном исследовании мы представили подробный анализ и прогнозирование выработки электроэнергии на ветряных электростанциях в Германии с использованием моделей машинного обучения Lasso, LightGBM и CatBoost. Для обработки данных использовался метод Feature Engineering, который позволил извлечь более подробные данные с дат, использованные для улучшения качества моделей. С помощью расширенного анализа данных (Extensive Data Analysis, EDA) мы определяем и разрабатываем запаздывающие и скользящие признаки из временного ряда производства энергии, исходя из того, что точные прогнозы могут значительно повысить стабильность энергетических систем, особенно в контексте растущей зависимости от возобновляемых источников энергии. Производительность каждой модели оценивается на основе показателей средней абсолютной ошибки (MAE), средней квадратичной ошибки (MSE) и корневой средней квадратичной ошибки (RMSE), при этом среди этих моделей CatBoost демонстрирует самую высокую точность по всем показателям. В заключение указываются возможности для дальнейших исследований, направленных на оптимизацию этих моделей и их адаптацию к другим регионам, подчеркивается комплексный и долгосрочный потенциал данного исследования в контексте энергетической сферы.

**Ключевые слова:** энергия ветра, прогнозирование, временные ряды, Lasso, LightGBM, CatBoost.

## Introduction

Renewable energy, especially wind power, has become a significant source of energy in the modern world. Countries around the world are actively investing in the development of wind power, as it offers a clean and sustainable source of energy, contributing to the reduction of greenhouse gas emissions. In this study, we focus on wind power in Germany, one of the leading countries in this field. Our goal is to develop a model for predicting electricity generation from wind farms in Germany. To achieve this goal, we will use a dataset containing information on power generation capacity and associated time stamps.

We will analyze the data to understand temporal trend patterns and identify possible outliers or anomalies in power generation. In addition, we will develop new features based on lags and autocorrelation that can help improve the quality of forecasting. We will consider several machine learning models as prediction methods, including Lasso[1], LightGBM[2], and CatBoost[3]. We will evaluate the performance of each model using different metrics and select the most effective model for our forecasting purposes. The results of this study can be useful to energy companies and regulators in helping them make informed decisions about planning and optimizing wind energy generation. More accurate forecasts will enable better management of energy production and ensure the stability of the energy system. In the following sections, we present details of the data analysis, a description of the methods and models used, forecasting results, and a discussion of the results.

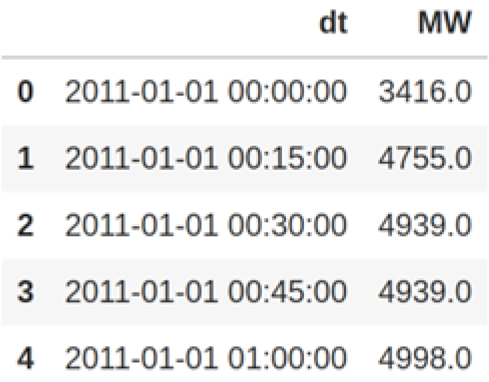
## Literature review

In this part of the study, a review of scientific papers devoted to time series forecasting was conducted. Scientific articles and publications devoted to methods and models of time series forecasting were studied. One of the significant studies in this field is the work of Tibshirani [1]. In his work, the author presents the Lasso (Least Absolute Shrinkage and Selection Operator) method, which allows reducing the dimensionality of the feature space and selecting the most important features for prediction. The author describes the properties and advantages of the Lasso method as well as its applicability in the context of time series forecasting. Another important study in this area was performed by Ke et al. [2]. In their work, they considered the application of the LightGBM model, which is a highly efficient gradient-based solver tree boosting. The authors describe the working principle of the LightGBM model, its advantages and potential in time series forecasting. Also an important study is the work of Prokhorenkova et al. [3]. In their work the authors present the CatBoost model, which has a unique ability to process categorical features without preprocessing. The authors investigate the properties of the CatBoost model and demonstrate its application to various tasks, including time series forecasting. These studies represent important contributions to the field of time series forecasting and will be used in this study to develop predictive models and evaluate their performance.

Tibshirani [1] identified the problems associated with the application of the Lasso method for time series prediction, including the choice of the optimal value of the regularization parameter and the stability of the model to the presence of strongly correlated predictors. The paper by Ke, G. et al. [2] notes the need for proper selection of the LightGBM model hyperparameters to achieve optimal performance, which may require significant computational resources. Prokhorenkova et al. [3] point out that using CatBoost to predict time series with categorical features requires careful adjustment of the learning rate and number of iterations, and may also require large amounts of memory and computing resources.

### Data

In this study, we used an extensive dataset of wind energy generation that includes more than 380,000 records. This amount of data provides sufficient statistical significance and allows for a more accurate analysis and prediction of the energy generation process. The data in demonstrated in **Figure 1** set consists of two columns. The first column contains timestamps that indicate the date and time of each record. The second column contains wind power generation capacity values in megawatts (MW).



| | dt | MW |
|---|---|---|
| **0** | 2011-01-01 00:00:00 | 3416.0 |
| **1** | 2011-01-01 00:15:00 | 4755.0 |
| **2** | 2011-01-01 00:30:00 | 4939.0 |
| **3** | 2011-01-01 00:45:00 | 4939.0 |
| **4** | 2011-01-01 01:00:00 | 4998.0 |

Figure 1 – The wind power producing data

Both columns are important variables for studying and analyzing the energy generation process. The data collection frequency is 15 minutes, which means that measurements were taken every 15 minutes. This allows us to account for changes in energy generation over short time intervals and identify temporal patterns. This extensive data set provides us with an opportunity to perform a deeper and more comprehensive analysis of the wind energy generation process and develop effective predictive models. This will help to optimize the energy production process, improve its stability and ensure more efficient use of wind resources.

### Main provisions. Methods and materials

In this section, a exploratory data analysis[14] of the wind energy data was performed using various methods and visualizations. Having large data, we divided them into columns for analysis to make it easier to highlight parameters. We divided the data by hours, then days, days of the week by another year, which allowed us to graph and analyze. This analysis allowed us to gain valuable insights and discover patterns related to wind energy generation.

In analyzing the annual generation capacity data, it was found that there was a significant increase in energy generation by the hour in 2017 and 2019, which is given in Figure 2[13]. This may indicate special events or factors that influenced the increase in energy production during these periods.

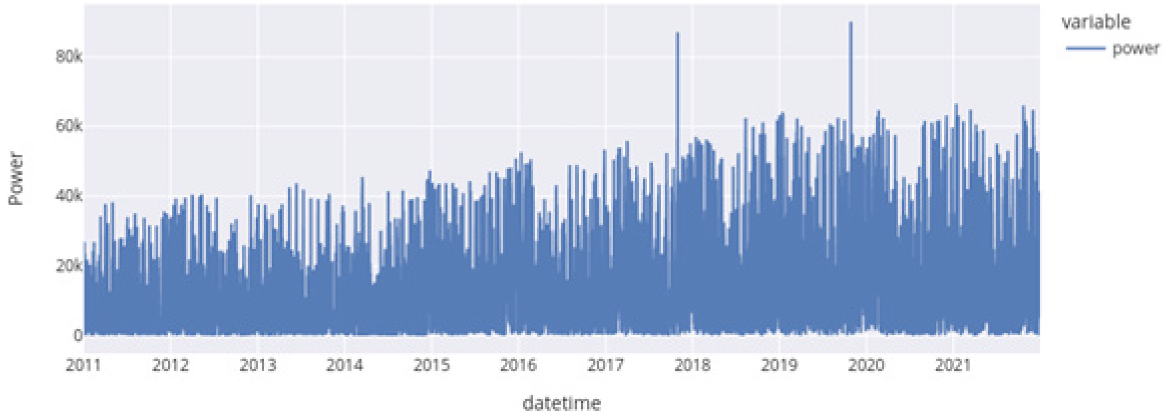


Figure 2 – Discretization of power generation up to 1 hour

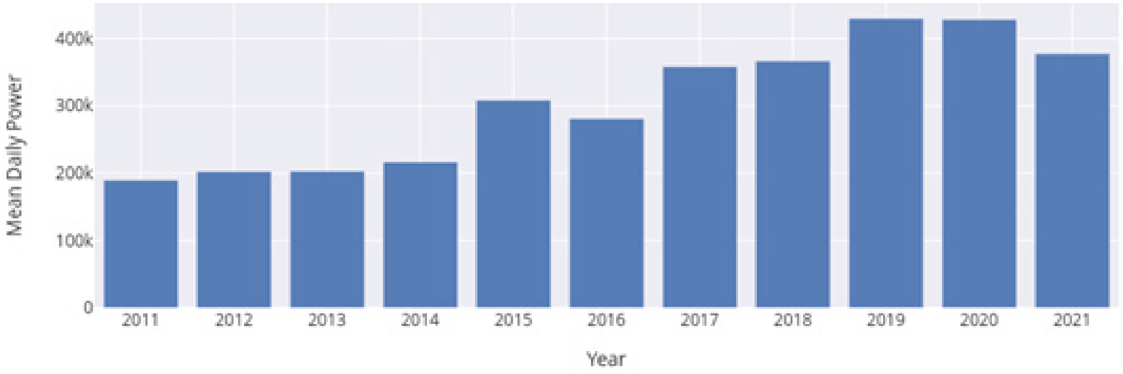


Figure 3 – Power generation trend across years

We also analyzed annual generation capacity data and found that 2019 and 2020 had the highest capacity. This may indicate increased wind farm activity and other factors contributing to increased power generation during these periods see Figure 3.

In Figure 4 we can observe that the total energy generation on Tuesday, Thursday, Friday, and Saturday was higher in 2021, while the total generation on Monday and Sunday was higher in 2019, which can be seen in. This may indicate differences in energy demand on different days of the week and the corresponding impact on wind generation.

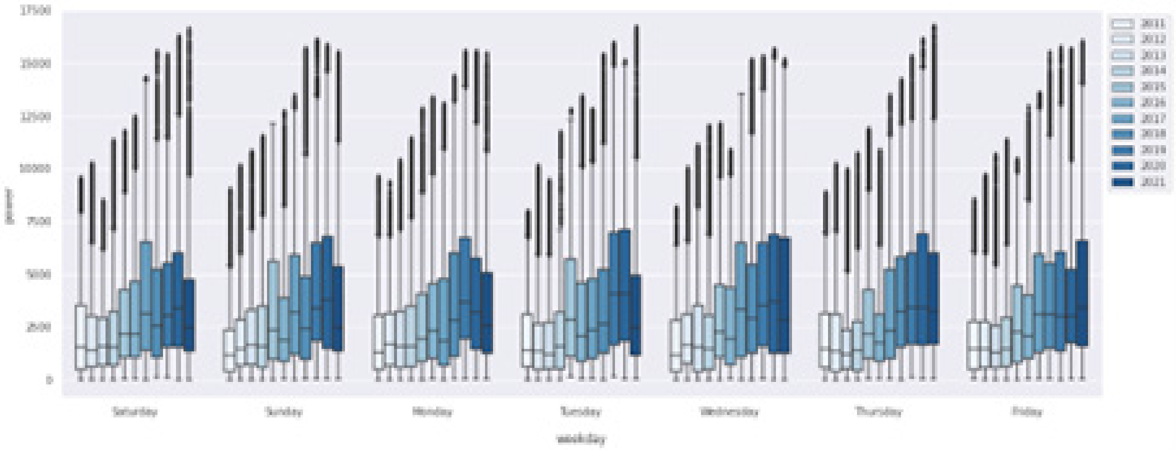


Figure 4 – Trends of week across years.

An analysis of the percentage growth of energy generation plotted that 2015 and 2017 saw the largest increase in generation from the previous year by 35%. This may indicate a significant change in the development of wind energy and its contribution to total energy generation can be seen in Figure 5[12].

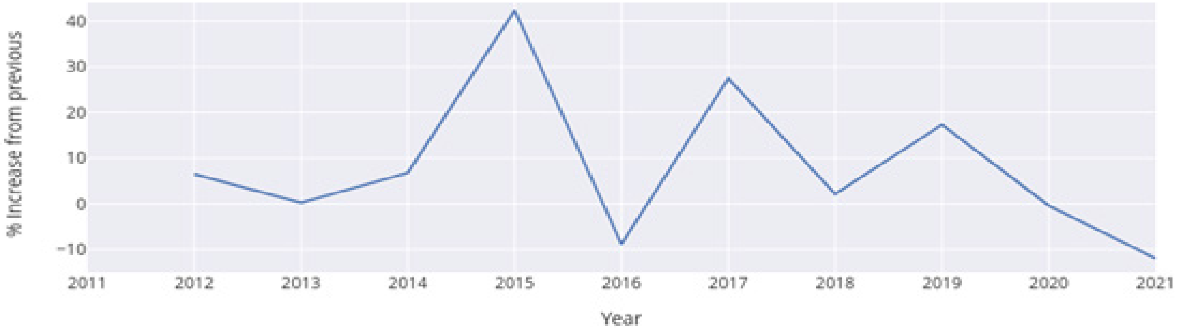
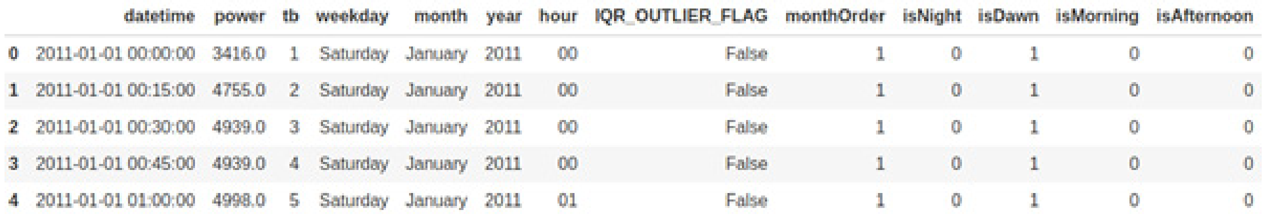


Figure 5 – Annual growth in electricity production

In addition, emissions in the data were evaluated and it was found that power values above 10,000 MW were emissions. These emissions could be the result of anomalous situations or errors in the data, one might consider.

We conducted a process of feature engineering[6][7] to improve the data set. Feature engineering is the creation of new features from existing data, in order to expand the information and capture important patterns and dependencies. Using datetime information we extracted several parameters described in Figure 6 such as the order of the month, time of day divided into categories, which describes that it is night, afternoon, sunset, morning, dawn.



| | datetime | power | tb | weekday | month | year | hour | IQR_OUTLIER_FLAG | monthOrder | isNight | isDawn | isMorning | isAfternoon |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2011-01-01 00:00:00 | 3416.0 | 1 | Saturday | January | 2011 | 00 | False | 1 | 0 | 1 | 0 | 0 |
| 1 | 2011-01-01 00:15:00 | 4755.0 | 2 | Saturday | January | 2011 | 00 | False | 1 | 0 | 1 | 0 | 0 |
| 2 | 2011-01-01 00:30:00 | 4939.0 | 3 | Saturday | January | 2011 | 00 | False | 1 | 0 | 1 | 0 | 0 |
| 3 | 2011-01-01 00:45:00 | 4939.0 | 4 | Saturday | January | 2011 | 00 | False | 1 | 0 | 1 | 0 | 0 |
| 4 | 2011-01-01 01:00:00 | 4998.0 | 5 | Saturday | January | 2011 | 01 | False | 1 | 0 | 1 | 0 | 0 |

Figure 6 – New features extraction

One of the main aspects of feature engineering in this paper is the creation of lagged features[8]. Lagged features allow us to account for consistent dependence and cyclicality in the time series. In this case, lagged features were created with lags of 1, 12, 24, 48, and 72 time blocks in the "power" variable. This allows the model to account for the influence of previous periods on the current power value. Additionally, rolling features, such as the rolling average, were selected. Rolling features are a power average over a certain time window. In this case, rolling averages with a window of 4 and 24 time blocks were chosen. These signs help to average temporal fluctuations and reveal general trends in the data. After feature engineering, a correlation analysis was performed, which allowed us to identify the most strongly associated features with the "power" variable. In particular, the attributes "lagged_power_1", "lagged_power_12", "rolling_4_power_mean" and "rolling_24_power_mean" connect high correlation with power. This indicates a significant influence of previous periods and averaged values on current wind power generation, observed in Figure 7.

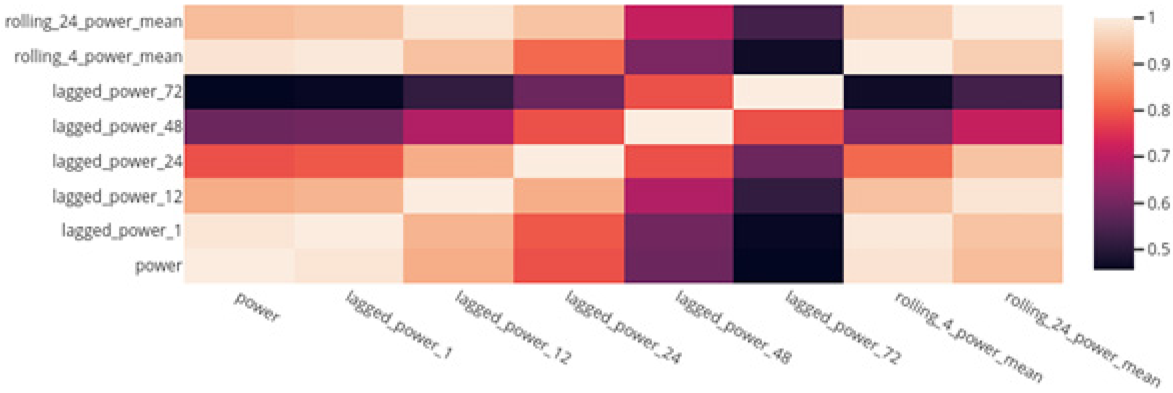
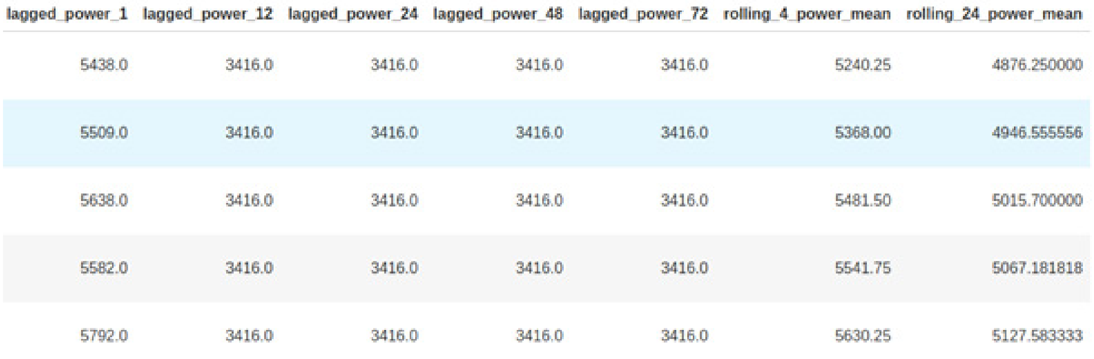


Figure 7 – Correlation(Heat) map

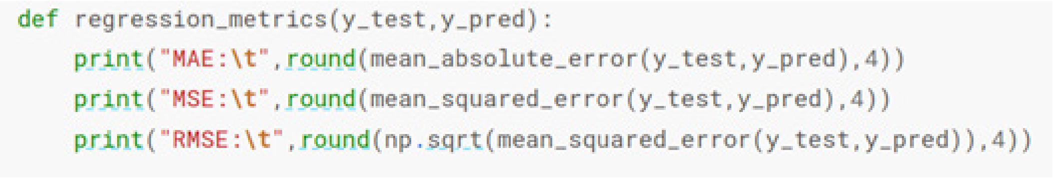
Thus, the process of feature engineering has expanded and enriched the original data set to include additional features that capture temporal dependencies and cyclicality in the time series. This allows the model to more accurately predict wind energy generation given the important characteristics and patterns highlighted in the data. For example, lagged features help account for the influence of previous periods on current wind generation. This is especially useful for analyzing time dependencies and cyclicality in the data. Creating lagged features with different lags, such as 1, 12, 24, 48 and 72 time blocks, allows the model to account for the influence of previous periods on the current generation capacity. In addition, additional rolling features, such as a rolling average, have been selected. Rolling features are averages of power values over a certain time window. In this case, we chose a rolling average with a window of 4 and 24 time blocks. These signs allow you to capture general trends and smooth out temporal fluctuations in the data. A correlation analysis was performed to assess the relationship between the signs and the "power" variable. The results observes that the traits "lagged_power_1", "lagged_power_12", "rolling_4_power_mean" and "rolling_24_power_mean" are highly correlated with power. This suggests that previous power values and averages play an important role in predicting current wind power generation. We have included the significant data in Table Figure 8.



| lagged_power_1 | lagged_power_12 | lagged_power_24 | lagged_power_48 | lagged_power_72 | rolling_4_power_mean | rolling_24_power_mean |
|---|---|---|---|---|---|---|
| 5438.0 | 3416.0 | 3416.0 | 3416.0 | 3416.0 | 5240.25 | 4876.250000 |
| 5509.0 | 3416.0 | 3416.0 | 3416.0 | 3416.0 | 5368.00 | 4946.555556 |
| 5638.0 | 3416.0 | 3416.0 | 3416.0 | 3416.0 | 5481.50 | 5015.700000 |
| 5582.0 | 3416.0 | 3416.0 | 3416.0 | 3416.0 | 5541.75 | 5067.181818 |
| 5792.0 | 3416.0 | 3416.0 | 3416.0 | 3416.0 | 5630.25 | 5127.583333 |

Figure 8 – Added Lagged features

Finally, the process of feature engineering has enriched the original data set with new features that account for temporal dependencies and cyclicality in the time series. his allows the model to better understand and predict wind energy generation by accounting for important characteristics and patterns identified in the data.



```
def regression_metrics(y_test,y_pred):
    print("MAE:\t",round(mean_absolute_error(y_test,y_pred),4))
    print("MSE:\t",round(mean_squared_error(y_test,y_pred),4))
    print("RMSE:\t",round(np.sqrt(mean_squared_error(y_test,y_pred)),4))
```

Figure 9 – Metrics of evaluation models

Various metrics are used to evaluate the performance of predictive models, including mean absolute error (MAE)[10], root mean square error (MSE), and root mean square error (RMSE)[11] observe in Figure 9. MAE measures the average absolute deviation between predicted and actual values and provides an estimate of the mean of the prediction error. MSE measures the root mean square deviation and pays more attention to large deviations. RMSE is the root of MSE and allows you to compare model performance on the raw data scale. The smaller the MAE, MSE, and RMSE values, the better the performance of the models in predicting power generation.

The Least Absolute Shrinkage and Selection Operator is a regression method used in statistics and machine learning to predict data. The main feature of Lasso is the use of L1-regularization, which helps to reduce the number of features in the model by zeroing out the weights of some features. This feature makes Lasso very useful when there are a large number of features and there is a need for feature selection. Lasso regression is formulated as follows:

For the data *(X, y)* where *X* - input data matrix of size *(n, p)* *(n* - the number of examples, *p* - number of features), and *y* - is a vector of output data of dimension *(n, 1),* the Lasso regression problem is to minimize the following loss function:

$$Loss = minimize\ ||\frac{1}{2n}y - X\beta||_2^2 + \lambda||\beta||_1$$

Where:

$-\ ||\cdot||_2$ denotes the Euclidean norm (L2-norm),

$-\ ||\cdot||_1$ denotes the L1-norm,

$-\beta$ - vector of regression coefficients of dimension (p, 1),

$-\lambda$- is a non-negative regularization parameter.

The second term in this function is the L1-regulator, which is controlled by the parameter $\lambda$. It penalizes large values of the coefficients $\beta$, which leads to their reduction or even zeroing. This ensures the selection of features in the Lasso-regression. Because of this property, Lasso-regression is often used when working with data with a large number of features, when it is necessary to simplify the model and make it interpretable.

Light Gradient Boosting Machine (LightGBM)[2] is a machine learning algorithm based on the method of gradient boosting. This algorithm was developed and introduced by Microsoft Research in 2017. The model differs from most other boosting algorithms in that it uses a "leaf-wise" learning strategy instead of the usual "level-wise" learning strategy, which allows it to achieve higher efficiency while maintaining model accuracy. The gradient-busting algorithm on which LightGBM is based can usually be described as follows:

Given:

- Learning sample $\{(x))_1, y_1), \ldots, (x_n, y_n)\}$,
- Loss function *L(y,F)*,
- number of iterations *M*.

For *m = 1 to M*:

1. Calculate pseudo-residuals:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

In the context of LightGBM, each $h_m(x)$ - is a decision tree, and these trees are built using a "leaf-wise" strategy. Instead of developing the tree level by level (level-wise), LightGBM chooses the leaf with the maximum loss to split and continues to split it, allowing more tree depth and providing more accuracy, while controlling overlearning. However, the "leaf-wise" strategy can lead to overtraining with a large number of leaves, so it is important to apply regularization by adding a summand to the loss function that increases with the number of leaves. In LightGBM this is achieved using the following condition for splitting:

$$Gain = \frac{1}{2}\left[\frac{\sum_i {G_i}^2}{\{\sum_i H_i + \lambda\}} - \frac{\{(\sum_i G_i + G_{new})^2\}}{\{\sum_i H_i + H_{new} + \lambda\}} - \frac{\{G_{new}^2\}}{\{H_{new} + \lambda\}}\right] - \gamma$$

Where:

- $G_i$ и $H_i$- gradients and hessians of the loss function,

- $G_{new}$ и $H_{new}$ - gradients and hessians added by the new sheet,

- λ is a regularization parameter controlling the magnitude of the penalty for increasing the number of leaves,

- $\gamma$ - additional regularization parameter, which adds a fixed penalty for each new splitting.

Thus, LightGBM strikes a balance between the speed of learning and the accuracy of predictions, allowing efficient models to be built even on large data sets.

CatBoost is a machine learning algorithm designed to handle categorical features efficiently. It uses gradient-based boosting and is based on decision tree learning. CatBoost[3] applies an ordered-target learning technique, which significantly improves the learning process compared to standard categorical feature learning approaches. Let us give you the formulas used in this algorithm. The basic equation for the gradient boosting model:

$$F_M(x) = \sum_{m=1}^{M} \rho_m h_m(x)$$

– *F_M(x)* - is the prediction of the model after *M* iterations,

– *h_m(x)* - the base algorithm (in our case a decision tree),

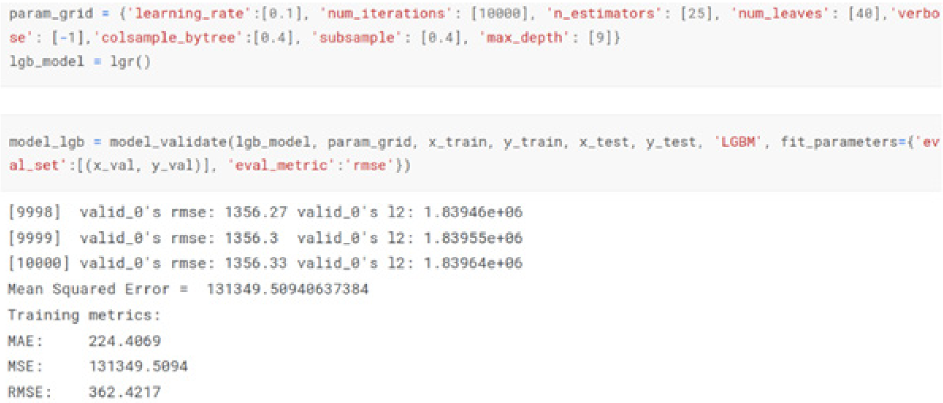
– *rho*$_m$ - the coefficient determining the contribution of each decision tree.

However, CatBoost uses a slightly modified version of the gradient-busting algorithm, which uses learning with ordered targets. When the tree is trained on a dataset, it treats objects in random order. When an object is predicted, it uses the average of its previous values to encode categorical variables. This is done to avoid leakage of target values, which is often the case when processing categorical features. As with most other boosting algorithms,

CatBoost uses a loss function that is minimized at each iteration. The loss function is chosen depending on the problem. For example, Logloss is often used for classification and RMSE for regression. CatBoost uses common regularization approaches, similar to those used in other gradient-busting algorithms, to avoid overtraining the model.

### Results and discussion

In the research paper, we calculated the results of the forecast models for the target variable - electricity power with a forecast for 8 time blocks ahead. The Lasso, LightGBM, and CatBoost models resulted in predictions for each of the Figure 10, Figure 11, and Figure 12 models.

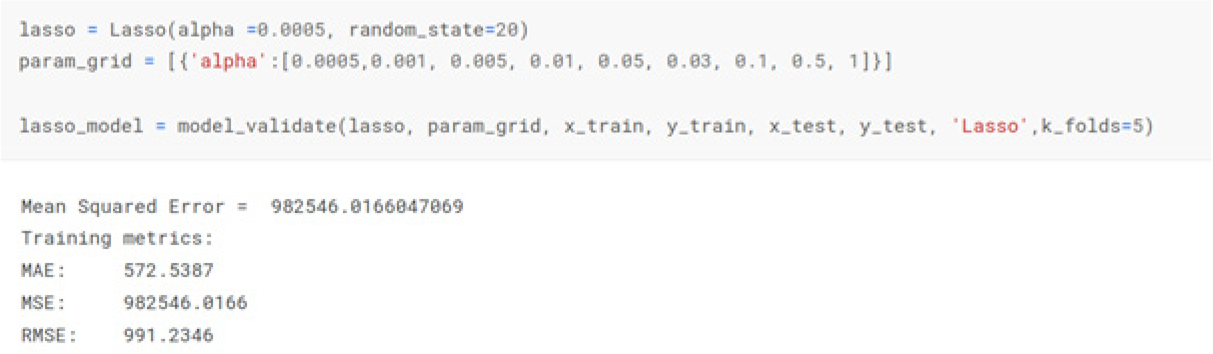


```
param_grid = {'learning_rate':[0.1], 'num_iterations': [10000], 'n_estimators': [25], 'num_leaves': [40],'verbo
se': [-1],'colsample_bytree':[0.4], 'subsample': [0.4], 'max_depth': [9]}
lgb_model = lgr()
```

```
model_lgb = model_validate(lgb_model, param_grid, x_train, y_train, x_test, y_test, 'LGBM', fit_parameters={'ev
al_set':[(x_val, y_val)], 'eval_metric':'rmse'})
```

```
[9998]  valid_0's rmse: 1356.27 valid_0's l2: 1.83946e+06
[9999]  valid_0's rmse: 1356.3  valid_0's l2: 1.83955e+06
[10000] valid_0's rmse: 1356.33 valid_0's l2: 1.83964e+06
Mean Squared Error =  131349.50940637384
Training metrics:
MAE:     224.4069
MSE:     131349.5094
RMSE:    362.4217
```

Figure 10 – Indicators of the Lasso model metrics

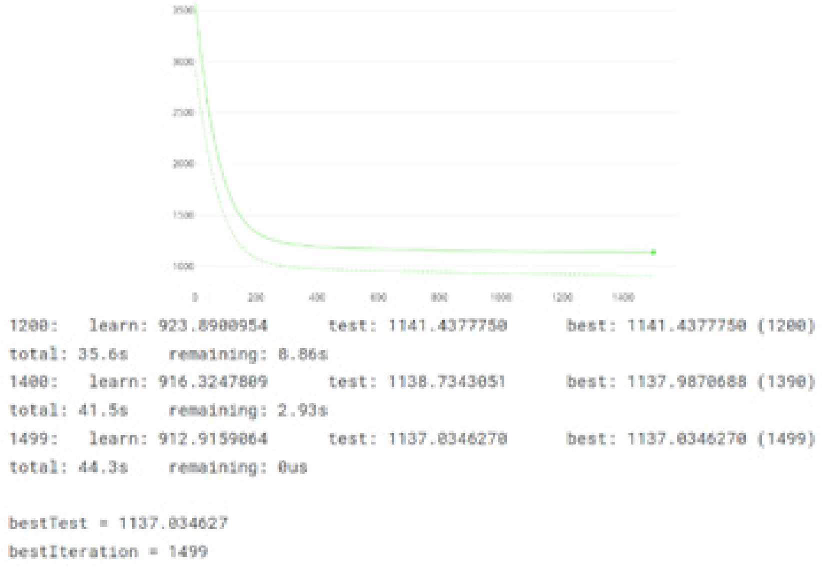


```
lasso = Lasso(alpha =0.0005, random_state=20)
param_grid = [{'alpha':[0.0005,0.001, 0.005, 0.01, 0.05, 0.03, 0.1, 0.5, 1]}]

lasso_model = model_validate(lasso, param_grid, x_train, y_train, x_test, y_test, 'Lasso',k_folds=5)
```

```
Mean Squared Error =  982546.0166047069
Training metrics:
MAE:      572.5387
MSE:      982546.0166
RMSE:     991.2346
```

Figure 11 – Indicators of the LightGBM model metrics



```
1200:    learn: 923.8900954      test: 1141.4377750      best: 1141.4377750 (1200)
total: 35.6s     remaining: 8.86s
1400:    learn: 916.3247809      test: 1138.7343051      best: 1137.9870688 (1390)
total: 41.5s     remaining: 2.93s
1499:    learn: 912.9159064      test: 1137.0346270      best: 1137.0346270 (1499)
total: 44.3s     remaining: 0us

bestTest = 1137.034627
bestIteration = 1499
```

Figure 12 – Catboost model metrics

Comparing the models, we can note that all three models observe different prediction accuracy. LightGBM demonstrates the best accuracy with the lowest MAE, MSE and RMSE, which indicates a lower prediction error compared to the other models. Lasso also presents itself as a quite reliable model with acceptable MAE and RMSE values. CatBoost, even though it has the highest RMSE value, still performed good performance and can be a useful forecasting tool. Fiqure 13 compares the results of the test and model values. In addition, consider Figure 14 demonstrated a time interval of 2 hours.

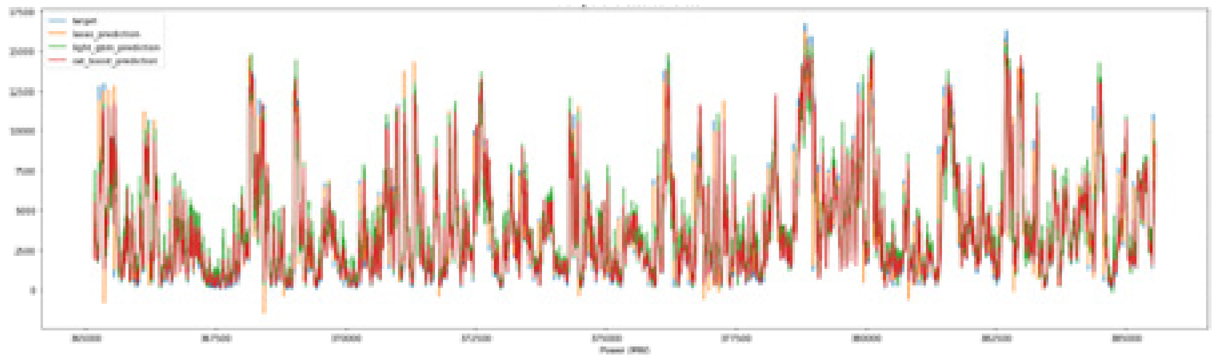


Figure 13 – Comparison plot of target values and models predicted values

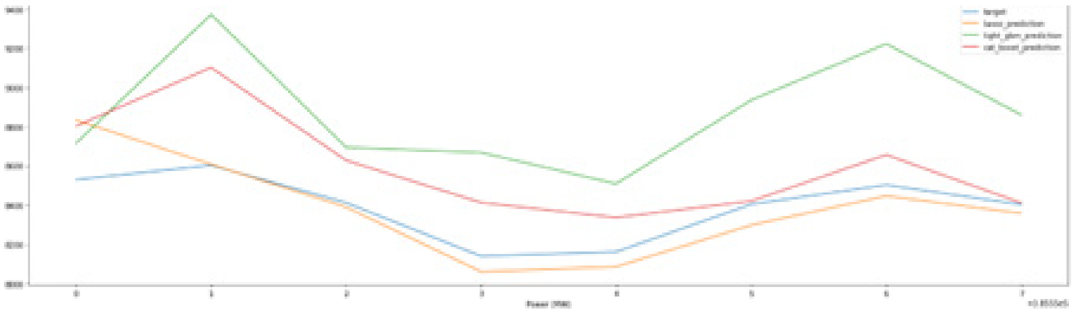


Figure 14 – Two hours interval prediction comparison

Discrepancies and variations between models may be due to different approaches and peculiarities of each algorithm. For example, the CatBoost model has the ability to process categorical features without preprocessing, which can be useful when working with electricity data. At the same time, the LightGBM model has high speed and works effectively with sparse data.

**Conclusion**

In this study, modeling was performed to predict electricity generation based on weather and timing data. Three machine learning models were used: Lasso, LightGBM, and CatBoost. Each model demonstrated its own characteristics and advantages in forecasting. As a result of the simulations performed, the predicted values for the target variable (power) with a prediction for 8 time blocks ahead were obtained. Analysis of the results illustrated that all three models demonstrated good performance in forecasting power generation. The Lasso model calculated a mean square error (MSE) of 982546 and a mean absolute error (MAE) of 572. The LightGBM model performed an MSE of 131349 and an MAE of 224. The CatBoost model demonstrated an MSE of 1137. When comparing these models, it can be noted that they all achieved good results in prediction, although with some discrepancies. Thus, the results of this study confirm the potential of machine learning models in predicting electricity generation. Further research and improvement of the models can lead to even more accurate predictions, which in turn can be useful for optimizing the planning and management of electricity generation.

**References**

1 Tibshirani R. (1996) Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society: Series B (Methodological), 58 (1), pp. 267–288.

2 Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q. & Liu T. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree, Advances in Neural Information Processing Systems, 30, pp. 3146–3154.

3 Prokhorenkova L., Gusev G., Vorobev A., Dorogush A. & Gulin A. (2018) CatBoost: unbiased boosting with categorical features, Advances in Neural Information Processing Systems, 31, pp. 6638–6648.

4 Pereira F., Portela F. & Neves J. (2021) Fraud detection in digital payments: A CatBoost approach, Journal of Computational Science, 53, 101344.

5 Ye M., Li X., Shao M., Li Y. & Liu X. (2021) An ensemble machine learning framework for custom.

6 García S., Luengo J., Herrera F. (2015) Data Preprocessing in Data Mining. Springer.

7 Wang L., Li L. & Khedr A.M. (2019) Feature Engineering and Selection for Time Series Forecasting: A Review, ACM Computing Surveys, 52(5), pp.1–37.

8 Godahewa R. & Samaraweera L. (2019) Time Series Feature Engineering: A Systematic Review, In International Conference on Advances in Computing and Data Sciences, pp. 571–581.

9 Hyndman R.J. & Athanasopoulos G. (2018) Forecasting: Principles and Practice. OTexts.

10 Chai T. & Draxler R.R. (2014) Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, Geoscientific Model Development, 7(3), pp.1247–1250.

11 Aldrin M. & Holden M. (2019) On the use of RMSE and MAE in model evaluation, Ocean Dynamics, 69(7), pp. 925–933.

12 VanderPlas J.T. (2016) Python Data Science Handbook: Essential Tools for Working with Data, O'Reilly Media.

13 Van den Bossche J. (2017) Interactive Data Visualization in Python With Plotly, Journal of Open Source Education, 20(2), 43.

14 Suresh H.P. & Mohan C.K. (2018) A Comprehensive Review of Predictive Data Mining Techniques for Credit Scoring in Banking Sector, Journal of King Saud University-Computer and Information Sciences, 30(3), pp. 360–374.

**Information on the authors**

**Knaytov Yernar Nurlanuly** (corresponding author)
Master student, Kazakh-British Technical University, 59, Tole bi street, Almaty, 050000, Kazakhstan
ORCID ID https://orcid.org/0009-0005-0003-669X
E-mail: y_knayatov@kbtu.kz

**Akzhalova Assel Zholdasovna**
Dr. Kazakh-British Technical University, Head of International project groups, Coordinator of SDG center, Professor of IT Faculty, PhD in Math.Modelling (RK), PhD in Computer Science (King's College London,UK), Kazakh-British Technical University, 59, Tole bi street, Almaty, 050000, Kazakhstan
ORCID ID 0000-0002-1141-7595
E-mail: a.akzhalova@kbtu.kz

**Sadok Ben Yahia**
Professor, Tallinn Univeristy of Technology, Tallinn, Estonia
ORCID ID: 0000-0001-8939-8948
E-mail: sadok.ben@taltech.e

**Авторлар туралы мәліметтер**

**Қнаятов Ернар Нұрланұлы** (корреспонденция авторы)
Қазақстан-Британ техникалық университеті, Магистратура студенті, Төле би көш., 59, 050000, Алматы қ., Қазақстан
ORCID ID https://orcid.org/0009-0005-0003-669X
E-mail: y_knayatov@kbtu.kz

**Акжалова Асель Жолдасовна**
Қазақстан-Британ техникалық университетінің докторы, Халықаралық жобалық топтардың жетекшісі, SDG орталығының үйлестірушісі, Ақпараттық технологиялар факультетінің профессоры, математикалық модельдеу бойынша PhD (ҚР), компьютерлік ғылымдар бойынша PhD (Лондон Корольдік колледжі, Ұлыбритания), Қазақстан-Британ техникалық университеті, Төле би көш., 59, 050000, Алматы қ., Қазақстан
ORCID ID 0000-0002-1141-7595
E-mail: a.akzhalova@kbtu.kz

**Садок Бен Яхиа**
Профессор, Таллин технологиялық университеті, Таллин қ., Эстония
ORCID ID: 0000-0001-8939-8948
E-mail: sadok.ben@taltech.ee

**Информация об авторах**

**Кнаятов Ернар Нурланулы** (автор для корреспонденции)
Магистрант, Казахстанско-Британский технический университет, ул. Толе би, 59, 050000, г. Алматы, Казахстан
ORCID ID https://orcid.org/0009-0005-0003-669X
E-mail: y_knayatov@kbtu.kz

**Акжалова Асель Жолдасовна**

Доктор Казахстанско-Британского технического университета, руководитель международных проектных групп, координатор центра SDG, профессор факультета информационных технологий, PhD по математическому моделированию (РК), PhD по компьютерным наукам (Королевский колледж Лондона, Великобритания), Казахстанско-Британский технический университет, ул. Толе би, 59, 050000, г. Алматы, Казахстан

ORCID ID 0000-0002-1141-7595

E-mail: a.akzhalova@kbtu.kz

**Садок Бен Яхиа**

Профессор, Таллинский университет технологий, г. Таллин, Эстония

ORCID ID: 0000-0001-8939-8948

E-mail: sadok.ben@taltech.ee