Hindawi
Geofluids
Volume 2021, Article ID 5520942, 10 pages
https://doi.org/10.1155/2021/5520942

Research Article

The Optimization Model for Reducing RON Loss in Gasoline Refining Process

Xuefei Lu,[1] **Xiaoyan Wang**,[1] **Yifang Yang**,[1] **and Jianan Xue**[2]

[1]*College of Sciences, Xi'an Shiyou University, Xi'an, Shaanxi 710065, China*
[2]*College of Mechanical Engineering, Xi'an Shiyou University, Xi'an, Shaanxi 710065, China*

Correspondence should be addressed to Xuefei Lu; luxuefei@xsyu.edu.cn

Received 18 January 2021; Revised 27 January 2021; Accepted 1 February 2021; Published 20 February 2021

Academic Editor: Chun Zhu

As gasoline is the main fuel of small vehicles, the exhaust emissions from its combustion will affect air quality. The focus of gasoline cleaning is to reduce the sulfur and olefin content in gasoline while maintaining its RON as much as possible. The reduction of RON will bring great economic losses to enterprises. Therefore, it is very important for petrochemical enterprises to construct a RON loss model in the gasoline refining process. The model construction, which reduces RON loss during gasoline refining, is the main question in this paper. By Python and SPSS software, we got two variable filtering methods: the random forest importance filtering and PCA filtering, and combined with SVR and random forest models, RON of the product and sulfur content were predicted. The filtering order of the original data by Excel and Python is maximum and minimum removal, 3σ criterion removal, deletion of too many sites in incomplete data, and filling of empty values in the mean within two hours. Several RON prediction models were established with the help of Python software, and the variables selected were compared by two filtering methods: one is the SVR model based on Gaussian, linear, polynomial, and Sigmoid kernel functions; the other is the random forest model. The sulfur content and RON prediction model was constructed, which use evaluation functions such as MSE, R^2, and RMSE to evaluate and sulfur content as the subject condition. We convert the problem into linear and nonlinear model variable optimization problems: the linear model is the variable selected by the SVR linear kernel function model and random forest; the nonlinear model is the combination of variables selected by the random forest model and random forest. Optimizing for each sample, the optimization method is to find the optimal solution for each variable and use the optimal method for each variable as the local optimal solution for the sample. The two models are evaluated from the perspectives of optimization degree, optimization rate, model running speed, etc.

1. Introduction

More than 95% of sulfur and olefins in finished gasoline come from catalytic cracking gasoline in our country. Therefore, the catalytic cracking gasoline must be refined to satisfy the gasoline quality requirements. RON is the most important indicator reflecting the combustion performance of gasoline and is used as the commercial brand name of gasoline, such as 89#, 92#, and 95#. Desulfurization and olefin reduction technology reduced RON of gasoline in modern catalytic cracking gasoline. However, the reduction of the RON will bring great economic losses to the enterprise. For 1-unit reduction for RON, the loss is equivalent to about 150 yuan/-ton. Taking a 1-million-ton/year catalytic cracking gasoline refining unit as an example, if the RON loss can be reduced by 0.3 units, its economic benefit will reach 45 million yuan [1]. Therefore, the RON loss model in the gasoline refining process is critical to petrochemical companies. At the same time, the reduction of RON not only brings huge economic benefits to petrochemical companies but also brings new opportunities and challenges to material science, engineering geology, and other energy disciplines [2–8]. Due to the complexity of the refining process and the diversity of equipment, by which operating variables (control variables) have a highly nonlinear and strongly coupled relationship with each other, there are relatively few variables in the traditional data association model, and mechanism modeling needs high analysis requirements of raw materials. The result cannot meet the needs of the industry.

According to industrial demand, the less the RON is reduced, the higher the economic benefits of the company. With the average loss of RON of existing petrochemical companies and related references, if the loss of RON can be controlled at 0.5-1, the economic benefits of enterprises will be very considerable.

Since the refining process of catalytic cracking gasoline is continuous, the operating variables are sampled every 3 minutes, the measurement of RON (dependent variable) is more troublesome, and it cannot be matched only twice a week. However, according to the actual situation, it can be considered that the measured value of RON is the comprehensive effect of the manipulated variable within two hours before the measurement time. Then, the average value of the manipulated variable within two hours of the pretreatment corresponds to the measured value of RON. The establishment of a model for reducing RON loss involves 7 raw material properties, 2 spent adsorbent properties, 2 regenerated adsorbent properties, 2 product properties, and other variables, as well as 354 other operating variables (a total of 367 variables) based on the sample data. The method of dimensionality reduction first and then modeling is often used in engineering technology applications, which is conducive to ignoring minor factors and discovering and analyzing the main variables and factors that affect the model.

2. Data Processing

Since most of the variable data of collecting raw data are normal, some data of each device has problems in some locations, some variables only contain data of a part time, and the data of some variables are all empty or part of data is empty. The quality of the data will directly affect the results of the research, so the original data must be processed first. The processing process is as follows:

(1) Converting the 2-D index into 1-D index, various properties are named: xx property_xx, e.g., raw material properties_sulfur content, product properties_RON, and regenerated adsorbent_coke, wt%. Due to the lack of Chinese names for data, English names are uniformly adopted

(2) We constructed a new sheet table named sample property, constructed a new sheet, and kept the original format. Then, the raw materials, products, spent adsorbent, and regenerated adsorbent were also copied to the corresponding position of the sample property table. After that, splitting the operating variable table into sample 285 and sample 313, the header of the new two tables is the second row of the operation variable table, such as time|S-ZORB.CAL_H2.PV|S-ZORB.PDI_2102.PV|…, and then, Python was used for data processing

(3) The data of sample 285 and sample 313 was imported, by the limit method of the maximum value of each column to filter, and some samples are removed that are not in this range. 0 is missing data, replaced with NA, deleting all columns with NA values. The average value was taken within two hours to fill in the missing values. Since the data are of two hours, the processed data was combined with appendix 1 with the method of mean fill

(4) The 3σ criterion is used to remove irregular values: suppose that the measured variable is measured with equal accuracy, the arithmetic mean x of $x_1, x_2, \cdots, x_n$ and residual error were got, and the standard error σ is calculated according to the Bessel formula. If the residual error $v_b (1 \leq b \leq n)$ of a certain measured value x_b satisfied $|v_b| = |x_b - x| > 3\sigma$, it is considered to be a bad value with a gross error value and should be eliminated [9]. The Bessel formula is as follows:

$$\sigma = \left[\frac{1}{n-1} \sum_{i=1}^{n} v_i^2 \right]^{1/2} = \left\{ \frac{\left[\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 / n \right]}{n-1} \right\}^{1/2} \tag{1}$$

3. Analysis

The establishment of a reducing RON loss model includes 7 raw material properties, 2 spent adsorbent properties, 2 regenerated adsorbent properties, 2 product properties, and another 354 operating variables (a total of 367 variables). The method of dimensionality reduction first and then modeling is helpful for ignoring secondary factors, discovering, and analyzing the main variables and factors that affect the model. The procedure can be divided into five parts: missing data processing, low variance filtering processing, correlation analysis, principal component analysis, and random forest feature selection.

3.1. Missing Data Processing. It can be known that the sample data values are randomly missing by data analysis; we need to reduce the dimensionality and filter the operating variables and process the missing data according to the data obtained after processing. There are three ways to deal with missing values: deleting data, data imputation, and no processing. Data imputation is adding the unknown value to the subjective estimate value, which will bring errors. The imputation methods include mean imputation, data imputation, similar mean imputation, maximum likelihood estimation, and multiple imputation [10]. The columns of missing value with more than 50% will be deleted in this paper, the remaining missing data is processed by mean interpolation, and the number of columns deleted here is 8 columns.

3.2. Low Variance Filtering Processing. Low variance filtering is similar to the method of missing value deletion, which assumes that the column with very small changes in the data column contains less information. Therefore, all columns with small variances are removed, and the data needs to be normalized first because of the correlation between variance and data range. It is determined to normalize the data, then delete the columns with data variance less than 0.1, and the number of columns deleted here is 34 columns in this paper.

3.3. Correlation Analysis. Correlation analysis studies the direction and closeness between variables. First, a correlation matrix can be obtained by calculating the Pearson correlation coefficient between 359 features; only one variable with a correlation greater than 0.9 is retained. The number of variables filtered out is 153, the number of repeated filtering variables is 5, and the number of remaining variables is 177.

3.4. Principal Component Analysis (PCA). PCA is a method of reducing the dimensionality of high-dimensional data, turning multiple variables into a few principal components, and removing noise at the same time. The purpose of the method is to use fewer features to explain most of the variation in the original data and to convert many highly correlated features into mutually independent or uncorrelated features. The idea is to select several new features that are more than the original features, which can explain the variation in most of the data; this is the so-called principal component [11].

3.4.1. The Principle of PCA. Supposed that we have *n*th sample and *m*th features, which can be denoted by$y_{11}, y_{12}, \cdots, y_{nm}$, it is more efficient to write them in the matrix form:

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{pmatrix}_{n\times m}, \tag{2}$$

where $y_{ij} (i = 1,\cdots,n,\ j = 1,\cdots,m)$ is the *i*th eigenvalue of the *j*th sample.

The basic implementation steps of PCA can be divided into data standardization, calculation of covariance matrix, calculation of eigenvalues and eigenvectors, calculation of principal component contribution rate and cumulative contribution rate, calculation of principal component load, calculation of principal component score, and operating variable weights. Meanwhile, the retention of several principal components depends on the cumulative contribution rate of the retained part. In order to ensure that the main information is not lost, the cumulative contribution rate of the retained principal components should be greater than 85%.

3.4.2. Implementation of PCA. The data after processing the missing data is analyzed by principal components by SPSS software, and 359-dimensional features are described by 24 principal components. The contribution rate and cumulative contribution rate of each principal component are shown in Table 1.

From Table 1, we can see that the cumulative contribution rate of the first 24 principal components reaches 85.268%, and the feature value of the 24th principal component is 1.920 > 1, which almost contains most of the information of 359-dimensional features.

As shown in Figure 1, which is made according to the contribution of the principal components to the feature value in Table 1, they are the corresponding relationship. We can see that the advantage of the gravel figure in the gentle curve can explain the change of the characteristic and draw the conclusion. In Figure 1, each feature is called a factor, and there are 359 features, that is, 359 factors. The eigenvalue of the 24th principal component has a larger decline compared with the previous eigenvalue, this eigenvalue is smaller, and the following eigenvalues do not change much, indicating that adding factors corresponding to the eigenvalue can only add very little information. Therefore, the first 24 principal components can cover 359-feature information according to Figure 1.

3.5. Random Forest Feature Selection. It is necessary to select features that have a greater impact on the result for modeling, when the number of features in the data set exceeds 300 dimensions. The feature selection method we choose is random forest.

The method of random forest to evaluate the importance of features is considering the contribution of each feature on each tree, taking the average value, and comparing the contribution of different features. The evaluation indicators of contribution include the Gini index (Gini) and the error rate of out-of-bag data (OOB) [12].

In general, the Gini value is used as the criterion for splitting nodes in the random forest model. In weighted random forest (WRF), the weight has two functions: the first is to select the split point to calculate the Gini value, which can be denoted as

$$\begin{aligned} i(N) &= \frac{\sum_{i=1}^{c} (n_i W_i)^2}{\sum_{i=1}^{c} n_i W_i}, \\ \Delta i &= i(N_{\mathrm{L}}) - i(N_{\mathrm{R}}), \end{aligned} \tag{3}$$

where N is an unseparated node; N_{L} and N_{R} are the left and right nodes after separation, respectively; W_i is the class weight of c samples; n_i is the number of various samples inside the node; and Δi is the reduction in impurity. The larger the value, the better the separation effect of the point; the second is that the class weight can be used to determine its class label in the terminal node; the expression is as follows:

$$\text{node class} = \arg\ \max_i (n_i W_i)\ (i = 1, 2, \cdots, C). \tag{4}$$

The Gini value is used as the evaluation index of the contribution rate in this paper, the importance score of the variable is denoted by VIM, and GI denotes the Gini value. Suppose that $X_1, X_2, \cdots, X_m$ are *m*th features, calculating the Gini index score VIM_j of each feature, which is the average change of split impurity for the *j*th feature in

Table 1: Contribution rate and cumulative contribution rate of each principal component.

Component	Initial eigenvalue			Extract the sum of squares and load		
	Total	Variance (%)	Cumulative (%)	Total	Variance (%)	Cumulative (%)
1	113.571	31.547	31.547	113.571	31.547	31.547
2	39.590	10.997	42.545	39.590	10.997	42.545
3	23.930	6.647	49.192	23.930	6.647	49.192
4	20.292	5.637	54.829	20.292	5.637	54.829
5	14.529	4.036	58.865	14.529	4.036	58.865
6	12.305	3.418	62.283	12.305	3.418	62.283
7	9.904	2.751	65.034	9.904	2.751	65.034
8	8.488	2.358	67.391	8.488	2.358	67.391
9	7.390	2.053	69.444	7.390	2.053	69.444
10	6.783	1.884	71.328	6.783	1.884	71.328
11	6.418	1.783	73.111	6.418	1.783	73.111
12	5.282	1.467	74.579	5.282	1.467	74.579
13	4.993	1.387	75.966	4.993	1.387	75.966
14	4.446	1.235	77.201	4.446	1.235	77.201
15	4.255	1.182	78.382	4.255	1.182	78.382
16	3.735	1.038	79.420	3.735	1.038	79.420
17	3.305	0.918	80.338	3.305	0.918	80.338
18	3.245	0.901	81.239	3.245	0.901	81.239
19	2.909	0.808	82.048	2.909	0.808	82.048
20	2.743	0.762	82.810	2.743	0.762	82.810
21	2.546	0.707	83.517	2.546	0.707	83.517
22	2.368	0.658	84.175	2.368	0.658	84.175
23	2.015	0.560	84.734	2.015	0.560	84.734
24	1.920	0.533	85.268	1.920	0.533	85.268

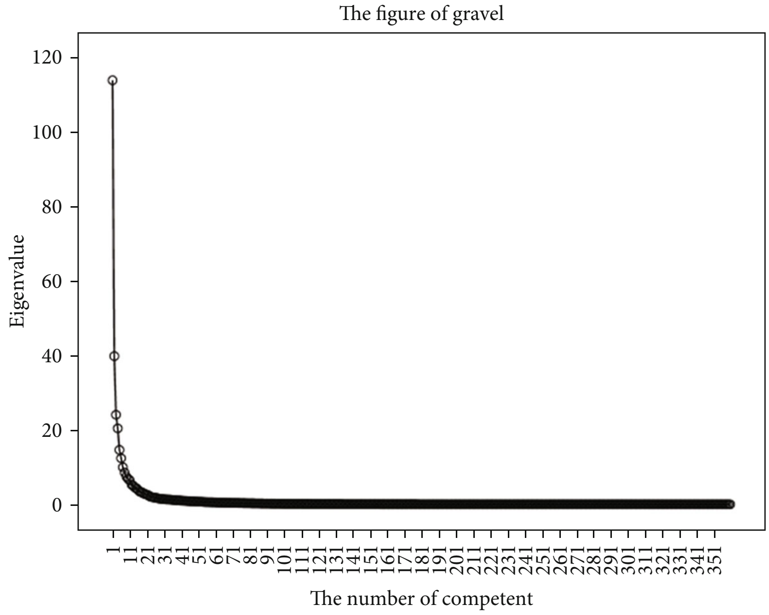

Figure 1: The figure of gravel.

Table 2: Operating variable weight of PCA.

No.	Operating variable	Weight	No.	Operating variable	Weight
1	S-ZORB.FT_9102.PV	0.0420	10	S-ZORB.PT_7103.DACA	0.0407
2	S-ZORB.FT_1204.TOTAL	0.0417	11	S-ZORB.PC_2902.DACA	0.0406
3	S-ZORB.PDT_2001.DACA	0.0416	12	S-ZORB.PDT_2605.DACA	0.0400
4	S-ZORB.SIS_PDT_2103B.PV	0.0409	13	S-ZORB.FC_2801.PV	0.0395
5	S-ZORB.TE_7102.DACA	0.0408	14	S-ZORB.TE_3112.DACA	0.0370
6	S-ZORB.SIS_LT_1001.PV	0.0408	15	S-ZORB.SIS_TEX_3103B.PV	0.0364
7	S-ZORB.SIS_PT_2703	0.0408	16	S-ZORB.PC_1603.PV	0.0360
8	S-ZORB.PT_7510.DACA	0.0408	17	S-ZORB.AT-0003.DACA.PV	0.0360
9	S-ZORB.PT_7107.DACA	0.0407	18	S-ZORB.CAL.SPEED.PV	0.0357

Table 3: Filtered variables of PCA.

No.	Variable	No.	Variable
1	Raw material properties_sulfur content (μg/g)	16	S-ZORB.TE_7102.DACA
2	Raw material properties_RON	17	S-ZORB.SIS_LT_1001.PV
3	Raw material properties_saturated hydrocarbon (v%)	18	S-ZORB.SIS_PT_2703
4	Raw material properties_olefin (v%)	19	S-ZORB.PT_7510.DACA
5	Raw material properties_aromatics (v%)	20	S-ZORB.PT_7107.DACA
6	Raw material properties_bromine value	21	S-ZORB.PT_7103.DACA
7	Raw material properties_density, 20°C	22	S-ZORB.PC_2902.DACA
8	Spent adsorbent properties_coke (wt%)	23	S-ZORB.PDT_2605.DACA
9	Spent adsorbent properties_S (wt%$'$)	24	S-ZORB.FC_2801.PV
10	Spent adsorbent properties_coke (wt)	25	S-ZORB.TE_3112.DACA
11	Spent adsorbent properties_S (wt%)	26	S-ZORB.SIS_TEX_3103B.PV
12	S-ZORB.FT_9102.PV	27	S-ZORB.PC_1603.PV
13	S-ZORB.FT_1204.TOTAL	28	S-ZORB.AT-0003.DACA.PV
14	S-ZORB.PDT_2001.DACA	29	S-ZORB.CAL.SPEED.PV
15	S-ZORB.SIS_PDT_2103B.PV		

Table 4: Operating variable feature importance of random forest.

No.	Operating variable	Feature importance	No.	Operating variable	Feature importance
1	S-ZORB.TC_2801.PV	0.002204	10	S-ZORB.PC_5101.PV	0.000973
2	S-ZORB.CAL_H2.PV	0.002075	11	S-ZORB.TE_7106.DACA	0.000964
3	S-ZORB.FC_1203.PV	0.001812	12	S-ZORB.AT_1001.DACA	0.000912
4	S-ZORB.TE_1106.DACA	0.001408	13	S-ZORB.FC_1102.PV	0.000878
5	S-ZORB.PDC_2702.DACA	0.001348	14	S-ZORB.FT_2502.DACA	0.000827
6	S-ZORB.SIS_TEX_3103B.PV	0.001344	15	S-ZORB.TE_5006.DACA	0.000804
7	S-ZORB.PC_2902.DACA	0.001028	16	S-ZORB.TC_2607.PV	0.000740
8	S-ZORB.LT_3801.DACA	0.000990	17	S-ZORB.ZT_2634.DACA	0.000717
9	S-ZORB.FT_1003.PV	0.000974	18	S-ZORB.PT_1501.PV	0.000675

all decision trees of random forest, the formula of the Gini index is as follows:

$$\mathrm{GI}_m = 1 - \sum_{k=1}^{|K|} p_{mk}^2, \tag{5}$$

where k are kth categories and p_{mk} represents the proportion of category k in node m. The importance of X_j on the node m, which is the change of the Gini index before and after the branch of node m, can be denoted by

$$\mathrm{VIM}_{jm}^{\mathrm{Gini}} = \mathrm{GI}_m - \mathrm{GI}_l - \mathrm{GI}_r, \tag{6}$$

Table 5: Filtered variables of random forest (RF).

No.	Filtered variable of RF	No.	Filtered variable of RF
1	Raw material properties_sulfur content (μg/g)	16	S-ZORB.PDC_2702.DACA
2	Raw material properties_RON	17	S-ZORB.SIS_TEX_3103B.PV
3	Raw material properties_saturated hydrocarbon (v%)	18	S-ZORB.PC_2902.DACA
4	Raw material properties_olefin (v%)	19	S-ZORB.LT_3801.DACA
5	Raw material properties_aromatics (v%)	20	S-ZORB.FT_1003.PV
6	Raw material properties_bromine value	21	S-ZORB.PC_5101.PV
7	Raw material properties_density, 20°C	22	S-ZORB.TE_7106.DACA
8	Spent adsorbent properties_coke (wt%)	23	S-ZORB.AT_1001.DACA
9	Spent adsorbent properties_S (wt%′)	24	S-ZORB.FC_1102.PV
10	Spent adsorbent properties_coke (wt)	25	S-ZORB.FT_2502.DACA
11	Spent adsorbent properties_S (wt%)	26	S-ZORB.TE_5002.DACA
12	S-ZORB.TC_2801.PV	27	S-ZORB.TE_5005.DACA
13	S-ZORB.CAL_H2.PV	28	S-ZORB.ZT_2634.DACA
14	S-ZORB.FC_1203.PV	29	S-ZORB.PT_1501.PV
15	S-ZORB.TE_1106.DACA		

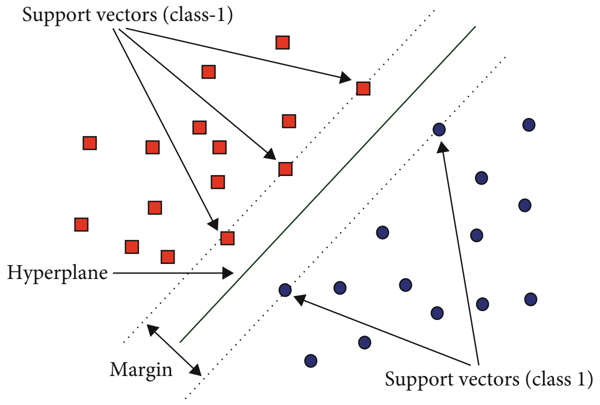

Figure 2: The figure of SVM.

Table 6: SVR+PCA model score.

Kernel function	Model score			
	MSE	R^2	MAE	RMSE
Linear	0.0497	0.9425	0.1691	0.2230
Polynomial	0.3390	0.6081	0.4189	0.5822
Gauss	0.0751	0.9132	0.2076	0.2740
Sigmoid	12.9687	-15.1201	2.1734	3.7375

Table 7: SVR+RF model score.

Kernel function	Model score			
	MSE	R^2	MAE	RMSE
Linear	0.0530	0.9387	0.1758	0.2302
Polynomial	0.1539	0.8221	0.2910	0.3923
Gauss	0.0987	0.8859	0.2374	0.3141
Sigmoid	1.3112	-0.5160	0.8102	1.1451

Table 8: PCA+RF model score.

Model	Model score			
	MSE	R^2	MAE	RMSE
PCA and RF	0.0406	0.9620	0.1480	0.2014

Table 9: RF model score.

Model	Model score			
	MSE	R^2	MAE	RMSE
RF	0.0407	0.9619	0.1520	0.2016

where GI_l and GI_r represent the Gini indices of the two new nodes after branching, respectively.

If the node of feature X_j in decision tree i belongs to set M, then the importance of X_j in the ith tree is

$$\mathrm{VIM}_{ij}^{\mathrm{Gini}} = \sum_{m\in M} \mathrm{VIM}_{jm}^{\mathrm{Gini}}. \tag{7}$$

Assuming there are n trees in the random forest, then

$$\mathrm{VIM}_{j}^{\mathrm{Gini}} = \sum_{i=1}^{n} \mathrm{VIM}_{ij}^{\mathrm{Gini}}. \tag{8}$$

Normalize the importance score:

$$\mathrm{VIM}_j = \frac{\mathrm{VIM}_j^{\mathrm{Gini}}}{\sum_{i=1}^{c} \mathrm{VIM}_i}. \tag{9}$$

Return the importance of features through the random forest in Sklearn.

The operating variable weights and filtered variable obtained by PCA, operating variables, and filtered variable of RF are shown in Tables 2–5. The conclusion is that the operating variables are selected differently, which leads to different final weights. The first ten variables selected by two methods are the same and different from the eleventh. However, according to industrial demand, the variables selected by the random forest have a more important and direct impact on RON loss. The reliability of the calculation results can also be seen in the subsequent model calculations.

4. Establishing Model

The establishment of a model for reducing RON loss prediction is based on the processing of the original data, filtering, and extracting of data by dimensionality reduction. With various theoretical models or mathematical methods for data analysis, then get the final prediction model results. Since the RON loss is calculated by the RON of raw material minus the RON of the product, it is more accurate to calculate the RON after predicting the RON of the product. There are two types of selected variables: the partial linear variables obtained by the PCA and partial nonlinear variables obtained by the RF. By combining two variables and different models, the best combination of variables and models can be obtained. Therefore, the filtered features and model selection play a decisive role in the final establishment of the RON loss prediction model.

4.1. Regression Model of Support Vector Machine (SVR). The support vector machine, denoted by SVM, is mainly applied in pattern recognition, classification, and regression analysis. As shown in Figure 2, 2-D data points of red and blue can be separated by a straight line, which is called a linearly separable problem in the pattern recognition; the black solid line is the dividing line, also known as the "decision surface." Each decision surface corresponds to a linear classifier [13]. SVM can be expressed as

$$\min_{w,b}\frac{1}{2}w^{T}w+C\cdot\sum_{n=1}^{N}\max\left(1-y_{n}\left(w^{T}z_{n}+b\right),0\right). \tag{10}$$

A general procedure for finding a decision function according to the given training sample $\{(x_1, y_1),\cdots,(x_n, y_n)\} \subset (X\times Y)$ when applying SVM for regression is denoted by SVR, where $x_i \in X = R^n$, $y_i \in Y = R^n$, $i = 1, \cdots, n$; the decision function can be expressed as

$$f(x)=wx+b, \tag{11}$$

where ω and b are undetermined model parameters, which can be obtained by fitting the data of $f(x)$ and y. In order to solve ω and b, the above problem is transformed into an optimization problem:

$$\min\frac{1}{2}\left(\|\boldsymbol{\omega}\|^{2}+C\sum_{i=1}^{l}\left(\xi_{i}+\xi_{i}^{*}\right)\right)$$
$$\text{s.t.}\begin{cases} f(\mathbf{x}_{i})-y_{i}\leq\epsilon+\xi_{i}, \\ y_{i}-f(\mathbf{x}_{i})\leq\epsilon+\widehat{\xi}_{i}, \end{cases} \quad \xi_{i}\geq0,\ \widehat{\xi}_{i}\geq0,\ i=1,2,\cdots,n. \tag{12}$$

Usually, equation (11) is not solved directly, by which the dual problem is introduced:

$$\max_{\alpha,\alpha^{*}}\sum_{i}^{n}y_{i}(\alpha_{i}^{*}-\alpha_{i})-\varepsilon(\alpha_{i}^{*}+\alpha_{i})-\frac{1}{2}\sum_{i}^{n}\sum_{j}^{n}(\alpha_{i}^{*}-\alpha_{i})\left(\alpha_{j}^{*}-\alpha_{j}\right)x_{i}^{T}x_{j}$$
$$\text{s.t.}\begin{cases} \sum_{i}^{l}(\alpha_{i}^{*}-\alpha_{i})=0, \\ 0\leq\alpha_{i},\ \alpha_{i}^{*}\leq C, \quad i=12,\cdots,n. \end{cases} \tag{13}$$

After obtaining α_i, if $0 < \alpha_i < C$, then $\xi_i = 0$, so

$$w=y_{i}+\varepsilon-\sum_{i}^{l}(\alpha_{i}^{*}-\alpha_{i})x_{i}^{T}x. \tag{14}$$

$f(x)$ can be expressed as

$$f(x)=\sum_{i=1}^{n}(\alpha_{i}^{*}-\alpha_{i})k(x,x_{i})+b, \tag{15}$$

where $k(x, x_i)$ is a kernel function; the linear kernel function formula is as follows:

$$\kappa(\mathbf{x},\mathbf{x}_{i})=\mathbf{x}_{i}^{T}\mathbf{x}. \tag{16}$$

The polynomial kernel function

$$\kappa(\mathbf{x},\mathbf{x}_{i})=\left(\mathbf{x}_{i}^{T}\mathbf{x}+\lambda\right)^{d},\quad \lambda\geq0. \tag{17}$$

The Gauss kernel function

$$\kappa(\mathbf{x},\mathbf{x}_{i})=\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_{i}\|^{2}}{2\sigma^{2}}\right). \tag{18}$$

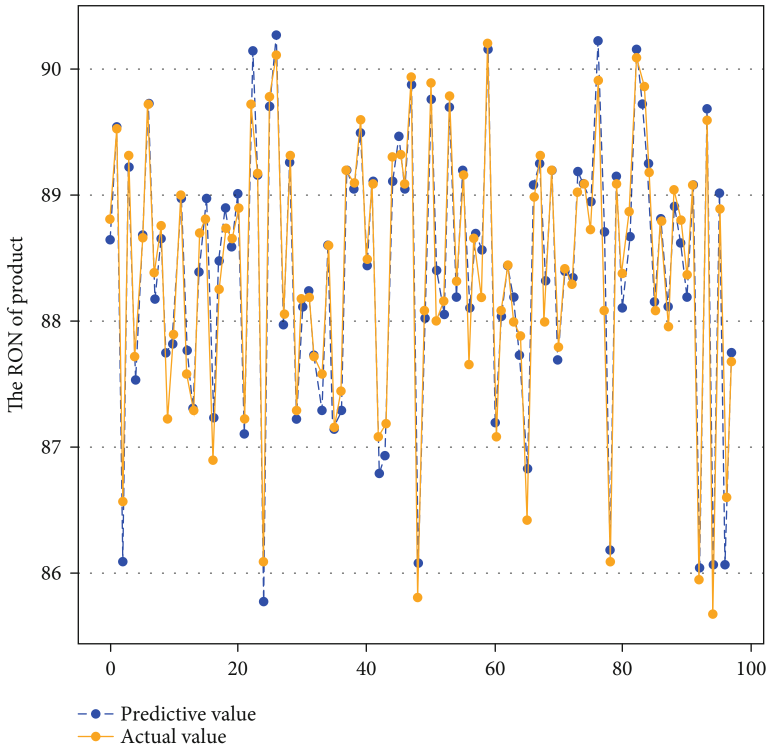

Figure 3: Fitting figure of PCA variable random forest model.

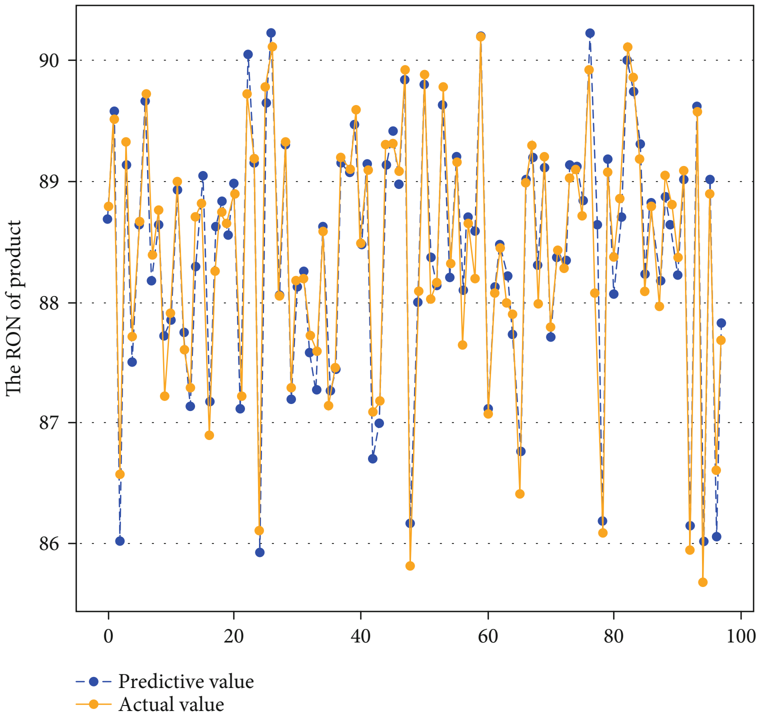

Figure 4: Fitting figure random forest variable random forest model.

The Sigmoid kernel function

$$\kappa(\mathbf{x}, \mathbf{x}_i) = \tanh\left(\beta \mathbf{x}_i^T \mathbf{x} + \theta\right)(\beta > 0, \theta < 0). \qquad (19)$$

4.2. Random Forest Model. Random forest is a relatively new machine learning model ensemble method, which is also called the nonlinear tree-based model [14, 15]. It is composed of a decision tree and bagging. The principle of random forest is to build a forest in a random way, which is a kind of cluster classification model. The decision trees that make up the random forest are not related to each other. After the random forest model is constructed, new samples are input into the model to be judged by decision trees.

We choose different errors to measure the deviation between the RON of the predicted product and the RON of the true value in the model; the selected errors are as follows [16, 17].

MSE (Mean Square Error) is used to measure the deviation between the RON of the predicted product and the RON of the true value in the model; MSE is close to 0, which means that the predictive ability of the model is better; on the contrary, it means that the predictive ability of the model is worse. We can use the following formula to calculate MSE:

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n} = \frac{1}{n}\sum_{i=1}^{m} w_i (y_i - y\wedge_i)^2. \tag{20}$$

The interval of R^2_score is [0,1], R^2_score = 1, which means that the predictive ability of the model is better. The formula of R^2_score is as follows:

$$R^2 = 1 - \frac{\left(\sum_{i=1}^{m} \left(y\wedge^{(i)} - y^{(i)}\right)^2\right)/m}{\left(\sum_{i=1}^{m} \left(y^{(i)} - \bar{y}\right)^2\right)/m} = 1 - \frac{\mathrm{MSE}(\hat{y}, y)}{\mathrm{Var}(y)}. \tag{21}$$

MAE is the average value of absolute errors, which can better reflect the actual situation of predicted value errors. The formula is as follows:

$$\mathrm{MAE}(X, h) = \frac{1}{m}\sum_{i=1}^{m} |h(x_i) - y_i|. \tag{22}$$

RMSE is the square root of MSE, which can be calculated by the following formula:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{\mathrm{SSE}}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{m} w_i (y_i - y\wedge_i)^2}. \tag{23}$$

5. Results and Conclusion

The problem of predicting the loss of RON is transformed into the problem of product RON, where the loss of RON equals RON of raw material minus RON of the product. According to the indicators in Tables 2 and 4, the following four models were established: the prediction model of product RON based on PCA and SVR [18–21], the prediction model of product RON based on RF and SVR, the prediction model product RON based on PCA and RF, and the prediction model product RON based on RF [22, 23].

The indicators of each model are shown in Tables 6–9.

We can see that the evaluation indicators of random forest are in the forefront according to Tables 6–9. Compared with MSE, R^2 of SVR+PCA and SVR+RF does not perform well in random forest. Therefore, the various evaluation indicators obtained by random forest to measure the range of RON reduction is more illustrative in the industry and more convenient in practice.

The variables selected by PCA and the variables selected by the random forest are similar in performance on the random forest model. The comparison between their respective predictions and the original values is as Figures 3 and 4.

We can get the conclusion that the PCA variable random forest model and the random forest variable random forest model have similar fitting results according to Figures 3 and 4, but the random forest variables are more in line with industrial needs and close to the variables required in the traditional octane number prediction formula. The variables selected by random forest establish a random forest model to predict the loss of octane number (RON) according to Figure 4.

Data Availability

The data used to support the results of this study are included within the manuscript.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was financially supported by the Shaanxi Provincial Natural Science Foundation Research Project (2017JQ4005) and Shaanxi Province Natural Science Basic Research plan (2020JM-543); we thank the support of these projects.

References

[1] Q. X. Meng and W. Wang, “A novel closed-form solution for circular openings in generalized Hoek-Brown media,” *Mathematical Problems in Engineering*, vol. 2014, Article ID 870835, 7 pages, 2014.

[2] L. Zhen, S. Liu, W. Ren, J. Fang, Q. Zhu, and Z. Dun, “Multiscale laboratory study and numerical analysis of water-weakening effect on shale,” *Advances in Materials Science and Engineering*, vol. 2020, Article ID 5263431, 14 pages, 2020.

[3] Q. Meng, H. Wang, M. Cai, W. Xu, X. Zhuang, and T. Rabczuk, “Three-dimensional mesoscale computational modeling of soil-rock mixtures with concave particles,” *Engineering Geology*, vol. 277, article 105802, 2020.

[4] Q. Meng, L. Yan, Y. Chen, and Q. Zhang, “Generation of numerical models of anisotropic columnar jointed rock mass using modified centroidal Voronoi diagrams,” *Symmetry*, vol. 10, no. 11, p. 618, 2018.

[5] Y. Wang, B. Zhang, S. H. Gao, and C. H. Li, “Investigation on the effect of freeze-thaw on fracture mode classification in marble subjected to multi-level cyclic loads,” *Theoretical and Applied Fracture Mechanics*, vol. 111, p. 102847, 2021.

[6] D. Yin, S. Chen, Y. Ge, and R. Liu, “Mechanical properties of rock-coal bi-material samples with different lithologies under uniaxial loading,” *Journal of Materials Research and Technology*, vol. 10, pp. 322–338, 2021.

[7] W. Wang, L. Li, W. Xu, Q. Meng, and J. Lv, "Creep failure mode and criterion of Xiangjiaba sandstone," *Journal of Central South University*, vol. 19, no. 12, pp. 3572–3581, 2012.

[8] D. Ren, H. Huang, J. Qi, and Z. Peng, "One-pot template-free cross-linking synthesis of SiOx–SnO2@C hollow spheres as a high volumetric capacity anode for lithium-ion batteries," *Energy Technology*, vol. 8, no. 7, article 2000314, 2020.

[9] C. Wu and H. Lin, "Research on early warning model of wind power tower based on conic exponential smoothing method," *Power System Protection and Control*, vol. 42, no. 9, pp. 81–85, 2014.

[10] S. Wang and H. Zhao, "Application of data mining technology in power station equipment failure analysis," *Software Guide*, vol. 15, no. 12, pp. 121–124, 2016.

[11] A. Li, Y. Chang, and X. Kong, "Comprehensive evaluation of development indexes of oilfield production area," *Journal of Southwest Petroleum University (Social Science Edition)*, vol. 14, no. 6, pp. 6–11, 2012.

[12] P. Zhou, Z. He, Y. Zhen, Y. Jing, and X. Y. Wang, "Modeling method of filtering rules for finding difference data subsets," *Software Engineering*, vol. 22, no. 11, pp. 1–7, 2019.

[13] W. J. Shu, *Research on Adaptive Rational Over-Limit Learning Machine for Electricity Price Forecasting*, Xiangtan University, 2018.

[14] Z. W. Liu, *Research on Borehole Trajectory Measurement and Tracking Method of Mine Horizontal Drilling Rig*, Taiyuan University of Technology, 2018.

[15] C. Wu and J. Luo, "Automatic identification of tax evasion based on random forest," *Software Guide*, vol. 17, no. 8, pp. 13–16, 2018.

[16] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, New York, NY, USA, 1995.

[17] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[18] X. D. Liu and Z. Q. Chen, "Face recognition based on support vector machines," *Mini-Micro Systems*, vol. 25, no. 12, pp. 2261–2263, 2004.

[19] W. Bledsoe, *Man-machine facial recognition*, Panoramic Research Inc., Palo Alto, CA, USA, 1966.

[20] N. Y. Deng and Y. J. Tian, *A New Method in Data Mining, Support Vector Machine*, Science Press, Beijing, 2004.

[21] S. Zhang, J. Liu, and J. W. Tian, "A SVM-based small target segmentation and clustering approach," in *Proceedings of 2004 International Conference on Machine Learning and Cybernetics (IEEE Cat. No.04EX826)*, pp. 3318–3323, Shanghai, China, August 2004.

[22] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[23] Y. J. Huang, X. F. Yao, and L. Yan, "The optimization study based on the target cascade method and genetic algorithm of multidisciplinary design," *Mechanical Design and Manufacturing*, vol. 9, pp. 39–41, 2010.