Hindawi
Journal of Advanced Transportation
Volume 2019, Article ID 4109148, 14 pages
https://doi.org/10.1155/2019/4109148

Research Article

A Machine Learning Method for Predicting Driving Range of Battery Electric Vehicles

Shuai Sun[1,2] **Jun Zhang**[3] **Jun Bi**[1,2] **and Yongxing Wang**[1,2]

[1]*School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*
[2]*MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, Beijing Jiaotong University, Beijing 100044, China*
[3]*Yunnan Travelsky Airport Technology Co. Ltd., Kunming 650200, China*

Correspondence should be addressed to Shuai Sun; 17114261@bjtu.edu.cn and Jun Bi; bilinghc@163.com

Received 24 September 2018; Revised 30 November 2018; Accepted 26 December 2018; Published 9 January 2019

Guest Editor: Mohammad H. Y. Moghaddam

It is of great significance to improve the driving range prediction accuracy to provide battery electric vehicle users with reliable information. A model built by the conventional multiple linear regression method is feasible to predict the driving range, but the residual errors between -3.6975 km and 3.3865 km are relatively unfaithful for real-world driving. The study is innovative in its application of machine learning method, the gradient boosting decision tree algorithm, on the driving range prediction which includes a very large number of factors that cannot be considered by conventional regression methods. The result of the machine learning method shows that the maximum prediction error is 1.58 km, the minimum prediction error is -1.41 km, and the average prediction error is about 0.7 km. The predictive accuracy of the gradient boosting decision tree is compared against that of the conventional approaches.

1. Introduction

With the rapid development of automobile industry and the continuous improvement of people's living standard, car ownership and sales continue to rise, which brings a series of energy and environment problems. In the face of increasing energy and environmental problems, the development of new energy vehicles has become a new trend in the automobile industry [1], and the battery electric vehicle (BEV) is the main force of new energy vehicles. However, BEVs have many disadvantages compared with conventional fuel vehicles; for example, the charging station has a sparse distribution, the charging time is too long, and the energy stored per unit of mass is lower in electrochemical batteries with respect to fossil fuels [2]. Besides, users of BEVs have a range anxiety problem that the residual power will be worried about not ensuring to reach the destination. All of these restrict the promotion and development of BEVs. The range anxiety is an easier case to be figured out than other issues in real-world application of BEVs [3]. Therefore, it is of great significance to increase the practicability and reliability of BEVs by improving the driving range prediction accuracy to provide users with reliable information [4].

Therefore, various mathematic methods have been used in the driving range prediction to improve the accuracy and the credibility of it. The driving mode was incorporated into the study of driving range; it indicates that the stable driving habit plays an important role in saving the battery power and extending the driving range [5]. Fuzzy Transform, a model-free method, was adapted to online use for the prediction of remaining range of an electric vehicle [6]. A simple feature-based linear regression framework modeling the distribution parameters was proved to be an efficient approach to compute probabilistic attainability maps and model a driver's route preferences for electric vehicles [4]. A multiobjective problem, the driving range prediction, with maximized electric motor efficiency and minimized energy consumption, was solved to get the optimal speeds, along with the total trip time corresponding to a predicted driving range [7]. In another study, the energy consumption was analyzed and it was found that the electric vehicle has the lower energy consumption in the lower speed and more frequent

Table 1: Statistical comparison of SOC before and after deletion operations.

Interval	Original frequency number	Frequency number after deletion	Original frequency	Frequency after deletion
(0,10]	40	10	0.0003	0.0001
(10,20]	134	54	0.0009	0.0004
(20,30]	1920	1438	0.0127	0.0113
(30,40]	7168	6130	0.0475	0.0481
(40,50]	14914	12638	0.0988	0.0992
(50,60]	18916	15854	0.1253	0.1244
(60,70]	23028	18954	0.1526	0.1487
(70,80]	27148	22468	0.1799	0.1763
(80,90]	24066	21354	0.1595	0.1676
(90,100]	33584	28528	0.2225	0.2239
Total	150918	127428	1	1

stops [1]. In addition, basing on the LR (linear regression) and SVR (support vector regression) and the neural network, genetic algorithm and fuzzy logic intelligent optimization methods were fused into driving range prediction model of energy consumption to improve the prediction accuracy [8]. A real-time method was proposed to estimate the continuous driving range, considering both the driving behavior and the steepness of the driving route [9]. In another research, the relationship between the energy consumption and the load (air conditioning, heating, etc.) was studied to put forward the prediction model in different load and different driving mode [10, 11].

Many studies used to take less factors into account when establishing the prediction model of driving range, which might lead to the poor applicability and prediction accuracy of the model. A series of energy equations based on linear models and Dijkstra's graph search algorithm were derived to calculate the driving range and the route minimizing energy consumption available to EVs based on the real-world traffic condition and topology of the road. However, weight, temperature, and many other parameters were not be included in this work [12]. The battery remaining discharge energy prediction technique was studied by an energy prediction method based on the coupled prediction of future energy-related variables, but the future temperature variation was not be considered [13].

In a word, there are many methods that can predict the driving range and many factors that affect the driving range prediction, but the current studies cannot take both the accuracy and comprehensiveness of them into account [14].

For the problem existing in the previous studies, basing on the gradient boosting decision tree (GBDT), a new driving range prediction method, the machine learning method novel in including a large number of feature variables [15], has been presented to improve both the applicability and the accuracy, considering the real-world working condition, the battery status, and the traffic environment. The organization of the study is as follows. The collecting and processing of data is simply introduced in Section 2. The conventional multiple linear regression model of driving range is established and verified in Section 3. The machine learning method for the prediction problem is presented in Section 4. To investigate the prediction performance of the proposed method, a comparative study is conducted in Section 5. Finally, conclusions are drawn in Section 6. The nomenclature of symbols and abbreviations in this study is shown in Nomenclature.

2. Data Collecting and Processing

2.1. Data Collecting. To begin with we will provide a brief introduction on the data collection. The real-time data of travel status and battery status, collected by vehicle-mounted information collection equipment, was sent to the remote data monitoring center every 5~10 seconds by GPRS wireless transmission network, and for storage. The research in this paper is based on the historical operation data of BEV, of which the type is E150EV produced by Baic New Energy Automobile Co., Ltd., rented and managed by a car-sharing company in Beijing. Of all the rented BEVs, No. 25 BEV, which has the longest running time and the largest data volume, is selected as the main research object. The discharge data, including 596 discharge process and 523,678 original data from March 1, 2015, to March 1, 2016, is extracted from the database, filtered, and processed.

2.2. Data Processing. The information about vehicle state and battery status transmits through the wireless network. In the process, the transmission can be affected by many factors, such as weather, building density, channel conflict, data stability, and so on. Therefore, there will be data losses and errors in the collected data.

For subsequent analysis and modeling, deletion has been operated on the attributes (*SOC*, current, voltage, speed, etc.) of repeated and error data. Table 1, for example, shows the results of frequency number and frequency before and after the deletion operation of *SOC*.

It can be seen from Table 1 that frequency number after the deletion operation is reduced, while frequency after the deletion operation is basically the same as the original frequency. The result proves that the error is generated with the random influence of the driving environment and the data acquisition device, rather than deliberately. As shown above, the frequency number of *SOC* between 0 and 20 is relatively scarce. Besides, previous studies find that the battery

Table 2: Analysis of the performance index of interpolation.

Index	*SOC*	*VTotal*	*MaxV*	*MinV*	*MaxT*	*MinT*	*Speed*	*TotalMile*
RMSE	0.178	3.2	0.0262	0.0251	3.2E-15	0.2331	15.21	0.381
RMSRE	0.0021	0.0098	0.0079	0.0067	1.12E-16	0.012	54.23	0.0001

Table 3: Correlation test result between parameters and driving range.

	No. 25 BEV		No. 15 BEV		No. 12 BEV	
	R	Sig.	R	Sig.	R	Sig.
SOC	-0.998	.000	-0.999	.000	-0.998	.000
Speed	0.029	.775	-0.172	.089	0.021	.791
VTotal	-0.946	.000	-0.826	.000	-0.861	.000
ITotal	0.145	.149	-0.241	.153	0.137	.159
MaxV	-0.827	.000	-0.789	.000	-0.830	.000
MinV	-0.850	.000	-0.808	.000	-0.858	.000
MaxT	0.816	.000	0.970	.000	0.826	.000
MinT	0.889	.000	0.963	.000	0.899	.000
EVD	0.063	.534	0.092	.377	0.081	.517
ETD	-0.163	.104	-0.159	.115	-0.173	.109

R represents Pearson's simple correlation coefficient between parameters and the distance range; Sig. represents the probability value p of t test statistics, and $\alpha = 0.05$.

performance is unstable when the *SOC* of the battery is less than 10%, which is easy to cause irreversible damage to the physical properties of the battery [11]. Battery performance is relatively stable only when *SOC* is above 15%. Therefore, the *SOC* should be greater than or equal to 20% in the calculation of the driving range prediction in this study.

To facilitate subsequent analysis and modeling, Lagrange interpolation method has been used to make up the data gaps, making sure each discharge process complete. To accurately determine the interpolation effect, the root mean square error and the relative error of root mean square are calculated, as shown in Table 2.

In addition, the processed data has been averaged, which conforms to the requirements of modeling for accuracy and standard.

3. Multiple Linear Regression Modeling

3.1. Correlation Analysis. Generally, there are many factors affecting the driving range of BEV under the actual working conditions, including the driver's own characteristics, the vehicle's own parameters, and the road environment, etc. However, only several items of data can be collected and used. Therefore, the performance parameters of battery (*SOC*, voltage, current, and temperature) and state parameters of the vehicle (speed) are chosen to be researched [5].

Considering that the data used in this paper is distance-dependent, Pearson's simple correlation coefficient is used to measure the strength of the correlation degree between the driving range and another variable. The definition of Pearson's simple correlation coefficient is shown in the following:

$$r = \frac{\sum_{i=1}^{n} (x_i - \hat{x})(y_i - \hat{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \hat{x})^2 \sum_{i=1}^{n} (y_i - \hat{y})^2}} \quad (1)$$

where n is the number of samples, x_i and y_i are the variable values of two variables, respectively, and $\hat{x}$ and $\hat{y}$ are the corresponding mean values, respectively. When $|r| > 0.8$, there is a strong linear correlation between the two variables; when $|r| < 0.3$, the linear correlation between the two variables is weak. Corresponding to Pearson's simple correlation coefficient are t test statistics, and its mathematical definition is as follows:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \quad (2)$$

when the probability value p of t-test statistics is less than the significance level α, the two variables are generally considered to have significant linear correlation. Otherwise, there is no significant linear correlation between the two variables.

No. 25 BEV discharge process data, from 09:45 to 14:30 on September 1, 2015, No. 15 BEV discharge process data, from 10:05 to 15:30 on August 11, 2015, and No. 12 BEV discharge process data, from 09:23 to 13:45 on July 20, 2015, are selected to calculate Pearson's simple correlation coefficient and the probability value p, reflecting the correlation between parameters and driving range in a numerical way, as shown in Table 3.

From Table 3, the driving range has a strongly linear relationship with *SOC*, total voltage, maximum cell voltage, minimum cell voltage, maximum cell temperature, and minimum cell temperature ($|R|^{25,15,12}_{SOC,VTotal,MaxV,MinV,MaxT,MinT} > 0.8$, $p^{25,15,12}_{SOC,VTotal,MaxV,MinV,MaxT,MinT} \ll 0.01$), respectively. There is no significant linear relationship between speed, total current, extreme voltage difference, extreme temperature difference, and driving range ($|R|^{25,15,12}_{Speed,ITotal,EVD,ETD} < 0.8$, $p^{25,15,12}_{Speed,ITotal,EVD,ETD} > 0.05$), respectively.

Table 4: Partial correlation test results under SOC is controlled.

	No. 25 BEV		No. 15 BEV		No. 12 BEV	
	R	Sig.	R	Sig.	R	Sig.
VTotal	-0.172	.088	-0.251	.067	-0.209	.045
MaxV	-0.100	.322	-0.214	.388	-0.301	.276
MinV	-0.267	.007	-0.321	.012	-0.491	.034
MaxT	0.408	.000	0.517	.000	0.559	.000
MinT	0.487	.000	0.629	.000	0.537	.000

R represents partial correlation coefficient between parameters and the distance range; Sig. represents the probability value p of t test statistics, and $\alpha = 0.05$.

3.2. Partial Correlation Analysis. In multivariate correlation analysis, Pearson's simple correlation coefficient, however, generally cannot truly reflect the correlation between variables. Because the relationship between variables is more complex at this time, it may be affected by more than one variable respectively. Currently, partial correlation coefficient is a better choice. Partial correlation coefficient reflects the degree of net correlation between variables.

When analyzing the partial correlation between variables x_1 and y, under the condition of controlling the linear action of x_2, the first-order partial correlation coefficient between x_1 and y is defined as follows:

$$r_{y1,2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{\left(1 - {r_{y2}}^2\right)\left(1 - {r_{12}}^2\right)}} \tag{3}$$

where r_{y1}, r_{y2}, and r_{12} is the correlation coefficient of y and x_1, is the correlation coefficient of y and x_2, and is the simple correlation coefficient of x_1 and x_2. The basic steps of partial correlation analysis are as follows: firstly, the null hypothesis is proposed; that is, the partial correlation coefficient between two populations is not significantly different from zero. Secondly, the test statistic of partial correlation analysis is t statistic, whose mathematical definition is shown in the following:

$$t = r\sqrt{\frac{n - q - 2}{1 - r^2}} \tag{4}$$

where r is the partial correlation coefficient, n is the sample number, q is the order number, and n-q-2 is the degree of freedom. Thirdly, calculate the observation value of the t-test statistic and the corresponding probability value p. Lastly, if the probability value p of the t-test statistic is less than the given significance level α, the null hypothesis should be rejected and the partial correlation coefficient of the two populations is significantly different from zero. Otherwise, it is considered that there is no significant difference between the partial correlation coefficient and zero of the two populations.

No. 25 BEV discharge process data, from 09:45 to 14:30 on September 1, 2015, No. 15 BEV discharge process data, from 10:05 to 15:30 on August 11, 2015, and No. 12 BEV discharge process data, from 09:23 to 13:45 on July 20, 2015, are selected to calculate the partial correlation coefficient and the probability value p, determining whether the correlation between each parameter and the driving distance is affected by other parameters.

From Table 3, SOC has the highest absolute value of the simple correlation coefficient, while total current, speed, extremum voltage difference, and extreme temperature difference have no significant linear relationship with the driving range, respectively. Therefore, SOC is selected as control variable, and partial correlation coefficients of total voltage, maximum cell voltage, minimum cell voltage, maximum cell temperature, and minimum cell temperature are calculated.

As can be seen from Table 4, the linear relationship between total voltage, maximum cell voltage and minimum cell voltage, and the driving distance is affected by SOC. Therefore, after controlling the variable SOC, total voltage, maximum cell voltage, and minimum cell voltage have no significant linear effect on the driving range ($|R|^{25,15,12}_{VTotal,MaxV,MinV} < 0.5$, $p^{25,15,12}_{VTotal,MaxV,MinV} > 0.05$). Correspondingly, there is a significant linear correlation between maximum cell temperature, minimum cell temperature, and the driving distance ($|R|^{25,15,12}_{MaxT,MinT} > 0.4$, $p^{25,15,12}_{MaxT,MinT} \ll 0.01$).

According to the above correlation analysis and partial correlation analysis, minimum cell temperature has the second highest correlation with the driving range, so it is selected as the control variable for the partial correlation test of maximum cell temperature and the driving range. From the partial correlation test results, the relationship between the driving distance and maximum cell temperature is affected by minimum cell temperature, and there is no significant linear correlation between them ($R^{25}_{MaxT} = -0.066$, $p^{25}_{MaxT} = 0.519 > 0.05$; $R^{15}_{MaxT} = -0.129$, $p^{15}_{MaxT} = 0.351 > 0.05$; $R^{12}_{MaxT} = -0.218$, $p^{12}_{MaxT} = 0.229 > 0.05$).

3.3. Variable Selection and Modeling. In multivariate linear regression analysis, it is very important to choose the right independent variables to enter the regression model to make it have better generalization ability and higher prediction accuracy. It is necessary that only independent variables that play a major role are retained and the average variation of the dependent variable is described with fewer independent variables. It can avoid the problem of overfitting and generalization ability reducing caused by the entry of all relevant variables into the model. Therefore, based on the correlation analysis and partial correlation analysis results, some parameters that have greater impact on the dependent variable can be considered and selected as the independent variables. On the contrary, other parameters that have little influence on the dependent variable can be ignored.

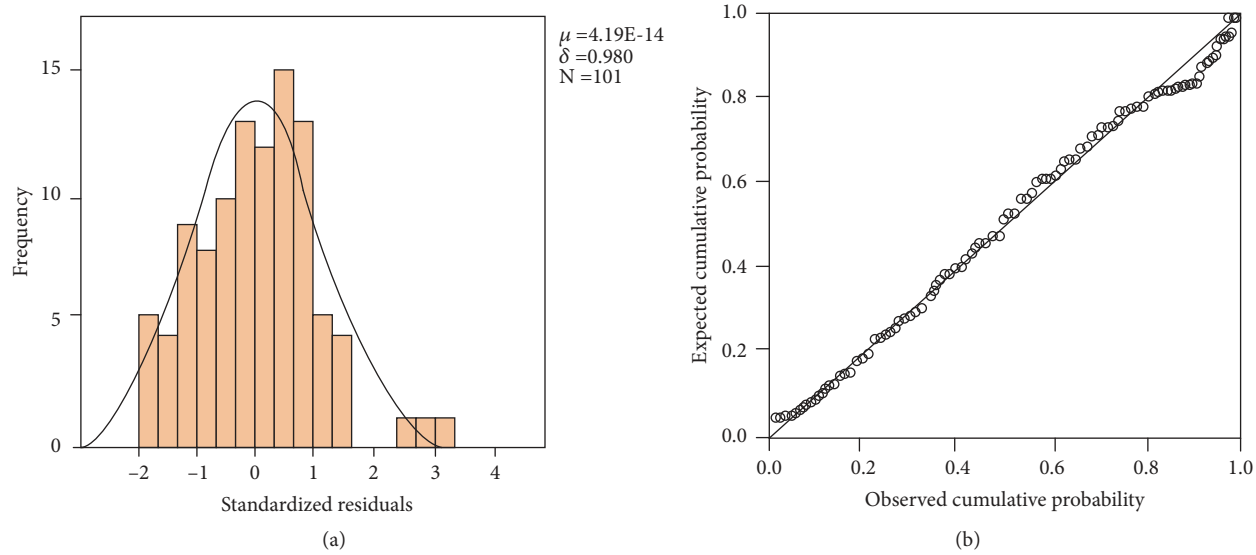

Figure 1: The statistical tests of standardized residuals.

In view of the above result of the correlation analysis and partial correlation analysis, *SOC* and minimum cell temperature have been selected into the variables of the model. The multiple linear regression model is as follows:

$$M = \beta_0 + \beta_1 s + \beta_2 t + \varepsilon \quad (5)$$

where M represents the driving range, the unit being km; s represents SOC, the value ranges from 20 to 100; t represents minimum cell temperature; β_0, β_1, β_2 are parameters to be measured; ε is the residual error.

3.4. Parameter Identification and Statistical Test. When the regression model is determined, it is necessary to use the collected data to identify unknown parameters in the model according to certain estimation criteria. The least square method is widely used to identify parameters because of its excellent properties. No. 25 BEV discharge process data, from 09:45 to 14:30 on September 1, 2015, are used as input, and the least square parameter identification has been performed. The parameter identification results (β_0 = 126.960, β_1 = -1.719, β_2 = 1.627) are introduced into (3), and the driving range prediction model is as follows:

$$M = 126.960 - 1.719s + 1.627t + \varepsilon \quad (6)$$

A variety of statistical tests are conducted to ensure that the model has good stability and generalization ability, and the results are obtained as shown in Figure 1.

As can be seen from Figure 1(a), the residual sequence of the model is basically normal distribution, with the mean value of 4.19E-14, which approximates 0, and the standard deviation is 0.98. Figure 1(b) shows that the residual distribution of the observed value is compared with the normal distribution, standardized residual distribution scatter is very close to the straight line, so that standardized residuals obey normal distribution with mean zero. According to the statistical test results, the goodness of fit is high ($\overline{R}^2$ = 0.996); the linear relationship between the driving range and the explained variables is significant (F = 14095.605, p<0.01); the linear relationship between the driving range and each of the explained variables (*SOC*, *MinT*) is significant (t_1 = 12.328, t_2 = -76.532, t_3 = 5.525, $p_{1,2,3}$ <0.01); there is no autocorrelation between residuals; the residual sequence is independent (DW = 2.23).

To sum up, the multiple regression model satisfies a series of requirements of statistical test, and the model can be used to predict and analyze.

3.5. Model Establishment and Verification. No. 25 BEV total 10 discharge process data, from September 2 to September 22, 2015, have been chosen to conduct the pretreatment and the least squares parameter identification, making the model have higher prediction precision and applicability. Then, the final model parameters were obtained in order: β_0 = 126.527, β_1 = -1.579, β_2 = 1.564. The final driving range prediction model is as follows:

$$M = 126.527 - 1.579s + 1.564t \quad (7)$$

where the value range of s is [20, 100].

No. 15 BEV discharge process data, on September 11, September 19, and September 28, 2015, have been selected to further verify the reliability and practicability of the model. The results of the residual error sequence are shown in Figure 2, and the statistical residual errors are shown in Table 6.

It can be seen from Table 5 that the residual error is between -3.6975 km and 3.3865 km, the mean absolute error is about 1.5 km, the root-mean-square error is less than 2 km, and the root-mean-square relative error is less than 0.5 km. Although it is feasible to predict the driving range by the multiple linear regression model, the residual errors are relatively large for real-world driving condition.

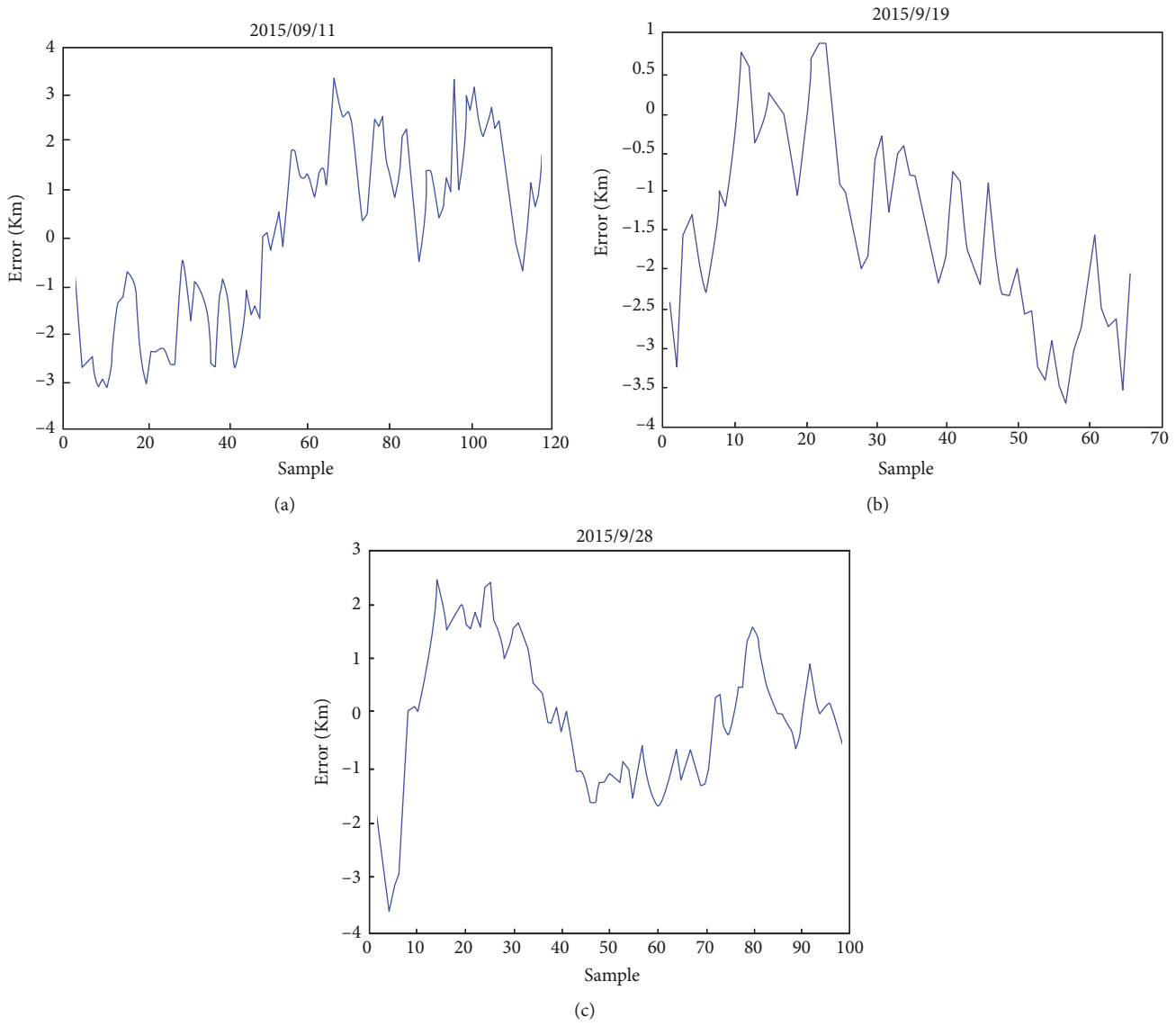

Figure 2: The residual error sequence chart of No. 15 BEV.

Table 5: Statistical error results of prediction.

Date	Maximum error	Minimum error	MAE	RMSE	RMSPE
2015/9/11	3.3865	-3.1664	1.6260	1.8575	0.229
2015/9/19	0.8749	-3.6975	1.5701	1.8523	0.418
2015/9/28	2.5925	-3.6078	1.4612	1.3026	0.377

4. Machine Learning Methods

4.1. Classification and Regression Tree. Decision tree is a kind of classification and regression method. Decision tree method generally includes three processes: the feature selection, the tree creation, and the tree pruning (remove fitting). It can summarize some good performance classification rules from training set, which not only can well fit the training data, but also can make well predictions to the unknown data.

Table 6: Prediction error of the GBDT model.

Date	Minimum error	Maximum error	MAE
2015/3/10	-0.72	1.49	0.61
2015/8/21	-0.59	1.58	0.82
2016/1/9	-1.41	1.52	0.76

Classification and Regression Tree (CART) was put forward by Breiman et al. in 1984, different with ID3 and

C4.5 classification tree whose none-leaf nodes have multiple branches, CART's none-leaf nodes only have two branches, and its output values of a leaf node are the mean of the sample label [16]. Therefore, the generation process is to construct the binary decision tree based on the training set recursively, and to prune the generated trees by using the loss function and validation set.

4.1.1. CART Generation. First, a training set D is given as follows:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \tag{8}$$

where y is a continuous variable. CART model generated under D is defined as follows:

$$f(x) = \sum_{i=1}^{M} c_i I \quad (x \in R_i) \tag{9}$$

From (9), CART divides its eigenspace into M units $R_1 R_2 \ldots, R_M$, and each unit corresponds to a fixed output value c_i. The generation process of CART can be expressed as follows.

Algorithm Framework 1: CART Generation

Input: A training set $\boldsymbol{D}$;
Output: CART $f(x)$;
Begin

In the characteristic space of D, each region is divided into two subregions recursively, and the optimal output value of each subregion is calculated, and the binary decision tree is constructed.

(1) Solve (10); select the optimal cut variable j and the optimal cut point s.

$$\min_{j,s} \left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right] \tag{10}$$

Equation (11) is the space value of R_1 and R_2:

$$\begin{aligned} R_1(j,s) &= \left\{ x \mid x^{(j)} \le s \right\}, \\ R_2(j,s) &= \left\{ x \mid x^{(j)} > s \right\} \end{aligned} \tag{11}$$

Iterate through the variables j, and then scan the cut point s orderly in the specified cut variable j, and select the value pair (j, s) to make sure that (8) is minimum.

(2) Figure out the corresponding optimal output value:

$$c_i = \frac{1}{N_i} \sum_{x_k \in R_i(j,s)} y_i, \quad x \in R_i, \ i = 1, 2 \tag{12}$$

(3) Continue to call steps (1) and (2) of the two subregions until the stop condition is satisfied.

(4) The input space of D is divided into M regions$R_1 R_2 \ldots, R_M$ and CART is generated (9).
End

By the CART generation algorithm, each time the recursive calculation, the optimal output value is generated from each division unit using the least square error criterion; that is, the optimal output value is the mean of all labels on the unit; a heuristic algorithm is used to solve the optimal cut variables and optimal cut points. The decision tree constructed from the above generation algorithm is called the least square CART.

4.1.2. CART Pruning. In view of the problem of overfitting in the CART generated above, the pruning operation is necessary. The CART pruning is cut from the bottom end of the decision tree to make it simple, so that the unknown data has better generalization ability and higher prediction accuracy.

In the pruning process, the loss function of subtree is calculated by the following:

$$\begin{aligned} C_\alpha(T) &= C(T) + \alpha |T|, \\ C(T) &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \end{aligned} \tag{13}$$

where T represents any subtree, $C(T)$ represents the square error of training data, $|T|$ is the number of leaf nodes of T, $\alpha(> 0)$ represents the fitting degree and the complexity of the model, $C_\alpha(T)$ represents the overall loss in the subtree under α, and the only optimal subtree for fixed α exists. The CART pruning algorithm is given as follows.

Algorithm Framework 2: CART Pruning

Input: $\boldsymbol{T_0}$ constructed from CART generation;
Output: the optimal CART $\boldsymbol{T_\alpha}$;
Begin

(1) Suppose $k = 0$, $T = T_0$;
(2) Suppose $\alpha = +\infty$;
(3) Calculate from the top down on $C(T_t)$, $|T|$, and

$$g(t) = \frac{C(t) - C(T_t)}{|T_t| - 1} \tag{14}$$

$$\alpha = \min(\alpha, g(t)) \tag{15}$$

(4) $g(t) = \alpha$ is pruned by the internal node t, the output value of the leaf node t is calculated by average method, and the tree T is obtained;
(5) $k = k + 1$, $\alpha_k = \alpha$, $T_k = T$;
(6) Determine whether T_k is composed of the root node and two leaf nodes, if it is, $T_k = T_n$; if not, go back to step (3);
(7) Based on the independent verification data set, the cross-validation method is used to select the optimal subtree

T_α in the subtree sequence $\{T_k\}$ $(k = 1, 2, ..., n)$ according to the square error.
End

In the above algorithm, $g(t)$ represents the decrease degree of the total loss function after pruning. It is indicated that (1) the size of the optimal subtree T_α is positively correlated with the size of α; (2) the subtrees in the corresponding subtree sequences $\{T_k\}$ $(k = 1, 2, ..., n)$ are nested by small increments α; (3) in the optimal subtree sequence, each subtree T_k corresponds to one α, so when the optimal subtree T_k is determined, the corresponding α is determined. When the pruning operation is completed, it is possible to integrate the new base learner into the existing GBDT model.

4.2. Gradient Boosting Decision Tree. The CART is used as the base learner in the gradient boosting decision tree (GBDT) [17]. For its excellent performance, GBDT is widely used in various fields of real life.

4.2.1. Estimation Function. The purpose of GBDT algorithm is to estimate the unknown function [18]. Since it is a kind of supervised learning, the prerequisite for learning is to have enough data sets with labels $(x_i, y_i)_{i=1}^N$, where N is the size of the sample set, $x_i = (x_i^{(1)}, x_i^{(2)}, ..., x_i^{(n)})^T$, y_i is the sample label. The purpose of supervised learning is to give an estimation function $\widehat{f}(x)$ to the real function $f : x \longrightarrow y$ and to minimize the loss function $L(y, \widehat{f}(x))$ to improve the accuracy of the prediction, as shown in the following:

$$\widehat{f}(x) = \arg\min_{f(x)} L\left(y, \widehat{f}(x)\right) \tag{16}$$

Equation (16) can also be written to the minimized expected loss form, as shown in the following:

$$\widehat{f}(x) = \arg\min_{f(x)} E_x\left[E_y\left[L\left(y, \widehat{f}(x)\right)\right] \mid x\right] \tag{17}$$

To materialize the target problem, the parameters θ of the search space are limited, as shown in the following:

$$\widehat{\theta} = \arg\min_{\theta} E_x\left[E_y\left[L\left(y, \widehat{f}(x, \theta)\right)\right] \mid x\right] \tag{18}$$

So far, no specific formal assumptions have been made on estimation functions and real functions. Moreover, in most cases, the problem described above does not have a closed form solution, so the recursive numerical process is usually optimized.

4.2.2. Optimization Method. At normal circumstances, the loss function adopted in optimizing is square loss function and index loss function; the general Boosting algorithm (such as AdaBoost) can achieve the goal of optimization. However, for general loss function, it is difficult to adopt common optimization methods. In response to this problem, Freidman proposed GBDT algorithm, using the value of the loss function in the negative gradient direction, as shown in (19), to approximate residuals and fit regression trees, improving the performance of the prediction model.

$$-\left[\frac{\partial L\left(y, f\left(x_i\right)\right)}{\partial f\left(x_i\right)}\right] \tag{19}$$

GBDT is an algorithm to recursively solve prediction model. In the beginning of each stage of solving, unperfect model, a very weak model, can be used only to predict the average of the training set; and then a better model can be got by adding an estimator $h(x)$ to $F_m(x)$, as shown in the following:

$$F_{m+1}(x) = F_m(x) + h(x) \tag{20}$$

According to the empirical risk minimization principle,

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, \gamma\right) \tag{21}$$

$$F_m(x) = F_{m-1}(x) + \arg\min_{h} \sum_{i=1}^{n} L\left(y_i, F_{m-1}\left(x_i\right) + h_m\left(x_i\right)\right) \tag{22}$$

Then, the gradient descent method is used to minimize the loss function, and the model is updated according to the following:

$$F_m(x) = F_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_{F_{m-1}} L\left(y_i, F_{m-1}\left(x_i\right)\right) \tag{23}$$

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}\left(x_i\right) - \gamma \nabla_{F_{m-1}} L\left(y_i, F_{m-1}\left(x_i\right)\right)\right) \tag{24}$$

To sum up, the algorithm framework of GBDT is as follows:

Algorithm Framework 3: GBDT

Input:

(i) A labeled training set D
(ii) Iterations M
(iii) The loss function $L(y, f)$
(iv) The base learner $h(x)$

Output: A prediction model $F_m(x)$
Begin
(1) Initialization model:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, \gamma\right) \tag{25}$$

(2) **For i = 1 to M, do**

Calculated pseudo residuals:

$$\gamma_{im} = -\left[\frac{\partial L\left(y, F\left(x_i\right)\right)}{\partial F\left(x_i\right)}\right]_{F(x)=F_{m-1}(x)} \quad i = 1, 2, \ldots, n \quad (26)$$

Obtain $h_m(x)$ using CART to fit pseudo residual, and calculate the weighted coefficient γ_m:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}\left(x_i\right) + \gamma h_m\left(x_i\right)\right) \quad (27)$$

Update model:

$$F_m\left(x\right) = F_{m-1}\left(x\right) + \gamma_m h_m\left(x\right) \quad (28)$$

(3) Get the final prediction model $F_m(x)$

End

In some cases, overfitting and prediction error bias may occur in the above algorithm. In general, the regularization technique can be used to reduce overfitting effect by controlling the fitting process, so the updating rules of the above algorithm are modified as follows:

$$F_m\left(x\right) = F_{m-1}\left(x\right) + \upsilon \cdot \gamma_m h_m\left(x\right), \quad 0 < \upsilon < 1 \quad (29)$$

where υ is called the "learning rate", which is the weight reduction coefficient of the base learner.

It has been found that a small learning rate ($\upsilon < 0.1$) can significantly improve the generalization ability of the model, but the disadvantage is that the number of iterations is increased. Overall, a regularized GBDT algorithm framework is adopted in the following modeling process.

5. GBDT Modeling

5.1. Data Integration and Feature Extraction. Above all, No. 25 BEV discharge data from March 1, 2015, to March 1, 2016, is selected as training set, and No. 15 BEV discharge process data on January, March, and August 2015, is selected as test set.

Considering the influence of the external environment on BEVs [19], weather information of Beijing urban area needs to be integrated in the training and test set, which comes from the national meteorological science data sharing service platform. The speed variable in the data is processed to average for its frequent change and nonlinear effect on the driving range, namely, the speed of the driving range for k corresponding to the average speed of driving range from 0 to k. In this way, the effect of average speed on the driving range is incorporated into the future prediction model.

The purpose of GBDT algorithm is to extract the structure and essence of the target problem from the original data set. To make the selected features well explain the current problem, the selection of features should meet the following requirements that can construct the prediction model with high efficiency and low consumption, improving prediction accuracy. In fact, the extraction of features is to select the optimal feature set for model training from the original feature set. Good features often improve the prediction accuracy of GBDT algorithm. According to the previous analysis and research, *SOC, MaxT, MinT, MaxV, MinV, TotalV, EDT, EDV, AveSpeed, TotalMile, Temper, Visibility,* and *Precip* are extracted to train and test the model.

5.2. Parameter Setting and Relative Importance Calculation. GBDT algorithm needs to set some key parameters, including each iteration step length, υ; loss function, L; maximum depth of tree, *MaxDepth*; number of iterations, N.

The specific steps of the parameter adjustment of GBDT algorithm are as follows.

(1) According to experience, the maximum depth of the tree is set to 10 (reference range for 6 to 20). Considering the accuracy requirement, the step length is set to 0.1; the loss function is set as the mean square error. Search for appropriate number of iterations within a range of 100 to 400.

(2) Then, the maximum depth of the tree and step length υ are detected and adjusted until the optimal parameters are found.

In practice, the input features rarely have the same correlation. In order to understand the size of contribution of each characteristic in driving range prediction, the relative importance of input variables need to be calculated. The calculation of global relative importance of features is as shown in the following:

$$F_j = \frac{1}{M}\sum_{m=1}^{M} \widehat{F}_j\left(T_m\right) \quad (30)$$

where M is the number of base learners. The importance of feature j in a single tree is as shown in the following:

$$\widehat{F}_j\left(T_m\right) = \sum_{t=1}^{N-1} i_t\left(v_t = j\right) \quad (31)$$

where N is the number of leaf nodes, N-1 is the number of non-leaf nodes, v_t is the characteristic associated with node t, and i_t is the reduction value of square loss after node t division. In short, the importance of a feature is the mean of its importance in all the basic learners.

5.3. Result

5.3.1. Model Establishment. According to the parameter setting method above, the statistical error results of the initial iteration are shown in Figure 3.

As shown in Figure 3, when the number of iterations is $[100, 300]$, the mean absolute error is rising, and the root mean square error is decreasing; when the number of iterations is greater than 300, the mean absolute error shows a downward trend, and the root mean square error is decreased after increasing trend. Since the maximum value of the mean absolute error is only 0.00466 from the minimum, considering the stability of the prediction model, the optimal iteration number is the number of iterations with the minimum root mean square error, 300. Then find the optimal maximum depth of the tree, and its statistical error is shown in Figure 4.

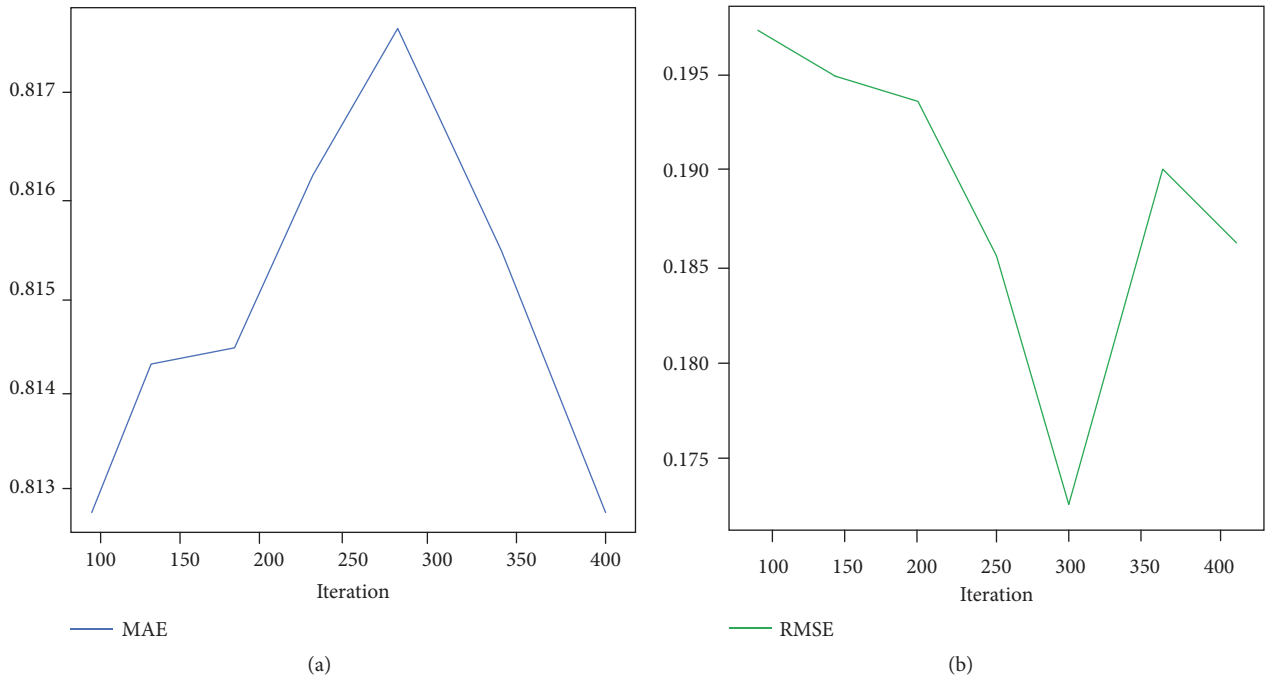

Figure 3: Statistical error results of iterations.

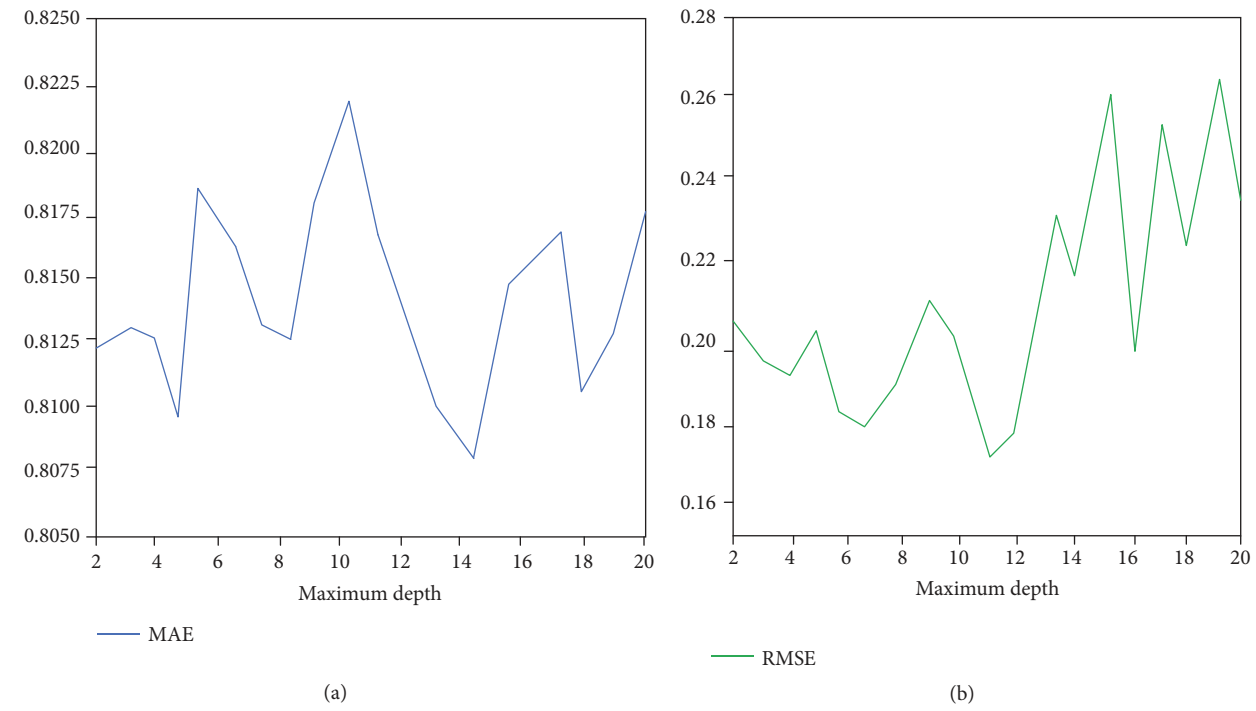

Figure 4: Statistical error results of maximum depth.

From Figure 4, as the maximum depth of the tree increases, the mean absolute error fluctuates. However, the difference between the maximum and minimum values of the mean absolute error is only 0.01212. To make the model have better robustness, the optimal maximum depth should be chosen according to the root mean square error. The minimum mean square error of 0.1733 corresponds to the maximum depth of 11, which is the optimal maximum depth of tree. Then, other optimal parameters are detected as $\upsilon = 0.05$ and $N = 300$.

Training the GBDT model with the optimal parameters, the results are shown in Figure 5.

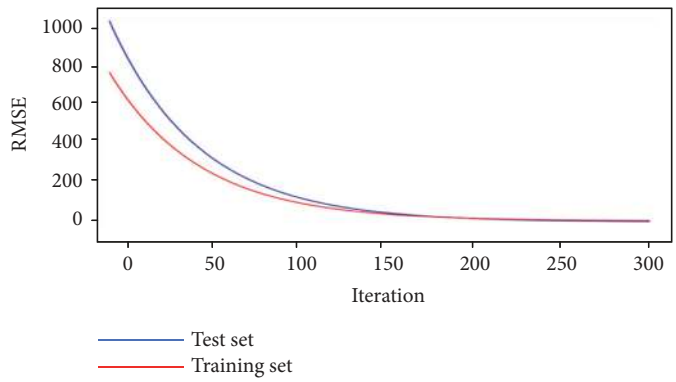

Figure 5: Results of model training.

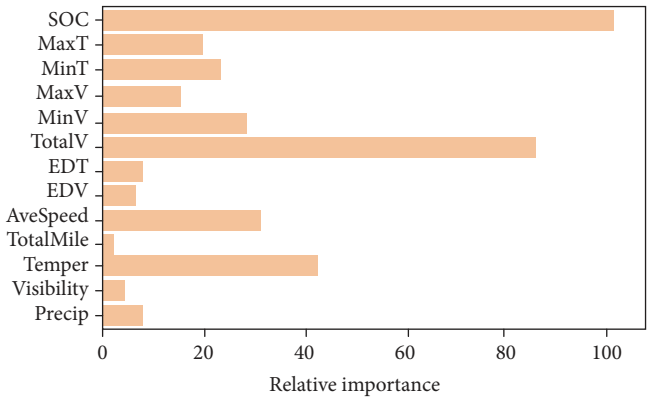

Figure 6: Relative importance of the feature.

Figure 5 shows that, in the beginning of the iteration, both the training set error and the test set error are large; the two errors decrease with the increase of the number of iterations; when the number of iterations reaches about 300, the two error curves basically coincide and stop changing. The error statistical results of the GBDT model are given as RMSE = 0.278, MAE = 0.813, maximum error = 1.61 and minimum error = -1.58.

According to (28), the relative importance of each feature is shown in Figure 6. It can be seen that *SOC* and *TotalV* are the key feature of GBDT model for driving range.

5.3.2. Model Verification. To verify the reliability of the GBDT prediction model, No. 12 BEV discharge process data on March 10, 2015, August 21, 2015, and January 9, 2016, are used for verification; the result is shown in Figure 7.

The results of the minimum error, the maximum error, and the mean absolute error obtained from the verification are shown in Table 6.

Table 6 shows that the maximum prediction error is 1.58 km, the minimum prediction error is -1.41 km, and the average prediction error is about 0.7 km.

5.4. Discussion. No. 15 BEV discharge process data on September 11, 2015, September 19, 2015, and September 28, 2015, have been selected as a data sample. To conduct comparison analysis, three methods, that is, GBDT, CART, and the multiple linear regression (MLR), are performed on the same data sample. The comparison results are shown in Table 7.

When data has many features and the relationships between them are complex, the idea of building a global model is difficult. One approach is to use conventional linear regression analysis to model; some variables will be excluded from the model for the multicollinearity between them. However, that does not mean the model ignores the global impact of other variables. The excluded variables still affect the model because of the existence of the remaining variables. The model established by traditional regression method can be used to predict and the results are reliable. Another approach is to use the decision tree to model; CART is a widely used decision tree. In the regression with CART, each node has a predicted value, which is equal to the average value of all samples belonging to the node. When branching, the best segmentation point of each threshold value of each attribute is selected, and the criterion to be measured is to minimize the mean variance. The value of this node is set as the average value of the training sample that falls on this node until it is indivisible or reaches a certain height or the attribute is used up or the mean square error does not decrease. The test samples are dropped according to the segmentation points during the training and fall to the leaf

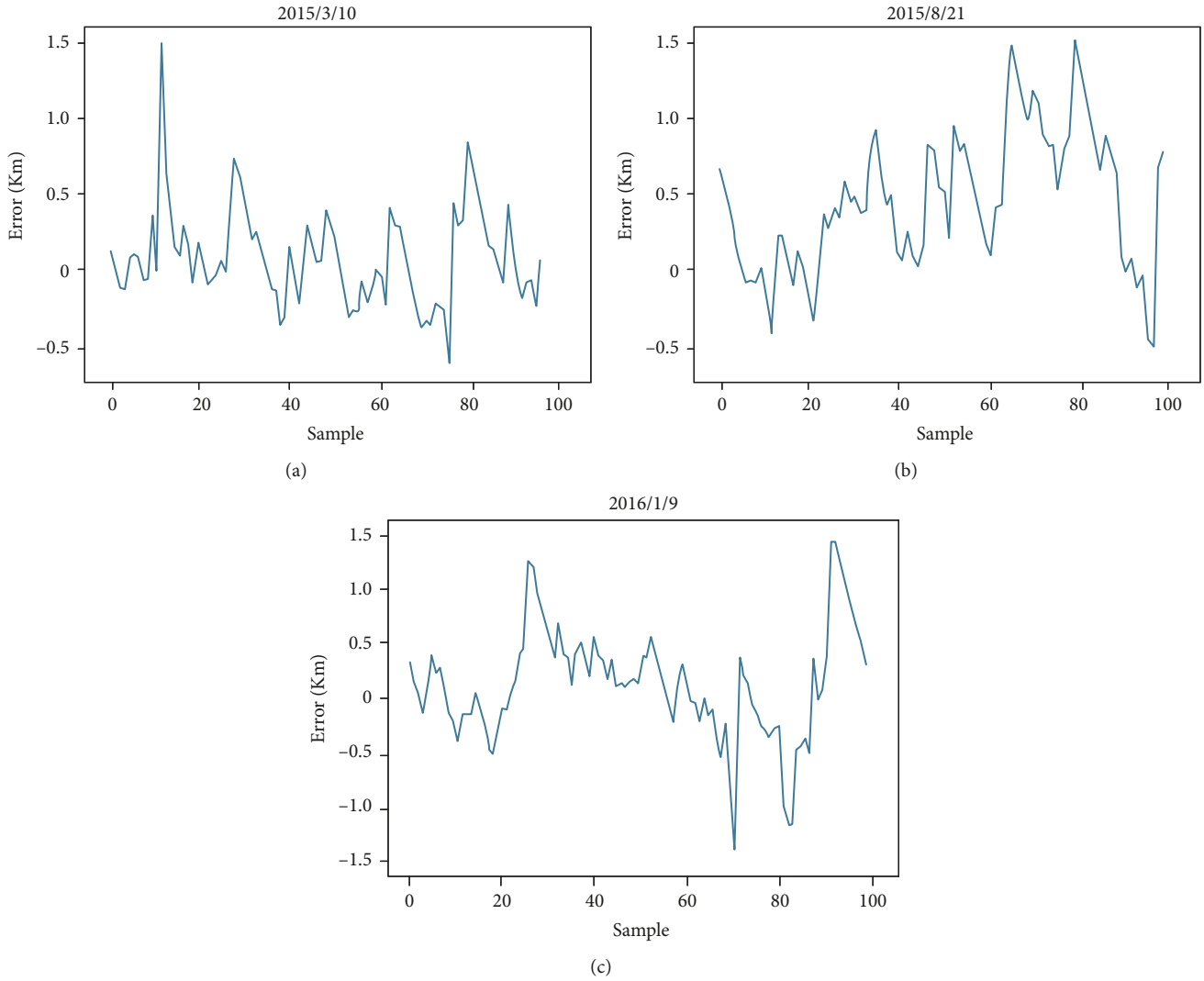

Figure 7: Residual error sequence chart of No. 12 BEV.

Table 7: Comparison results of predictive performance.

Data	Maximum error			Minimum error			MAE		
	GBDT	CART	MLR	GBDT	CART	MLR	GBDT	CART	MLR
2015/9/11	1.234	2.034	3.3865	-0.98	-1.684	-3.1664	0.678	1.237	1.6260
2015/9/19	1.459	1.497	0.8749	-1.023	-1.767	-3.6975	0.719	1.116	1.5701
2015/9/28	1.501	1.972	2.5925	-1.237	-1.791	-3.6078	0.821	1.084	1.4612

nodes. The average value of the leaf nodes is the predicted value. Moreover, GBDT is an algorithm based on CART and is an iterative tree. Obviously, all variables are considered by GBDT and CART in the process of modeling. Although conventional linear regression only introduces some variables in the final model, it is also established based on all variables. Therefore, it is no problem to compare their prediction results.

It can be seen from Table 7 that the predictive performance (maximum error, minimum error, and mean absolute error) of GBDT is overall better than that of CART. GBDT produces a weak classifier through multiple iterations, each iteration produces a weak classifier. Each classifier trains based on the residual of the classifier in the previous round and continuously improves the accuracy of the final classifier by reducing the deviation. Furthermore, even if GBDT and CART takes as many as 13 parameters into consideration, the predictive performance (maximum error, minimum error, and mean absolute error) of MLR model is the worst of them. Therefore, it is apparent that the GBDT model has higher prediction accuracy and reliability. It indicates that the GBDT model of driving range has better predictive performance

and can better meet the requirements of real-world driving conditions.

6. Conclusions

In recent years, the number of BEVs is increasing gradually, but the problem of inaccurate residual power display has been restricting the promotion and the use of BEVs. The purpose of this study is to solve the problem of "range anxiety" caused by battery performance and other factors by predicting the BEV driving range. Many studies usually take less factors into account when establishing the prediction model of driving range, which may lead to the poor applicability and prediction accuracy of the model. In this study, a prediction model for BEV driving range based on machine learning has been established. The study is innovative in its application of machine learning method, GBDT algorithm, which includes a very large number of feature variables that cannot be considered by conventional regression methods. Moreover, the study is novel in its accuracy and reliability of a prediction model for BEV driving range.

As the GBDT model belongs to the black box algorithm, it can only give the importance distribution of the feature variables but cannot specify the interconnection and interaction between the feature variables. In future studies, there is a lot of research space for the correlation of variables within the model. The prediction model proposed in this study can meet the requirements of actual working conditions, but it needs to be further optimized to improve the prediction accuracy in the future. For instance, the cloud computing can be applied in the task of modeling, which is responsible for irregularly training, to obtain more accurate prediction model.

Nomenclature

BEV:	Battery electric vehicle
GBDT:	Gradient boosting decision tree
CART:	Classification and regression tree
MLR:	Multiple linear regression
MAE:	Mean absolute error
RMSE:	Root mean square error
RMSPE:	Root mean square percent error
SOC:	State-of-charge
MaxT:	Maximum cell temperature
MinT:	Minimum cell temperature
MaxV:	Maximum cell voltage
MinV:	Minimum cell voltage
TotalV:	Battery set total voltage
EDT:	Extreme temperature difference
EDV:	Extreme voltage difference
AveSpeed:	Average speed of BEV
TotalMile:	Total driving range of BEV
Temper:	Environment temperature
Visibility:	Horizontal visibility
Precip:	Amount of precipitation.

Data Availability

Real-world operation BEV data used to support the findings of this study have not been made available because we signed a confidential agreement with Yi Weixing (Beijing) Technology Co. Ltd., and all operation data related to commercial secrets is not suitable for disclosure. However, the authenticity and validity of all the data can be guaranteed.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is supported by National Key R&D Program of China under Grants no. 2018YFC0706005 and no. 2018YFC0706000.

References

[1] X. Yuan, L. Li, H. Gou, and T. Dong, "Energy and environmental impact of battery electric vehicle range in China," *Applied Energy*, vol. 157, pp. 75–84, 2015.

[2] M. Ceraolo and G. Pede, "Techniques for estimating the residual range of an electric vehicle," *IEEE Transactions on Vehicular Technology*, vol. 50, no. 1, pp. 109–115, 2001.

[3] A. Shekhar, V. Prasanth, P. Bauer, and M. Bolech, "Generic methodology for driving range estimation of electric vehicle with on-road charging," in *Proceedings of the IEEE Transportation Electrification Conference and Expo, ITEC '15*, June 2015.

[4] P. Ondruska and I. Posner, "Probabilistic attainability maps: Efficiently predicting driver-specific electric vehicle range," in *Proceedings of the 25th IEEE Intelligent Vehicles Symposium, IV '14*, pp. 1169–1174, June 2014.

[5] C. Bingham, C. Walsh, and S. Carroll, "Impact of driving characteristics on electric vehicle energy consumption and range," *IET Intelligent Transport Systems*, vol. 6, no. 6, pp. 29–35, 2012.

[6] I. Cunningham and K. Burnham, "Online use of the fuzzy transform in the estimation of electric vehicle range," *Measurement and Control*, vol. 46, no. 9, pp. 277–282, 2013.

[7] W. Vaz, A. K. R. Nandi, R. G. Landers, and U. O. Koylu, "Electric vehicle range prediction for constant speed trip using multi-objective optimization," *Journal of Power Sources*, vol. 275, pp. 435–446, 2015.

[8] A. Bolovinou, I. Bakas, A. Amditis, F. Mastrandrea, and W. Vinciotti, "Online prediction of an electric vehicle remaining range based on regression analysis," in *Proceedings of the 2014 IEEE International Electric Vehicle Conference, IEVC '14*, December 2014.

[9] C. K. Wai, Y. Y. Rong, and S. Morris, "Simulation of a distance estimator for battery electric vehicle," *Alexandria Engineering Journal*, vol. 54, no. 3, pp. 359–371, 2015.

[10] E. Kim, J. Lee, and G. S. Kang, "Real-time prediction of battery power requirements for electric vehicles," in *Proceedings of the ACM/IEEE International Conference on Cyber-Physical Systems*, IEEE, 2013.

[11] G. Wager, J. Whale, and T. Braunl, "Driving electric vehicles at highway speeds: the effect of higher driving speeds on energy consumption and driving range for electric vehicles in australia," *Renewable and Sustainable Energy Reviews*, vol. 63, pp. 158–165, 2016.

[12] M. Neaimeh, G. A. Hill, Y. Hübner, and P. T. Blythe, "Routing systems to extend the driving range of electric vehicles," *IET Intelligent Transport Systems*, vol. 7, no. 3, pp. 327–336, 2013.

[13] G. Liu, M. Ouyang, L. Lu, J. Li, and J. Hua, "A highly accurate predictive-adaptive method for lithium-ion battery remaining discharge energy prediction in electric vehicle applications," *Applied Energy*, vol. 149, no. 1, pp. 297–314, 2015.

[14] X. Yuan, C. Zhang, G. Hong, X. Huang, and L. Li, "Method for evaluating the real-world driving energy consumptions of electric vehicles," *Energy*, vol. 141, pp. 1955–1968, 2017.

[15] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 308–324, 2015.

[16] B. Trawiński, M. Smetek, Z. Telec, and T. Lasota, "Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms," *International Journal of Applied Mathematics and Computer Science*, vol. 22, no. 4, pp. 867–881, 2012.

[17] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[18] Q. M. J. L. Ying Cao, "Advance and prospects of adaBoost algorithm," *Acta Automatica Sinica*, vol. 39, no. 6, pp. 745–758, 2013.

[19] M. K. M. A. Mathias Petzl, "Lithium plating in a commercial lithium-ion battery - A low-temperature aging study," *Journal of Power Sources*, vol. 275, pp. 799–807, 2015.