JOIN (Jurnal Online Informatika)
p-ISSN: 2528-1682, e-ISSN: 2527-9165
Volume 9 Number 2 | June 2024: 169-178
DOI: 10.15575/join.v9i1.1324

CatBoost Optimization Using Recursive Feature Elimination

Agus Hadianto[1], Wiranto Herry Utomo[2]
[1,2]Master of Informatics, President University, Indonesia

Article Info

Article history:

Received March 27, 2024
Revised May 27, 2024
Accepted June 2, 2024
Published August 24, 2024

Keywords:

CatBoost
Feature Selection
RFE

ABSTRACT

CatBoost is a powerful machine learning algorithm capable of classification and regression application. There are many studies focusing on its application but are still lacking on how to enhance its performance, especially when using RFE as a feature selection. This study examines the CatBoost optimization for regression tasks by using Recursive Feature Elimination (RFE) for feature selection in combination with several regression algorithm. Furthermore, an Isolation Forest algorithm is employed at preprocessing to identify and eliminate outliers from the dataset. The experiment is conducted by comparing the CatBoost regression model's performances with and without the use of RFE feature selection. The outcomes of the experiments indicate that CatBoost with RFE, which selects features using Random Forests, performs better than the baseline model without feature selection. CatBoost-RFE outperformed the baseline with notable gains of over 48.6% in training time, 8.2% in RMSE score, and 1.3% in R^2 score. Furthermore, compared to AdaBoost, Gradient Boosting, XGBoost, and artificial neural networks (ANN), it demonstrated better prediction accuracy. The CatBoost improvement has a substantial implication for predicting the exhaust temperature in a coal-fired power plant.

Corresponding Author:

Wiranto Herry Utomo,
Master of Informatics, Faculty of Computing, President University
Jl. Ki Hajar Dewantara, RT.2/RW.4, Mekarmukti, Cikarang Utara, Bekasi Regency, Indonesia 17530
Email: Email: wiranto.herry@president.ac.id

1. INTRODUCTION

Decision-tree-based methods have been widely recognized for their effectiveness in handling large datasets, with the added benefit of significantly reducing training convergence time [1]. The trees within the forest exhibit distinct characteristics, so forming a diverse assemblage that together outperforms any individual tree. One of the algorithms discussed in the literature is the CatBoost algorithm [2]-[3], which represents an improved iteration of the gradient-boosting decision tree method and employs binary decision trees as its foundational predictors. CatBoost –an algorithm based on decision trees – demonstrates high suitability for machine learning problems involving categorical and heterogeneous data [4]. As a result, CatBoost has been widely utilized in numerous research works to address categorization issues. The utilization of this technique is observed in various disciplines, including finance [5]-[6], medical [7]-[8], failure prediction [9], fraud detection [10]-[13], psychology [14], anomaly detection [15], cyber-security [16], material engineering [17], biochemistry [18]-[19], biology [20]-[21], and numerous other domains.

The primary focus of the development of CatBoost has been on enhancing its ability to handle categorical features. Additionally, CatBoost has been widely employed in regression problems as well. It is applied to forecast daily diffuse horizontal sun radiation [22], determine the isothermal compressibility of ionic liquids [23], predict the compressive strength of concrete and enhance production processes [24], weather forecasting model for short-term predictions [25], prediction of water evapotranspiration [26], forecast in the medium- and long-term power load [27], estimations of sea surface pCO2 in the North Atlantic region [28], estimation of the daily dew point temperature [29], predicting the quantity of aboveground biomass (AGB) in forested areas [30], medical [31], addressing the issue of short-term voltage stability (STVS) [32].

The research proves that CatBoost is a robust machine learning algorithm that demonstrates success in classification and regression tasks. A significant research gap is evident during the literature review, indicating the need for further investigation in a specific area. It has been noted that much prior research has primarily focused on how CatBoost model is performed compared to other machine learning algorithms for certain regression predictive tasks. They do not focus on how to improve CatBoost itself using feature selection so that it can have optimal performance and predictive power. In particular to RFE application for feature selection, there has been limited investigation into the implementation of RFE, which heavily relies on the learning algorithm employed as the objective function. Only [30] and [31] have concluded that RFE successfully impacts CatBoost performance. However, they do not state how RFE works with different learning algorithms as its objective function. This gap highlights the necessity for further research to fully comprehend the specifics and consequences of using various learning algorithms in the RFE. Furthermore, in terms of applying CatBoost-RFE to predictive modeling, no studies have been conducted thus far utilizing CatBoost-RFE as a predictive model for exhaust gas emissions in coal-fired boilers. The literature review highlights the gaps in knowledge, underscoring the need to investigate these unexplored aspects to progress the field and generate novel insights for the existing body of knowledge.

This study proposes an integrated feature selection using RFE to enhance CatBoost regression models. The contribution lies in integrating various RFE combinations with the learning algorithms as the objective function, leading to improved model performance. The use of RFE in the preprocessing pipeline of CatBoost regression models is not novel. However, this study extends the existing research by examining various learning algorithms wrapped within RFE as a black-box model for RFE and comparing each impact on CatBoost performance, thereby enhancing the prediction accuracy and generalization capabilities of CatBoost regression models. In addition, the novelty lies in using CatBoost-RFE for boiler exhaust temperature prediction, which has not yet been explored in current studies. The proposed RFE feature selection approach offers a novel contribution by addressing the limitations of previous studies focusing solely on applying RFE with a single learning algorithm. This study also explores the combined impact of outlier detection and removal using iForest [33] and RFE as feature selection in providing a more comprehensive and effective preprocessing of CatBoost regression models. The novelty lies in the synergistic effect of integrating iForest-RFE to improve CatBoost performance. It is expected to provide valuable insights to researchers and practitioners in predictive modeling for energy efficiency in coal-fired power plants.

2. METHOD

2.1. Data Collection and Preprocessing

A total of 42 operating and control parameters of a coal-fired power plant, as shown in Table 1, have been chosen as predictive variables with 2882 records at typical boiler loading levels ranging from 40% to 100% during stable process conditions. Data preparation is the initial phase in ensuring data quality before it is utilized for model development. In the preparation phase, data cleansing is performed to address missing values, duplicates, and anomalies. The initial stage involves addressing the issue of missing and duplicate values. In this study, we excluded data entries that had missing or duplicated values. To identify data anomalies, we engage in outlier detection using Isolation Forest (iForest) to produce high-quality data for modeling.

2.2. Feature Selection using RFE

This study uses RFE and several other machine learning algorithms as the learning algorithm or estimator. They are Random Forest (RFE-RFR), SVR (RFE-SVR), Decision Tree (RFE-DTR), Lasso Regression (RFE-Lasso), and Ridge Regression (RFE-Ridge). During the feature selection stage with RFE, we designed that the dataset would undergo three distinct scenarios for the number of features to be selected as follows:

(i) RFE Feature selection with 40 selected features
(ii) RFE Feature selection with 35 selected features
(iii) RFE Feature selection with 30 selected features

Each of the RFE feature selection scenarios will produce two (2) datasets: a training-validation dataset and a test dataset from RFE combinations with five (5) different learning algorithms (estimators). Therefore, the total dataset produced from all three (3) schemes will be thirty (30) datasets. In addition, there are two (2) datasets: the training-validation dataset and the test dataset as a baseline that has not undergone any feature selection process. Finally, in this study, thirty-two (32) datasets are involved in the modeling process.

Table 1. Selected operating parameters

No.	Variables	Code	No.	Variables	Code
1	FLUEGAS TEMP	X1	22	NO PORT RR R AIR FLW	X22
2	CF A COAL FLW	X2	23	ECON INL FW PRESS	X23
3	CF B COAL FLW	X3	24	TBN INL MS PRESS	X24
4	CF C COAL FLW	X4	25	MAIN STM TEMP	X25
5	CF D COAL FLW	X5	26	HRH STM PRESS	X26
6	CF E COAL FLW	X6	27	HRH A STM TEMP	X27
7	CF F COAL FLW	X7	28	ECON INL FW FLW	X28
8	PULV A OTL TEMP	X8	29	FDF A AIR VOLUME FLW	X29
9	PULV B OTL TEMP	X9	30	PAF A AIR VOLUME FLW	X30
10	PULV C OTL TEMP	X10	31	GEN LOAD OPUT	X31
11	PULV D OTL TEMP	X11	32	HSE LOAD	X32
12	PULV E OTL TEMP	X12	33	ECON OTL GAS TEMP	X33
13	PULV F OTL TEMP	X13	34	RAPH OTL SECA TEMP	X34
14	TOTAL AIR FLW	X14	35	PULV A PA FLW	X35
15	PULV A PA TEMP	X15	36	PULV B PA FLW	X36
16	PULV B PA TEMP	X16	37	PULV C PA FLW	X37
17	PULV C PA TEMP	X17	38	PULV D PA FLW	X38
18	PULV D PA TEMP	X18	39	PULV E PA FLW	X39
19	PULV E PA TEMP	X19	40	PULV F PA FLW	X40
20	PULV F PA TEMP	X20	41	COAL FLW DMN	X41
21	NO PORT FR L INL AIR FLW	X21	42	TOT COAL FLW SP	X42

Since the boiler operation includes a control system designed to respond based on the loading level, multiple parameters are expected to show a significant correlation. Therefore, the problem of multicollinearity is disregarded. The data processing procedures of the coal-fired boiler significantly impact the accuracy of the model's predictions. The wide range of boiler operating parameters can cause significant variations in their magnitudes, leading to a decrease in the precision of the model. As a result, the model's capacity to accurately represent the relationships between the variables will be diminished. Therefore, it is essential to standardize the initial dataset to reduce the influence of variations in magnitude within the target parameters before starting the modeling process. The z-score method is employed for data pre-processing, as illustrated in equation (1). The Z-score method is a statistical technique employed to standardize and compare data points within a dataset. It quantifies the number of standard deviations by which a data point deviates from the mean of the dataset. It also determines the position of a data point concerning the mean, indicating whether it is above or below the mean and by how many standard deviations. A positive Z-score indicates that the data point is positioned above the mean, whereas a negative Z-score indicates that it is below the mean. This method facilitates the comparison of data points from disparate datasets with varying scales and distributions by standardizing them onto a common scale.

$$z_i^* = \frac{z_i - \mu}{\sigma} \tag{1}$$

where z_i^*represents the parameter value after normalization by the z-score and z_i is the original process data. μ is the mean of the sampled data, σ is the standard deviation, and i is the number of samples.

2.3. Modeling Development

A crucial stage in machine learning practice is selecting the optimal model from several candidates, which involves evaluating each model using an appropriate error measure. Following the implementation of RFE with different learning algorithms as the estimator, the regression model is built using the CatBoost algorithm with the hyperparameter setting, as shown in Table 2. CatBoost is an open-source and publicly available implementation of the gradient-boosting decision tree approach, which has been improved and perfected. In this study, we employed the 10-fold cross-validation technique for each CatBoost model. Following the conclusion of the comprehensive training procedure, we evaluate the models employing distinct test datasets corresponding to each different RFE combination. The testing dataset functions as a set of unseen data that will clarify the model's performance during the evaluation phase.

Table 2. Applied CatBoost's Hyperparameter Setting in Orange Data Mining Software.

No.	Hyperparameter Setting	Values
1	No. of trees	300
2	Learning rate	0.5
3	Limit depth of Individual trees	6
4	Fraction of features for each tree	1
5	Regularization	1.5

2.4. Evaluation Metrics

The models were evaluated by assessing the coefficient of determination (R^2) and root mean squared error (RMSE) on distinct test datasets for each combination of feature selection using RFE. Chicco et al. [34] have suggested that the R2 may provide more meaningful information than RMSE in the context of regression analysis evaluation. Additionally, R2 was proposed as a standardized metric for assessing regression analysis in several scientific disciplines. More intuitively, it is possible to express R^2 as a percentage, while the measures of RMSE have arbitrary ranges. R^2 and RMSE can be expressed below in equations (2) and (3).

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\bar{Y} - Y_i)} \tag{2}$$

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2} \tag{3}$$

where $\bar{Y}$ is the mean of actual value, m signifies the number of samples in the dataset, X_i is the predicted i^{th} value by the models, and Y_i is the actual i^{th} value.

3. RESULT AND DISCUSSION

3.1. Preprocessing Results

Following the completion of the data cleaning process, we removed duplicates and values that were absent from the dataset. After completing the data cleaning procedure, iForest was used to locate and remove any statistically significant outliers from the dataset. To enhance the quality of the dataset before its incorporation into the models, the contamination limit of 10% was determined. It was determined that there were only 2594 entries left in the sample after the outliers were removed. In the context of a regression task, the purpose of this study was to conduct an exhaustive investigation to determine the impact of RFE on the efficiency of CatBoost. To enhance the dataset's quality, we carried out extensive preprocessing processes before beginning the training of the model, including outlier detection using iForest with a threshold of 10% to identify and eliminate any data anomalies.

3.2. Feature Selection Results

The raw data contains features as described in Table 1. They are variables of the operating and control parameters of a coal-fired power plant recorded in a plant historical database. For this regression task, the target feature of the dataset is Flue gas Temperature (X1). During the feature selection process, we used Recursive Feature Elimination (RFE) would be employed as the feature selection method with various combinations of other learning algorithms as follows:

(i) Random Forest (RFE-RFR)
(ii) SVR (RFE-SVR)
(iii) Decision Tree (RFE-DTR)
(iv) Lasso Regression (RFE-Lasso)
(v) Ridge Regression (RFE-Ridge)

Meanwhile, the number of important features to be selected by RFE as important features follows the three scenarios below:

(i) Feature selection with 40 selected features

(ii) RFE Feature selection with 35 selected features

(iii) RFE Feature selection with 30 selected features

The feature selection process was conducted using five different methods: Recursive Feature Elimination with Support Vector Regression (RFE-SVR), Ridge Regression (RFE-Ridge), Random Forest Regression (RFE-RFR), Lasso Regression (RFE-LASSO), and Decision Tree Regression (RFE-DTR). Table 3 demonstrates the RFE results, where 40 important features were selected for predicting the target feature (X1). The features excluded by each RFE method are as follows: X21 was omitted by RFE-SVR, X40 by RFE-Ridge, X42 by RFE-RFR, X2 by RFE-Lasso, and X6 by RFE-DTR. This variation in excluded features highlights the differing criteria and assumptions inherent in each RFE method, underscoring the importance of a comprehensive approach to feature selection.

Table 3. RFE feature selection with 40 selected features

Feature Selection	Selected Features
RFE-SVR	X2; X3; X4; X5; X6; X7; X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X22; X23; X24; X25; X26; X27; X28; X29; X30; X31; X32; X33; X34; X35; X36; X37; X38; X39; X40; X41; X42
RFE-Ridge	X2; X3; X4; X5; X6; X7; X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X21; X22; X23; X24; X25; X26; X27; X28; X29; X30; X31; X32; X33; X34; X35; X36; X37; X38; X39; X41; X42
RFE-RFR	X2; X3; X4; X5; X6; X7; X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X21; X22; X23; X24; X25; X26; X27; X28; X29; X30; X31; X32; X33; X34; X35; X36; X37; X38; X39; X40; X41
RFE-LASSO	X3; X4; X5; X6; X7; X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X21; X22; X23; X24; X25; X26; X27; X28; X29; X30; X31; X32; X33; X34; X35; X36; X37; X38; X39; X40; X41; X42
RFE-DTR	X2; X3; X4; X5; X7; X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X21; X22; X23; X24; X25; X26; X27; X28; X29; X30; X31; X32; X33; X34; X35; X36; X37; X38; X39; X40; X41; X42

In the refined feature selection process, 35 critical features were identified and are detailed in Table 4. This selection was conducted using various Recursive Feature Elimination (RFE) methods paired with different regression models, which revealed notable variations in the features deemed non-essential by each method. Specifically, the RFE-Support Vector Regression (RFE-SVR) excluded features X21, X22, X25, X27, X36, and X41. In contrast, the RFE-Ridge Regression (RFE-Ridge) omitted X21, X25, X28, and X39 through X41. Meanwhile, the RFE-Random Forest Regression (RFE-RFR) identified X2, X5, X6, X26, X37, and X42 as non-important. The RFE-Lasso Regression (RFE-Lasso) excluded a broader range, from X2 to X7. Finally, the RFE-Decision Tree Regression (RFE-DTR) found X2, X5, X6, X26, X39, and X42 to be non-essential.

Table 4. RFE feature selection with 35 selected features

Feature Selection	Selected Features
RFE-SVR	X2; X3; X4; X5; X6; X7; X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X23; X24; X26; X28; X29; X30; X31; X32; X33; X34; X35; X37; X38; X39; X40; X42
RFE-Ridge	X2; X3; X4; X5; X6; X7; X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X22; X23; X24; X26; X27; X29; X30; X31; X32; X33; X34; X35; X36; X37; X38; X42
RFE-RFR	X3; X4; X7; X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X21; X22; X23; X24; X25; X27; X28; X29; X30; X31; X32; X33; X34; X35; X36; X38; X39; X40; X41
RFE-LASSO	X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X21; X22; X23; X24; X25; X26; X27; X28; X29; X30; X31; X32; X33; X34; X35; X36; X37; X38; X39; X40; X41; X42
RFE-DTR	X3; X4; X7; X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X21; X22; X23; X24; X25; X27; X28; X29; X30; X31; X32; X33; X34; X35; X36; X37; X38; X40; X41

In the feature selection scenario with 30 selected features, as shown in Table 5, the integration of RFE with various regression models highlighted significant differences in the features chosen. The RFE with Support Vector Regression (RFE-SVR) identified a comprehensive set of features, excluding X8, X21, X22, X25, X26, X27, X28, X32, X36, X40, and X41. These omissions indicate that the remaining features are crucial for accurately predicting the target variable within the SVR framework. Similarly, the application of RFE with Ridge regression (RFE-Ridge) resulted in a selection of 30 features, omitting X8, X21, X22, X25, X26, X27, X28, X32, X39, X40, and X41. The congruence between the RFE-SVR and RFE-Ridge selections underscores the robustness of the retained features across different regression algorithms, reaffirming their importance in predictive modeling tasks. Integrating Random Forest Regression (RFR) within the RFE framework presented a nuanced selection, excluding X2, X4, X5, X6, X23, X26, X37, X38, X39, X40, and X42. This demonstrates the algorithm-specific nuances in feature selection, highlighting the distinct predictors deemed relevant by the RFR model. The combination of RFE with Lasso regression excluded a range of features from X2 to X12, indicating a different subset of predictors as crucial within the Lasso framework. Lastly, RFE with Decision Tree Regression (RFE-DTR) omitted features X3 to X7, X26, X31, X36, X39, X40, and X42, further showcasing the unique selection criteria of the Decision Tree model.

Table 5. RFE feature selection with 30 selected features

Feature Selection	Selected Features
RFE-SVR	X2; X3; X4; X5; X6; X7; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X23; X24; X29; X30; X31; X33; X34; X35; X37; X38; X39; X42
RFE-Ridge	X2; X3; X4; X5; X6; X7; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X23; X24; X29; X30; X31; X33; X34; X35; X36; X37; X38; X42
RFE-RFR	X3; X7; X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X21; X22; X24; X25; X27; X28; X29; X30; X31; X32; X33; X34; X35; X36; X41
RFE-LASSO	X13; X14; X15; X16; X17; X18; X19; X20; X21; X22; X23; X24; X25; X26; X27; X28; X29; X30; X31; X32; X33; X34; X35; X36; X37; X38; X39; X40; X41; X42
RFE-DTR	X3; X8; X9; X10; X11; X12; X13; X14; X15; X16; X17; X18; X19; X20; X21; X22; X23; X24; X25; X27; X28; X29; X30; X32; X33; X34; X35; X37; X38; X41

These findings highlight the importance of using multiple RFE methods to capture a broad spectrum of important features, ensuring a comprehensive and robust feature selection process for predictive modeling.

3.3. Baseline Model Performance

Throughout this study, we constructed several CatBoost models using 10-fold cross-validation. These models are constructed based on the RFE combination applied, which was RFE with various learning algorithms. Additionally, we utilized separate datasets for training and testing.

Table 6. Baseline model performance (without feature selection)

Training Time (sec.)	RMSE	R^2
10.482	1.543	0.925

The baseline model uses a preprocessed dataset where there is no RFE feature selection is applied. It is identical CatBoost model using same hyperparameter in Table 2. The baseline CatBoost model's performance which is shown in Table 6, served as the point of reference for assessing RFE-improved CatBoost models. The time required for training this model was 10.482 seconds, the RMSE was 1.543, and the R^2 score was 0.925.

3.4. Performance Evaluation of CatBoost Models

3.4.1. Analysis of Feature Selection Impact

In this study, we analyze the impact of RFE in combination with Random Forest, SVR, Decision Tree, Lasso Regression, and Ridge Regression on the time required for training and the accuracy of predictions. A comprehensive analysis of the CatBoost algorithm is carried out with the purpose of analyzing the influence that several different feature selection methods have on the effectiveness of the model. The research investigated the effects of different RFE techniques on the time required for training, RMSE, and R2 score gains compared to the baseline dataset, which did not include feature selection. Table 6 shows the result of changes in training time, RMSE, and R^2, while Table 7 shows the result of modeling using various RFE feature selection scenarios.

Table 7. Model improvement percentage at various feature selection combinations

Feature Selection	Selected Features	Training Time (sec.)	RMSE	R^2
RFE-RFR	40	42.8	4.731	0.757

Feature Selection	Selected Features	Training Time (sec.)	RMSE	R^2
RFE-SVR	40	-4.1	4.083	0.649
RFE-DTR	40	43.6	3.564	0.649
RFE-LASSO	40	-44.3	2.139	0.432
RFE-RIDGE	40	-150.8	0.648	0.108
RFE-RFR	35	48.6	8.231	1.297
RFE-DTR	35	48.4	4.083	0.649
RFE-RIDGE	35	7.1	2.787	0.432
RFE-SVR	35	12.5	2.398	0.432
RFE-LASSO	35	13.7	0.518	0.108
RFE-LASSO	30	1.0	7.907	1.297
RFE-DTR	30	54.0	6.740	1.081
RFE-SVR	30	21.4	3.046	0.541
RFE-RFR	30	53.0	2.852	0.541
RFE-RIDGE	30	18.1	2.528	0.432

When the experiment was carried out using RFE with a selection of 40 features. It is noteworthy that the combination of RFE-RFR and CatBoost brought about a considerable decrease in the amount of time required for training by 42.8%. Both improvements were negligible. The RFE-DTR likewise displayed good results, with a reduction in training time of 43.6%, a substantial drop in RMSE by 3.6%, and an increase in R^2 by 0.65%. These findings are comparable to those that were seen in the previous approach.

In contrast, the RFE-LASSO, RFE-SVR, and RFE-Ridge did not indicate any improvement in the time required for training. In particular, the RFE-Ridge model had a training period that was 150% longer than the baseline model. In the experiment that used 35 distinct features, the results showed clear trends across a variety of feature selection methods. Notably, using RFE with Random Forest Regression in conjunction with CatBoost results in a decrease of 48.6% in the amount of time necessary for training. At the same time, it is associated with a substantial improvement in RMSE by 8.2% and a significant increase in R^2 by 1.3%. Applying RFE with Decision Tree Regression reveals a considerable reduction in the length of the training process, equivalent to a drop of 48.41%. This is similar to the previous example. This decrease in training time is complemented by good results in both RMSE, which displays an increase of 4.1%, and R^2, which reveals an improvement of 0.6%. Both of these improvements are a result of better training. The other three RFE versions, which are referred to as RFE-SVR, RFE-Lasso, and RFE-Ridge, exhibited slight improvements in terms of training time, as well as in terms of assessment metrics such as RMSE and R^2 values. The results of our last experiment, which used RFE with a selection of 30 features and its influence on the CatBoost model's performance, illustrate considerable differences between the various strategies for selecting features. A significant improvement of 53.9% in the time required for the training process is achieved by the combination of CatBoost and RFE-DTR, which uses Decision Tree Regression. RMSE and R^2 values have seen significant increases, with the former increasing by 6.7% and the latter by 1.1%, respectively.

Out of the three different scenarios that were taken into consideration for feature selection, it was discovered that RFE-RFR, which had 35 features that were chosen, showed the most promising outcomes for the CatBoost model and demonstrated the most optimum trade-off between the amount of time needed for training and the gains in RMSE score. When compared to the baseline, the experimental findings show that the amount of time spent training has decreased by 48.6%, the RMSE score has improved by 8.2%, and the R^2 score has improved by 1.3%.

3.4.2. Analysis of CatBoost Performance

From the point of view of model performance, three CatBoost model candidates have shown optimal results in terms of R^2 and RMSE scores, as well as training length. As seen in Table 7, with a low RMSE score of 1.416 and an excellent R^2 score of 0.937, the RFE-RFR, which used 35 features, had the most ideal performance. It is important to note that the training period for this specific model, which is 5.391 seconds, is relatively short compared to the baseline. Compared to the baseline, the RFE-LASSO technique, which uses a 30-feature selection, achieved the second-greatest degree of performance with an RMSE score of 1.421 and an R^2 score of 0.937. However, the training length of the model is 10.4 seconds, which is much longer than the training period of the RFE-RFR model on average.

Table 8. The result of modeling with various RFE combinations

Feature Selection	Selected Features	Training Time (sec.)	RMSE	R^2
RFE-RFR	40	5.993	1.470	0.932
RFE-SVR	40	10.91	1.480	0.931
RFE-DTR	40	5.917	1.488	0.931
RFE-LASSO	40	15.13	1.510	0.929
RFE-RIDGE	40	26.29	1.533	0.926
RFE-RFR	35	5.391	1.416	0.937
RFE-DTR	35	5.408	1.480	0.931
RFE-RIDGE	35	9.733	1.500	0.929
RFE-SVR	35	9.171	1.506	0.929
RFE-LASSO	35	9.051	1.535	0.926
RFE-LASSO	30	10.38	1.421	0.937
RFE-DTR	30	4.826	1.439	0.935
RFE-SVR	30	8.236	1.496	0.930
RFE-RFR	30	4.924	1.499	0.930
RFE-RIDGE	30	8.582	1.504	0.929

Finally, the RFE-DTR approach, which used a collection of 30-feature, demonstrated the third greatest degree of performance, as seen in Table 8. This was shown by an RMSE value of 1.439 and a significantly raised R^2 value of 0.935, the latter of which was much higher than the baseline. Additionally, it is important to point out that the model being considered has the shortest training time, which is 4.83. This makes it more efficient than both RFE-RFR and RFE-LASSO. Although there were marginal improvements in training durations for all feature selection approaches, the RFE-RFR approach with 35 features showed notable potential by exhibiting improved prediction accuracy while maintaining reasonable training efficiency. The results highlight the significance of carefully choosing feature selection methods that align with the demands of the predictive modeling objective while considering the balance between the time required for training and the performance of the model.

Table 9. Applied hyperparameters for each model

Model	Hyperparameter Setting
CATB	No. of trees: 300, learning rate: 0.5, max. depth. Ind. Trees: 6, fraction of features for each tree: 1
XGB	No. of trees: 300, learning rate: 0.5, regularization:1, max. depth. Ind. Trees: 6, fraction of features for each tree/level/split: 1
GBM	No. of trees: 300, learning rate: 0.5, max. depth. Ind. Trees: 6, smallest subset: 2, fraction of training instances: 1
ADB	No. of estimator: 300, learning rate: 0.5, regression loss function: linear
ANN	Neurons in hidden layers: 64/128, Activation: ReLu, Solver: Adam, max iter: 300

Table 10. Comparison of CatBoost performance to other well-known regression models.

Regression Model	Training Time (sec.)	RMSE	R^2
CATB	5.391	1.416	0.937
XGB	6.428	1.587	0.921
GBM	72.815	1.634	0.916
ADB	91.703	1.722	0.907
ANN	26.897	4.524	0.358

3.4.3. Comparative Analysis

To demonstrate the effectiveness of feature selection and enhancement in the performance of CatBoost, we compare the CatBoost model constructed using RFE and Random Forest Regression with other widely recognized machine learning techniques. In comparing CatBoost with various regression models, we set specific hyperparameter configurations, detailed in Table 9. The hyperparameter setup was implemented using the Orange Data Mining software for each predictor. The regression models being examined exhibit an identical feature selection scenario, specifically the RFE combined with Random Forest Regression as the learning algorithm, with 35 features. As indicated in Table 10, it is apparent that CatBoost demonstrates a notably improved performance compared to other algorithms concerning RMSE and R^2 values. CatBoost demonstrated superior performance to other boosting algorithms and artificial neural networks (ANN) across many performance measures, including training time, RMSE, and R^2 scores. The findings suggest that CatBoost is a very appropriate machine learning technique for regression tasks involving datasets of limited size.

4. CONCLUSION

This study conducted a comprehensive experimental investigation to enhance the performance of CatBoost for regression tasks. By integrating Recursive Feature Elimination (RFE) with various learning algorithms, including Decision Tree, Random Forest, Support Vector Regression (SVR), LASSO regression, and Ridge regression, alongside outlier detection and removal using iForest, a more robust preprocessing approach was achieved. Notably, when Random Forest Regression was utilized within the RFE-iForest framework, significant enhancements in CatBoost models were observed.

The results demonstrate the superiority of CatBoost-RFE with Random Forest regressor over other well-known machine learning algorithms such as AdaBoost, XGBoost, Gradient Boosting, and Artificial Neural Networks (ANN). Specifically, the CatBoost-RFE with Random Forest regressor exhibited the highest level of predictive capability, particularly in predicting boiler flue gas exit temperature in coal-fired plants.

Future research endeavors could explore automated search techniques for RFE to determine the optimal number of features, especially when dealing with larger datasets.

REFERENCES

[1] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

[2] S. Karimi, J. Shiri, and P. Marti, "Supplanting missing climatic inputs in classical and random forest models for estimating reference evapotranspiration in humid coastal areas of Iran," *Comput Electron Agric*, vol. 176, 2020, doi: 10.1016/j.compag.2020.105633.

[3] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features support," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.11363

[4] J. T. Hancock and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J Big Data*, vol. 7, no. 1, p. 94, Dec. 2020, doi: 10.1186/s40537-020-00369-8.

[5] "comparison-between-xgboost-lightgbm-and-catboost-using-a-home-credit-dataset".

[6] Y. Xia, L. He, Y. Li, N. Liu, and Y. Ding, "Predicting loan default in peer-to-peer lending using narrative data," *J Forecast*, vol. 39, no. 2, pp. 260–280, Mar. 2020, doi: 10.1002/for.2625.

[7] P. S. Kumar, A. K. K, S. Mohapatra, B. Naik, J. Nayak, and M. Mishra, "CatBoost Ensemble Approach for Diabetes Risk Prediction at Early Stages," in *2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology(ODICON)*, IEEE, Jan. 2021, pp. 1–6. doi: 10.1109/ODICON50556.2021.9428943.

[8] Y. Rathod *et al.*, "Predictive Analysis of Polycystic Ovarian Syndrome using CatBoost Algorithm," in *2022 IEEE Region 10 Symposium (TENSYMP)*, IEEE, Jul. 2022, pp. 1–6. doi: 10.1109/TENSYMP54529.2022.9864439.

[9] S. Ben Jabeur, C. Gharib, S. Mefteh-Wali, and W. Ben Arfi, "CatBoost model and artificial intelligence techniques for corporate failure prediction," *Technol Forecast Soc Change*, vol. 166, p. 120658, May 2021, doi: 10.1016/j.techfore.2021.120658.

[10] N. Nguyen *et al.*, "A Proposed Model for Card Fraud Detection Based on CatBoost and Deep Neural Network," *IEEE Access*, vol. 10, pp. 96852–96861, 2022, doi: 10.1109/ACCESS.2022.3205416.

[11] S. Hussain *et al.*, "A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection," *Energy Reports*, vol. 7, pp. 4425–4436, Nov. 2021, doi: 10.1016/j.egyr.2021.07.008.

[12] R. Punmiya and S. Choe, "Energy Theft Detection Using Gradient Boosting Theft Detector With Feature Engineering-Based Preprocessing," *IEEE Trans Smart Grid*, vol. 10, no. 2, pp. 2326–2329, Mar. 2019, doi: 10.1109/TSG.2019.2892595.

[13] K. M. Ghori, A. Rabeeh Ayaz, M. Awais, M. Imran, A. Ullah, and L. Szathmary, "Impact of Feature Selection on Non-technical Loss Detection," in *2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*, IEEE, Mar. 2020, pp. 19–24. doi: 10.1109/CDMA47397.2020.00009.

[14] A. Sau and I. Bhakta, "Screening of anxiety and depression among seafarers using machine learning technology," *Inform Med Unlocked*, vol. 16, p. 100228, 2019, doi: 10.1016/j.imu.2019.100228.

[15] J. Nayak, B. Naik, P. B. Dash, S. Vimal, and S. Kadry, "Hybrid Bayesian optimization hypertuned catboost approach for malicious access and anomaly detection in IoT nomalyframework," *Sustainable Computing: Informatics and Systems*, vol. 36, p. 100805, Dec. 2022, doi: 10.1016/j.suscom.2022.100805.

[16] N. Bakhareva, A. Shukhman, A. Matveev, P. Polezhaev, Y. Ushakov, and L. Legashev, "Attack Detection in Enterprise Networks by Machine Learning Methods," in *2019 International Russian Automation Conference (RusAutoCon)*, IEEE, Sep. 2019, pp. 1–6. doi: 10.1109/RUSAUTOCON.2019.8867696.

[17] Y. Wang, X. Huang, X. Ren, Z. Chai, and X. Chen, "In-process belt-image-based material removal rate monitoring for abrasive belt grinding using CatBoost algorithm," *The International Journal of Advanced Manufacturing Technology*, vol. 123, no. 7–8, pp. 2575–2591, Dec. 2022, doi: 10.1007/s00170-022-10341-w.

[18] M. Ou, P. Cui, J. Pei, Z. Zhang, and W. Zhu, "Asymmetric Transitivity Preserving Graph Embedding," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 1105–1114. doi: 10.1145/2939672.2939751.

[19] H.-C. Yi, Z.-H. You, and Z.-H. Guo, "Construction and Analysis of Molecular Association Network by Combining Behavior Representation and Node Attributes," *Front Genet*, vol. 10, Nov. 2019, doi: 10.3389/fgene.2019.01106.

[20] F. Lin, E.-M. Cui, Y. Lei, and L. Luo, "CT-based machine learning model to predict the Fuhrman nuclear grade of clear cell renal cell carcinoma," *Abdominal Radiology*, vol. 44, no. 7, pp. 2528–2534, Jul. 2019, doi: 10.1007/s00261-019-01992-7.

[21] A. A. Kolesnikov, P. M. Kikin, and A. M. Portnov, "DISEASES SPREAD PREDICTION IN TROPICAL AREAS BY MACHINE LEARNING METHODS ENSEMBLING AND SPATIAL ANALYSIS TECHNIQUES," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-3/W8, pp. 221–226, Aug. 2019, doi: 10.5194/isprs-archives-XLII-3-W8-221-2019.

[22] J. Fan, X. Wang, F. Zhang, X. Ma, and L. Wu, "Predicting daily diffuse horizontal solar radiation in various climatic regions of China using support vector machine and tree-based soft computing models with local and extrinsic climatic data," *J Clean Prod*, vol. 248, p. 119264, Mar. 2020, doi: 10.1016/j.jclepro.2019.119264.

[23] E. B. Postnikov, B. Jasiok, and M. Chorążewski, "The CatBoost as a tool to predict the isothermal compressibility of ionic liquids," *J Mol Liq*, vol. 333, p. 115889, Jul. 2021, doi: 10.1016/j.molliq.2021.115889.

[24] A. N. Beskopylny *et al.*, "Concrete Strength Prediction Using Machine Learning Methods CatBoost, k-Nearest Neighbors, Support Vector Regression," *Applied Sciences*, vol. 12, no. 21, p. 10864, Oct. 2022, doi: 10.3390/app122110864.

[25] D. Niu, L. Diao, Z. Zang, H. Che, T. Zhang, and X. Chen, "A Machine-Learning Approach Combining Wavelet Packet Denoising with Catboost for Weather Forecasting," *Atmosphere (Basel)*, vol. 12, no. 12, p. 1618, Dec. 2021, doi: 10.3390/atmos12121618.

[26] G. Huang *et al.*, "Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions," *J Hydrol (Amst)*, vol. 574, pp. 1029–1041, Jul. 2019, doi: 10.1016/j.jhydrol.2019.04.085.

[27] W. Xiang, P. Xu, J. Fang, Q. Zhao, Z. Gu, and Q. Zhang, "Multi-dimensional data-based medium- and long-term power-load forecasting using double-layer CatBoost," *Energy Reports*, vol. 8, pp. 8511–8522, Nov. 2022, doi: 10.1016/j.egyr.2022.06.063.

[28] H. Sun, Y. Chen, L. Li, and B. Zhao, "Estimating Sea Surface pCO2 in the North Atlantic based on CatBoost," 2021, doi: 10.20944/preprints202104.0065.v1.

[29] F. Yao, J. Sun, and J. Dong, "Estimating Daily Dew Point Temperature Based on Local and Cross-Station Meteorological Data Using CatBoost Algorithm," *Computer Modeling in Engineering & Sciences*, vol. 130, no. 2, pp. 671–700, 2022, doi: 10.32604/cmes.2022.018450.

[30] M. Luo *et al.*, "Combination of Feature Selection and CatBoost for Prediction: The First Application to the Estimation of Aboveground Biomass," *Forests*, vol. 12, no. 2, p. 216, Feb. 2021, doi: 10.3390/f12020216.

[31] N. H. M. Khalid, A. R. Ismail, N. A. Aziz, and A. A. A. Hussin, "Performance Comparison of Feature Selection Methods for Prediction in Medical Data," 2023, pp. 92–106. doi: 10.1007/978-981-99-0405-1_7.

[32] R. Zhu, G. Ciren, B. Tang, and X. Gong, "Power system short-term voltage stability assessment based on improved CatBoost with consideration of model confidence," *Energy Sci Eng*, vol. 11, no. 2, pp. 783–795, Feb. 2023, doi: 10.1002/ese3.1362.

[33] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *2008 Eighth IEEE International Conference on Data Mining*, IEEE, Dec. 2008, pp. 413–422. doi: 10.1109/ICDM.2008.17.

[34] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, p. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.