International Journal of Advances in Data and Information Systems
Vol. 5, No. 2, October 2024, pp. 157~172
ISSN: 2721-3056, DOI: 10.59395/ijadis.v5i2.1339

Machine Learning Algorithms for Prediction of Boiler Steam Production

Duan Lianzhai[1], Rusdianto Roestam[2], Tjong Wan Sen[3], Hasanul Fahmi[4], Ong ChungKiat[5], Dian Tri Hariyanto[6]
[1,2,3,4] Master of Science in Information, Faculty of Computing, President University, Indonesia
[5]PT. Lontar Papyrus Pulp & Paper Industry, Jambi Province, Indonesia
[6]Department of Information System, Faculty of science and technology, Jambi University, Indonesia

Article Info

Article history:

Received Sep 01, 20xx
Revised Oct 09, 20xx
Accepted Oct 30, 2024

Keywords:

Boiler efficiency
Steam prediction
Machine learning
Data processing
Modeling

ABSTRACT

The continuous increase in global electricity demand has resulted in boiler power plants becoming a significant energy source. The production of steam is a principal indicator of boiler efficiency, and the accurate prediction of steam production is paramount importance for the enhancement of boiler efficiency and the reduction of operational costs. In this study employs a boiler dataset with a steam production capacity of 420 tons per hour. A total of 25 independent variables were extracted from the original 39 variables through data processing and feature engineering for the purpose of prediction analysis. Subsequently, 8 machine learning models were used for modeling predictions. Grid search cross-validation was employed in order to optimise the performance of the model. The models were analysed and assessed using the Mean Squared Error (MSE) metrics. The results show that random forest achieves the highest accuracy among the 8 single models. Based on 8 models, New Bagging ensemble model is proposed, which combined predictions from 8 single models, demonstrated the optimal overall fit and the lowest MSE, achieved the purpose of the research. The present study demonstrates the ability to analyse and predict complex industrial systems with machine learning algorithms, and provides insights into the use of machine learning algorithms for industrial big data analytics and Industry 4.0. Further work could explore using larger datasets and deep learning to make predictions more accurate.

Corresponding Author:

Duan Lianzhai,
Master of Science in Information, Faculty of Computing, President University
Jl. Ki Hajar Dewantara, Kota Jababeka, Cikarang Baru, Bekasi, Indonesia
Email: duan.lianzhai@student.president.ac.id

1. INTRODUCTION

In recent years, with the continuous development of big data and artificial intelligence technology, automation and intelligence have shown greater advantages in the boiler operation process, providing a new research direction for boiler steam production modeling and prediction[1],[2]. Scholars have begun to model the data of the boiler combustion process based on machine learning algorithms[3],[4],[5]. In the boiler operation process, the control and data acquisition system acquire a large amount of working condition data and operating parameters, and this data information can reflect the specific operating conditions and results produced in the boiler combustion process, reflecting the trend of dynamic changes in operating parameters along with time[6]. By collecting a large amount of production data, using statistics, machine learning and

other methods to find the laws contained in the data, then optimize boiler operating parameters to improve the boiler's steam production has become a more popular modeling method[7],[8]. Machine learning modeling can also enable the model to learn and improve itself. These advantages of machine learning have led to new breakthroughs in boiler steam production prediction[9].

In present study, the machine learning model is used to predict the boiler steam production, The main purpose is to improve the boiler combustion efficiency through the machine learning method[10]. Through the prediction model to get the key process parameters, engineers can quickly target the time period of low combustion efficiency, focusing on analyzing and adjusting the key influencing factors, to speed up the processing of low combustion efficiency, and improve the overall process level. Secondly, The model's steam prediction enables energy consumption assessment and optimization, as well as providing guidance on real-time operations[11]. it plays an important role in enterprise production scheduling, energy saving, and capital arrangements. Based on steam production forecasts, enterprises can adjust the operating load of each boiler to meet user demand and reduce operating costs. Reasonable arrangements for boiler start-up, shutdown, and overhaul time can reduce the loss of abnormal shutdowns. Advance planning for each boiler's water, coal, electricity, and other resources can reduce resource wastage[12]. Advance procurement of reserves of coal and other fuel can prevent sudden changes in the market causing shortages of resources and capital turnover issues for the enterprise.

2. RESEARCH METHOD

2.1 Data Processing and Feature Engineering

2.1.1 Sources and types of data

The object of this case is a coal-fired boiler with a capacity of 420 t/h in PT. Lontar. The operating data of the boiler running at 60%-100% of the rated load are selected as the analyze the dataset, and the dataset has been desensitized [13],[14].

As shown in the results below Table 2.1, the statistical values of all the variables are counted. Total 39 variables, independent variable from P0 to P37, dependent variable is targeting predicted variable. The train data size: (2888,39), the test data size: (1925,38). The statistical values including the mean, standard deviation, minimum value, lower quartile, median, upper quartile and maximum value. In the case of the target variable, for example, 50% of the values fall between -0.35 and 0.79, with a maximum of 2.54 and a minimum of -3.04. The mean value is 0.13 and the standard deviation is 0.98.

Table 2.1 Boiler parameter data statistics

	P0	P1	P2	P3	P4	P5	P6	P7	P8	P9	...	P29	P30	P31	P32	P33	P34	P35	P36	P37	target
mean	0.12	0.06	0.29	-0.07	0.01	-0.56	0.18	0.12	0.18	-0.17	...	0.10	0.06	0.13	0.02	0.01	0.01	0.20	0.03	-0.13	0.13
std	0.93	0.94	0.91	0.97	0.89	0.52	0.92	0.96	0.90	0.95	...	1.06	0.90	0.87	0.90	1.01	1.00	0.99	0.97	1.02	0.98
min	-4.34	-5.12	-3.42	-3.96	-4.74	-2.18	-4.58	-5.05	-4.69	-12.89	...	-2.91	-4.51	-5.86	-4.05	-4.63	-4.79	-5.70	-2.61	-3.63	-3.04
25%	-0.30	-0.23	-0.31	-0.65	-0.38	-0.85	-0.31	-0.30	-0.16	-0.39	...	-0.66	-0.28	-0.17	-0.41	-0.50	-0.29	-0.20	-0.41	-0.80	-0.35
50%	0.36	0.27	0.39	-0.04	0.11	-0.47	0.39	0.34	0.36	0.04	...	-0.02	0.05	0.30	0.04	-0.04	0.16	0.36	0.14	-0.19	0.31
75%	0.73	0.60	0.92	0.62	0.55	-0.15	0.83	0.78	0.73	0.04	...	0.75	0.49	0.64	0.56	0.46	0.27	0.60	0.64	0.50	0.79
max	2.12	1.92	2.83	2.46	2.69	0.49	1.90	1.92	2.24	1.34	...	4.58	2.69	2.01	2.40	5.46	5.11	2.32	5.24	3.00	2.54

7 rows × 39 columns

2.1.2 Outlier Detection and Handling

Box plot as one of the tools to describe the statistics, its function has a unique, intuitive and clear identification of outliers in the batch of data, a batch of outliers in the data is worth attention, ignoring the existence of outliers is very dangerous, without elimination of the outliers included in the process of calculating and analyzing the data, the results will bring about a negative impact; pay attention to the emergence of outliers, and analyze the reasons for their emergence, often become a discovery of problems and an opportunity to improve decision making. The distribution of the variables is observed by plotting a box-and-line diagram for all variables, which is shown in Figure 2.1 for 39 variables:

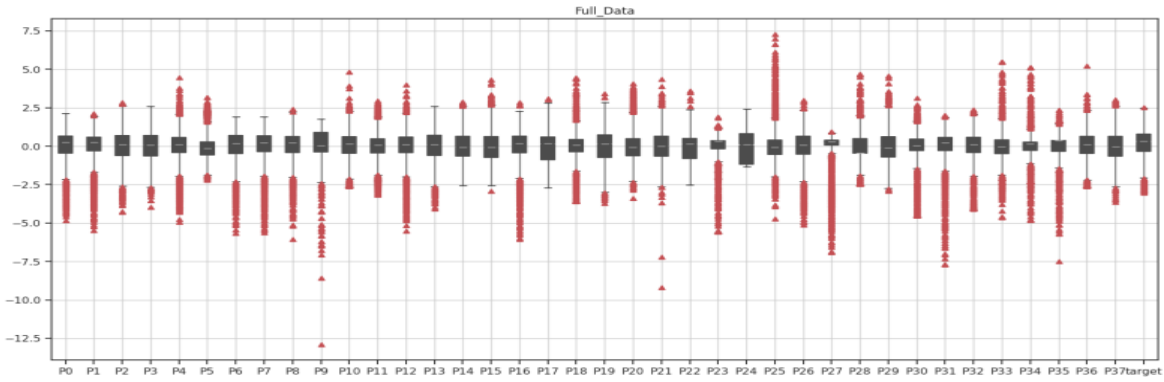

Figure 2.1 Train.txt Box plots of all variables

According to Figure 2.1, the mean values of almost all variables are swaying around 0. Each point in red in the figure represents an outlier, and the discretization of the data distribution of P9, P25, P33, and P34 is particularly prominent, which speculates that there are still many outliers in the dataset, and due to the large number of outliers, it is not appropriate to use a box-and-line diagram here to directly eliminate the outliers, and instead identify the outliers through the subsequent Instead, the outliers are identified by the model after feature screening.

2.1.3 Exploration of variable distribution patterns

The distribution patterns of the variables are of great significance to the study. By observing the distribution of the variables, we can get a general idea of the concentration trend and the degree of variability and skewness of the variables, because many machine learning models require the variables to satisfy the assumption of normality, and even if there is no mandatory requirement of normality, the prediction accuracy of the model built on a dataset with a better distribution pattern will be higher and more stable [15]. Since the ultimate goal of this study is to predict the value of the "target" variable in test.txt, we need to compare the distributions of the variables in the two datasets before modeling with the data in train.txt, so that the feature variables in train.txt, which are manually screened and entered into the model training, should be more accurate and stable. The purpose is to make the feature variables in train.txt, which are manually screened and entered into the model training, as close as possible to the distribution of the corresponding feature variables in test.txt, otherwise the inconsistency of the data will affect the prediction accuracy of the constructed model. The distribution of the variables in the two datasets is shown in Figure 2.2 (due to the large number of variables, it is inconvenient to show them one by one, and only the distribution of individual representative variables is listed for reference).

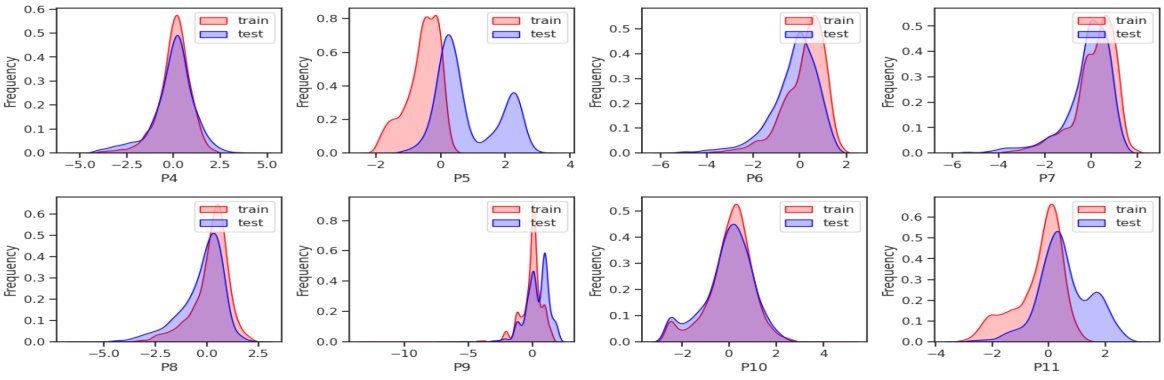

Figure 2.2 The distribution of the variables in the two datasets (part)

By comparing and analyzing the distribution characteristics of the variables on train.txt and test.txt, it can be found that the distributions of the six variables P5, P9, P11, P17, P22, and P28 on the two datasets have relatively large differences, and in order to avoid the overfitting problem of the training model due to the inconsistency of the distribution of the data, in the need of the purpose of the study, it is considered that these six characteristics will be excluded from the data analysis and model construction in this step. In order to avoid the problem of overfitting due to the

inconsistency of the data distribution, these 6 features were excluded from the subsequent data analysis and model construction, and these 6 variables were no longer included in this step. The other remaining 32 features have consistent distributions and are retained.

2.1.4 Data pre-processing methods

In real industrial data, the data we get may be dirty and contain a lot of missing values, a lot of noise and outliers. It may be due to equipment failure or manual input errors, which is very unfavorable to the training of the model and plays a big role in interfering with the final prediction results. So, it is very necessary to carry out data cleaning, because a good data cleaning work can make the data standardized, orderly and neat, and finally provide it to machine learning models and data mining to use. Data cleansing can effectively deal with different issues of data, mainly accuracy, applicability, timeliness, consistency, and authority. The problems we face are different, and there are different ways to deal with different problems. The data studied in this study does not have missing values, but there are a lot of outliers and noise [16].

2.1.5 Missing value processing

There are quite a few outliers in many independent variables in the original data. For example, in the data preprocessing of this study, it is found that the dosage of a raw material or parameter changes caused by a sudden increase or decrease in the amount of steam, at this time it can be considered that the sample belongs to the outliers, which may cause interference in the final prediction of the model, and at this time we can choose to delete the sample. Because the sample data set is not large enough, so directly through each variable by three times the standard deviation of the way to eliminate outliers will cause a large loss of sample data, so this study adopts the ridge regression algorithm to filter out the outliers that exist in the data. Multiple linear regression has a major disadvantage is very sensitive to outliers, because of its loss function for the actual value and the predicted value of the square term, when there is a special deviation from the center of the distribution of the sample points, these outliers will be strong to pull the regression curve to their own side, which has a great impact on the fitting effect of the model, while the ridge regression algorithm in the least squares regression loss function on the basis of the increase in the penalty term to make it more robust characteristics. making it more robust and not as sensitive to the presence of outliers as multivariate linear regression. [17]The distribution of outliers removed by ridge regression in the data is shown in Figure 2.3:

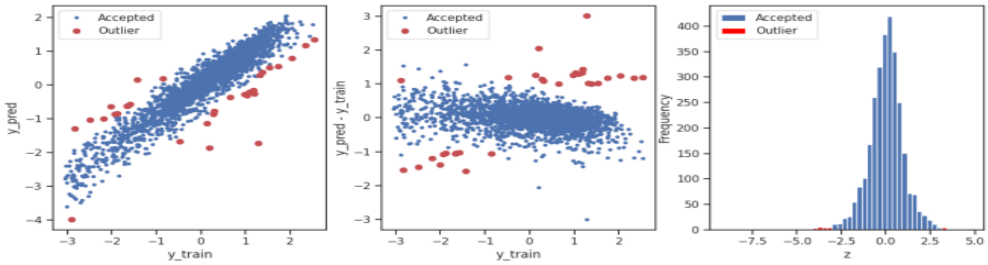

Figure 2.3 Distribution of outliers in the dataset

As shown in Figures 2.3, the horizontal axis of the left graph represents the actual value of the target variable "target", and the vertical axis represents the predicted value of the ridge regression, and the outliers marked in red are all in the position of deviating from the 45-degree line. to 0.5, while the deviation of the outliers is far away; the right graph represents the position of the outliers in the distribution graph of the target variable, which are basically at the most boundary of the target variable values. With 1 to 2888 representing the serial number of each sample point, the method eliminates a total of 31 sample points, accounting for 1% of the total number of sample points, and their serial numbers are [321, 348, 376, 777, 884, 1145, 1164, 1310, 1458, 1466, 1484, 1523, 1704, 1874, 1879, 1979, 2002, 2279, 2528, 2620, 2645, 2647, 2667, 2668,2669, 2696, 2767, 2769, 2807, 2842, 2863]. It is thus argued that by the ridge regression algorithm rejects anomalies to good effect.

2.1.6 Data standardization and normalization

The purpose of the data standardization transformation reduces the data to a smaller interval by a certain proportion so that different variables can be analyzed and compared equally

after standardization; for gradient descent, standardization can achieve the effect of accelerated convergence and does not change the distribution of the original data. Min-Max standardization makes the data samples in the interval of [0, 1], as in the following equation:

$$x^* = \frac{x-\min}{\max-\min} \qquad (2\text{-}1)$$

The Z-Score is standardized so that the variables conform to a standard normal distribution with 0 as the mean and 1 as the standard deviation, as in the following equation:

$$x^* = \frac{x-\mu}{\sigma} \qquad (2\text{-}2)$$

where the mean of the sample feature is μ and the standard deviation of the sample feature is σ.

Since the distribution of each variable is different, this study adopts the more flexible Box-Cox to find the optimal parameter corresponding to each variable, and transforms the variables to normality. The distribution of the transformed variables is compared with that before the transformation as shown in Figure 2.4 (part):

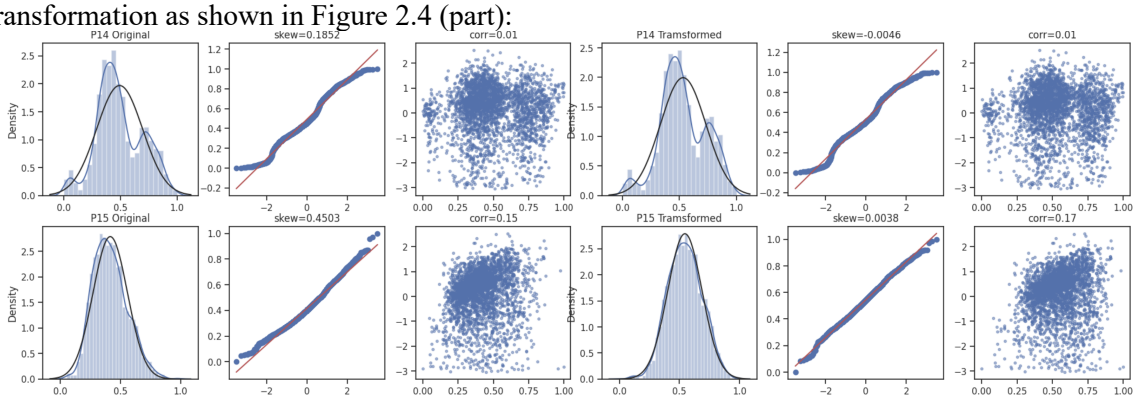

Figure 2.4 Comparison of distribution before and after transformation of independent variables

According to Figure 2.4, the left two columns are the density curve distribution and Quantile-Quantile Plots(QQ plot) of the variables before Box-Cox transformation, and the right two columns are the corresponding plots after the transformation. Both the density curve distribution and the QQ plot show that the normality of the variables has been significantly changed after the Box-Cox transformation, especially in the QQ plot, the distribution of the cumulative probability of the independent variable has been better coincided with the 45-degree curve.

2.1.7 Feature Selection and Engineering

Feature engineering ultimately determines the degree of model fitting, the selection of fewer features leads to poorer model learning ability, resulting in underfitting and ultimately poorer prediction accuracy; if more features are selected, it leads to overlearning ability of the model, resulting in overfitting, which performs well in the training set samples, but performs poorly in the prediction of unknown samples.

Pearson's correlation coefficient is a statistic used to reflect the degree of linear correlation between two variables. The correlation coefficient is denoted by r, which describes the degree of linear correlation between two variables, with larger absolute values of r indicating a stronger correlation. By analyzing the correlation coefficients of the variables in the train.txt data set, we can not only explore which independent variables are strongly correlated with the dependent variable from the preliminary descriptive analysis, but also find out the degree of correlation between the independent variables. The heat map of the distribution of correlation coefficients between variables is shown in Figure 2.5.

The colors in the figure represent the correlation, in which the darker the color means the lower the correlation. Through the figure, we can find that some features have high correlation, such as boiler bed temperature and boiler pressure, furnace bed temperature and furnace bed pressure, although there is a certain degree of correlation, but through the logic of judgment does not affect this part of the features into the model, so do not deal with. The following is the feature

variables and target variables are all put together to do a correlation coefficient matrix. The results are shown in Figure 2.6.

Figure 2.5 Heat map of correlation coefficients between variables

According to Figure 2.6, for the dependent variable "target", the independent variables P0, P1, P8, P27, P31 have strong correlation with it, and these independent variables may have strong influence on the dependent variable. The relationship between P14, P21, P23, P25, P32, P33, P34, P35 and the dependent variable is completely negligible, and the training ability of the models used in this study is relatively good, so the correlation threshold set in this study does not need to be too strict, which is 0.1, and all the independent variables whose correlation with the target variable is less than 0.1 in absolute value are eliminated. After this elimination process, 25 independent variables were retained for modeling predictions. For the independent variables, for example, there is a strong linear correlation between P0 and P1, P8, P27, P31, P12, and between P4 and P12, P15, P29.

	P0	P1	P2	P3	P4	P6	P7	P8	P10	P12	...	P29	P30
P0	1.000000	0.908607	0.463643	0.409576	0.781212	0.189267	0.141294	0.794013	0.298443	0.751830	...	0.302145	0.156968
P1	0.908607	1.000000	0.506514	0.383924	0.657790	0.276805	0.205023	0.874650	0.310120	0.656186	...	0.147096	0.175997
P2	0.463643	0.506514	1.000000	0.410148	0.057697	0.615938	0.477114	0.703431	0.346006	0.059941	...	-0.275764	0.175943
P3	0.409576	0.383924	0.410148	1.000000	0.315046	0.233896	0.197836	0.411946	0.321262	0.306397	...	0.117610	0.043966
P4	0.781212	0.657790	0.057697	0.315046	1.000000	-0.117529	-0.052370	0.449542	0.141129	0.927685	...	0.659093	0.022807
P6	0.189267	0.276805	0.615938	0.233896	-0.117529	1.000000	0.917502	0.468233	0.415660	-0.087312	...	-0.467980	0.188907
P7	0.141294	0.205023	0.477114	0.197836	-0.052370	0.917502	1.000000	0.389987	0.310982	-0.036791	...	-0.311363	0.170113
P8	0.794013	0.874650	0.703431	0.411946	0.449542	0.468233	0.389987	1.000000	0.419703	0.420557	...	-0.011091	0.150258
P10	0.298443	0.310120	0.346006	0.321262	0.141129	0.415660	0.310982	0.419703	1.000000	0.140462	...	-0.105042	-0.036705
P12	0.751830	0.656186	0.059941	0.306397	0.927685	-0.087312	-0.036791	0.420557	0.140462	1.000000	...	0.666775	0.028866
P13	0.185144	0.157518	0.204762	-0.003636	0.075993	0.138367	0.110973	0.153299	-0.059553	0.098771	...	0.008235	0.027328
P14	-0.004144	-0.006268	-0.106282	-0.232677	0.023853	0.072911	0.163931	0.008138	-0.077543	0.020069	...	0.056814	-0.004057
P15	0.314520	0.164702	-0.224573	0.143457	0.615704	-0.431542	-0.291272	0.018366	-0.046737	0.642081	...	0.951314	-0.111311
P16	0.347357	0.435606	0.782474	0.394517	0.023818	0.847119	0.752683	0.680031	0.546975	0.025736	...	-0.342210	0.154794
P18	0.148622	0.123862	0.132105	0.022868	0.136022	0.110570	0.098691	0.093682	-0.024693	0.119833	...	0.053958	0.470341
P19	-0.100294	-0.092673	-0.161802	-0.246008	-0.205729	0.215290	0.158371	-0.144693	0.074903	-0.148319	...	-0.205409	0.100133

Figure 2.6 Correlation Matrix (part)

For the result of steam quantity prediction, using a single variable of the sample does not give good prediction results, for example, it is unreasonable to rely only on combustion raw materials, boiler temperature or boiler pressure, because the influence on the final prediction result includes not only the above factors, but also various adjustable parameters of the boiler, such as primary air, secondary air and so on. So we need to transform and construct the data features, the

method used in this study is to add or multiply several features, this method is called feature combination. A good combination of features plays an obvious role in the prediction of the result.

In the present study features with high correlation are combined to construct combined features. Based on a priori knowledge and data visualization observation, features P0 (boiler bed pressure) and P1 (boiler bed temperature) are summed; features P2 (feedwater volume) and P10 (combustion feed volume) are summed; and features P5 (furnace temperature) and P9 (furnace pressure) are summed. Based on the a priori knowledge that the relationship between each characteristic variable and the steam quantity is not linear, and that there may be quadratic, cubic, or higher order relationships, the squared term characteristics are constructed for characteristics P0 (boiler bed pressure), P1 (boiler bed temperature), P2 (feedwater quantity), P5 (furnace temperature), P9 (furnace pressure), and P10 (combustion feed quantity).

2.2 Machine Learning Algorithm Selection and Model Construction

In the present study, K-Nearest Neighbors(KNN) model, Support Vector Regression(SVR) model and statistical models such as Ridge, Lasso, Elastic Net as well as fusion models such as Gradient Boosting Decision Tree(GBDT), eXtreme Gradient Boosting(XGBoost) and Random Forest are selected for modeling and analyzing the processed data[18], [19]. Based on the fact that all the models used in this study need to be tuned, the training set is first divided by cross-validation to find the optimal parameters of the model and to determine the best model for each algorithm, and at the same time, for individual models with only a single parameter (e.g., Ridge, Lasso), we visualize the change of their modeling effect with the confidence interval of the parameter. Based on the determined optimization parameters, the model is built and the target values of the test set are predicted, and the scatter plots, residual plots, and histograms of the actual and predicted values of the target variables are plotted [20].

In the this study, the Mean Square Error (MSE) and Root Mean Square Error(RMSE) is used to evaluate the predictive effectiveness of the model. The mean square error (MSE) is a convenient measure of the "average error". It is the expected value of the square of the difference between the estimated and true values of a parameter. RMSE is square root of MSE, compared with the MSE, the RMSE can better reflect the generalization ability of the model on different data sets. The formula for the MSE and RMSE is as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2 \qquad (2-4)$$

$$RMSE = \sqrt{MSE} \qquad (2-5)$$

Where f_i is the predicted value and y_i is the true value.

2.2.1 Model parameter tuning

In the this study, through the method of grid search model tuning, through the built-in library of machine learning, according to the model of each parameter, based on the negative value of the mean square error of cross-validation and the fluctuation amplitude of the case of conversion to get a score value, score and root mean square error is positively correlated with each other, and the smaller the score means that the model under the parameter of the model is more effective [21], [22].

(1) KNN model cross-validation, each fold with all the 25 features retained for KNN model training, the value of K from 5 to 80, with the increase of the number of nearest neighbors K, the model effect is gradually good, in K = 10 to reach the best, at this time the score value is 0.717.

(2) Ridge modeling and cross-validation, each fold with all the 25 retained features for Ridge model training, the second-order regular term coefficient u value range from 0 to 1, the step size is 0.01, through the operation, the parameter u value of 0.25 when the model reaches the best effect, the value of the score is 0.888.

(3) Lasso modeling and cross-validation, each fold with all the retained 25 features for Lasso model training, the first-order regular term coefficient y value range from 0 to 1, step size of

0.01, through the operation, the value of the parameter y is 0.0014 when the model effect reaches the best, at this time, the value of the score is 0.888.

(4) Elastic Net modeling and cross-validation, Elastic Net model training with all the retained 25 features per fold, the model sparsity w value range from 0 to 0.5, the step size is 0.001, the first-order regular term coefficients into the range of 0 to 0.2, the step size is 0.001, through the lattice search, the value of parameter w for the 0.002, 1 for the best model effect is achieved when the value of w is 0.002 and 1 is 0.01, and the score value is 0.887 at this time.

(5) Random Forest modeling and cross-validation, each fold with all the 25 retained features for Random Forest model training, using grid search for the number of trees in the range of [10, 150] for the tuning parameter, the maximum depth of each tree in the range of [4, 15] for the tuning parameter, the results show that the number of trees is 100, the depth of the tree for the best results of the model for the depth of the tree is 8. The results show that the model works best when the number of trees is 100 and the depth of the tree is 8, and the score value is 0.981 at this time.

(6) SVR modeling was cross-validated, and the SVR model was trained with all the retained 25 features per fold, and the parameter C in the model was parameterized in the range of [0.1, 1], with a step size of 0.1, and the results showed that the model was most effective when the parameter C was 0.5, and the score value at this time was 0.33.

(7) GBDT modeling and cross-validation, GBDT model training with all the retained 25 features per fold, using grid search for the number of trees in the range of [10, 150] for the reference, step size of 10, the iteration speed of the parameters in the model, i.e., the learning rate of leaning rate in the range of [0.1,1] for the reference, step size of 0.01, the results show that The model works best when the number of trees is 90 and the leaning rate is 0.17, and the score value is 0.906 at this time.

(8) XGBoost modeling and cross-validation, XGBoost model training with all the retained 25 features per fold, the number of trees is tuned in the range of [10, 150] using the grid search with a step size of 10, and the iteration speed of the parameters in the model, i.e., the learning rate, leaning rate, is tuned in the range of [0.1, 1] with a step size of 0.01. The results show that the model works best when the number of trees is 80 and the leaning rate is 0.19, and the score value is 0.944 at this time.

2.2.2 Modelling analysis

After determining the optimal parameters of each model, the models were built based on the optimal parameters and all the training sets, and a total of nine models were built in this experiment, which are KNN model, SVR model, Ridge regression, Lasso regression, Elastic Net, Random Forest, GBDT, XGBoost, and Bagging prediction model trained based on multiple base models. Bagging prediction model trained on multiple base models.

The first 8 models have been introduced, and the Bagging prediction model used in the present study, i.e., bootstrap aggregating, is a method that generates multiple different datasets from the original dataset through resampling techniques, and then trains multiple base models. These base models are trained on their respective training datasets, and finally the predictions of each model are combined to make the final prediction. The main goal of Bagging is to reduce the variance of the model predictions and improve the robustness of the model. The primary learner of the fusion model is selected from the first eight models, and the secondary learner uses the traditional multiple regression model. The Bagging prediction model building process in the present study is as follows:

(1) Input training and test sets, and eight primary learners.

(2) First for each primary learner, the model is trained based on the complete training set data and the determined optimal parameters.

(3) For the trained model, facilitate each sample in the training set to get the predicted value corresponding to each model under the sample, thus obtaining 8 predicted values, and then use each predicted value as an independent variable and the target real value corresponding to each sample in the training set as the target variable to build a multiple regression fusion model.

(4) Steps 1~3 have trained the fusion model, next to predict the samples on the test set using the fusion model, you need to get the values corresponding to the 8 independent variables of

the test set corresponding to the fusion model. In fact, the values of these 8 independent variables are the predicted values of these 8 primary learners on the test set.

(5) The predicted values of the eight models on the test set are used as input variables for the trained multiple regression fusion model, which then predicts the predictions for each test set sample.

The following are the prediction results of the test set corresponding to the nine models. The prediction effect of the models is evaluated according to the RMSE and MSE, and the scatter plots, residual plots, and histograms of the actual values and the predicted values of the samples in the test set are plotted. The sample points labeled in red in the three plots represent the samples where the prediction results deviate from the real values, corr represents the correlation between the predicted values and the actual values, and std resid represents the RMSE.

KNN Model: As shown in Figure 2.7, the distribution of the error of the target variable is more discrete, the correlation coefficient of the predicted value and the true value is 0.847, and the RMSE is 0.524, because the value of the target variable is generally small, so the standard deviation of the error of the results of this model is relatively large. Although the linear relationship between the predicted and true values can be seen more clearly, the distribution of these two values on the sample points is more discretized on a 45° straight line, at which point 22 sample points fall outside the triple standard deviation (0.524).

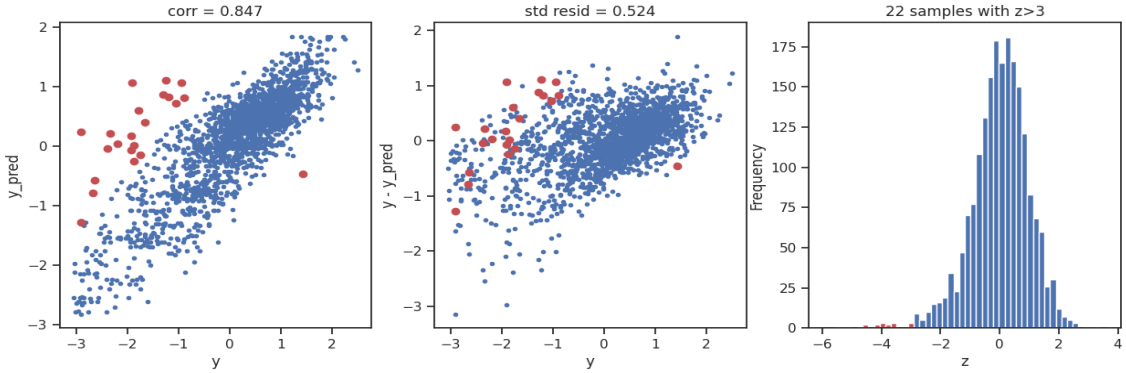

Figure 2.7 Prediction analysis of KNN model

Ridge Model: As shown in Figure 2.8, the error distribution of the target variable has improved significantly compared with the KNN model, the correlation coefficient of the predicted value and the true value is 0.942, and the RMSE is 0.33, the correlation has been strengthened and the root mean square error has been narrowed accordingly. The distribution of the predicted and true values on the sample points is more clustered on a 45° line, and only 7 sample points fall outside the triple standard deviation (0.33) at this point.

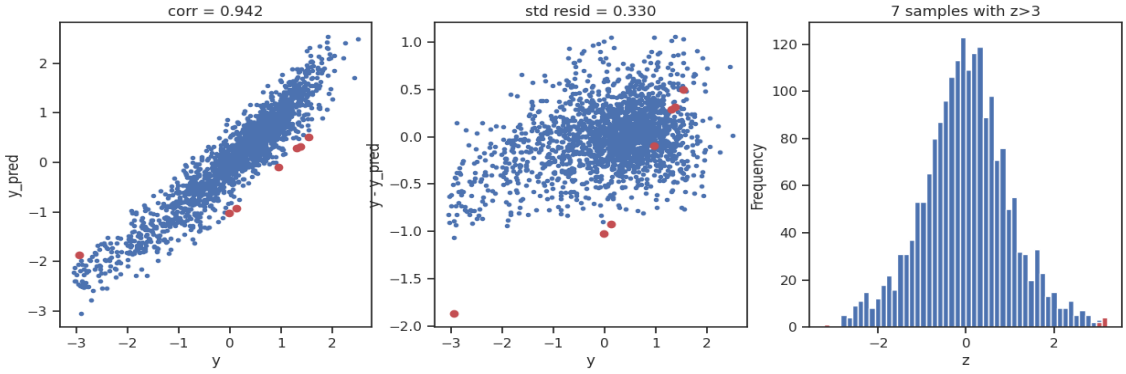

Figure 2.8 Prediction analysis of the Ridge model

Lasso Model: As shown in Figure 2.9, the error distribution of the target variable is better than that of the KNN model, but compared with Ridge, which belongs to the same regression model, there is not much difference, probably because of the sparsity of Lasso's features caused some loss of data, the correlation coefficient of the predicted value and the true value is 0.942, and

the RMSE is 0.33. The distribution of predicted value and true value in the sample points is in the 45° straight line, and it seems to show some nonlinear relationship. The clustering effect on the straight line is average, and also seems to show some non-linear relationship, although at this point only 7 sample points fall outside the triple standard deviation (0.33), but their standard deviation itself is not much different from Ridge.

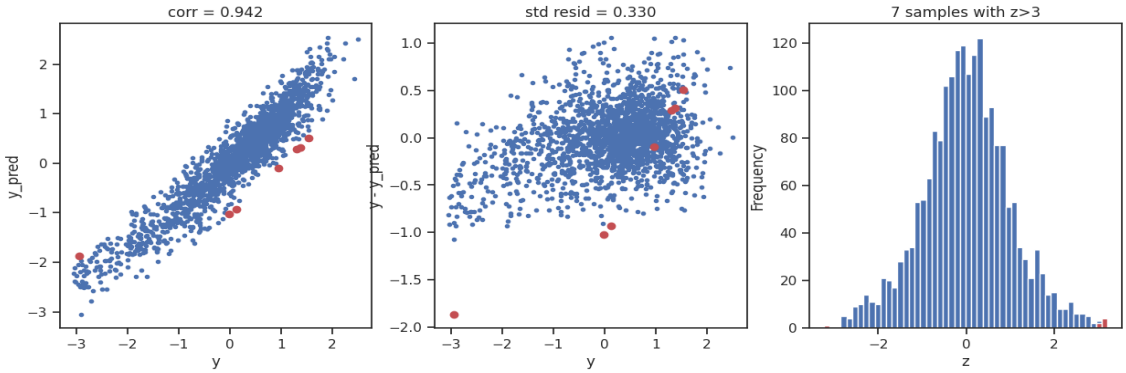

Figure 2.9 Prediction analysis of Lasso model

Elastic Net Model: As shown in Figure 2.10, there is almost no difference between the error distribution of the target variable and that of Ridge except for a few points. Combined with the Modelling results of Ridge and Lasso, it shows that this study is more suitable to use the regression model of the second-order regular penalty term, and the Elastic Net regression model is only closer to the results of the Ridge model in the process of parameterization.

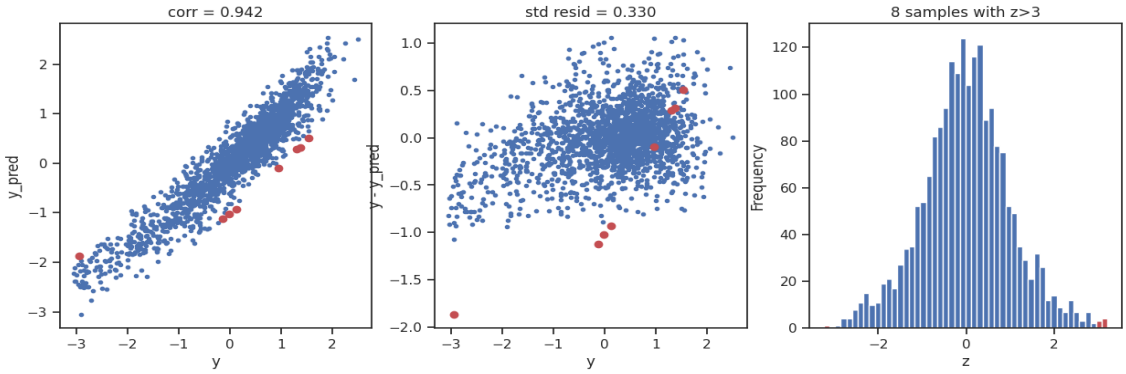

Figure 2.10 Prediction analysis of the Elastic Net model

Random Forest Model: As shown in Figure 2.11, the error distribution of the target variable has further improved relative to the Ridge model, with the correlation coefficient between the predicted and true values reaching 0.991, and the root mean square error shrinking to 0.135. The clustering of the distribution of the predicted and true values at the sample points on a 45° line is already very clear, with 27 sample points falling outside the triple standard deviation (0.135). The distribution of the predicted and true values on the 45° line is now clearly clustered.

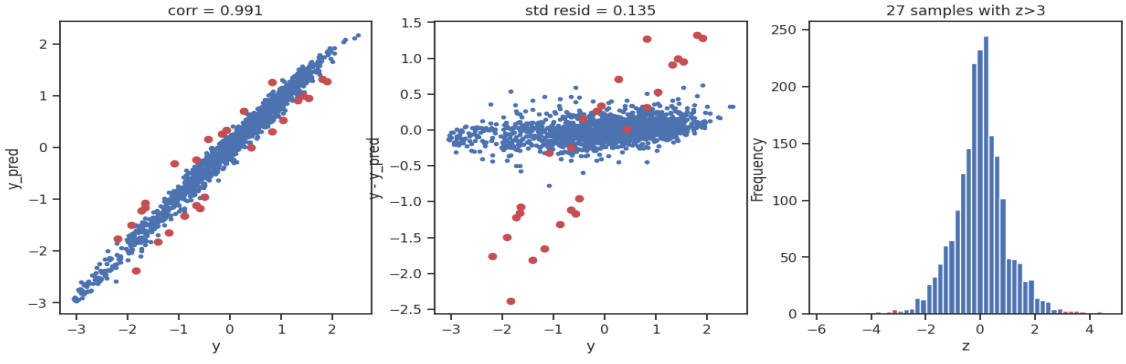

Figure 2.11 Prediction analysis of the Random Forest model

SVR Model: As shown in Figure 2.12, The SVR result is not as good as the random forest, the correlation coefficient of the predicted value and the true value is 0.703, and the RMSE is 0.702. The distribution of predicted value and true value in the sample points is better clustered on the 45° straight line, and it is more neatly and uniformly distributed in the vicinity of the straight line, but the overall effect is not as good as the Random Forest. And with 32 sample points falling outside the triple standard deviation (0.702).

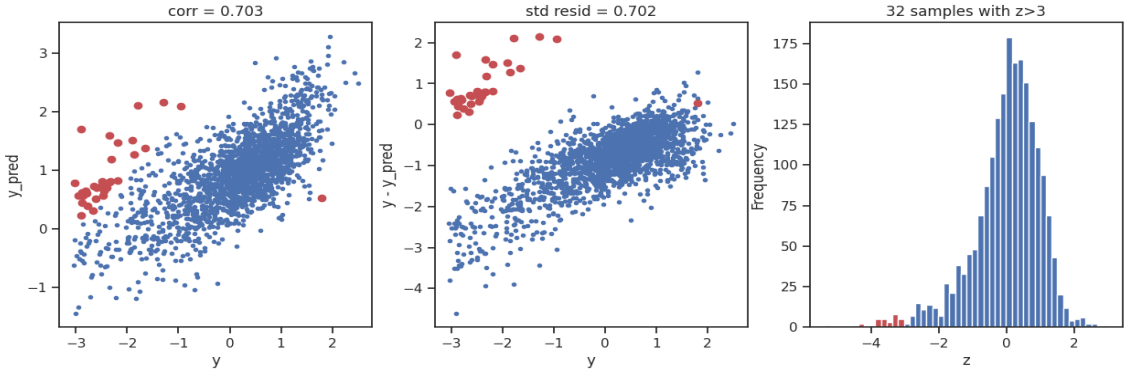

Figure 2.12 Prediction analysis of the SVR model

GBDT Model: As shown in Figures 2.13, the error distribution of the target variable is further improved over the results of Random Forest, but not by much, with a correlation coefficient between the predicted and true values of 0.954, and the RMSE is 0.303. At this point, only 11 sample points fall outside the triple standard deviation (0.303).

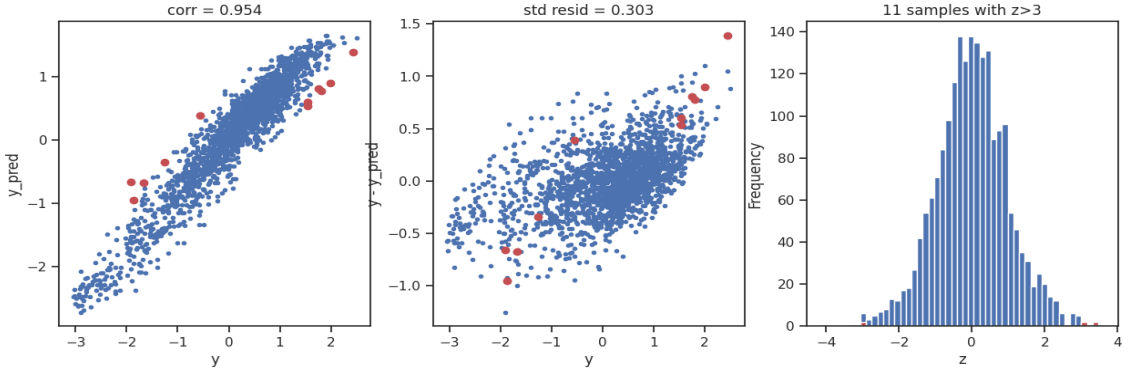

Figure 2.13 Prediction analysis of the GBDT model

XGBoost Model: As shown in Figure 2.14, the model fits the best relative to the other primary models, and the improvement is more obvious, the correlation coefficient of the predicted and true values is 0.972, which is 0.018 higher than that of the GBDT, and the RMSE is 0.233, which is 0.07 smaller than that of the GBDT. The distribution of the predicted and true values in the sample points on a straight line at 45° is very good, and there are still 22 sample points falling outside the triple standard deviation (0.233).

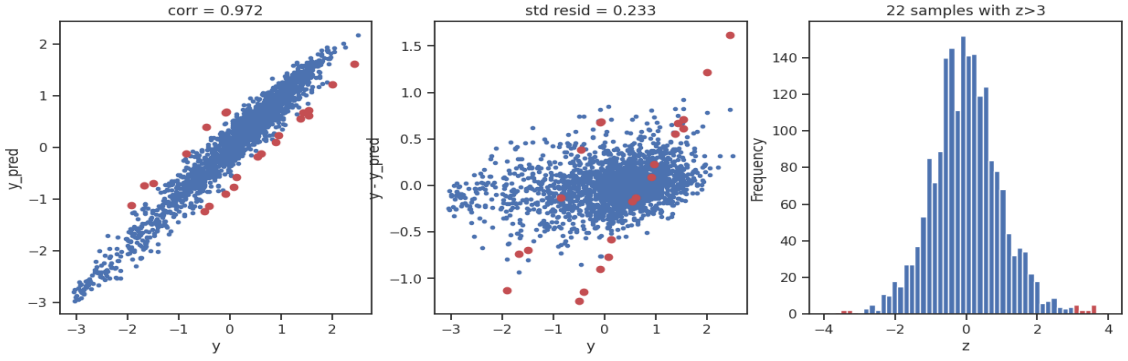

Figure 2.14 Prediction analysis of the XGBoost model

Figure 4.11 Comparison the RMSE for all single models as follows:

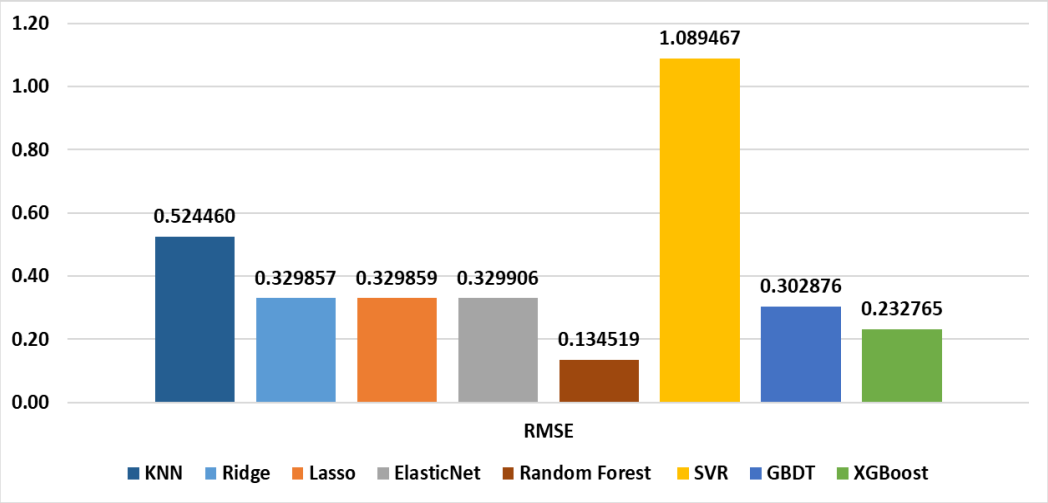

Figure 4.11 Comparison of RMSE of Single Models

As shown in the modelling results Figure 4.11, The top three rankings of the single model training respectively: Random Forest, XGBoost and GBDT; SVR's performance is the worst.

Bagging Model:

```
def model_predict(test_data,test_y=[],stack=False):
        i=0
        y_predict_total=np.zeros((test_data.shape[0],))
        for model in opt_models.keys():
                if model!="LinearSVR" and model!="KNeighbors":
                        y_predict=opt_models[model].predict(test_data)
                        y_predict_total+=y_predict
                        i+=1
                if len(test_y)>0:
                        print("{}_mse:".format(model),mean_squared_error(y_predict,test_y))
        y_predict_mean=np.round(y_predict_total/i,3)
        if len(test_y)>0:
                print("mean_mse:",mean_squared_error(y_predict_mean,test_y))
        else:
                y_predict_mean=pd.Series(y_predict_mean)
                return y_predict_mean
model_predict(X_valid,y_valid)
```

```
Ridge_mse: 0.14302516289475417
Lasso_mse: 0.1430259432271954
ElasticNet_mse: 0.14310838437990978
LinearSVR_mse: 0.14310838437990978
KNeighbors_mse: 0.14310838437990978
GradientBoosting_mse: 0.1496173753253325
XGB_mse: 0.1410394807996852
RandomForest_mse: 0.14116457810553631
mean_mse: 0.12967109342560554
```

Figure 4.9 Multi-model Bagging method prediction

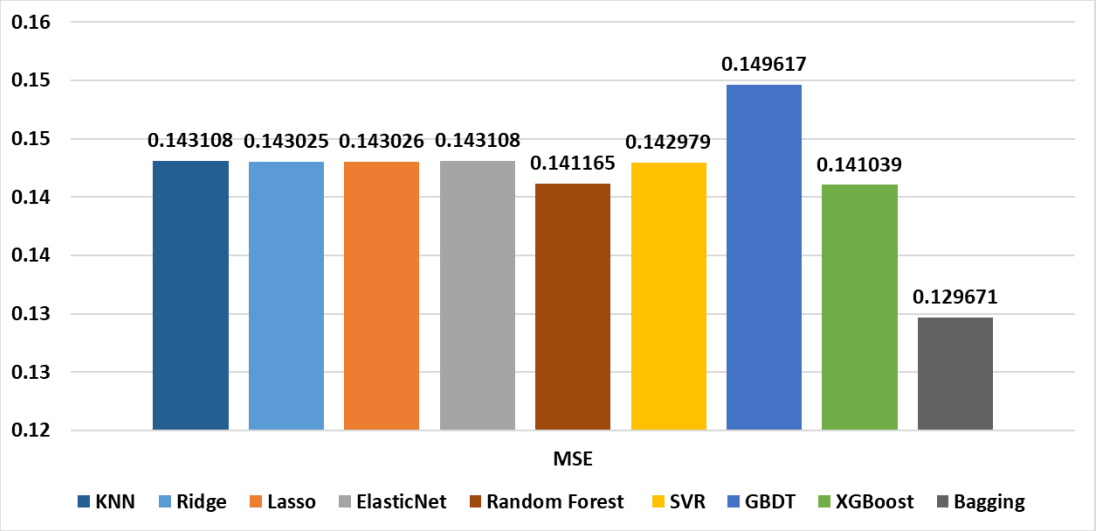

Figure 4.10 Comparison of MSE of Multi-model Bagging models

As shown in Figure 4.9 and Figure 4.10, the model achieved the best fit relative to all the previous models with a root mean square error of 0.1297. It can be seen that the MSE predicted by model fusion is the smallest and the prediction performance is optimal.

3. RESULTS AND DISCUSSION

Combining the characteristics of each model and the specific experimental sessions, the following comparisons were made between the models.

(1) Lasso regression: Lasso regression is similar to Ridge regression in that a penalty term is added to the objective function, which reduces the influence of similar features and thus reduces the complexity of the model. Lasso regression is different from Ridge regression, Lasso use the absolute value of the deviation as a penalty term, resulting in sparse coefficients, from the experimental results of the mean squared error 0.1093 better than Ridge regression (Kernel Ridge), from the experimental results of the analysis of the industrial steam data part of the characteristics of the role of the comparative similarity of the experimental results caused by the interference, so that to the data to carry out the characteristics of the selection is necessary, but the characteristics of the selection of the inappropriate However, improper selection of features may reduce the complexity of the model, thus reducing the performance ability of the model, thus reflecting the advantages of Lasoo regression.

(2) Ridge regression. If there is a very strong covariance between the characteristic variables, it will have a great impact on the regression analysis, Ridge is a program to solve the covariance, and its method is to add a square deviation factor. From the analysis of experimental prediction results, its prediction accuracy is not as good as Lasso regression, and part of the variable parameters in the Lasso regression model in the experiment is 0, which fully indicates that there are a large number of covariate variables in the sample features after the construction of the features, and it can be concluded that the selection of the features before the training of the model is necessary, and it can provide good modeling ideas for the training of the subsequent other models.

(3) ElasticNet regression: ElasticNet regression combines the features of Lasso and Ridge, and from the experimental results, it can be seen that its prediction results are better than the former two. Thus, it can be shown that there are interference features and co-linear features in the sample.

(4) SVR regression. Support vector regression steam prediction is significantly better than multivariate linear regression, because SVR can solve the problem of machine learning in small samples; it can improve the generalization ability; it can solve high-dimensional, nonlinear problems. However, the model is sensitive to missing data and memory consumption, difficult to interpret, the training time is longer than linear regression, and the adjustment is more annoying, in the choice of kernel function radial basis function performance is better. And the model is poorly

interpretable, and it is difficult to give reasonable and clear parameter suggestions when optimizing the boiler parameters, but the model can be considered in the evaluation of energy consumption.

(5) XGBoost: XGBoost steam prediction is better than multivariate linear regression and interpretability is stronger than SVR, but the training time of the model is longer, and interpretability is not as easy to understand as linear regression. It can be considered that XGBoost as a prediction model if the industrial data magnitude is not very large because its only drawback is that the training time is longer.

(6) Random Forest: Random Forest 's steam prediction is the best among all the single models; Random Forest 's speed is about 10 times higher than XGBoost, and its memory consumption is about 3 times lower than XGBoost without reducing the accuracy. The training time is significantly lower than XGBoost, its stability and reliability are still to be tested. It can be considered that Random Forest is going to be the main prediction model, because its prediction accuracy is relatively high, not under the large volume of data, the training time is acceptable, and the model is more interpretable.

(7) Bagging: Through the comparison of the present study, Bagging prediction model is the machine learning model that is most suitable for predicting steam production.

4. CONCLUSION

With the development of science and technology and economy, industrial sensors are becoming more and more extensive, including various switch sensors, pressure sensors, flow sensors, temperature sensors, monitoring sensors and so on, which provide solutions for the collection of power plant operation data. An accurate boiler steam production prediction model provides a reference basis for boiler design and development, and at the same time, it plays a crucial role in optimizing boiler parameters, improving the overall process level, and reducing labor costs [23]. How to be able to dig out the effective information hidden in the data and improve the accuracy of steam quantity prediction has become a new research topic. It requires expertise and background in thermal engineering and machine learning, two completely unrelated disciplines, so there is a lot of attention but relatively few researchers [24].

Machine learning was born from the theory of pattern recognition that computers can learn from themselves when they are not programmed for a specific task, and scientists are interested in whether artificial intelligence can acquire this self-learning ability from massive amounts of data. Based on machine learning and data-driven ideas, it provides a new optimization idea for power boiler optimization [25]. In the this study, each link of steam prediction is analyzed experimentally, model construction and prediction link are the focus of present study, based on the algorithmic aspects of present study, The current study mainly carries out some analysis and improvement of this link of model fusion, the main work of this study is shown as follows:

(1) The features related to boiler steam production prediction data were analyzed, and the features were combined and constructed based on a priori knowledge.

(2) We model and predict machine learning algorithms multivariate linear regression (Lasso, Ridge, ElasticNet), support vector regression SVR, and tree models XGBoost, GBDT and Random Forest, then analyze and experimentally validate the strengths and weaknesses of each of them in our experiments.

(3) The improved model integration method proposed in this study has achieved very good results in steam quantity prediction for boiler. Machine learning compared to the traditional empirical method, data analysis method, greatly reducing the experimental cycle and experimental costs, accelerate the cycle of equipment development and the fastest to find the optimal parameters can be put into practical applications, to accelerate the improvement of boiler efficiency plays a role.

Machine learning methods to predict the steam production to improve the power plant efficiency is a research direction of practical significance, which is one of the ways to create green energy for human beings. Based on the research in this study, more detailed research can be carried out in the future from the following aspects:

(1) Collect more power boiler operation data, not only limited to the dozen variables of the research data in the present study, so that more accurate regression prediction models can be constructed [26].

(2) In recent years, computer hardware technology has made rapid development, a large amount of data can be processed, deep learning shows its own advantages, crushing traditional machine learning in some aspects, but the research data in the this study is too small due to the sample size, easy to overfitting under the training of the deep learning model, but with the continuous accumulation of industrial data, we can try to modeling prediction of deep learning methods, such as Multilayer Perceptron [27].

(3) Machine learning algorithms have a very broad prospect in the optimization of power plant, the next step can be based on the model established by machine learning combined with expert knowledge for the development of industrial software platforms to analyze the actual operation data on site. In addition, machine learning classification algorithms have very good application prospects in boiler operation fault diagnosis[28], state detection[29], operating conditions, etc., and it is worthwhile to study them in combination with practical engineering.

REFERENCES

[1] Y. Meng *et al.*, "Application of Machine Learning in Industrial Boilers: Fault Detection, Diagnosis, and Prognosis," Oct. 01, 2021, *John Wiley and Sons Inc*. doi: 10.1002/cben.202100008.

[2] J. Liang, H. Guo, K. Chen, K. Yu, C. Yue, and Y. Ma, "A Survey on Intelligent Optimization Approaches to Boiler Combustion Optimization," *CAAI Artificial Intelligence Research*, p. 9150014, Dec. 2023, doi: 10.26599/air.2023.9150014.

[3] C. Bisset, P. V. Z. Venter, and R. Coetzer, "A systematic literature review on machine learning applications at coal-fired thermal power plants for improved energy efficiency," *International Journal of Sustainable Energy*, vol. 42, no. 1, pp. 845–872, 2023, doi: 10.1080/14786451.2023.2244618.

[4] J. C. Eslick *et al.*, "Predictive Modeling of a Subcritical Pulverized-Coal Power Plant for Optimization: Parameter Estimation, Validation, and Application," 2022.

[5] A. A. M. Rahat, C. Wang, R. M. Everson, and J. E. Fieldsend, "ORE Open Research Exeter A NOTE ON VERSIONS Data-Driven Multi-Objective Optimisation of Coal-Fired Boiler Combustion Systems," 2018. [Online]. Available: http://hdl.handle.net/10871/33710

[6] İ. Çetiner and H. Çetiner, "Developing real-Time Boiler Control Algorithm for Fuel Consumption Savings," *El-Cezeri Journal of Science and Engineering*, vol. 9, no. 2, pp. 853–868, 2022, doi: 10.31202/ecjse.1020132.

[7] M. Jiang, H. Yu, M. Jin, I. Nakamoto, G. T. Tang, and Y. Guo, "Research on Boiler Energy Saving Technology Based on Internet of Things Data," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 28, no. 2, pp. 296–302, Mar. 2024, doi: 10.20965/jaciii.2024.p0296.

[8] R. Mikulandrić, D. Lončar, D. Cvetinović, and G. Spiridon, "Improvement of existing coal fired thermal power plants performance by control systems modifications," *Energy*, vol. 57, pp. 55–65, Aug. 2013, doi: 10.1016/j.energy.2013.02.033.

[9] K. Sujatha, N. Pappa, K. Senthil Kumar, and U. Siddharth Nambi, "Monitoring power station boilers using ANN and image processing," in *Advanced Materials Research*, 2013, pp. 1154–1159. doi: 10.4028/www.scientific.net/AMR.631-632.1154.

[10] T. Sudhakar, Dr. B. A. Prasad, and Dr. K. P. Rao, "Analysis of Process Parameters to Improve Power Plant Efficiency," *IOSR Journal of Mechanical and Civil Engineering*, vol. 14, no. 01, pp. 57–64, Jan. 2017, doi: 10.9790/1684-1401025764.

[11] H. Aiki, K. Saito, K. Domoto, and H. Hirahara, "Boiler Digital Twin Applying Machine Learning KAZUTAKA OBARA *5 SOICHIRO SAHARA *6."

[12] Y. El Kihel, A. El Kihel, A. Bakdid, and H. Gziri, "IJTPE Journal BOILER EFFICIENCY OPTIMIZATION USING ARTIFICIAL INTELLIGENCE AND RSM RESPONSE SURFACE METHOD," *International Journal on "Technical and Physical Problems of Engineering" (IJTPE) Issue*, vol. 49, pp. 85–89, 2021, [Online]. Available: www.iotpe.com

[13] M. Osmić and I. Buljubašić, "THE INFLUENCE OF THE OPERATING PARAMETERS OF THE STEAM BOILER AT THE HEIGHT OF BED OF FUEL ON THE GRATE," 2017. [Online]. Available: https://www.researchgate.net/publication/318394572

[14] H. M. U. Ayub, M. Rafiq, M. A. Qyyum, G. Rafiq, G. S. Choi, and M. Lee, "Prediction of Process Parameters for the Integrated Biomass Gasification Power Plant Using Artificial Neural Network," *Front Energy Res*, vol. 10, Jun. 2022, doi: 10.3389/fenrg.2022.894875.

[15] K. Zhang, F. Li, C. Yi, and L. Huang, "Modelling and Optimization of Boiler Steam Temperature System Based on Neural Network and Genetic Algorithms," in *IOP Conference Series: Earth and Environmental Science*, IOP Publishing Ltd, May 2021. doi: 10.1088/1755-1315/772/1/012042.

[16] H. Salim, Kh. F. Sultan, and R. Jawad, "Comparison between PID and Artificial Neural Networks to Control of Boiler for Steam Power Plant," *Journal of Engineering Sciences*, vol. 6, no. 1, pp. e10–e15, 2019, doi: 10.21272/jes.2019.6(1).e2.

[17] M. Cfp, "APPLICATION OF MACHINE LEARNING ALGORITHMS IN BOILER PLANT ROOT CAUSE ANALYSIS: A CASE STUDY ON AN INDUSTRIAL SCALE BIOMASS UNIT CO-FIRING SUGARCANE BAGASSE AND FURFURAL RESIDUE AT EXCESSIVE FINAL STEAM TEMPERATURES," 2018. [Online]. Available: https://cabidigitallibrary.org

[18] H. Wang, G. Zhang, Y. Huang, and Y. Zhang, "Study on boiler's comprehensive benefits optimization based on PSO optimized XGBoost algorithm," in *E3S Web of Conferences*, EDP Sciences, May 2021. doi: 10.1051/e3sconf/202126101027.

[19] R. Shohet, M. S. Kandil, and J. J. McArthur, "Machine learning algorithms for classification of boiler faults using a simulated dataset," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Oct. 2019. doi: 10.1088/1757-899X/609/6/062007.

[20] H. Pan, W. Zhong, Z. Wang, and G. Wang, "Optimization of industrial boiler combustion control system based on genetic algorithm," *Computers and Electrical Engineering*, vol. 70, pp. 987–997, Aug. 2018, doi: 10.1016/j.compeleceng.2018.03.003.

[21] W. M. Ashraf *et al.*, "Artificial Intelligence Modeling-Based Optimization of an Industrial-Scale Steam Turbine for Moving toward Net-Zero in the Energy Sector," *ACS Omega*, vol. 8, no. 24, pp. 21709–21725, Jun. 2023, doi: 10.1021/acsomega.3c01227.

[22] Y. Ma, S. Liu, S. Gao, C. Xu, and W. Guo, "Optimizing boiler combustion parameters based on evolution teaching-learning-based optimization algorithm for reducing NOx emission concentration," *Mathematical Biosciences and Engineering*, vol. 20, no. 11, pp. 20317–20344, 2023, doi: 10.3934/mbe.2023899.

[23] H. Szczepaniuk and E. K. Szczepaniuk, "Applications of Artificial Intelligence Algorithms in the Energy Sector," Jan. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/en16010347.

[24] S. B. Savargave and M. J. Lengare, "INTELLIGENT MODELING AND OPTIMIZATION OF BOILER DESIGN."

[25] B. D. Ross-Veitía *et al.*, "Machine learning regression algorithms to predict emissions from steam boilers," *Heliyon*, vol. 10, no. 5, Mar. 2024, doi: 10.1016/j.heliyon.2024.e26892.

[26] S. B. Savargave and M. J. Lengare, "Modeling and Optimizing Boiler Design using Neural Network and Firefly Algorithm," *Journal of Intelligent Systems*, vol. 27, no. 3, pp. 393–412, Jul. 2018, doi: 10.1515/jisys-2016-0113.

[27] Z. Tang, H. Dong, C. Zhang, S. Cao, and T. Ouyang, "Deep Learning Models for SO2Distribution in a 30 MW Boiler via Computational Fluid Dynamics Simulation Data," *ACS Omega*, vol. 7, no. 46, pp. 41943–41955, Nov. 2022, doi: 10.1021/acsomega.2c03468.

[28] A. Mukherjee *et al.*, "Award#:DE-FE0031768 PI: Debangsu Bhattacharyya a Other Key Persons from WVU: Other Key Persons from EPRI: Boiler Health Monitoring Using a Hybrid First Principles-Artificial Intelligence Model Motivation: Flexible Operation and Extended Life," 2020. [Online]. Available: www.caiso.com.

[29] W. Chen and G. Liu, "Computational intelligence approach for NOx emissions minimization in a 30 MW premixed gas burner," 2020.