Journal of Organizational and End User Computing
Volume 35 • Issue 3

The Use of an Internet of Things Data Management System Using Data Mining Association Algorithm in an E-Commerce Platform

Guan Wang, Macao University of Science and Technology, Macao
Xuan Zhang, Monroe College, USA
Yifan Gao, University of Wisconsin-Madison, USA
Austin Lin Yee, Peking University, China
Xue Wang, Nanning College of Technology, China*

ABSTRACT

The development of e-commerce has greatly changed the development of social retail formats. Business-to-consumer (B2C) e-commerce model is important. Due to the characteristics of high consumer trust and commodities dominated by electronic products and brand commodities, the income and profits generated are also very considerable. Therefore, the major e-commerce giants have increased the development of B2C formats. Logistics service capability and level have become an important driving force for the development of B2C e-commerce. How to optimize the inventory of B2C e-commerce and realize the organic balance between the economy and service capacity of the whole logistics chain has become a very urgent problem faced by major e-commerce giants. From the perspective of big data, first, the overview of the dataset used is analyzed based on the real operation data of a business to consumer (B2C) e-commerce platform.

KEYWORDS

B2C E-Commerce Platform, Data Mining, Demand Prediction, Feature Engineering

INTRODUCTION

The development of e-commerce is based on the popularization of the Internet. By fully using the fast Internet and the security and reliability of network support, it realizes the rapid interaction of information flow and capital flow with consumers and reaches the transaction intention. Finally, the physical goods are delivered to consumers by express and other logistics methods, and the consumers confirm the receipt and complete the final transaction (Wu et al., 2017). Among the three flows of

DOI: 10.4018/JOEUC.322553 *Corresponding Author

capital flow, information flow and logistics, logistics has the most significant time delay. Under the Business to Consumer (B2C) e-commerce mode, the achievement rate of e-commerce to consumer demand directly determines whether it can gain an advantage in the competition. The interaction and transfer of information flow and capital flow can be completed instantly, but the delivery of physical goods must beachieved through logistics. Therefore, under the B2C e-commerce mode, logistics has become a vital link affecting the competitiveness of e-commerce enterprises. B2C e-commerce platform needs to purchase, store, conduct warehouse management, packaging, circulation, processing and distribution of goods, and finally deliver the goods to consumers to complete the completion of a transaction (Zhang, 2020; Abdulkareem et al., 2021). While the business scale of B2C e-commerce is expanding rapidly, it is not impossible to set up a general warehouse to ensure high logistics service quality, and all requirements are delivered from the general warehouse. It is necessary to set up sub-center warehouses in the areas involved in the business to ensure high logistics efficiency and service quality (Biagi and Falk, 2017; Jannach et al., 2017; Vinodhini and Chandrasekaran, 2017).

The e-commerce platform based on B2C mode can conduct direct transactions with consumers, greatly reducing the level of intermediate retailers and effectively reducing the "bullwhip effect", but this does not mean that the change information of consumer demand can be easily obtained. In particular, in the modern era of the short life cycle of consumer goods and rapid market change, consumers' demand is affected by multiple factors. It fluctuates wildly, bringing certain pressure and challenges to the inventory management of e-commerce (Xiong et al., 2020). Li and Huang (2019) pointed out that data mining technology refers to using computers as tools. The statistical learning model is constructed and applied to "new" data for prediction and analysis based on data. Its essence is consistent with statistical and machine learning (Li and Huang, 2019). However, e-commerce has the advantage of big data, which can mine consumers' consumption records, transaction browsing behavior and other data. Shahrel et al. (2021) explored the application of the time series model in e-commerce platforms, and the final error is less than 20%. However, there is an apparent deficiency in its research, that is, the time series model only analyzes the historical demand data of consumers, while the multi-dimensional data characteristics of consumers have not been fully utilized. Besides, the predicted value of this research method also lags in time (Shahrel et al., 2021; Tabassum et al., 2022; Savoli & Bhatt, 2022; Ngassam et al., 2022).

Disciplines such as big data, machine learning and data mining emphasize practicality, engineering and application. Here, the existing models and strategies on inventory optimization, demand forecasting and replenishment strategies based on data mining and other different schemes are analyzed based on the actual operation of big data of a B2C e-commerce platform. The real data are used to test the effect of varying demand forecasting schemes on inventory optimization. Besides, the research innovation is to apply the data mining method to the demand prediction of B2C e-commerce. It makes a detailed analysis from the basic model training to the model training based on data classification. The constructed fusion model training based on data classification achieves high prediction accuracy. Numerical analysis shows that the effect of the fusion model based on data classification is significantly better than the prediction ability of a single model, and proves the feasibility of the fusion model scheme in B2C e-commerce demand prediction.

PRINCIPLE OF CHARACTERISTIC ENGINEERING AND DEMAND FORECASTING MODEL

This section analyzes the demand forecasting model and the final feature engineering scheme used in detail and completes the feature engineering work. This part of the work is among the most critical links in big data-based demand forecasting. The reason is that the independent variables and data quality calculated by feature engineering directly determine the upper limit of the model's prediction performance. Although feature engineering belongs to the engineering link in the whole data mining process, it is an essential step in the entire data mining process and a hot issue in academic circles.

Analysis of Feature Engineering

Feature engineering is based on the exploration, analysis and understanding of datasets to process the original datasets and convert them into data that can be input into models and trained. Therefore, it is necessary to deeply understand the original data set and master specific data mining knowledge and technology (Oakden-Raayner et al., 2017; Oh et al., 2017). In developing and verifying the demand forecasting model, feature engineering is the most extended link. Still, it is also the most critical link to determine the upper bound of the generalization ability of the demand forecasting model.

According to the characteristics of the data used, the feature engineering scheme is determined through multiple iterative design, test and verification (Jiang et al., 2020). Time and identification (ID) of stock keeping unit (SKU) are taken as data granularity and combined with sliding time window to process the original dataset and finally derive the dataset for the prediction model. Using the ID of SKU to distinguish the sample data can carry out model training on finer data granularity, and determine the commodity ID and its related feature data through the mapping relationship between commodity ID and SKU ID, which is conducive to constructing coarse-grained feature data and facilitating secondary processing. Figure 1 displays the specific feature logic diagram:

The feature data used are categorized into class A and B features according to the feature engineering logic proposed above. The feature data of class A features are already available in the original dataset. They do not need subsequent processing, while the feature data of class B features are obtained after extraction and corresponding calculation from the original dataset based on business logic relationships (Bi and Wang, 2019; Howe, 2018; Viegas et al., 2017). Figure 2 presents an explanation of the meaning of class A and class B features:

Principle of the Basic Demand Prediction Model

For B2C e-commerce, there are many kinds of goods on sale, and there are significant differences in the purpose, category and price of different goods. The influencing factors of different commodity demands are various. Even the same influencing factors will affect the demand of different commodities differently. Therefore, using the same model to predict the demand of all commodities is not in line with the actual operation of B2C e-commerce. The following four basic models are adopted, and their basic principles will be briefly introduced below.

(1) Linear regression model

Figure 1.
Schematic diagram of feature engineering logic

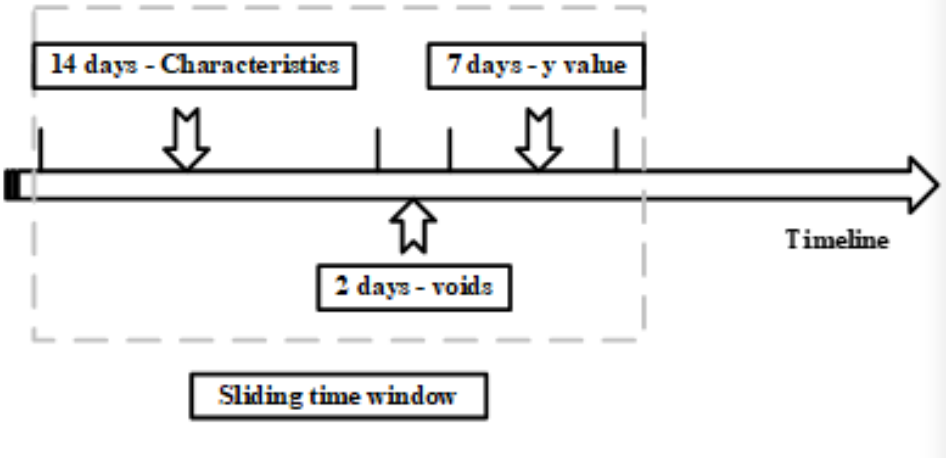
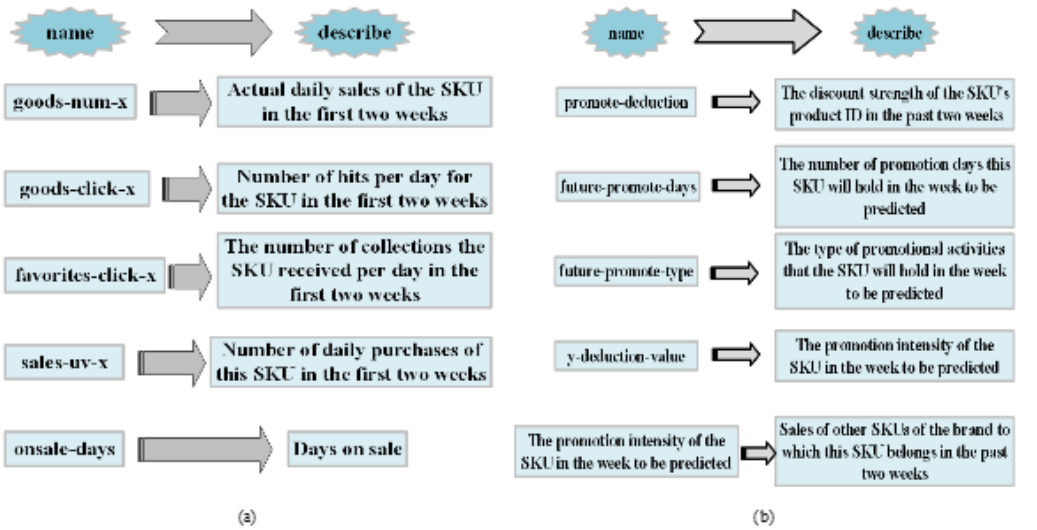


Figure 2.
Schematic diagram of characteristic indicators (a: Class A features; b: Class B features)



The linear regression model is a widely used data mining model that belongs to the supervised regression prediction model (Kayanan and Wijekoon, 2017; Iwasaki and Tsubaki, 2020). The training process is as follows:

n samples have been given, which are described by m attributes of $x = \left(x_1; x_2; x_3; \ldots; x_m\right)$. The attributes of n samples are linearly combined by using the linear regression model to obtain a function that can effectively predict, as shown in the following equations (1) and (2):

$$f\left(x\right) = w^T x + b \tag{1}$$

$$w = \left(w_1; w_2; w_3; \ldots; w_m\right) \tag{2}$$

In (1), b is an offset term.

The values of parameters w and b are obtained by minimizing mean square error (MSE). Finally, the linear model is solved, and its evaluation indexes are MAE and MSE. The following equation (3) is the solution method of MSE.

$$MSE = \frac{1}{n}\left\|Xw - y\right\|_2^2 \tag{3}$$

The advantages of using a linear regression model for analysis are a short time for solution, simple form, and relatively easy to establish model. However, it can only fit the linear relationship among variables but not the nonlinear one (Zhu et al., 2021; Kar et al., 2019; Chung et al., 2022).

(2) Decision tree model

Decision trees can be classified and grouped according to different decision problems. In the construction of classification decision tree, the cutting features are mainly selected according to the

indexes in information entropy theory such as information gain and gain rate; during the construction of regression decision tree, the characteristic variables to be segmented based on MSE minimization and corresponding segmentation points are mainly synthesized for determination (Wang et al., 2018; Niu and Chen, 2019). The regression decision tree is adopted here, and its algorithm process is as follows:

First, the training dataset D is input, the characteristic variable j to be segmented and the corresponding segmentation point s are determined, and the following equation (4) is solved:

$$\min_{j,s}\left[\min_{C_1}\sum_{x_i\in R_1(j,s)}\left(y_i-c_1\right)^2+\min_{C_2}\sum_{x_i\in R_2(j,s)}\left(y_i-c_2\right)^2\right] \tag{4}$$

The tuple $\left(j,s\right)$ is determined so that the value obtained by equation (4) above is the minimum of all possible values. The region is divided according to the determined tuple $\left(j,s\right)$ and the output value of each region is calculated. The following equations (5) - (7) are the specific process:

$$R_1\left(j,s\right)=\left\{x\middle|x^{(j)}\le s\right\} \tag{5}$$

$$R_2\left(j,s\right)=\left\{x\middle|x^{(j)}> s\right\} \tag{6}$$

$$\hat{c}_m=\frac{1}{N_m}\sum_{x_i\in R_m(j,s)}y_i, x\in R_m, m=1,2 \tag{7}$$

The two divided sub-areas must continue the above operations until they reach the preset conditions such as tree depth and leaf node number (Pourabdollahi et al., 2018; Maharlouei et al., 2019; Uma, 2020).

The input feature space is categorized into M regions such as $R_1,R_2,\ldots,R_M$, and finally a regression tree $f\left(x\right)$ is generated, as shown in equation (8):

$$f\left(x\right)=\sum_{m=1}^{M}\hat{c}_m I\left(x\in R_m\right) \tag{8}$$

(3) Random forest model

The parallel training method is adopted when the random forest model trains multiple independent decision trees. When the model is adopted to predict the data, the methods used for different decision problems are also different. When the final category for classification problems is determined, the method used is “majority voting”; for the regression problem, the prediction results of all regression trees should be calculated and averaged (Zhan et al., 2018; Brokamp et al., 2018) to output the final model. The algorithm process is as follows:

First, the training set shown in the following equation (9) is input:

$$D=\left\{\left(x_1,y_1\right),\left(x_2,y_2\right),\ldots,\left(x_n,y_n\right)\right\} \tag{9}$$

Before training, the number of logarithms T, the number of randomly taken samples m, the characteristic quantity d used in random sampling and the maximum depth h of each tree should be determined in advance (Shi et al., 2018; Calle et al., 2021; Qiu et al., 2020).

According to the D, m, d, h and other hyperparameters determined in above, the MSE based on the base learner: regression decision tree $f\left(x\right)$ is trained; parallel training of T decision trees is conducted; the average value is calculated after the prediction results of T trees are calculated, and the final output is determined as equation (10):

$$H\left(x\right)=\frac{1}{T}\sum_{j}^{T}f_{j}\left(x\right) \tag{10}$$

(4) Neural network model

In model fitting, neurons, as a key part of the neural network model, can nonlinear map the relationship among data. Moreover, a plurality of neural network layers are set, and a plurality of network nodes are set in the neural network of each layer to enable interactive connection among data. The data input to each network node will pass through the nonlinear mapping of neurons before output. Currently, the commonly used neurons mainly include Sigmoid function, tanh function and ReLU function (Chen, 2019). Equations (11) - (13) are the functional expressions:

$$g\left(x\right)=\frac{1}{1+e^{-x}} \tag{11}$$

$$g\left(x\right)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{12}$$

$$g\left(x\right)=\begin{cases}x,x>0\\0,x\leq 0\end{cases} \tag{13}$$

The sigmoid function can map any numerical value in real number space to the interval (0,1) through nonlinear mapping. It is easy to solve the derivative function; tanh function can compress the logarithm into the interval (-1, +1); when the ReLU function is employed for calculation, if the input is positive, the problem of gradient disappearance does not need to be considered in the solution process. Thereby, the derivation is relatively simple and the calculation time is short. ReLU function is selected for demand prediction according to the prediction data and tasks (Larson et al., 2018; Liao et al., 2018; Pedersen et al., 2017).

DATA RESULT ANALYSIS

Data Profiling

This section mainly introduces the operation data of a B2C e-commerce platform used in subsequent data mining. The dataset includes the operation and sales of B2C e-commerce companies in the zone B market from June 2019 to June 2020. Figure 3 is an overview of the dataset.

(1) Information table of consumption behavior

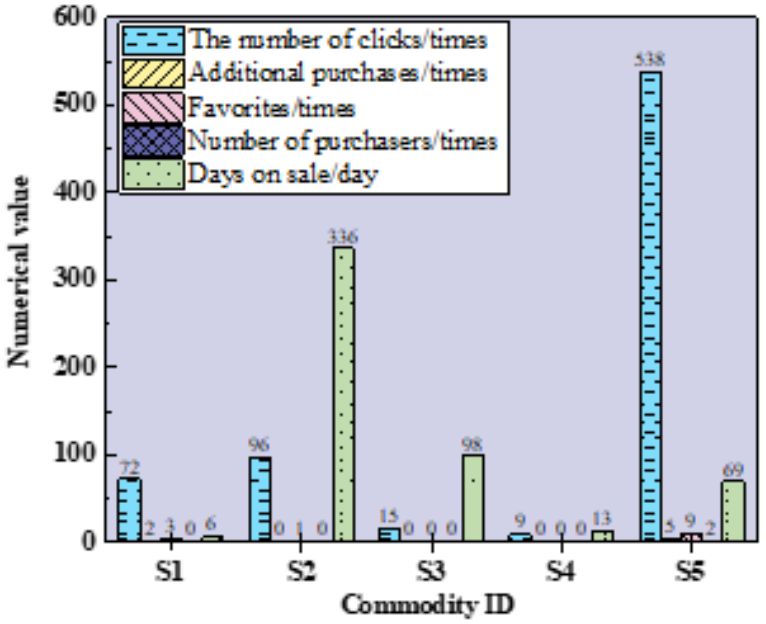
Figure 3.
An overview of dataset

1	Consumption behavior information table	It records consumer click, structure, purchase and other behaviors
2	Commodity category information table	Hierarchical division of commodity categories
3	Commodity sales data sheet	Daily sales data for each category
4	Product relationship mapping table	Mapping relationship between each type of product and subclass
5	Price list of commodity promotion	Promotion time, price and other information of each kind of goods
6	Product relationship mapping table	Time, type and rhythm of e-commerce platform promotion activities

This table mainly records the number of clicks, additional purchases, favorites, number of purchasers, days on sale in the statistical day and other information of consumers on each commodity ID. Figure 4 shows the number of additional purchases and number of purchasers of goods S1, S2, S3, S4 and S5 on August 23, 2019:

Figure 4 reveals that on August 23, 2019, the number of clicks, additional purchases and purchasers of commodity S5 are the largest, and the number of days on sale of commodity S2 is the longest. The dataset contains more than 35 million pieces of data, including the number of clicks, additional

Figure 4.
Information diagram of commodity consumption behavior



purchases, number of purchasers and days on sale of more than 400000 commodities from June 2019 to June 2020. These data can be used to predict the shopping tendency of consumers in the future.

(2) Price list of commodity promotion

This table mainly records the promotion time, list price and sale price of promoted goods. Figure 5 shows the list price, sale price and promotion time of goods H1, H2, H3, H4 and H5 in the statistical day.

Figure 5(b) displays that the promotion time of commodity H1 not shown in the figure. This is because commodity H1 has no promotion activities within the investigated time, so the sale price of commodity H1 in Figure 5(a) is 0. When the commodity sale price dataset is used, attention should be paid to this problem to avoid the improper use of the data. The dataset contains more than 1.04 million pieces of data, which can be adopted to predict the promotion strength and promotion time of goods.

(3) Commodity sales data table

This table mainly records the SKU sales volume, sales price, tag price and other information corresponding to the commodity. Figure 6 displays the sales volume, tag price and sales price of commodities F1, F2, F3, F4 and F5 on July 18, 2019.

Figure 6 presents that the difference between the sales price and the tag price of commodity H3 is the largest, and the sales volume is also the highest. The dataset contains more than 7.3 million pieces of data. The actual sales volume of each type of commodity is recorded every day. The historical sales volume of commodities can be analyzed according to these data and used as the source data of feature engineering.

(4) Product relationship mapping table

In the actual management and sales process of an e-commerce product platform, usually, a product ID will correspond to the ID of multiple SKUs (for example, the product ID of certain pants

Figure 5.
Commodity sale price information chart (a: commodity list price and sale price; b: commodity promotion time)

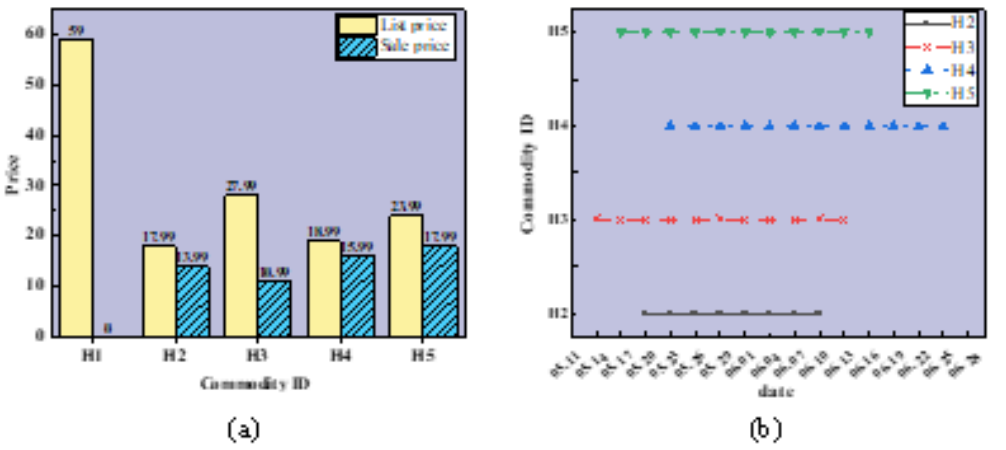


Figure 6.
Commodity sales data chart

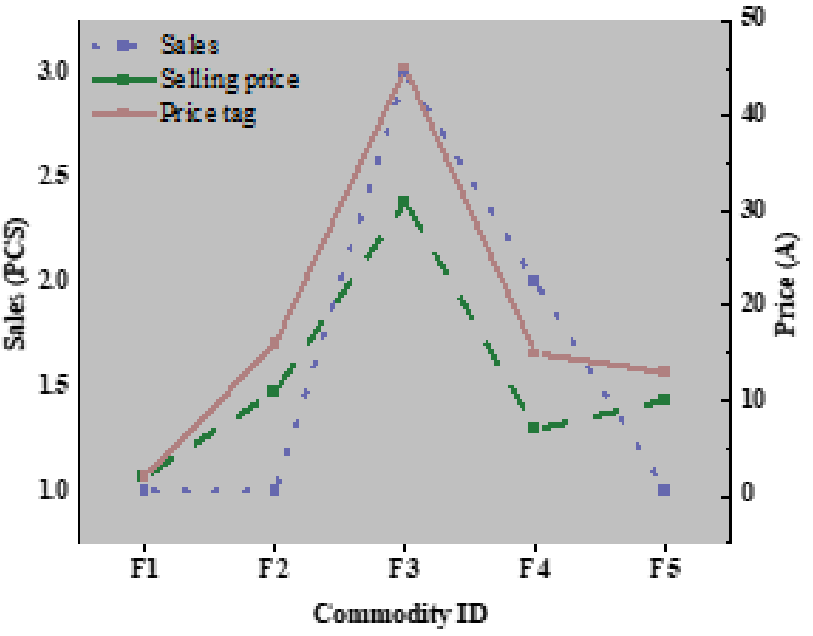
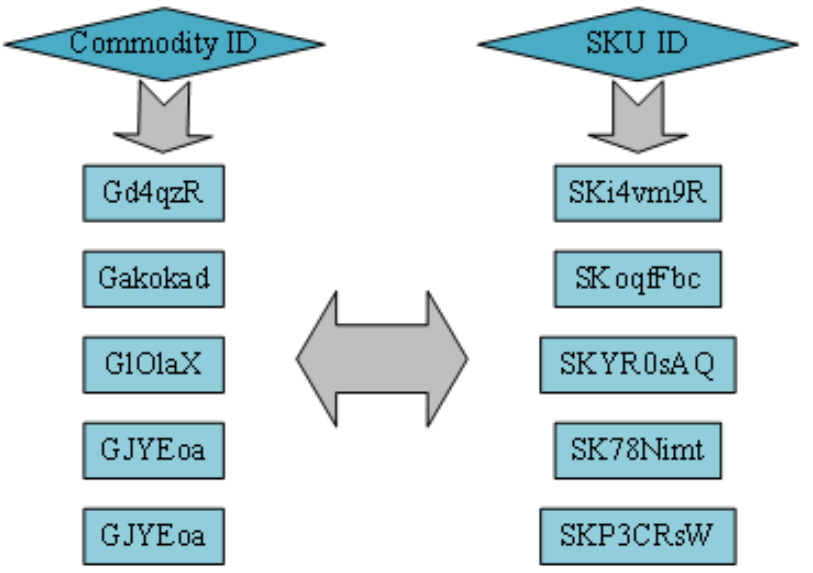


Figure 7.
Product relationship map



is Gpants-011. Because the pants have a combination of colors and sizes, multiple SKU IDs will identify multiple styles of the pants). In Figure 7, the commodity IDs of five commodities and their corresponding SKU IDs are described.

(5) Platform promotion activity table

This table mainly records the promotional activities of commodities on the B2C e-commerce platform at different times, including activity date, activity type and rhythm. The activity type belongs to the classification variable and is the number used to represent the activity type; activity rhythm is the expression of activity intensity, which has a size relationship. Figure 8 displays the type and rhythm of promotional activities for a commodity at different times in 2019.

Data Results of Feature Engineering

With "sku-name" + "data-data" as the joint primary key, days on sale, discount strength, consumer collection and purchase behavior information as the feature independent variable, and "y-value" as the dependent variable, Figure 9 displays the feature results of SKUs with the values of "SK8WNkCU" and "SKD3gYGy".

In Figure 9, the primary keys "data-date" corresponding to SKUs with the value of "SK8WNkCU" are 20190801 and 20190808, respectively, and the primary keys "data-data" corresponding to SKUs with the value of "SKD3gYGy" are 20200103, 20200110, 20200117, 20200124 and 20200131, respectively. The existing data dimensions are fully mined and utilized according to understanding the B2C e-commerce platform and analysing and exploring relevant data. The final feature dataset used contains 16000 SKU information of more than 6000 commodity IDs, a total of 680000 data records. This dataset is employed for the final model training. The dataset contains numerical variables and secondary variables such as activity intensity.

Numerical Analysis of Basic Model Training and Performance

680000 feature samples are randomly divided into 600000 train set samples and 80000 test set samples to evaluate the performance of the four basic models selected in 2.2. Figure 10 displays the performance indexes of the four basic models selected in section 2.2:

Next, the prediction performance of the four basic models mentioned in section 2.2 will be analyzed based on other data values such as model performance indexes and feature importance in Figure 10.

(1) Linear regression model

Figure 8.
Promotion activity data of a commodity platform (a: activity type; b: activity rhythm)

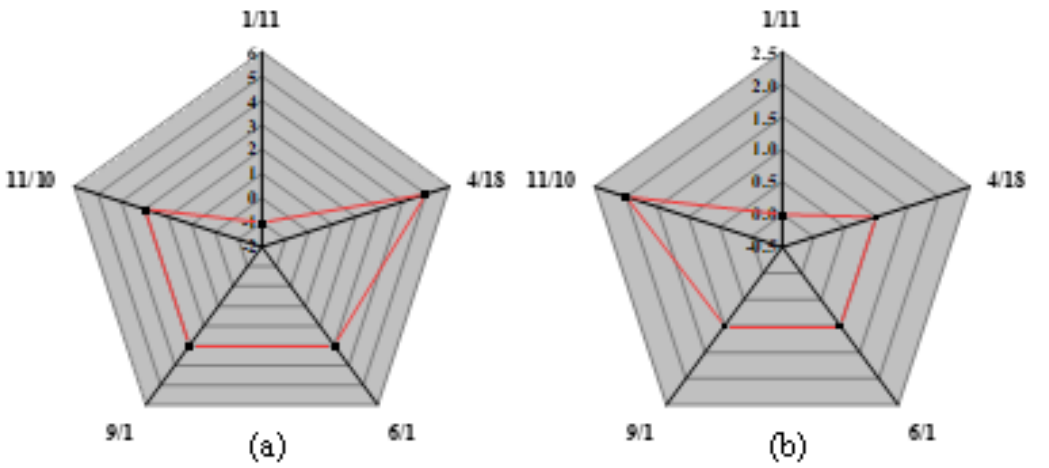
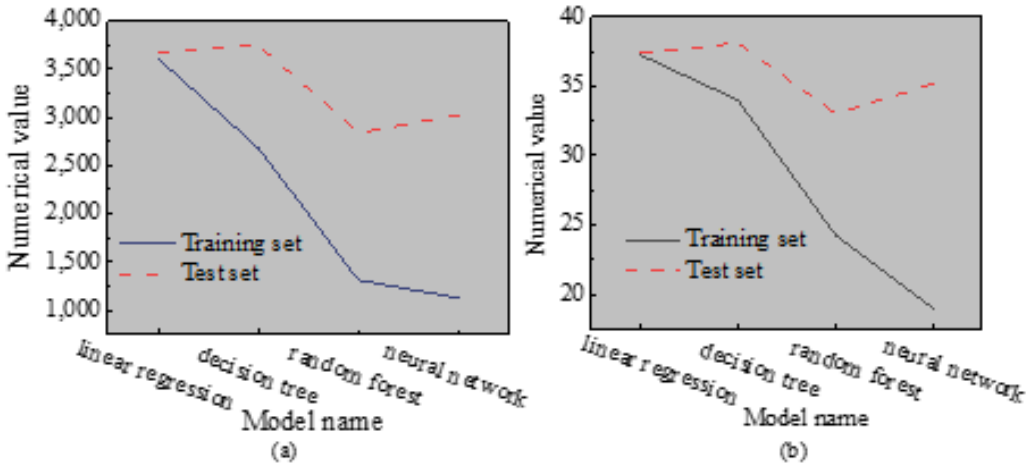


Figure 9.
SKU feature results (a: sales of other SKUs with the same commodity ID; b: information such as collection and purchase)

Sku_name	data_date	onsale_days	Promote_deduction	Count_promote_days	Previous_price_deduction	y_deduction_value	y_value
SK8WNkCU	20190801	134	0	0	0.43	0.52	110
SK8WNkCU	20190808	135	0	0	0.43	0.52	120
SKD3gYGy	20200103	464	0	0	0.48	0.47	240
SKD3gYGy	20200110	465	0	0	0.48	0.44	137
SKD3gYGy	20200117	466	0	0	0.47	0.45	135
SKD3gYGy	20200124	467	0.42	1	0.5	0.45	232
SKD3gYGy	20200131	468	0.42	8	0.48	0.46	131

Figure 10.
Performance index diagram of basic model based on full train set (a: MSE; b: MAE)



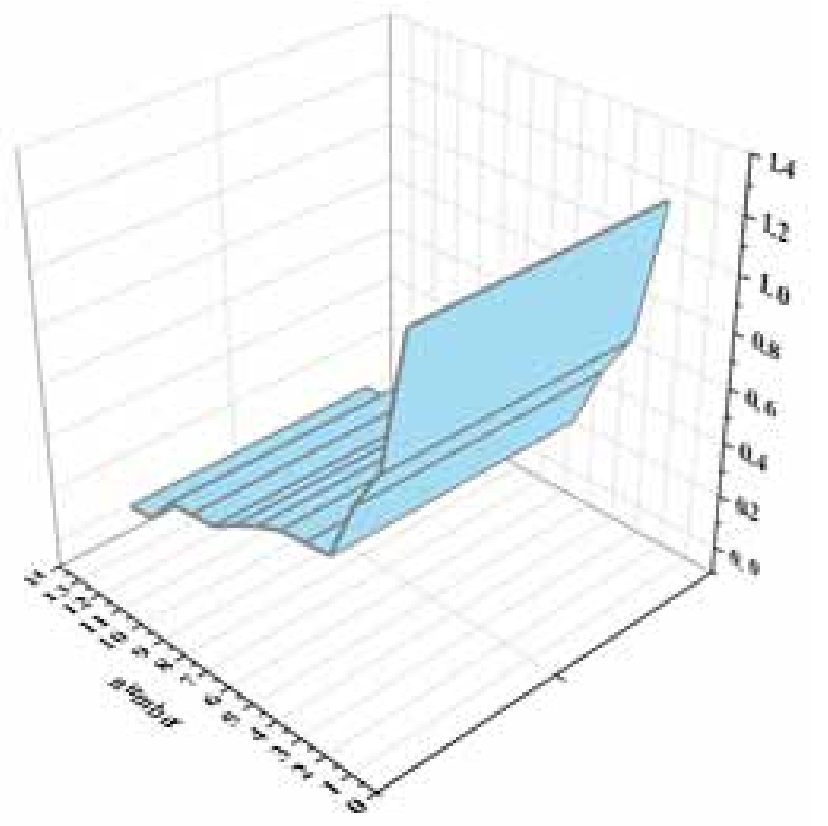
The indexes MSE and MAE of the linear regression model in Figure 10 show that the overfitting phenomenon is not serious for the linear regression model because the number of samples selected is relatively large and the correlation among various variables is not strong.

There are relatively multiple characteristic variables in the linear regression model. Statistics is primarily made on the more important 14 characteristic variables. Figure 11 displays the result.

The feature names corresponding to numbers 1-14 in Figure 11 are the historical discount strength of other SKUs with the same commodity ID, the sales volume on days 14, 13, 12, 10, 11 and 9, the collection times on day 14, the sales volume on days 8, 5, 6 and 7, the collection times on day 13

Figure 11.

Statistical chart of important characteristics of linear regression model



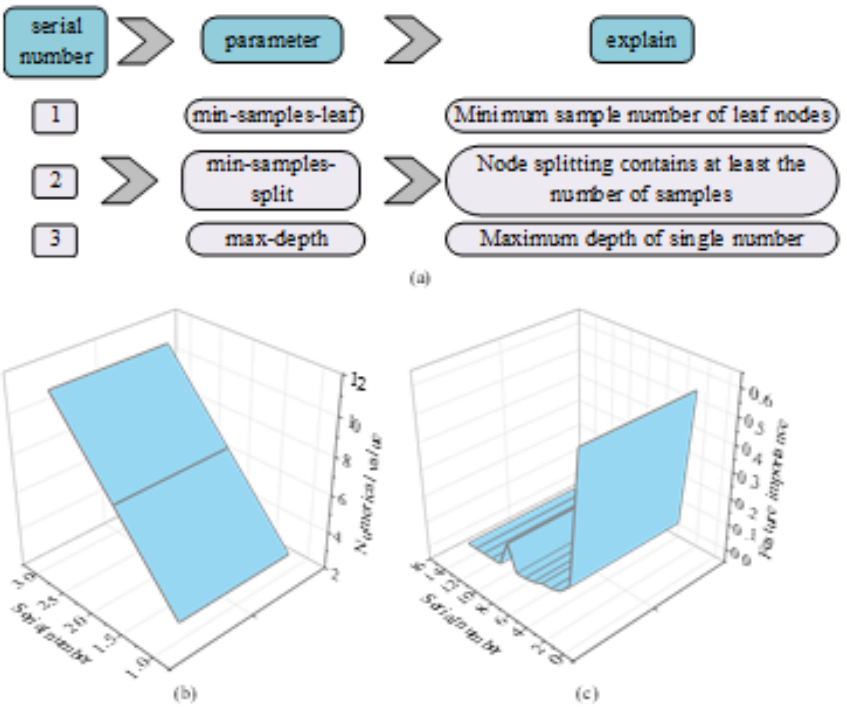
and the number of customers on day 14. According to the pre-planned promotion activities of the e-commerce platform, the "discount strength of this SKU in the next week" and "historical discount strength of other SKUs with the same commodity ID" are calculated through feature engineering. The results show that there is a strong positive effect among variables. Moreover, the historical sales volume of goods will exert a certain impact on the future sales volume; the closer the time between the historical sales volume and the future sales volume is, the greater the impact is. The relationship between the 14 variables shown in the figure and the future sales of goods is a positive correlation. The model's parameter values are consistent with the actual role of the marketing activities designed by the e-commerce platform, which assists in verifying the correctness of the model.

(2) Decision tree model

Figure 10 suggests that the training error of the decision tree model is smaller than that of the linear regression model, indicating that the fitting effect of the decision tree model is better when the nonlinear fitting is carried out for this problem. However, the test error of the decision tree model is larger than the training error, indicating that the model may have an overfitting phenomenon.

The hyperparameters of the decision tree model are set and trained to obtain 14 characteristic data. These 14 characteristic data effectively reduce MSE, as shown in Figure 12.

Figure 12.
Decision tree model (a: hyperparameter setting diagram; b: hyperparameter setting value; c: feature importance diagram)



The feature names corresponding to numbers 1-14 in Figure 12(c) are respectively the sales volume on the 14th and 13th day of history, the discount strength of promotion activities in the next week, the discount strength in the first two weeks of history, the sales volume on the 10th day of history, the purchases on the 14th day of history, the sales volume on the 11th and 12th days of history, the purchases on the 8th day of history, the sales volume on the 1st day of history, the days on sale, the additional purchases on the 9th day of history, the number of clicks on the 1st day of history and purchases on the 10th day of history.

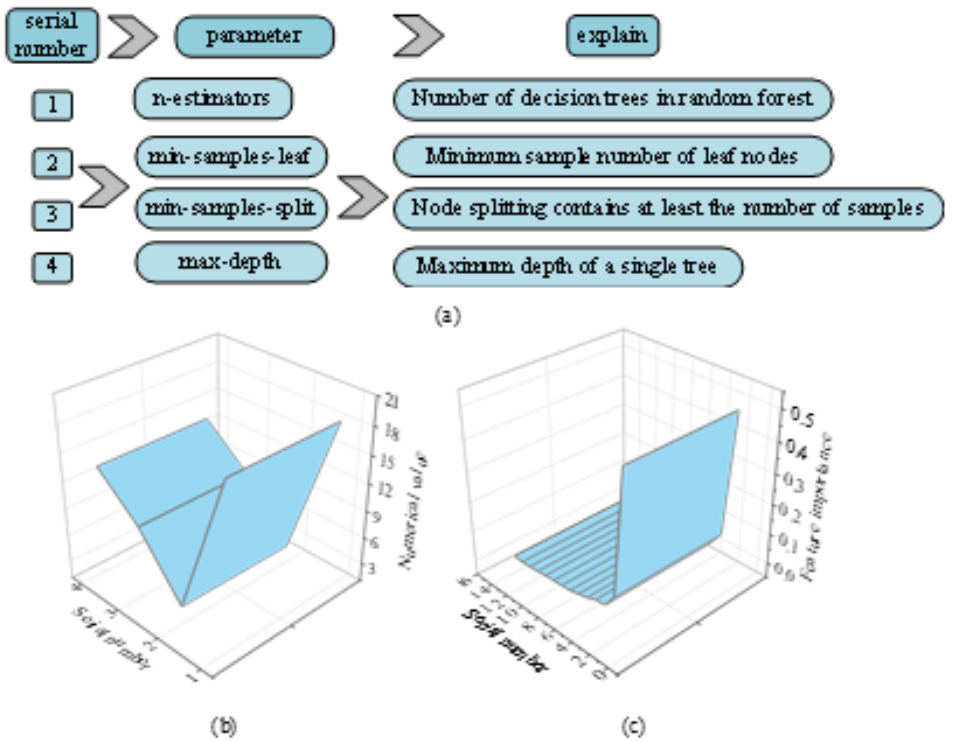
The feature importance in the figure is calculated based on the feature's contribution in minimizing MSE. The comparison between the linear regression model and the decision tree model suggests that a large part of the important features in the model are consistent, which further explains the role of these important features in the prediction process.

(3) Random forest model

Figure 10 reveals that the MSE and MAE indexes of the random forest model are significantly better than the first two models. The hyperparameters of the decision tree model are set and trained to obtain 14 characteristic data, which can significantly improve the prediction effect of the random forest model, as shown in Figure 13.

The feature names corresponding to numbers 1-14 in Figure 13(c) are respectively the sales volume on the 14th and 13th day of history, the discount intensity of promotion activities in the next week, the discount intensity of the first two weeks of history, the sales volume on the 10th, 11th and 12th day of history, the purchases on the 13th and 14th day of history, the sales volume on the 9th day of history, the days on sale, the purchases on the 8th day of history, the sales volume on the 8th

Figure 13.
Random forest model (a: hyperparameter setting diagram; b: hyperparameter setting value; c: feature importance diagram)



day of history, and the additional purchases on the 14th day of history. The calculation method of feature importance in the figure averages the T decision trees in the model. The comparison between the decision tree model with the random forest model reveals that the importance characteristics in the model are only different in ranking.

(4) Neural network model

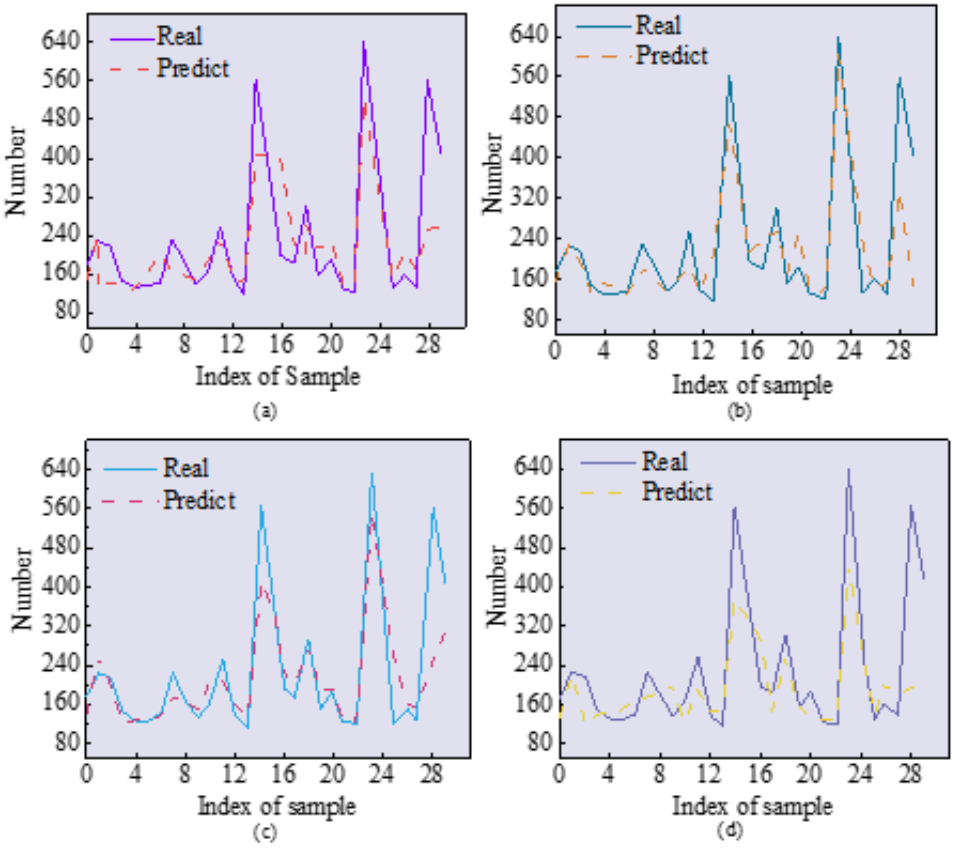
First, the preset hyperparameter range of the number of hidden layers and nodes in each hidden layer is searched. Then, cross-validation is performed on the train set. Finally, the number of hidden layers and nodes in each hidden layer in the optimal hyperparameters combination are 9 and 32, respectively.

(5) Visual analysis of prediction effect of the basic data model

The change data of the actual sales volume of SKU with ID "SKDeZhdu" in 30 weeks are obtained by random sampling in the dataset. The above four models are used for prediction, and the prediction effect is visually analyzed. Figure 14 displays the analysis results.

Figure 14 displays that the four basic models selected have good prediction effects and no obvious lag phenomenon exists. The relative error caused by the prediction model is small when the actual sales volume is small. When the actual sales volume of goods is high, the prediction effect of the decision tree model is the best compared with the other three models. Overall, the difference

Figure 14.
Effect diagram of basic model prediction (a: linear regression model; b: decision tree model; c: random forest model; d: neural network model)



between the actual value and the predicted value of the random forest model is relatively small, no matter when the sales volume is low or high.

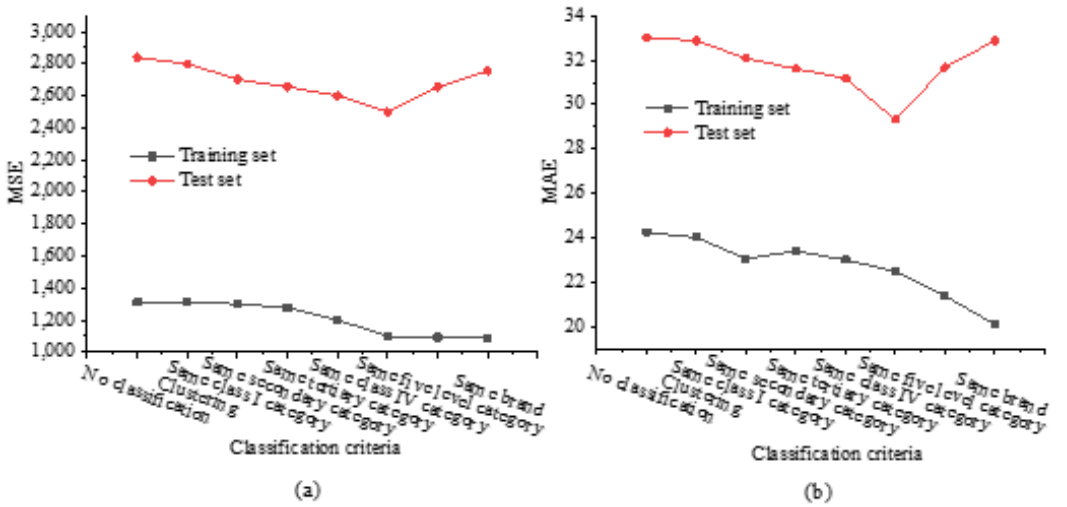
Performance Analysis of Demand Forecasting Model Based on Data Classification

The random forest model with the best effect in the section "Numerical Analysis of Basic Model Training and Performance" is adopted as the Model for algorithm planning training. Meanwhile, the performance improvement range parameter $\alpha = 10$ is set. Figure 15 is a statistical chart of the performance of a random forest model based on data classification:

Figure 15 displays that the effect of the model training algorithm based on data classification is better than the prediction ability of the model trained with complete data on the test set. Meanwhile, the following laws are found. When the total amount of data is determined, the overall MSE and MAE of the training set show a gradual downward trend with the increasing categories of data. However, the overall MSE and MAE of the test set show a downward trend and an upward trend. It shows that the data are divided into too many subsets, and the model trained based on a single data subset tends to overfit, which leads to the gradual decline of the prediction effect of the model on the test set. Under the random forest algorithm, the optimal classification standard of sample data should be "belonging

Figure 15.

Model performance analysis chart based on data classification



to the same secondary category", which can divide the whole dataset into 201 sub-data. Under this standard, the random forest performance achieves the best prediction accuracy.

CONCLUSION

Under the background of the vigorous rise of the scientific and technological revolution, e-commerce makes full use of the Internet's rapid, fast, safe and reliable characteristics to realize the flow of funds and information with consumers. The inventory management level of the e-commerce platform will directly impact enterprises' profitability and customer satisfaction. Thereby, to improve the inventory management efficiency of the e-commerce platform, the prediction of future demand, the core of B2C e-commerce inventory management optimization, is deeply discussed and analyzed. First, the dataset overview and the final feature engineering scheme are analyzed in detail. After calculation, the characteristic independent variables and dependent variables are obtained, which directly impacts the prediction performance of the model and is an essential link in the data mining process. Then, the basic models such as linear regression, decision tree, random forest, and neural network models are trained and analyzed based on the data samples obtained by feature engineering. With the change data of the actual sales volume of SKU with ID "SKDeZhdu" in 30 weeks as an example, the above four models are used for prediction. The results show no obvious lag in the prediction effect of the four basic models. The relative error caused by the prediction model is small when the actual sales volume is small. Moreover, the prediction effect of the random forest model is the best when a single basic model is trained based on full data.

REFERENCES

Abdulkareem, K. H., Mohammed, M. A., Salim, A., Arif, M., Geman, O., Gupta, D., & Khanna, A. (2021). Realizing an effective COVID-19 diagnosis system based on machine learning and IOT in smart hospital environment. *IEEE Internet of Things Journal*, *8*(21), 15919–15928. doi:10.1109/JIOT.2021.3050775 PMID:35782183

Bi, X., & Wang, H. (2019). An enhanced high-order Boltzmann machine for feature engineering. *Engineering Applications of Artificial Intelligence*, *78*, 37–52. doi:10.1016/j.engappai.2018.10.011

Biagi, F., & Falk, M. (2017). The impact of ICT and e-commerce on employment in Europe. *Journal of Policy Modeling*, *39*(1), 1–18. doi:10.1016/j.jpolmod.2016.12.004

Brokamp, C., Jandarov, R., Hossain, M., & Ryan, P. (2018). Predicting daily urban fine particulate matter concentrations using a random forest model. *Environmental Science & Technology*, *52*(7), 4173–4179. doi:10.1021/acs.est.7b05381 PMID:29537833

Calle, J. L. P., Ferreiro-González, M., Ruiz-Rodríguez, A., Barbero, G. F., Álvarez, J. Á., Palma, M., & Ayuso, J. (2021). A methodology based on FT-IR data combined with random forest model to generate spectralprints for the characterization of high-quality vinegars. *Foods*, *10*(6), 1411. doi:10.3390/foods10061411 PMID:34207095

Chen, E. Y. (2019). A neural network model of cortical information processing in schizophrenia II-role of hippocampal-cortical interaction: A review and a model. *Canadian Journal of Psychiatry*, *40*(1), 21–26. doi:10.1177/070674379504000107 PMID:7874671

Chung, J., Lee, Y., Kim, J., Jung, C., & Kim, S. (2022). Soil moisture content estimation based on sentinel-1 sar imagery using an artificial neural network and hydrological components. *Remote Sensing (Basel)*, *14*(3), 465. doi:10.3390/rs14030465

Howe, D. G. (2018). A statistical approach to identify, monitor, and manage incomplete curated data sets. *BMC Bioinformatics*, *19*(1), 1–12. doi:10.1186/s12859-018-2121-6 PMID:29609549

Iwasaki, M., & Tsubaki, H. (2020). A bivariate generalized linear model with an application to meteorological data analysis. *Statistical Methodology*, *2*(3), 175–190. doi:10.1016/j.stamet.2005.03.002

Jannach, D., Ludewig, M., & Lerche, L. (2017). Session-based item recommendation in e-commerce: On short-term intents, reminders, trends and discounts. *User Modeling and User-Adapted Interaction*, *27*(3-5), 351–392. doi:10.1007/s11257-017-9194-1

Jiang, L., Zhao, Y., Golsanami, N., Chen, L., & Yan, W. (2020). A novel type of neural networks for feature engineering of geological data: Case studies of coal and gas hydrate-bearing sediments. *Geoscience Frontiers*, *11*(5), 1511–1531. doi:10.1016/j.gsf.2020.04.016

Kar, N. B., Das, S., Ghosh, A., & Banerjee, D. (2019). Fuzzy linear regression model on mulberry silk cocoon characteristics. *Research Journal of Textile and Apparel.*, *23*(3), 201–211. doi:10.1108/RJTA-03-2019-0012

Kayanan, M., & Wijekoon, P. (2017). Performance of Existing Biased Estimators and the respective Predictors in a Misspecified Linear Regression Model. *Open Journal of Statistics*, *7*(05), 876–900. doi:10.4236/ojs.2017.75062

Larson, D. B., Chen, M. C., Lungren, M. P., Halabi, S. S., Stence, N. V., & Langlotz, C. P. (2018). Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*, *287*(1), 313–322. doi:10.1148/radiol.2017170236 PMID:29095675

Li, C., & Huang, Z. (2019). Subsidy strategy of pharmaceutical e-commerce platform based on two-sided market theory. *PLoS One*, *14*(10), e0224369. doi:10.1371/journal.pone.0224369 PMID:31671133

Liao, Y., Kodagoda, S., Wang, Y., Shi, L., & Liu, Y. (2018). Place classification with a graph regularized deep neural network. *IEEE Transactions on Cognitive and Developmental Systems*, *9*(4), 304–315. doi:10.1109/TCDS.2016.2586183

Maharlouei, N., Jafarzadeh, F., & Lankarani, K. B. (2019). Factors affecting recovery during the first 6 months after hip fracture, using the decision tree model. *Archives of Osteoporosis*, *14*(1), 1–6. doi:10.1007/s11657-019-0611-4 PMID:31183563

Ngassam, R. G. N., Ung, L., Ologeanu-Taddei, R., Lartigau, J., Demoly, P., Bourdon, I., & Chiriac, A. M. (2022). An Action Design Research to Facilitate the Adoption of Personal Health Records: The Case of Digital Allergy Cards. *Journal of Organizational and End User Computing*, *34*(4), 1–18. doi:10.4018/JOEUC.288551

Niu, F., & Chen, L. (2019). Forecasting of landslide stability based on gradient boosting decision tree model. *International Core Journal of Engineering*, *5*(11), 42–48.

Oakden-Rayner, L., Carneiro, G., Bessen, T., Nascimento, J. C., Bradley, A. P., & Palmer, L. J. (2017). Precision radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework. *Scientific Reports*, *7*(1), 1648. doi:10.1038/s41598-017-01931-w PMID:28490744

Oh, H., Jung, J. H., Jeon, B. C., & Youn, B. D. (2017). Scalable and unsupervised feature engineering using vibration-imaging and deep learning for rotor system diagnosis. *IEEE Transactions on Industrial Electronics*, *65*(4), 3539–3549. doi:10.1109/TIE.2017.2752151

Pedersen, T. S., Nielsen, K. M., Hindsborg, J., Reichwald, P., Vinther, K., & Izadi-Zamanabadi, R. (2017). Predictive functional control of superheat in a refrigeration system using a neural network model. *IFAC-PapersOnLine*, *50*(1), 43–48. doi:10.1016/j.ifacol.2017.08.008

Pourabdollahi, Z., Karimi, B., Mohammadian, A. K., & Kawamura, K. (2018). Shipping chain choices in long-distance supply chains: Descriptive analysis and decision tree model. *Transportation Research Record: Journal of the Transportation Research Board*, *2410*(1), 58–66. doi:10.3141/2410-07

Qiu, G., Bao, Y., Yang, X., Wang, C., Ye, T., Stein, A., & Jia, P. (2020). Local population mapping using a random forest model based on remote and social sensing data: A case study in Zhengzhou, China. *Remote Sensing (Basel)*, *12*(10), 1618. doi:10.3390/rs12101618

Savoli, A., & Bhatt, M. (2022). Chronic Patients' Emotions Toward Self-Managing Care IT: The Role of Health Centrality and Dependence on IT. *Journal of Organizational and End User Computing*, *34*(4), 1–14. doi:10.4018/JOEUC.288550

Shahrel, M. Z., Mutalib, S., & Abdul-Rahman, S. (2021). PriceCop-Price Monitor and Prediction Using Linear Regression and LSVM-ABC Methods for E-commerce Platform. *International Journal of Information Engineering & Electronic Business*, *13*(1), 1–14. doi:10.5815/ijieeb.2021.01.01

Shi, K., Qiao, Y., Zhao, W., Wang, Q., Liu, M., & Lu, Z. (2018). An improved random forest model of short-term wind-power forecasting to enhance accuracy, efficiency, and robustness. *Wind Energy (Chichester, England)*, *21*(12), 1383–1394. doi:10.1002/we.2261

Tabassum, K., Shaiba, H., Essa, N. A., & Elbadie, H. A. (2022). An Efficient Emergency Patient Monitoring Based on Mobile Ad Hoc Networks. *Journal of Organizational and End User Computing*, *34*(4), 1–12. doi:10.4018/JOEUC.289435

Uma, K. V. (2020). C5. 0 Decision Tree Model Using Tsallis Entropy and Association Function for General and Medical Dataset. *Intelligent Automation & Soft Computing, 26*(1).

Viegas, R., Salgado, C. M., Curto, S., Carvalho, J. P., Vieira, S. M., & Finkelstein, S. N. (2017). Daily prediction of ICU readmissions using feature engineering and ensemble fuzzy modeling. *Expert Systems with Applications*, *79*, 244–253. doi:10.1016/j.eswa.2017.02.036

Vinodhini, G., & Chandrasekaran, R. M. (2017). A sampling based sentiment mining approach for e-commerce applications. *Information Processing & Management*, *53*(1), 223–236. doi:10.1016/j.ipm.2016.08.003

Wang, J., Li, P., Ran, R., Che, Y., & Zhou, Y. (2018). A short-term photovoltaic power prediction model based on the gradient boost decision tree. *Applied Sciences (Basel, Switzerland)*, *8*(5), 689. doi:10.3390/app8050689

Wu, S. J., Chiang, R. D., Chang, S. H., & Chang, W. T. (2017). An interactive telecare system enhanced with IoT technology. *IEEE Pervasive Computing*, *16*(3), 62–69. doi:10.1109/MPRV.2017.2940967

Xiong, X., Yuan, F., Huang, M., Cao, M., & Xiong, X. (2020). Comparative evaluation of web page and label presentation for imported seafood products sold on Chinese e-commerce platform and molecular identification using DNA barcoding. *Journal of Food Protection*, *83*(2), 256–265. doi:10.4315/0362-028X.JFP-19-309 PMID:31961225

Zhan, Y., Luo, Y., Deng, X., Zhang, K., Zhang, M., Grieneisen, M. L., & Di, B. (2018). Satellite-based estimates of daily NO2 exposure in China using hybrid random forest and spatiotemporal kriging model. *Environmental Science & Technology*, *52*(7), 4180–4189. doi:10.1021/acs.est.7b05669 PMID:29544242

Zhang, X. (2020). Application of NB-IoT technology in urban lighting system. *International Core Journal of Engineering*, *6*(4), 246–251.

Zhu, G., Zhang, S., Bian, Y., & Hursthouse, A. S. (2021). Multi-linear regression model for chlorine consumption by waters. *Environmental Engineering Research*, *26*(4), 200402. doi:10.4491/eer.2020.402