Sales Prediction Based on Machine Learning Scenarios

Qirui Mao [1, *]

[1]University of Illinois at Urbana Champaign, Gies college of business, Champaign, United State

*Corresponding author: qiruim2@illinois.edu

Abstract: With the development of technology, business analysis plays a crucial role among companies. Companies prefer to manage their operation by using high-tech rather than the traditional method. Sale is one of the important parts of the operation of a company which also determines the company's profit and management behavior. On this basis, business analytics becomes a new area suitable for promoting the progress of companies. The sale prediction based on machine learning gets popular among them. The article is trying to introduce the mainstream algorithm and models of machine learning that are used for prediction and the process of how to analyze a certain part of a company using the machine learning method. Decision tree and Neural networks are two main algorithms that will be mentioned in the following article, each algorithm will present a basic mathematics approach that is easy to read. In addition, the application of sales prediction of how to apply the machine learning method to real-world examples will be the last part of the article. These results give a brief and plain understanding of sales prediction based on machine learning to the people who first contact this field, which shed light on guiding further exploration of sales prediction.

Keywords: Sales prediction, machine learning, business analysis.

1. Introduction

The occurrence of artificial intelligence makes machine learning becoming the main method that solves the AI problem. Last century, Rosenblatt, from Cornell University, create a model that mimics the human brain to solve problems [1]. Supervise learning, unsupervised learning, and Reinforcement learning are three main parts of machine learning. It expects to learn from given simples. Unsupervised learning expects to learn from the simple data without the training process. Reinforcement learning expects to maximize the result by executing given the circumstance. In short, machine learning simulates human learning ability via learning the given samples and experience. Machine learning is a branch of AI. It can learn from the data sample to find the law that can be used for deduction and decision-making.

In the 1980s, machine learning is emerging but not so popular. Between 1990 and 2010, it develops rapidly with the birth of new algorithms and models, leading to practical use. After 2012, reinforcement learning develops quickly and solve many current AI problems and boost the development of related industries [2]. Based on statistics of literature on business analytics, the amount of keywords business analytics increased exponentially. Sales prediction is one of the business analytics that increases the efficiency of an organization which is supported by machine learning techniques. Since sales prediction is interdisciplinary, a master of machine learning is critical to the business field [3]. Sale prediction is becoming one of the most important parts of companies' plannings. A survey, which focuses on 175 Midwest merchants, shows that 65% of them consider the sales prediction as the way to the success of their companies. Besides, 28% of them deem that the sales prediction is not necessary [4].

Big data become more important with a focus on the reformation of the Internet industry. The complicated thing is the prediction with the business sales, relating to business operation. Considering only the machine learning models, it is easy to predict the sales precisely, but it is difficult to increase the profit by only predicting the sales. Sales prediction is one part of business analytics, it can help companies to deal with short-term operational management. Since machine learning can learn from big data automatically without intervening with humans, it can help companies to make decisions by using it settled model. Machine learning is good in many fields, especially in commerce. Traditional

sales prediction should cost lots of human labor and money to build and maintain the system, and different departments should cooperate with each other to make sure that everyone is on the page. In that case, the process of sales prediction is inefficient and time-consuming. However, the usage of a machine learning model can help to target the issue of companies more precisely and effectively. It lowers the cost of companies and the labor costs because every staff can use the system to do the jobs according to the increase in the efficiency of using systems.

In addition, Sales prediction can predict the data given previously dataset to categorize the goods which can cater to consumers' behavior. On this basis, it can convey the correct and directional information to customers and effectively market the commodities. Nevertheless, due to the complicated reality and scarcity of the dataset, it is difficult to precisely predict the sales. In general, most sales predictions focus on total goods rather than specific goods. According to the previous test, the decision tress model had both advantages and disadvantages when predicting sales [5]. This research is going to introduce the mainstream algorithms in machine learning that are popular in business analytics, specifically in sales prediction. The motivation of this article is to give knowledge to the people who first contact with machine learning and want to have a brief understanding of sales prediction. The article is a review work, related to other's experimental works. The first part of the article introduces the algorithms – decision trees and neural networks - which may be used in sales predictions. The second part of the article introduces the evaluation metrics including RMSE to give the understanding of judging the models. The Last part of the article is to use examples and process work to introduce the process of sales prediction.

2. Algorithms

2.1 The Decision Tree

The decision tree is one of the machine learning approaches. The decision has algorithms of ID3, C4.5, and C5.0. The Decision Tree is basically a tree-like structure that can divide one node into two or more sub-nodes. Each sub-nodes represent the previous sub notes' results. The end nodes represent the result of one classification [6]. A Decision Tree is a common approach to classification based on supervised learning. Supervised learning is that given samples, each sample has one attribute and one result, which means the results of classification are known. Given a Decision Tree via learning samples, it is supposed to make a prediction on the new dataset. A simple example to explain the construction of a Decision Tree is given as follows. Assume ten samples in total (number of students), each sample has scores, attendance, speech, and homework submission. The result is to justify whether each sample is a good student or not based on the previous four attributes. For simplification, it is assumed that the Decision is binary. A sketch of the decision tree is given in Fig. 1.

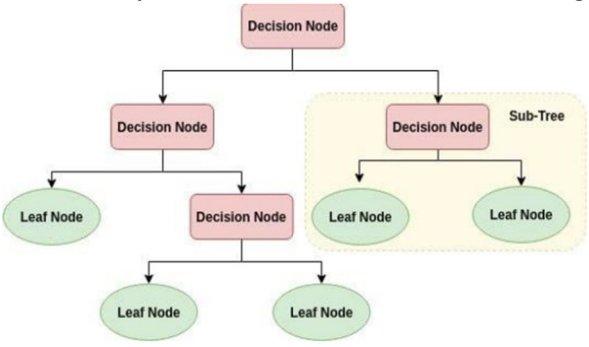

Figure 1. A sketch of Decision tree

The generation of the Decision Tree is mainly divided into two steps: Split of nodes and Determination of threshold. Primarily, if one node fails to judge a segment, the node should split into two sub-nodes or n sub-nodes. Subsequently, one chooses an appropriate threshold to limit the training error [7]. ID3 is one of the decision Trees, the principle of it is increasing the entropy to

determine the node, which needs to split. For a dataset, the lower the entropy, the better the result of classification.C4.5 is the revised version of ID3 in order to prevent overfitting because ID3 trains the dataset for extremely high accuracy [8]. Even if the training dataset can offset the error rate via numerous splits, the ID3, through overfitting, will increase the error rate for a new dataset because the new dataset is different from the training set [7]. A Decision Tree is supposed to make predictions on the new dataset by training the given datasets rather than training a specific dataset. Therefore, training datasets deeply will cause overfitting.

2.2 Classification and regression Tree (CART)

CART is an unconventional method to analyze the dataset. CART predicts y based on the given x on conditional. The model uses binary means to partition the dataset. Given a dataset, the model will use a greedy algorithm to grow a new tree and then prune the tree in case of a specific division of CART, which will cause overfitting. The greedy algorithm maximizes to fit the standard node by growing a tree based on split criteria sequentially. The new approach that has been put forward is the Bayesian approach [9]. The Bayesian rule in statistics means the standard approach that adjusts the judgment subjectively on relating probability distribution gave the observation. When the number of analytical samples approaches the population, the probability of incident occurrence among samples tends to be that of the population. The origin of the CART tree is the process of constructing a recursive binary decision tree, which is schematically shown in Fig. 2. By use of minimization squared error criterion for regression trees, generation of binary tree based on that criterion of Gini index. Gini Index states the purity of a branch from two randomly selected samples with different types. In the case of CART, the squared error is the sum of squares of deviations between the output value and true value of samples. In a classification problem, assume there is a K class, the probability of sample which belongs to the K class is p_k, the Gini index of the probability distribution defined as [10]:

$$Gini(p) = \sum_{k=1}^{k} P_{k(1-p_k)} \quad (1)$$

The Gini Index indicates the uncertainty of the samples. When Gini gets larger means the uncertainty of the sample gets larger as well.

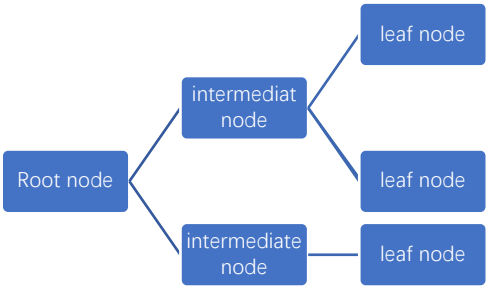

Figure 2. A sketch of CART

2.3 Random Forest

Random forest is the selection of random attributes in process of Decision Tree training. Specifically, Random Forest selects the best attribute among all the attributes at that point. Random Forest consists of numbers of Decision Trees that are not related to each other [7]. In classification, each tree from Random Forest will judge and classify the new sample, resulting in a unique result in its own classification. Random Forest regards the result of classification with more segments as result. Generally, there are four steps to constructing Random Forest. The first is to select one sample with replacement from the dataset with a sample size of N. The selected N samples are used to train the Decision Tress as the sample of nodes. Second, when each sample has M attributes, the Decision Tree will select the ‘m’ from ‘M’ attributes which satisfying m is less than M. Every split of a node will continue, according to process two, until it stops [6]. Depending on partition and ending standards, decision trees can be divided into categorical and regression processes. Logistic regression is the classic categorical. The way to determine the split node, based on the split criteria the result of

optimization, is the key issue of decisions trees. Entropy is the common split standard of splitting into tasks of classification. In each internal node of decision trees, the formula given:

$$E = -\sum_{i=1}^{c} p_i log p_i \quad (2)$$

Where c is the unique number of classifications and pi is the probability of given pi. The mean squared error can test the split standard [11].

2.4 Boosted trees

Gradient boosting is part of an ensemble, which is collecting the weak learning modes together, which expect to build a model and predict it at first. The fundamental of boosting is the concept that there is a polynomial learning algorithm that can learn it and has a high accuracy rate, then the concept is strong learnable. If there is a polynomial learning algorithm that can learn it, and the accuracy rate is very low, then the concept is weakly learnable. In this case, given weakly learning algorithms, whether the model improves it to be a strong learning model is critical. When the result comes out, it will calculate the residuals and those will become the new training samples for the next text. The gradient boosting will ask for a new model to fit the residuals and so on. Finally, the sum of the models in the process will be the final models. The model tries to lower the function loss, which means the model is renovating, making the loss function descending in the direction of its gradient.

XGBoosted tree consists of three phases: data processing, feature selection, and training process. In the phase of data processing, it analyzes the original dataset by deleting defaulted data and preliminary processing. In the phase of feature selection, the technique of feature engineering is used to train and predict the features based on one-hot coding. In the phase of training and prediction, XgBoosted model will adopt the selected feature and level up the accuracy of prediction by adjusting the parameters of the model. The XGBoosted model itself will finalize the prediction with the high accuracy model [12]. Fig. 3 presents the processing diagram for XGBoosted Tree.

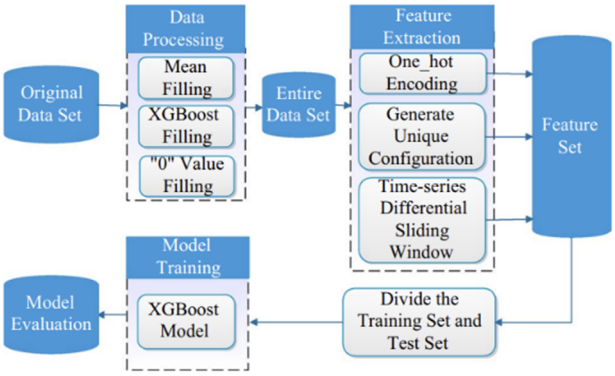

Figure 3. System architecture of XGBoosted tree [12].

Light Gradient Boosting Machine is a framework for achieving the GBDT algorithm. LightGBM supports highly efficient parallel training with faster speed, lower consumption of memory, and higher accuracy. The occurrence of LightGBM is to solve the problem when Gradient Boosting Decision Tree, using iterative training to achieve the optimal model, solving the tons of data, making GBDT a better-fit use on sale prediction [13]. LightGBM uses the growth strategy of Lead-wise. In case of the strategy, the model finds the leaf with the largest split gain, then continue to split, etc. The advantage of the Leaf-wise is the better accuracy in given conditions of numbers of the split.

2.5 Neural Network

The neural network is the statistical or mathematical model that mimics the structure and function of the neural network of animals, with a hierarchical directed graph with several nodes in each layer. With the development of the technology of computer processing, the neural network model can support the processing of a large number of features in the machine learning model [14]. The brain is the main part of the Central Nervous System and humans receive the new information by activating

each layer of neurons. The model of a Neural network is the analogy of a human brain that can study and differentiate the new information. With mimicry of the brain, the neural network is a general model that can find a way to study and process the new data. By using the neural network, users should input a series of values of features and learn what categories the new feature should be classified by calculating the process. In addition, networks have the ability to train the new data and predict them for the future by learning representative datasets. The networks can solve the problem for basic images for a short time and complex images for a long time.

Basically, input, hidden, and output is the three general layers of neural networks. All the independent variables form the input layer, and the dependent variable y forms the output layer which usually exists in one neuron. For the hidden layer, it is a set of nodes that each node in it receives weighted signals from all the nodes in the previous layer and transforms on the activation function [15].

Convolutional Neural Net (CNN) is one of the most influential revolutions in the field of computer vision. CNN uses the convolutional mathematical calculation in the structure of neural networks instead of the matrix multiplier of traditional neural networks. Since CNN uses the two-dimensional data structure, compared to other machine learning structures, it can have a better understanding of recognition and speech [16]. CNN consists of convolutional layers, pooling layers, and dense layers. The convolutional layer will convolute the image data that is transported to neural networks which are scanned by grids to a cubic structure containing three grams of red, yellow, and green. There are hundreds of neurons can recognize the information in reality and get the activation image when networks apply neurons to images. The specific neurons will then activate the area of feature.

The pooling layer will replace each grid area to the maximum value in feature graph. The reason is to reduce the overfitting by ignoring the exact position of a certain feature in a given area. The output layer will transport the data from the last layer to the fully connected part, including a dense layer that projects the output from neurons and apply the softmax function to the output. The softmax function is to achieve the classification of images through neural networks by returning a feature list of the probability of each image, which has a total probability of one. For higher accuracy, there should be more layers of convolution and pooling finishing a complete feature iteration and then transport back to calculate the weights on each feature by the decrease of stochastic gradient [17].

3. Evaluation Metrics

The way evaluate a predictive model is important for choosing a suitable model for sales prediction in a certain field. For each evaluation function, y represents the real value of the model whereas ŷ represents the value of prediction. There are some main evaluation indexes that can compare the regression models. The first one is R-Squared (R^2) which can be described as:

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=0}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=0}^{n}(y_i-\bar{y})^2} \tag{3}$$

R^2, also known as the coefficient of determination, reflects the change degree of dependent variable y as independent variable ŷ changes. The numerator represents the total error of the testing model whereas the denominator is the total error of the prediction. If the value is closer to 1, it means the good fitting of the model. Another group of metrics is Mean Absolute Error (MAE) and Mean Squared Error (MSE), which means the expected value of absolute loss of error and the expected value of squared loss of error, respectively:

$$MAE(y,\hat{y}) = \frac{1}{N}\sum_{i=0}^{n}|y_i - \hat{y}_i| \tag{4}$$

$$MSE(y,\hat{y}) = \frac{1}{N}\sum_{i=0}^{n}\left||y_i - \hat{y}_i|\right|_2^2 \tag{5}$$

Sometimes the Root mean Squared Error (RMSE) will be used, which is the squared root of MSE.

Different situations are suitable for different evaluation metrics except for R^2, which is "the lower, the better". If each metric is used separately, MSE and RMSE put a great value on the squared value between real and predictive value, whereas MAE focuses on the outlier. MSE will exemplify the big

error and train the deep neural networks. Since the linear regression tries to minimize the function loss, comparing each model, the lower the MSE value represents the high accuracy of the dataset. MAE will be functional when comparing different models because the value itself does not make sense. RMSE will increase with the increase of the distribution of error frequency. The significance of the square root sign is to make the data and result of the error at the same level. Since MSE is sensitive to the outlier error, it can reflect the accuracy of the measurement. It is worth noting that MSE validates a computable loss function quickly so that it can provide the extended measurement of error [18]. More practically, by mixing the use of both MAE and RMSE, it can show the degree of dispersion of sample errors. For example, if RMSE is far less than MAE, it shows a tiny distinction between different samples [19].

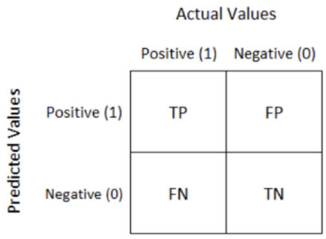

Figure 4. A sketch of the confusion matrix.

As for classifier, Receiving Operator Characteristic (ROC) and AUC will be used. The confusion matrix is the basic concept of ROC. Positive, Negative, True, and False are four factors of the matrix. Assume the prediction category 1 positive, prediction category 2 negative. Assuming true and false situations, a sketch of the confusion matrix is presented in Fig. 4. Subsequently, elicit two concepts of true positive rate (TP Rate), false positive rate (FP Rate):

$$TP\ Rate = \frac{TP}{TP + FN}, FP\ Rate = \frac{FP}{FP + TN} \tag{6}$$

The meaning of TP Rate is the proportion of predictive value 1 given actual value is 1. The meaning of FP Rate is the proportion of predictive value 0 given the actual value is 0. For the ROC curve, the vertical line is the true positive rate whereas the horizontal line is the false positive rate. Obviously, the area of the ROC curve is less than 1. When TP rate = FP rare, there is a 45-degree slope across the ROC curve, meaning that 50% is correct for the predictive positive samples. When the ROC curve is under the 45-degree slope, it indicates the accuracy of the classifier is worse than the random classifier that TP rate equals to FP rate, and vice versa. The expectation is that the ROC curve line can move up left to achieve a more accurate classifier. The limit of the ROC curve is not explicit to show the consequence of the classifier though it can reflect it at the same level. The introduction of the AUC curve, the area under the ROC curve, is to indicate the ability to classify of ROC curve. The classification will be better than the random classifier if AUC is between 0.5 and 1 [20].

4. Application in sales prediction

The machine learning algorithm is based on computers and mathematics. By applying it to real-world business, there are four steps to get involved: demand, explore, develop, and evaluate. The first step is to define the prediction object and know the demand of the object. Sales prediction needs an expectation, which determines the time for the target object. Generally, if a company wants to predict the sales for one month need the data for the previous two years. In specific, the short-term prediction will affect the storage or production of the company whereas long-term precision will affect the operation of the management. The importance of long-term and short-term predictions should be determined by the industry. For the food industry, food's shelf life is not long compared to electronic devices, so it is sensitive to the storage level. If the company fails to predict the demand for a certain food in a short period, the result will probably be a waste and have a negative impact on the company's

cost (Grigorios, 2018). According to the demand of the company, the manager can build a model to construct a system that leads to appropriate sales and storage measurements.

Subsequently, to include collecting data, exploring data, and preprocessing the data. When an e-commerce company wants to predict sales, more previous data will get more accurate results. Data exploration is the way to find the law of the data and prepare for feature selection on models. Usually, some early-stage companies do not have a complete database, the primary data is chaotic and low-qualified.

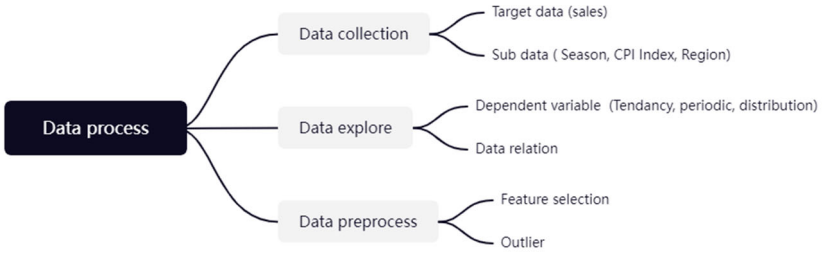

Figure 1. The procedure of data process

One important step in it is to determine which models should be used in the certain cases, the degree of model complexity, and interpretability is important because the simpler the model, the more efficient will the system be. The collation of the indicator system for sales forecast is mainly determined by business logic and the data of the commodity to be predicted. For example, for the sales forecast of a certain commodity in e-commerce, it can be concluded from business analysis that inventory factors, price factors, word-of-mouth factors, holiday factors, promotional advertising factors, and news hotspot factors, and the determination of the forecast period requires the business department to combine with the enterprise. It depends on the overall supply chain capabilities and the historical data of the inventory itself.

By reviewing sales prediction among companies, machine learning will be more focused on inter-relation between two variables. For sales prediction, companies should use the variables relating to changes in sales. In addition, to evaluate the model, the main purpose is to increase the profit of the companies, for the limit of predicting solely on sales, the companies should also consider all aspects of the companies' operation.

5. Conclusions

In summary, this paper summarizes machine learning algorithms, evaluation metrics, and steps for sale prediction in business analytics. Specifically, decision trees and neural networks are the two main algorithms mentioned in the articles. These two models are basic methods of machine learning. The decision tree is a graphical method given the known probability of situations to evaluate the feasibility of decision-making. The neural network is a simulated human brain consisting of tremendous neurons learning non-linear models. These two methods all supervise learning skills, which means learning to get a classifier that can give the correct classification of newly emerged objects given determined classification. The evaluation metrics, it is an important part of machine learning. Judging the quality of things requires certain criteria, and judging the quality of a classification system naturally requires certain evaluation methods. Different tasks have different criteria, even one machine learning model also has different criteria. The evaluation metric mentioned above is some of the popular and practical metrics that are often used in sales prediction. Since the train data will be split into the training set, developing set, and test set, the accuracy is the most intuitive data that can evaluate the feasibility of the model. Nonetheless, as there are many evaluation metrics that are only the introduction, the way to determine which evaluation metrics are suitable for a certain model should be deer time case by case, but usually, these evaluation metrics data will all be shown in the table when finishing the test. The last part of the article is the steps for sales prediction based on machine learning. Since sales

prediction is a part of business analytics and a part of computer science, it requires the people who should be in command of business knowledge and machine learning skills to properly conduct the sales prediction process. The article introduces the basic steps and processes for sales prediction, giving a basic understanding of sales prediction and how does it link to machine learning. For limitation, this article is a review work, which means all the data and examples stem from others' works. The author itself cannot assure the accuracy of the data and examples. All the results and conclusions come from others' work, and the author's opinion is based on these results and conclusions. The future of machine learning is going to make computer has their own consciousness, which can make the computer learn and understand computer. For sales prediction, in the future, more companies are expected to use machine learning skills to manage their operation, e.g., sales prediction to scientifically promote the profit of the companies. The meaning of writing this article is to give a basic understanding of machine learning algorithms and evaluations to people who first contact this area. These results offer a guideline for future study focusing on sales prediction.

References

[1] Fradkov A L. Early history of machine learning[J]. IFAC-PapersOnLine, 2020, 53(2): 1385-1390.

[2] Foote K D. A brief history of machine learning[J]. DATAVERSISTY,[Disponible en ligne: https://www. dataversity. net/a-brief-history-of-machine-learning/],[Date de la dernière consultation: 15 Novembre 2019], 2019.

[3] Yin J, Fernandez V. A systematic review on business analytics[J]. Journal of Industrial Engineering and Management, 2020, 13(2): 283-295.

[4] Dalrymple D J. Sales forecasting practices: Results from a United States survey[J]. International journal of Forecasting, 1987, 3(3-4): 379-391.

[5] Xu Shanshan. "Research on Machine Learning-Based Commodity Sales Prediction." Collection 4 (2019).

[6] Abdulkareem N M, Xia A M. Machine learning classification based on Radom Forest Algorithm: A review[J]. International Journal of Science and Business, 2021, 5(2): 128-142.

[7] Li Y, Jiang Z L, Yao L, et al. Outsourced privacy-preserving C4. 5 decision tree algorithm over horizontally and vertically partitioned dataset among multiple parties[J]. Cluster Computing, 2019, 22(1): 1581-1593.

[8] Sarker I H, Colman A, Han J, et al. Behavdt: a behavioral decision tree learning to build user-centric context-aware predictive model[J]. Mobile Networks and Applications, 2020, 25(3): 1151-1161.

[9] Chipman H A, George E I, McCulloch R E. Bayesian CART model search[J]. Journal of the American Statistical Association, 1998, 93(443): 935-948.

[10] Li M. Application of CART decision tree combined with PCA algorithm in intrusion detection[C]. 2017 8th IEEE international conference on software engineering and service science (ICSESS). IEEE, 2017: 38-41.

[11] Schonlau M, Zou R Y. The random forest algorithm for statistical learning[J]. The Stata Journal, 2020, 20(1): 3-29.

[12] Xia Z, Xue S, Wu L, et al. ForeXGBoost: passenger car sales prediction based on XGBoost[J]. Distributed and Parallel Databases, 2020, 38(3): 713-738.

[13] Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree[C]. Advances in Neural Information Processing Systems. 2017: 3146-3154.

[14] Sharkawy A N. Principle of neural network and its main types[J]. Journal of Advances in Applied & Computational Mathematics, 2020, 7: 8-19.

[15] Miller A S, Blott B H. Review of neural network applications in medical imaging and signal processing[J]. Medical and Biological Engineering and Computing, 1992, 30(5): 449-464.

[16] O'Shea K, Nash R. An introduction to convolutional neural networks[J]. arXiv preprint arXiv:1511.08458, 2015.

[17] Yamashita R, Nishio M, Do R K G, et al. Convolutional neural networks: an overview and application in radiology[J]. Insights into imaging, 2018, 9(4): 611-629.

[18] Hu Z, Jin Y, Hu Q, et al. Prediction of fuel consumption for enroute ship based on machine learning[J]. IEEE Access, 2019, 7: 119497-119505.

[19] Chai T, Draxler R R. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature[J]. Geoscientific model development, 2014, 7(3): 1247-1250.

[20] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves [C]. Proceedings of the 23rd international conference on Machine learning. 2006: 233-240.