Research on sales strategy of electric vehicle target customers based on machine learning algorithm

Zhichong Li *

College of data science, University of Taiyuan University of Technology of China, Jinzhong, 030600, China

* Corresponding Author Email: 305229019@qq.com

Abstract. Due to the influence of various subjective and objective factors, there is great uncertainty in consumers' willingness to purchase new energy vehicles. In this study, it is hoped that based on customer satisfaction and personal characteristic information, the main factors affecting whether the target customers are willing to purchase new energy vehicles are explored, so that corresponding sales strategies can be formulated. In this study, two different embedding methods based on penalty terms and tree models are used for feature selection, the former using three models LR, LASSO and SVM, and the latter using RF and LightGBM models for a total of five models for machine learning to find out the relevant features that affect the sales of different brands, and the voting method is applied to the selected features, and the results are found. The battery technology performance of electric vehicles, comfort, annual mortgage of target customers and the ratio of auto loan to annual household income have a significant impact on the sales of all three brands. In addition, affordability, safety and customer's work situation also had different degrees of influence on sales of the different brands. A multilayer perceptron (BP neural network) prediction model is built for each brand's target customers to predict their purchase intention. The prediction accuracy of the network is improved by balancing the positive and negative samples so that the ratio of the two types of samples in the dataset is approximately 1:1. The test set is also used for real-time testing to determine the fit status of the dataset. The accuracy of the final model was above 80% on both the training and test sets, and the model predicted the 1st, 5th, 6th, 7th, 12th, and 12th samples. The model predicted the purchase intention of the 1st, 5th, 6th, 7th, 12th and 13th customers.

Keywords: Sales strategy; Feature engineering; LightGBM; Multi-objective planning; Random forest

1. Introduction

In recent years, China has proposed to achieve the goal of "carbon peak" and "carbon neutral" as soon as possible, and vigorously developing new energy electric vehicles is considered an effective way to solve the energy and environment problems, which has a broad market prospect. General Secretary Xi Jinping also emphasized that "the development of new energy vehicles is the only way to become an automobile power" when he visited Shanghai Automotive Group in 2014. However, compared with traditional cars, people still have certain doubts about the batteries of new energy vehicles, so the marketing needs to make scientific decisions and develop corresponding marketing strategies for target customers to increase the purchase rate and create revenue for the company.

A car company has newly launched three brands of electric vehicles, namely, joint venture brand, independent brand and new power brand. In order to study consumers' willingness to purchase the company's electric vehicles and develop corresponding sales strategies for scientific decision making, the sales department invited 1964 target customers to experience the three brands of electric vehicles. This study uses the knowledge of mathematical models to solve the following problems.

1. To explore the factors that may affect the sales of different brands of electric vehicles based on the purchase of electric vehicles by the target customers in the experience campaign.

2. Combine the results of the previous studies to build a customer mining model for different brands of electric vehicles, evaluate the excellence of the model, and determine the likelihood that a given target customer will purchase an electric vehicle.

2. Model building and solving

2.1. Problem 1

The filtering and wrapping methods in feature selection methods are carried out independently in the feature selection process and model training process, while the embedding method integrates these two processes and carries out feature selection while learning, therefore, in order to make the results more representative and facilitate model training, the following two embedding methods are used in this paper.

One is the feature selection method based on penalty term. In this paper, we use the penalty term feature selection method based on SVM model, LASSO model, and logistic regression model to select features. In this paper, we use the penalty term feature selection method based on SVM model, LASSO model and logistic regression model to select features.

Second, the feature selection method of the tree model, which can be used to calculate the importance of features and therefore can be used to remove irrelevant features. In this paper, the random forest model and the tree model feature selection method of the LightGBM model are used to select features. Then a threshold value is set according to the feature importance degree to get the top ranked features for each algorithm. Finally, a voting method is used on the results selected by the five models to filter out the features with greater than or equal to 2 votes as the main factors affecting the sales of the brand [1,8].

2.1.1 Logistic regression modeling

Logistic regression (LR) is a powerful statistical method for representing a binomial outcome in terms of one or more explanatory variables [2-5]. It measures the relationship between class-dependent variables and one or more independent variables, which obey a cumulative logistic distribution, by using a logistic function to estimate probabilities. It is defined by the following equation [3-5]:

$$h(x) = \frac{1}{1 + \exp(-x)} \tag{1}$$

This function is defined over the entire domain of real numbers, has a value domain of (0, 1), and is a monotonically increasing function. According to the requirements for the distribution function, this function can be used as the distribution function of the random variable x, i.e.

$$p(x \leq z) = h(z) \tag{2}$$

2.1.2 LASSO regression modeling

The full name of LASSO is Least absolute shrinkage and selection operator, which is a compression estimator that obtains a more refined model by constructing a penalty function that makes it compress some regression coefficients, i.e., force the sum of the absolute values of the coefficients to be less than a fixed value, while setting some regression coefficients to zero. Thus retaining the advantage of subset shrinkage, it is a biased estimator dealing with data with complex covariance. lASSO is the addition of an lp parametrization as a penalty constraint to the calculation of RSS minimization. lp parametrization has the advantage of shrinking some coefficients to be estimated exactly to zero when λ is sufficiently large. By cross-validation method: for a given value of λ, cross-validation is performed and the value of λ with the smallest cross-validation error is selected, and then the model is simply refitted with the full data according to the obtained value of λ [6,7].

$$\beta_{Lasso} = argmin_{\beta \in R^d} \left(\left|\left|Y - X\beta\right|\right|^2 + \lambda \sum_{j=1}^{d} \left|\beta_j\right| \right) \tag{3}$$

where t corresponds to λ one-to-one and is the adjustment coefficient.

2.1.3 SVM model building

SVM support vector machines are mainly targeted at binary linear classifiers by determining a hyperplane such that the distance from the data points to the hyperplane is maximized, which is a quadratic programming problem [8-10]. Since finding its appropriate mapping for any dataset is difficult, it is usually chosen from the commonly used kernel functions [11]. The commonly used kernel functions are polynomial kernel functions, Gaussian functions, linear functions, etc. The constraint for the classification hyperplane [8] is

$$y_i(w^T x_i + b) \geq 1 \tag{4}$$

The objective function is for the hyperplane to be sufficiently far from the two types of samples, which can be written as

$$\frac{1}{2}\left|\left|w\right|\right|.pow(2) \tag{5}$$

If the slack variable ξi and the penalty factor C are added to penalize the samples that violate the inequality, the following optimization problem can be obtained.

$$min\frac{1}{2}w^T w + c\sum_{i=1}^{l}\epsilon_i \tag{6}$$

$$y_i(w^T x_i + b) \geq 1 - \epsilon_i \tag{7}$$

$$\epsilon_i \geq 0, i = 1,2,\ldots,l \tag{8}$$

2.1.4 Random forest model building

A random forest is an algorithm that uses multiple decision trees to train and predict samples [12,13]. The random forest algorithm is an algorithm that contains multiple decision trees, and its output is determined by the species tree of the output category of the individual decision trees. Its advantages lie in most of the data, it has better classification, can handle high-dimensional features, is less prone to overfitting, has faster model training for large data, can evaluate the importance of variables in deciding the category, and is adaptable to the data set, can handle both discrete and continuous data, and the data set does not need to be purposely normalized. Random forest is an integrated learning model based on the idea of Bagging, by building multiple learners to accomplish the learning task together [14]. The principle of this model Bagging algorithm is similar to voting, where one weak learner is trained using one training set at a time, and n weak learners are trained based on different training sets after n random draws with put-backs. For the classification problem, the final prediction is based on the voting of all the weak learners, and the final prediction is based on the "majority rule". For the regression problem, the average of all the learners is taken as the final result. The statistical gain of the Gini coefficient is used in the decision tree. The Gini coefficient indicates the degree of variation of samples in the dataset, and the larger the Gini coefficient, the more diverse the dataset is, i.e., it indicates that there are multiple classification results, and the more diverse the current characteristics of the dataset is, i.e., the more impure it is, i.e., it may be a multi-classification problem [9].

The specific formula is as follows.

$$Gini(D) = 1 - \sum_{k=1}^{|y|} p_k^2 \tag{9}$$

The feature with the smallest Gini coefficient is generally selected as the initial root node. The selection of the root node is performed in the actual project based on the defined classification results by counting the features and categories of the training samples, and deciding the selection of the root node at each subsequent step based on the gain of the Gini coefficient of the features relative to the initial category. The Gini index when used as a basis for judgment for each feature needs to be known before calculating the Gini coefficient gain by the following formula.

$$Gimi_{index}(D,a) = \sum_{v=1}^{v} \frac{|D^v|}{|D|} Gini(D^v) \tag{10}$$

Genie gain is:

$$\mathrm{Gini(D,a)} = \mathrm{Gini(D)} - \mathrm{Gini_{index(D,a)}} \tag{11}$$

In this paper, we run the program to continuously classify the internal nodes until the final classification result is displayed. At this point, the training of the samples is completed, and we make predictions based on the current model when new samples come in for testing. In practical engineering it is necessary to control the depth of the decision tree to prevent overfitting situations, i.e., the training samples can perform well, but the accuracy of the test samples is not guaranteed, which can easily occur when the training set is large.

2.1.5 LightGBM model building

LightGBM is a distributed gradient boosting framework based on decision tree algorithm. It has the advantage of reducing the use of data for memory, ensuring that a single and its use as much data as possible without sacrificing speed, while reducing the cost of communication, improving the efficiency when multi-stage parallelism, and achieving linear speedup in computation [10].

2.1.6 Results of selecting features based on tree model feature method

In this paper, the random forest model and the tree model feature selection method of LightGBM model are used to select features, and the calculation results are shown in Tables 1.

Table 1. Features selected by the penalty term-based embedding method

Feature selection method	Select results		
	Brand 1	Brand 2	Brand 3
LR Logistic Regression	a1, a2, a3, b4, b5, B15, B16, B17	a1, a2, a3, a4, a5,a7, B11, B15, B16, B17	a1, a2, a3, a5, a6, B15, B16
SVM	a1, a2, a3, b3, b6, B10, B16, B17	a1, a3, a5, a6, a8, b8, B10, B15, B16, B17	a1, a2, a3, a7, B2, B6, B16
LASSO	a1, a2, a3, a6, b6, B12, B16, B17	a1, a3, a5, a6, a7, B10, B11, B13, B16, B17	a1, a2, a5, b6, B10, B16, B17
RF	a1, a2, a3, B2, B15, B16, B17	a1, a2, a3, a4, a5, a7, B13, B15, B16, B17	a1, a2, a3, a5, a3, a6, a8, B16
LightGBM	a1, a4, a5, B6, B7, B10, B16, B17	a1, a3, a7, B1, B11, B13, B14, B15, B16, B17	a2, B16

The figure 1 shows the graph of the number of occurrences of the features obtained by voting on each feature using five models. In this paper, the features with occurrences greater than or equal to 2 are selected as important features and are considered to have an impact on each of the three brands of electric vehicles.

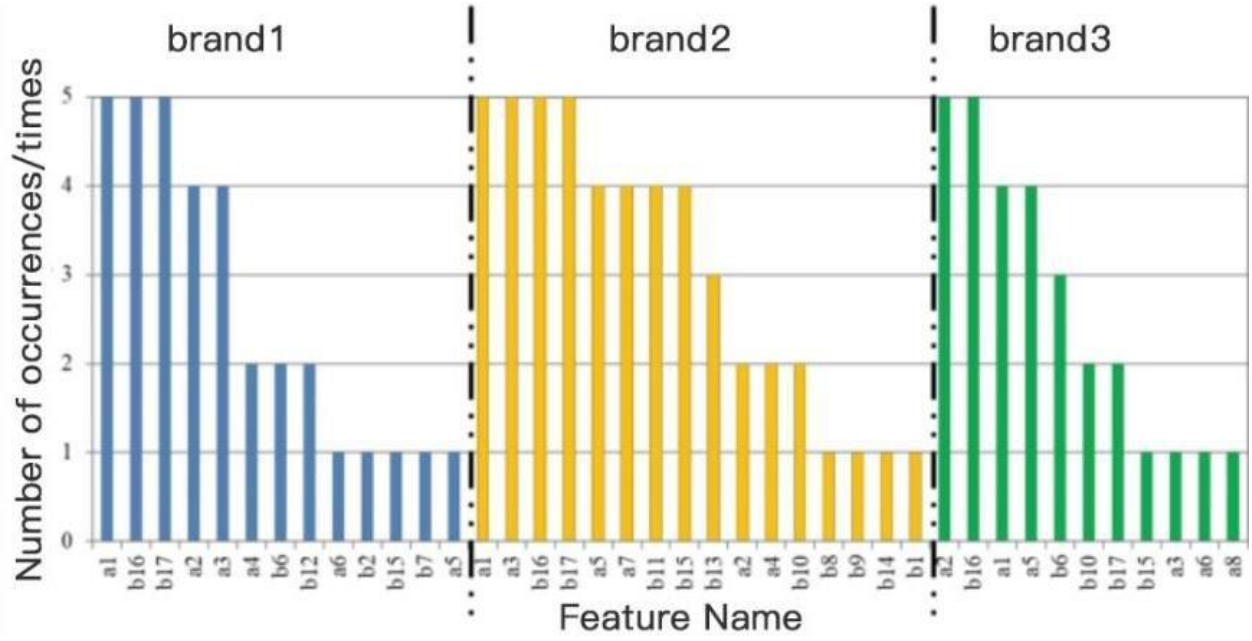

Figure 1. Number of feature occurrences

As can be seen from the figure, for the three brands, the battery technical performance, comfort, target customer's annual mortgage, and car loan as a percentage of annual household income are two or more for EVs, indicating that these four indicators have a greater influence on customers' purchase situation. In addition to the above four indicators, the factors affecting the purchase of Brand 1 include economy, safety, driving control, and the position of target customers; economy, safety, power, exterior and interior, target customers' work status (including the number of years of work and the nature of the unit), annual household income, and disposable annual income are the main factors affecting the purchase of Brand 2 by target customers; the main factors affecting the sales of Brand 3 are power, target customers' annual mortgage, and the percentage of annual household income. The main influences on brand 3 sales are dynamics, marital and family status of the target customer, and length of employment [11] [12].

2.2. Problem 2

2.2.1 Customer mining model based on multi-layer perceptron

Based on Problem 1, we use a multi-layer perceptron (BP neural network) to build a mining model for different brands of customers and then predict their purchase intention. Since the prediction is done separately for different brands of EVs, the multilayer perceptron model is also trained under the training set of different brands, so a total of 3 multilayer perceptron models $f_1(x)$, $f_2(x)$, and $f_3(x)$ are trained to predict the purchase intention of customers for brands 1, 2, and 3, respectively.

2.2.2 Multi-layer perceptron model

The multilayer perceptron model used in this problem is based on the single-layer perceptron model described above, with two hidden layers of 4 neurons each and input data with 25 features (a1~a8, B1~B17) and direct quantization of the categories corresponding to the category variables. The structure of double hidden layer neural network is shown in Figure 2.[13]

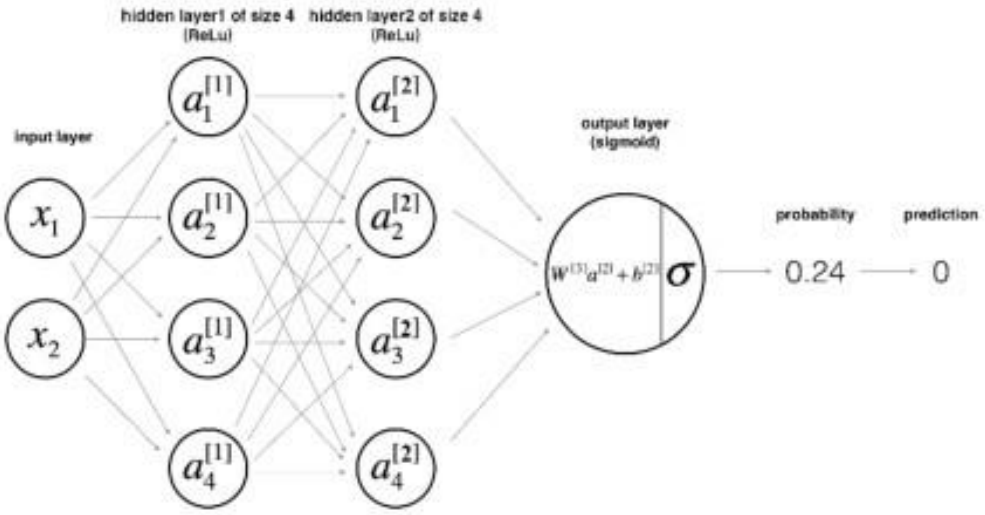

Figure 2. The structure of double hidden layer neural network

The specific steps of parameter initialization, forward propagation, loss function calculation and parameter update of the double hidden layer neural network are almost the same as those of the single hidden layer, while the activation function in the middle layer uses the ReLu activation function, and the layer-by-layer back propagation formula from the last output layer to the first hidden layer is given here, where the capital letters represent the matrices.[14]

$$\begin{cases} dA^{[3]} = \dfrac{1-Y}{1-A^{[3]}} - \dfrac{Y}{A^{[3]}} \\ dZ^{[3]} = dA^{[3]} * \dfrac{1}{1+e^{-z^{[3]}}} * \left(1 - \dfrac{1}{1+e^{-z^{[3]}}}\right) \\ dW^{[3]} = \dfrac{1}{N} dZ^{[3]}.A^{[2]} \\ db^{[3]} = \dfrac{1}{N}\sum_{i=1}^{N} dZ^{[3](i)} \\ dA^{[2]} = W^{[3]^T}.dZ^{[3]} \\ dZ_i^{[2]} = \begin{cases} dA^{[2]} & Z^{[i]} > 0 \\ 0 & Z^{[i]} \le 0 \end{cases} (i = 1,2 \dots n1) \\ dW^{[2]} = \dfrac{1}{N} dZ^{[2]}.A^{[1]} \\ db^{[2]} = \dfrac{1}{N}\sum_{i=1}^{N} dZ^{[2](i)} \\ dA^{[1]} = W^{[2]^T}.dZ^{[2]} \\ dZ_i^{[1]} = \begin{cases} dA^{[1]} & Z^{[i]} > 0 \\ 0 & Z^{[i]} \le 0 \end{cases} (i = 1,2 \dots n2) \\ dW^{[1]} = \dfrac{1}{N} dZ^{[1]}.X \\ db^{[1]} = \dfrac{1}{N}\sum_{i=1}^{N} dZ^{[1](i)} \dfrac{1}{1+e^{-Z^{[3]}}} \end{cases} \tag{12}$$

The process of back propagation is similar to that of the previous single hidden layer. It should be noted that * represents multiplication and · represents matrix multiplication. n1 is the number of neurons in the first hidden layer, and n2 is the number of neurons in the second hidden layer.

2.2.3 Solution of multi-layer perceptron model

1) Training test data pre-processing

If the input customer feature data is not normalized, the neural network training effect is very poor and the classifier is prone to failure. Each feature of the customer is normalized separately, and the normalization is done by

$$X^* = \frac{X - \mu}{\sigma} \tag{13}$$

In order to determine the training and test sets, the data set needs to be partitioned, and it is important to ensure that in the training set, the proportion of samples of each class is approximate (1:1), otherwise the neural network is more inclined to predict all the input samples as the class with more samples in the training set.

A macroscopic presentation of the distribution of the dataset is first made (Table 2).

Table 2. Distribution of the data set

	Number of people buying cars	Number of people who did not purchase a car	Purchase/Not purchased
Brand 1	23	533	0.0432
Brand 2	65	1208	0.0538
Brand 3	11	124	0.0887

Obviously, the distribution of the two types of samples is extremely unbalanced, and the number of samples that purchased cars is much smaller than the number of samples that did not purchase cars. Here, the method of replicating the sample of customers who bought cars is used to expand the data set, and then the data set is disrupted, which can alleviate the problem of unbalanced positive and negative samples to some extent.

The specific algorithm flow is as follows.

Let a be the number of samples in category A (purchased), b be the number of samples in category B (not purchased).

$$n = \frac{[b - mod(a, b)]}{a} \tag{14}$$

n is the result of the operation of dividing b by a. The samples of category A are copied n+1 times, so that category A has a(n+1) samples and a (n +1) b. The last a(n +1) -b samples of category A are deleted, and the number of samples of category A and category B is 1:1. The number of samples of category A and category B is 1:1, and the equalization is achieved. It is proved that the equalization of samples can greatly improve the prediction accuracy of the network. In order to avoid the occurrence of overfitting and underfitting, in addition to the training set, we also need to use the test set for real-time testing to determine the accuracy of the network.

In order to avoid the occurrence of overfitting and underfitting, in addition to training with the training set, a test set is also required for real-time testing to determine the fitting status of the data set (Table 3).

Table 3. Table of positive and negative sample proportions of the original data set

	Training set		Test set	
	Purchase	Not purchased	Purchase	Not purchased
Brand 1	526	526	7	7
Brand 2	1193	1193	15	15
Brand 3	120	120	4	4

3) Model solving results

The following are the iterative curves of the accuracy of each of the three brands of neural networks during the training process, respectively (Figure 3).

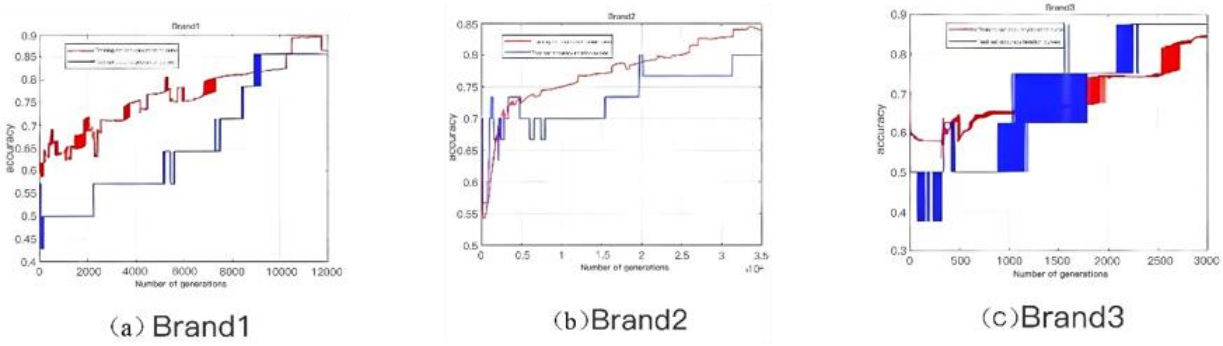

Figure 3. Iteration Curve

It can be seen that the training results are still relatively good, and there is no overfitting or underfitting, and the model converges to a relatively good state. The accuracy of the final model on the training and test sets is shown in Table 4 below.

Table 4. prediction accuracy of the model in the training and test sets

	Training Set Accuracy	Test Set Accuracy
Brand 1	86.4068%	85.7143%
Brand 2	83.9899%	80.0000%
Brand 3	84.5833%	87.5000%

It can be seen from the above table that the accuracy of the training set and the test set is high, which can be used for prediction. Then the target customers are predicted, and the first customer (brand 1), the fifth customer (brand 1), the sixth customer (brand 2), the seventh customer (brand 2), the 12th customer (brand 3) and the 13th customer (brand 3) are willing to buy new energy vehicles.

3. Conclusions

3.1. Model Advantages

1. Machine learning algorithms should be used, which are highly interpretable.
2. When performing feature screening, different operations are performed for different categories, making the data more convincing.
3. Apply multiple machine learning algorithms for prediction, making the results more accurate.

3.2. Model Disadvantages

1. ignoring the connection existing between different choices will lead to some errors.
2. The generalization ability is weak when applying BP neural network due to less data.

References

[1] REN H Y, LIANG Y, ZUO P X. Analysis of influencing factors of breastfeeding based on logistic regression and decision tree model[J]. Chinese Journal of Health Statistics,2019,36(4):532-534.

[2] Wubalem A, Meten M, 2020. Landslide susceptibility mapping using information value and logistic regression models in Gon-cha Siso Eneses area, northwestern Ethiopia. SN Applied Sciences,2(5):807. doi:10.1007/s42452-020-2563-0

[3] THORESEN M. Logistic regression-applied and applicable[J]. Tidsskr Nor Largeforen,2017.137:1

[4] Caigny A. D, Coussenment K,Bock K.W.D.A new hybrid classification algorithm for customer chum prediction based on logistic regression and decision trees[J].European Journal of Operational Research,2018,269(2):760-772.

[5] GENG K H, LI Z M. Prediction model of financial failure based on the principle components-logistic regression method[J], Journal of HebeiUniversity of Technology,2005,34(2):53-57.

[6] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. J Royal StatSocSer B Methodol, 1996,58(1):267-288.

[7] TIBSHIRANI R, SAUNDERS M, ROSSET S, et al. Sparsity and smoothness via the fused lasso[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology),2005, 67:91-108.

[8] Liang W, Luo S, Zhao G, et al. Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms[J]. 2020.

[9] Arvind, Rantan R. Identifying traffic of same keys in crypto-graphic communications using fuzzy decision criteria and bit_plane measures[J]. International Journal of System As-surance Engineering and Management,2020,11(2):466-480.

[10] Wang B, Wu P, Chen Q, et al. Prediction and Analysis of Train Passenger Load Factor of High-Speed Railway Based on LightGBM Algorithm[J]. Journal of Advanced Transportation, 2021,2021(4):1-10.

[11] LI H, WU T T, YANG D L, et al. Decision tree model for predicting in-hospital cardiac arrest among patients admitted with acute coronary syndrome[J]. Clinical Cardiology,2019,42(11):1087-1093

[12] ZHAO Z Q, ZHENG M. Apply classification tree to automatically screen some potential interaction factors in logistic regression[J]. Chinese Journal of Health Statistics,2007.24(2):114-116.

[13] LIU H, HAN J, XI L. Agricultural product price forecast based on wavelet transform and BP neural net-work[J]. China Agricultural Informatics,2019,31(6):85-92.

[14] Kulbear. Building your Deep Neural Netwoek–Step by Step.http:\\github.com/Kullbear/deep-learning-coursera. (2017)