Online sales prediction approach using methodology of CRISP-DM

Chunyang Wang[1], *Alena* Stupina[1,2*], and *Sergey* Bezhitskiy[2]

[1]Siberian Federal University, Vuzovsky Lanc,3, Krasnoyarsk, 660025, Russia
[2]Reshetnev Siberian State University of Science and Technology, System Analysis and Operation Research, 31, Krasnoyarsky Rabochy Av., Krasnoyarsk. Krasnoyarsk, 660037, Russia

Abstract. This article studies the sales forecasting problem in the field of e-commerce. Based on the CRISP-DM methodology, innovative data mining technology is used to construct a variety of forecasting models, and is compared and optimized. This article improves the quality and quantity of sales forecasts and provides enterprises with more accurate and effective decision support. In terms of modeling optimization in this article, data mining models such as random forest, support vector machine, and neural network are used for comprehensive prediction, and comparative analysis is conducted with the classic multiple linear regression model. Through model evaluation and optimization, this paper achieved better prediction performance and accuracy. This research has certain theoretical significance and practical value, and provides new ideas and methods for the marketing decisions and business development of e-commerce enterprises.

1 Introduction

In the field of e-commerce, China's major e-commerce platforms have recently shown a trend of slowing revenue growth, which indicates that the period of rapid growth in traffic and Internet dividends has passed [1]. However, with the advancement of information technology, the quality and quantity of data generated by e-commerce platforms have improved significantly. In the sales forecast of e-commerce platforms, the ability to quickly obtain feedback and make decisions is particularly important[2]. Accurate sales forecast is not only an important part of supply chain management, but the forecast results will have a profound impact on the overall production organization of the enterprise. At present, the industry lacks a systematic and scientific sales forecasting method, and the forecast errors are large[3,4]. To this end, we are committed to achieving higher quality and higher quantity sales forecast goals based on data and algorithms.

This article will use the CRISP-DM methodology as a framework to guide the implementation of prediction tasks. We take a closer look at how to use data mining tools. The methodology breaks down data mining projects into phases such as business understanding, data understanding, data preparation, modeling, evaluation, and deployment. At the same time, we use behavioral analysis methods to interpret the mining results in detail

* Corresponding author: h677hm@gmail.com

to provide companies with strong marketing support. In the modeling process, we will adopt the causal prediction method and construct three data mining models: random forest, support vector machine and neural network for comprehensive prediction. In addition, this article will also compare with the classic multiple linear regression model to verify the effectiveness of the selected model. Through this series of methods, we hope to provide enterprises with more accurate and effective sales forecast models to promote their marketing decisions and business development.

2 Predictive modeling theory based on data mining

2.1 Predictive target analysis

CRISP-DM first focuses on understanding business needs and clarifying mining goals. It is necessary to determine the measurement scale and influencing factors of the prediction object, and consider environmental factors. Before implementation, the problem boundaries, goals, constraints, etc. must be clarified, and data sources and quality requirements must be determined to lay the foundation for subsequent mining[5]. In short, it is necessary to understand the requirements from a business perspective, clarify the goals and constraints, and ensure the availability and quality of data to achieve successful data mining.

2.2 Predictive data analysis

In the second step of CRISP-DM, we focus on analyzing the data in depth to enhance our understanding of the data. Data understanding aims to discover the knowledge inherent in data through preliminary and exploratory analysis[6]. Exploratory analysis uses tools such as charts, tabulations, and summary statistics to display the distribution of variables, reveal relationships between variables, and calculate summary statistics[7]. Further analysis of the data can provide a deeper understanding of variable characteristics and the connections between different features, especially the correlation between attributes. Quantify the relationship through correlation coefficient testing, laying a solid foundation for subsequent modeling and analysis [8].

2.3 Forecast data preparation

2.3.1 Preprocessing of prediction data

Raw data may not be directly used for modeling due to various factors, such as anomalies or errors in data collection, storage, and extraction processes, as well as data anomalies in software operations. Therefore, data preprocessing becomes particularly critical. Preprocessing includes:

- Data cleaning: In order to ensure data quality, it is necessary to clean, delete duplications, fill in gaps, screen valid data, identify anomalies, and correct inconsistencies in statistical calibers. Among them, dealing with missing and outliers is the most complex, but the methods are similar [9]. You can choose to delete data containing missing values or fill them with a specific method, such as zero filling or mean filling.
- Regularization transformation: In order to avoid the influence of measurement units and meet the special requirements of certain algorithms, standardization or transformation processing is required. For example, genetic algorithms require

inputs to be between 0 and 1, and clustering algorithms are sensitive to data distance [10]. Common methods include normalization transformation and interval transformation, which make all data attributes have the same weight and within a small or common range.

2.3.2 Select explanatory variables

In academic research, selecting explanatory variables is a key step, which aims to find the smallest subset of features that can effectively explain the target. This process is called feature selection in the field of data mining. Feature selection plays an important role in data preprocessing. Usually after obtaining the data, we will perform feature selection first and then train the learning model. There are two main reasons for this: First, in actual tasks, due to the requirements of theoretical analysis and data storage, we may find a large number of variables related to the prediction target, which may lead to excessive computing power requirements and even trigger modeling disasters. Therefore, it is very meaningful to screen out important and independent variables from many variables[11]. Secondly, removing irrelevant features can reduce interference during model training, thereby reducing the difficulty of the learning task, leaving key factors that may make the model perform better. Through feature selection, we are able to improve the explanatory power and predictive accuracy of the model while reducing computational complexity and the risk of overfitting.

2.4 Predictive model building

Based on the characteristics of e-commerce sales data and its changing business scenarios, causal prediction is regarded by the academic community as a more effective method. However, the performance of machine learning models is highly dependent on the quality of the applied data, so we need to estimate in the early stage which algorithm may increase empirical risks.

In this study, we adopted diversified modeling strategies to enhance explanations in complex business environments. In the modeling process, we selected three mainstream machine learning algorithms with regression prediction capabilities, namely support vector machine, random forest and artificial neural network. Next, we will give a brief overview of these algorithms.

2.4.1 Linear regression

Linear regression is a statistical method that aims to predict or explain changes in the dependent variable by fitting a linear relationship between the independent variable and the dependent variable. It estimates the parameters of the model by minimizing the sum of squares of prediction errors. By further extending this method, a multiple linear regression model is obtained. In this model, a single y variable will establish a linear relationship with multiple (p-1) x variables. Such models can better capture the multi-variable relationships in complex phenomena, improving the accuracy of predictions and the explanatory power of the model[14]. The calculation method of this method is the following formula (1).

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_{\mathrm{i}} \tag{1}$$

where is the β coefficient, the slope of the line; ϵ is the redundancy, the intercept of the line.

2.4.2 Random Forest Regression

Random forest is an ensemble learning technology based on decision trees and bootstrap aggregating. This technology builds and integrates the prediction results of a large number of decision trees to enhance the accuracy of prediction and the robustness of the model, especially when dealing with complex regression problems[15]. The final prediction result is determined through a majority voting mechanism. The calculation method of this method is the following formula (2).

$$\hat{Y} = \phi_{T,P}(X) = \frac{1}{B}\sum\nolimits_{\mathrm{b}=1}^{B} \phi_{T_b,m}(X) \tag{2}$$

where *T* represents the entire training set. T_b represents a randomly selected subset (b = 1…, B) from the training set *T*.

2.4.3 SVM Regression

The support vector machine regression machine is a kind of machine that improves the generalization ability of the learning machine by seeking to minimize the structured risk, minimizing the empirical risk and confidence range, so as to obtain good statistics even with a small statistical sample size. purpose of the rule. The model uses the kernel function to map the input space to the feature space and find the optimal hyperplane for regression prediction, aiming to minimize the prediction error and maximize the model generalization ability. The calculation method of this method is the following formula (3).

$$f(x) = \sum_{n=1}^{N} (\alpha_n - \alpha_n^*)(x_n{}', x) + b \tag{3}$$

where f(x) is the objective function and x n is a multivariate set of N observations.

2.4.4 Neural Network

The neural network algorithm is a computing model that simulates the connection of neurons in the human brain. It uses the connection relationships between a large number of neurons to adjust the weight parameters in the network through learning and training to achieve tasks such as classifying and predicting input data. The algorithm has powerful nonlinear fitting capabilities and self-learning capabilities, can handle complex nonlinear relationships, and adaptively extract features in the data [16,17].

The neural network prediction algorithm process includes initializing parameters, forward propagation to calculate prediction results, back propagation to calculate gradients, updating network parameters, and iterative optimization. By continuously learning and adjusting parameters, neural networks can extract features from data and make accurate predictions. Result analysis

2.4.5 Model comparison strategy

In the field of machine learning, algorithms can be divided into supervised learning algorithms and unsupervised learning algorithms based on whether they carry target tags. Clustering problems use unsupervised learning algorithms, while classification and regression problems use supervised learning algorithms. For supervised learning algorithms, the two basic strategies in model selection are empirical risk minimization and structural risk minimization.

The empirical risk function is the average loss of the model on the training set. When the training set, hypothesis space and loss function are determined, the empirical risk is also determined. The empirical risk minimization strategy believes that the model with the smallest empirical risk is the optimal model. When the number of samples is sufficient, this strategy can ensure good learning effects and is therefore widely adopted in practical applications. For example, maximum likelihood estimation is an example of empirical risk minimization.

However, when the sample size is small and lacks representativeness, minimizing empirical risk may lead to poor learning results and cause overfitting problems. For this reason, the structural risk minimization strategy emerged, which comprehensively considers the empirical risk and confidence range to prevent overfitting. Currently, structured risk-based model selection methods include regularization and cross-validation. Compared with regularization, cross-validation is more popular because it is simple to operate, effective and easy to understand.

Cross-validation divides the training data set into two parts when there are sufficient samples: one part is used as the training set and the other part is used as the test set. The training set is used for model training, and the test set is used for model selection. In practical applications, the commonly used cross-validation method is the K-fold cross-validation method. This method first randomly divides the data into K subsets of the same size and non-overlapping, then selects (K-1) subsets as the training set, and the remaining one subset as the test set. After repeating K times, the model with the smallest average prediction error is selected as the final selection. This strategy effectively reduces the complexity of model selection while maintaining the model's generalization ability.

2.4.6 Average error function

In academic research, evaluating the error of prediction models is a crucial step. At present, the mainstream model fitting test indicators are mainly divided into two categories: one is the interpretability comparison type, and the other is the residual comparison type. Among the comparative indicators of interpretability, the correlation coefficient, coefficient of determination R^2 and adjusted coefficient of determination are widely used. They can effectively measure the proportion of the difference in the value of the predictor variable caused by the explanatory variable. Among the residual comparison indicators, the average absolute residual (MAE), the sum of square errors (MSE), the root mean square error (RMSE) and the average absolute percentage error (MAPE) are often used to measure the prediction error of the model.

In a business environment, many organizations tend to use MAPE to evaluate the accuracy of forecasts because MAPE expresses the error as a percentage, which is easier to understand and does not require knowing the specific quantity of the measurement object when conveying information. However, in the process of studying the literature, we found that a few researchers have misunderstandings about the use of R^2. As a key indicator in linear regression, we will explain R^2 in detail below.

We choose the root mean square error (RMSE) as the comparison index of model prediction error because it is sensitive to prediction accuracy and has a wide range of applications [18]. In addition, the coefficient of determination R^2 is an important measure to evaluate the goodness of fit of the linear regression equation. The basic principle is that it measures the proportion of the difference in the values of the predictor variable that is due to the explanatory variable [19]. Therefore, the larger the part that the regression line can explain, the better the fitting effect of the regression line is. This measurement method provides us with an intuitive and effective way to evaluate model fitting, thereby improving the prediction performance and accuracy of the model. Figure 1 is a demonstration diagram of the calculation formula (4) below.

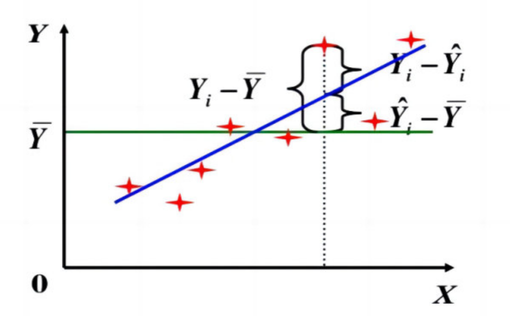


Fig. 1. Linear dispersion decomposition diagram.

The calculation method of R^2 is shown in the following formula (4).

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{4}$$

The issue of adaptability of historical data of predictor variables can be determined by systematically comparing the prediction accuracy of the training set and the prediction set[20]. To this end, we will use Mean Absolute Percentage Error (MAPE, Mean Absolute Percentage Error) as a quantitative indicator to comprehensively evaluate the feasibility of the prediction solution. The calculation method of MAPE is shown in the following formula (5).

$$MAPE = \frac{1}{N} \sum_{i=1}^{n} \left| P_{\text{ei}} \right| \tag{5}$$

Among them $P_{ei} = \frac{e_i}{y_i} = \frac{Y_i - F_i}{Y_i}$, Y i is the actual value of the i-th item, and F i is the predicted value of the i-th item.

2.4.7 Summary of results

Result analysis, as the output link of prediction activities, covers the solution results of the model and the in-depth and multi-angle interpretation of the model results. We can deepen our business understanding from the following perspectives: First, by comparing the results of model feature selection, we can deeply understand the impact of each factor on sales. At the same time, comparing the explanation degree or information gain of the regression model can help us identify the influence of different explanatory variables on the prediction target. Second, the process of selecting the optimal model can also reflect data characteristics. For example, if a linear model achieves better prediction accuracy, this may indicate that the influence pattern is more linear. In addition, comparing the forecast accuracy can reflect the company's market prediction ability and room for improvement, and provide valuable reference for refined operations. Finally, even model parameters may reveal characteristics of the data that can help deepen business understanding.

3 E-commerce sales forecast application examples

In the practice process, this article will use the CRISP-DM methodology as the framework process to carry out prediction work. Figure 2 is a schematic diagram of the cross-industry data mining process.

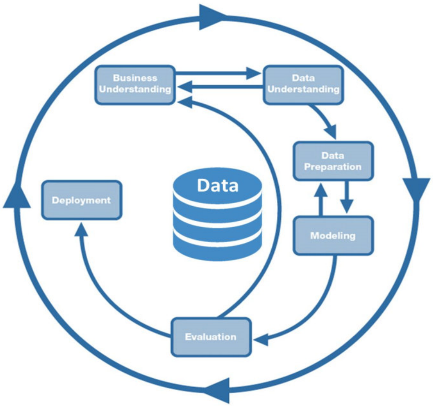


Fig. 2. Schematic diagram of the cross-industry data mining process.

3.1 Overview of the current situation of the enterprise

A recent communication with a large Chinese e-commerce company revealed that its forecasting process mainly relies on the experience and judgment of employees. This method is both cumbersome and lacks the stability of forecast results. In the current business environment, as market competition intensifies, how companies can allocate resources more efficiently, achieve increased revenue, reduce expenditure and green operations has become a key factor in determining the success or failure of the company. To address this challenge, it is particularly important to introduce more scientific and efficient prediction models and methods. Prediction models based on data and algorithms can reduce the interference of human experience and improve the accuracy and stability of predictions. In addition, there have been many studies in the academic community exploring how to use advanced artificial intelligence and machine learning technologies to optimize resource allocation and improve corporate operating efficiency. This will not only help enhance the competitiveness of enterprises in the market, but also promote their transformation to a greener and more sustainable business model.

3.2 Predictive target analysis

Understanding of e-commerce sales forecast: E-commerce platform sales forecast is a macro-level collective forecast. From the perspective of forecast time, the daily sales volume in the next month is a short-term accurate forecast, which requires data in days[21]. Combining the knowledge about consumption behavior and online marketing, we can summarize all the factors that affect the short-term sales of e-commerce, As shown in Table 1 below.

Table 1. Factors affecting e-commerce sales.

Source of influence	Influencing factors	Impact duration
Electronic business platform	product richness	length
Electronic business platform	product quality	length

Electronic business platform	website advertising	length
Electronic business platform	product price	short
Electronic business platform	fall	short
Electronic business platform	full discount	short
Electronic business platform	coupon	short
Electronic business platform	in stock	length
Electronic business platform	logistics	length
Electronic business platform	competitors	length
External environment	festival	length
External environment	social contact	length
External environment	week	short
External environment	weather	short
Consumer	consumption intention	length

The influencing factors of sales volume during the sales period are shown in the table. Since different influencing factors cannot directly measure the impact on forecast sales, these possible short-term influencing factors need to be quantified [22]. For factors that cannot be quantified directly, you can try to use indirect indicators to measure them. Specifically, influencing factors can be reflected through indicators such as amount, number of people, frequency, and number. The data types are mainly numerical values, ratios, and label types. The quantitative results are shown in Table 2 below.

Table 2. Quantification of factors affecting e-commerce sales.

Source of influence	Influencing factors	Quantitative indicators
Electronic business platform	product richness	SKU type
Electronic business platform	product quality	customer reviews
Electronic business platform	website advertising	advertising costs
Electronic business platform	product price	price
Electronic business platform	fall	direct discount amount/number of orders
Electronic business platform	full discount	full discount amount/number of orders
Electronic business platform	coupon	coupon discount amount/number of orders
Electronic business platform	in stock	inventory quantity/inventory amount
Electronic business platform	logistics	delivery time
External environment	competitors	inventory levels
External environment	festival	holiday tags
External environment	social contact	evaluate
External environment	week	week label
External environment	weather	temperature/weather tags
Consumer	purchase Intention	favorites/shopping cart

3.3 E-commerce sales related data analysis

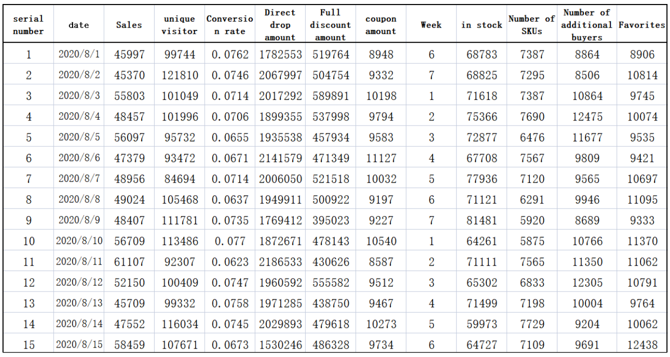
3.3.1 Data acquisition

In order to obtain the data required for this study, we mainly used two methods: on the one hand, we obtained relevant data through the internal business analysis system database of e-commerce companies; on the other hand, we extracted the required data from publicly available databases on the Internet[23]. The statistics of the final fields we obtained are shown in Table 3.

Table 3 Data set of factors affecting e-commerce sales.

Source of influence	Influencing factors	Quantitative indicators
Electronic business platform	product richness	SKU type
Electronic business platform	consumption intention	number of favorites/number of items
Electronic business platform	consumption intention	number of additional purchases/number of items
Electronic business platform	fall	direct discount amount/order/sales
Electronic business platform	full discount	full discount amount/order/sales
Electronic business platform	coupon	coupon discount amount/order/sales
Electronic business platform	in stock	inventory
External environment	festival	holiday tags
External environment	week	week label

In this study, we recorded sales volume, page views, promotion data, and inventory data at the SKU level. At the same time, for more refined analysis, labels such as holidays and working days are added at the daily data level. We record the number of collections and the number added to the shopping cart on a daily level. In terms of the time span of the data, the time span of sales data and view data is from January 2020 to July 2021, while the inventory data, collection data and data added to the shopping cart cover the time period from August 2020. Until August 2021. After screening, we finally selected August 2020 to July 2021 as the modeling time range. All types of data during this period are fully recorded, which facilitates our in-depth analysis. To achieve this, we process current data and organize it into daily data. During this process, we obtained a total of 353 pieces of valid data, part of which is shown in Figure 3.



serial number	date	Sales	unique visitor	Conversion rate	Direct drop amount	Full discount amount	coupon amount	Week	in stock	Number of SKUs	Number of additional buyers	Favorites
1	2020/8/1	45997	99744	0.0762	1782553	519764	8948	6	68783	7387	8864	8906
2	2020/8/2	45370	121810	0.0746	2067997	504754	9332	7	68825	7295	8506	10814
3	2020/8/3	55803	101049	0.0714	2017292	589891	10198	1	71618	7387	10864	9745
4	2020/8/4	48457	101996	0.0706	1899355	537998	9794	2	75366	7690	12475	10074
5	2020/8/5	56097	95732	0.0655	1935538	457934	9583	3	72877	6476	11677	9535
6	2020/8/6	47379	93472	0.0671	2141579	471349	11127	4	67708	7567	9809	9421
7	2020/8/7	48956	84694	0.0714	2006050	521518	10032	5	77936	7120	9565	10697
8	2020/8/8	49024	105468	0.0637	1949911	500922	9197	6	71121	6291	9946	11095
9	2020/8/9	48407	111781	0.0735	1769412	395023	9227	7	81481	5920	8689	9333
10	2020/8/10	56709	113486	0.077	1872671	478143	10540	1	64261	5875	10766	11370
11	2020/8/11	61107	92307	0.0623	2186533	430626	8587	2	71111	7565	11350	11062
12	2020/8/12	52150	100409	0.0747	1960592	555582	9512	3	65302	6833	12305	10791
13	2020/8/13	45709	99332	0.0758	1971285	438750	9467	4	71499	7198	10004	9764
14	2020/8/14	47552	116034	0.0745	2029893	479618	10273	5	59973	7729	9204	10062
15	2020/8/15	58459	107671	0.0673	1530246	486328	9734	6	64727	7109	9691	12438

Fig. 3. Some examples of raw data.

3.3.2 Exploratory analysis of e-commerce sales

The realization of e-commerce sales targets can be broken down into traffic targets and conversion rate targets, which are also the most important reference indicators used by e-commerce companies to measure operating conditions in practice. From Figure 4, you can see the line chart of sales volume, page views and conversion rate [24]. Combined with the communication with the employees of the enterprise sales department, the following judgments can be made:

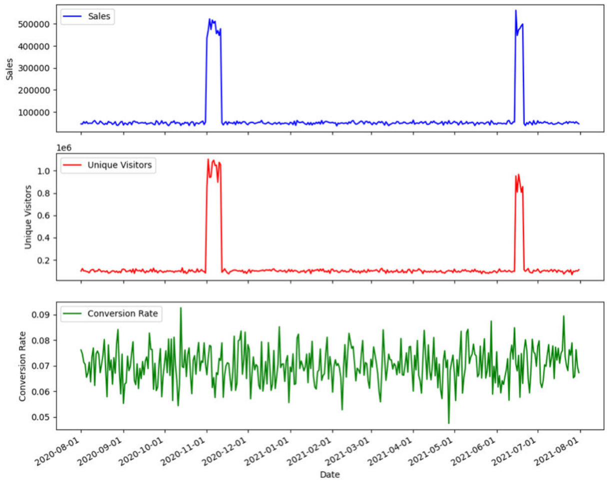


Fig. 4. Line chart of sales, pageviews and conversion rate.

According to a comprehensive analysis of the e-commerce's historical sales data, its overall performance showed a stable trend, and no significant cyclical fluctuations were observed. However, in the sales data, we find two significant peaks. After in-depth communication with the company's sales department, we learned that the occurrence of these peaks is closely related to major promotions every year. At these specific points in time, e-commerce platforms will adopt extremely favorable promotion strategies, which is largely influenced by the traditional festival atmosphere and the long-term consumption habits of consumers. However, these sales spikes are not sustainable and are driven by short-term promotions rather than long-term steady-state demand.

3.3.3 Feature exploration analysis

Scatter plots are a very visual and informative chart type that are ideal for exploring relationships between different variables. In this experiment, we use Python to visualize scatter plots. Fig.5 not only shows a scatter plot, but also plots a linear and smooth fitted curve, with the histogram placed along the main diagonal[25]. From the scatter plot, we can see that there is a very high linear correlation between sales and traffic. Furthermore, sales are significantly and positively related to both the amount of direct discounts and the amount of full reductions. Interestingly, inventory volume and SKU count exhibit significant non-normality, indicating that various determinants may influence these two volumes.

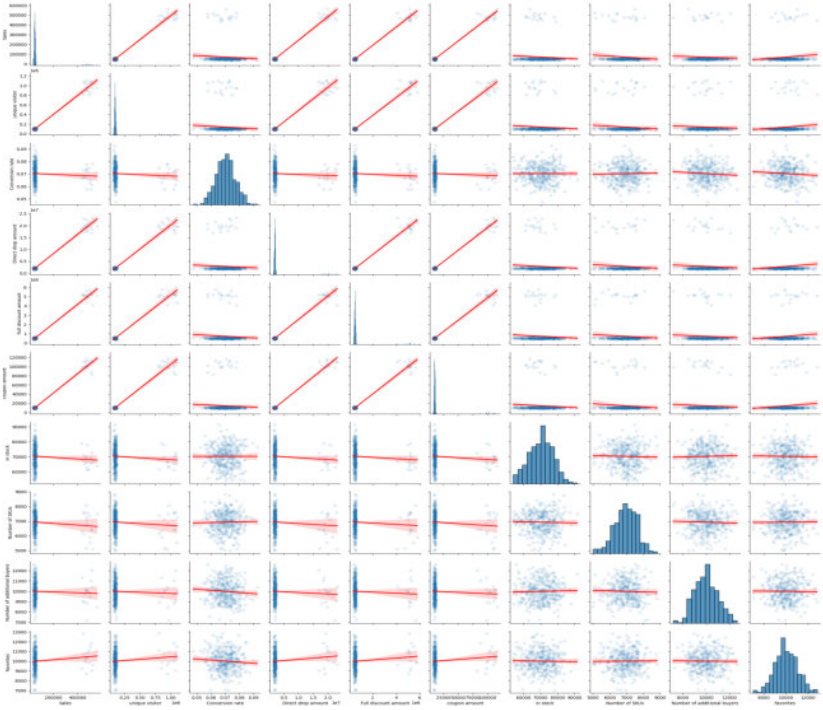
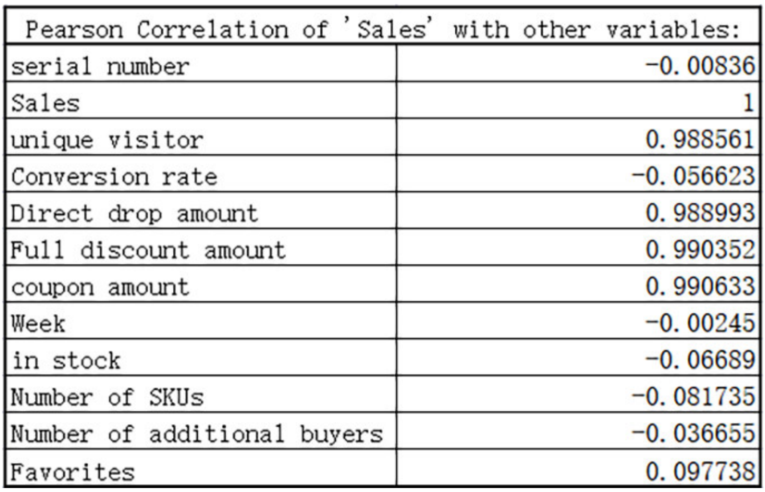


Fig. 5. Scatter plot of sales volume versus other attributes.

In order to accurately determine the correlation between sales data and other variables, we examined the specific value of the Pearson correlation coefficient [26]. Figure 6 shows that there is a significant correlation between the currently selected variables, and the number of views, conversion rate, promotion method and sales show a very high correlation.



Pearson Correlation of 'Sales' with other variables:	
serial number	-0.00836
Sales	1
unique visitor	0.988561
Conversion rate	-0.056623
Direct drop amount	0.988993
Full discount amount	0.990352
coupon amount	0.990633
Week	-0.00245
in stock	-0.06689
Number of SKUs	-0.081735
Number of additional buyers	-0.036655
Favorites	0.097738

Fig. 6. Correlation coefficient between sales volume and other attributes.

Draw a box plot of the factor variables in the feature to observe. From Figure 7, we can see that the statistical effect of each day of the week is not obvious.

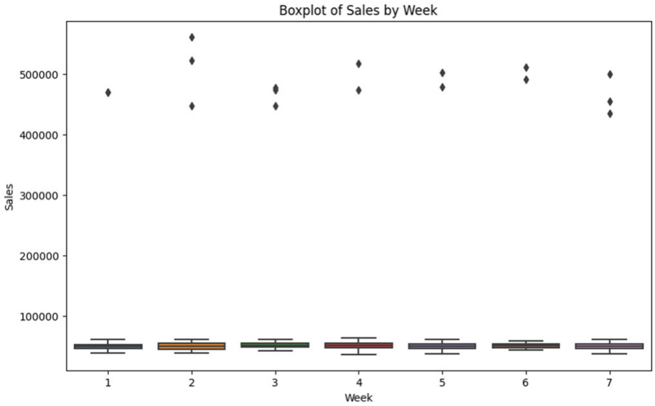


Fig. 7. Boxplot plotted by week.

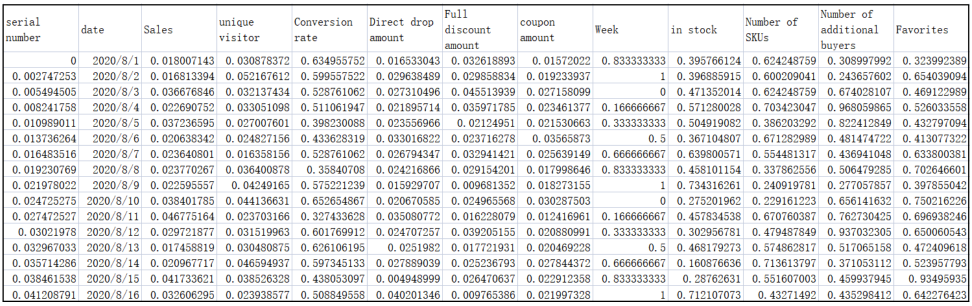
3.4 E-commerce sales data preparation

3.4.1 Data preprocessing

In order to achieve more reasonable data mining results, massive data must be preprocessed to build accurate data mining models. First, perform data cleaning, which removes data that does not meet requirements, such as errors or incomplete data in the data source. Second, deal with outliers. After communicating with the company's business department, we confirmed that the missing values did not cause an exception and that the company's system was operating normally during the period. Therefore, for non-abnormal missing values, we choose to retain the original state of the data and adopt a zero-filling strategy for processing.

Identifying and handling outliers is the most complex. Previous exploratory data analysis found that sales data showed a clear bimodal distribution. After further investigation, it was confirmed that the peak occurs mainly on major promotion days. Given the research objectives of this article, we eliminate data during promotion days as outliers.

Data normalization also plays a key role in data preprocessing. In the sales forecast in this article, a neural network model needs to be established. To meet the input requirements of the neural network model, we normalize the data to the range of 0 to 1. This preprocessing method provides a more accurate and reliable data basis for subsequent data mining work. Figure 8 shows some of the data after normalization.



serial number	date	Sales	unique visitor	Conversion rate	Direct drop amount	Full discount amount	coupon amount	Week	in stock	Number of SKUs	Number of additional buyers	Favorites
0	2020/8/1	0.018007143	0.030878372	0.634955752	0.016533043	0.032618893	0.01572022	0.833333333	0.395766124	0.624248759	0.308997992	0.323992389
0.002747253	2020/8/2	0.016813394	0.052167612	0.599557522	0.029638489	0.029858834	0.019233937	1	0.396885915	0.600209041	0.243657602	0.654039094
0.005494505	2020/8/3	0.036676846	0.032137434	0.528761062	0.027310496	0.045513939	0.027158099	0	0.471352014	0.624248759	0.674028107	0.469122989
0.008241758	2020/8/4	0.022690752	0.033051098	0.511061947	0.021895714	0.035971785	0.023461377	0.166666667	0.571280028	0.703423047	0.968059865	0.526033558
0.010989011	2020/8/5	0.037236595	0.027007601	0.398230088	0.023556966	0.02124951	0.021530663	0.333333333	0.504919082	0.386203292	0.822412849	0.432797094
0.013736264	2020/8/6	0.020638342	0.024827156	0.433628319	0.033016822	0.023716278	0.03565873	0.5	0.367104807	0.671282989	0.481474722	0.413077322
0.016483516	2020/8/7	0.023640801	0.016358156	0.528761062	0.026794347	0.032941421	0.025639149	0.666666667	0.639800571	0.554481317	0.436941048	0.633800381
0.019230769	2020/8/8	0.023770267	0.036400878	0.35840708	0.024216866	0.029154201	0.017998646	0.833333333	0.458101154	0.337862556	0.506479285	0.702646601
0.021978022	2020/8/9	0.022595557	0.04249165	0.575221239	0.015929707	0.009681352	0.018273155	1	0.734316261	0.240919781	0.277057857	0.397855042
0.024725275	2020/8/10	0.038401785	0.044136631	0.652654867	0.020670585	0.024965568	0.030287503	0	0.275201962	0.229161223	0.656141632	0.750216226
0.027472527	2020/8/11	0.046775164	0.023703166	0.327433628	0.035080772	0.016228079	0.012416961	0.166666667	0.457834538	0.670760387	0.762730425	0.696938246
0.03021978	2020/8/12	0.029721877	0.031519963	0.601769912	0.024707257	0.039205155	0.020880991	0.333333333	0.302956781	0.479487849	0.937032305	0.650060543
0.032967033	2020/8/13	0.017458819	0.030480875	0.626106195	0.0251982	0.017721931	0.020469228	0.5	0.468179273	0.574862817	0.517065158	0.472409618
0.035714286	2020/8/14	0.020967717	0.046594937	0.597345133	0.027889039	0.025236793	0.027844372	0.666666667	0.160876636	0.713613797	0.371053112	0.523957793
0.038461538	2020/8/15	0.041733621	0.038526328	0.438053097	0.004948999	0.026470637	0.022912358	0.833333333	0.28762631	0.551607003	0.459937945	0.93495935
0.041208791	2020/8/16	0.032606295	0.023938577	0.508849558	0.040201346	0.009765386	0.021997328	1	0.712107073	0.43271492	0.435298412	0.642276423

Fig. 8. Data after normalization.

3.5 Construction of e-commerce sales forecast model

When building the prediction model, we used three currently mainstream machine learning algorithms with regression prediction capabilities, namely support vector machines, random forests and artificial neural networks. At the same time, in order to compare the effects, we also introduced a multiple linear regression model. This model is recognized to be effective and highly interpretable, but its disadvantage is that it can only handle linear relationships [27]. In addition, we also discussed the comprehensive forecasting method. This method is currently attracting much attention, but due to the diversity of its combination methods, it is impossible to exhaust them one by one, and it is difficult to determine in advance which combination method is the best. Therefore, this article finally selected the comprehensive prediction method.

In the model selection stage, we conducted a comprehensive and rigorous comparison and evaluation, and selected linear regression models, support vector regression models, and random forest regression models from many prediction models. Support vector machines, in particular, can flexibly implement linear and nonlinear regression by carefully selecting kernel functions, such as "linear basis kernel" or "radial basis kernel", demonstrating their strong adaptability and accuracy. This selection process ensures that our model selection is comprehensive, objective, and in-depth, with the goal of providing the best predictive performance for the study. Figure 9 shows an example of the model used in this article.

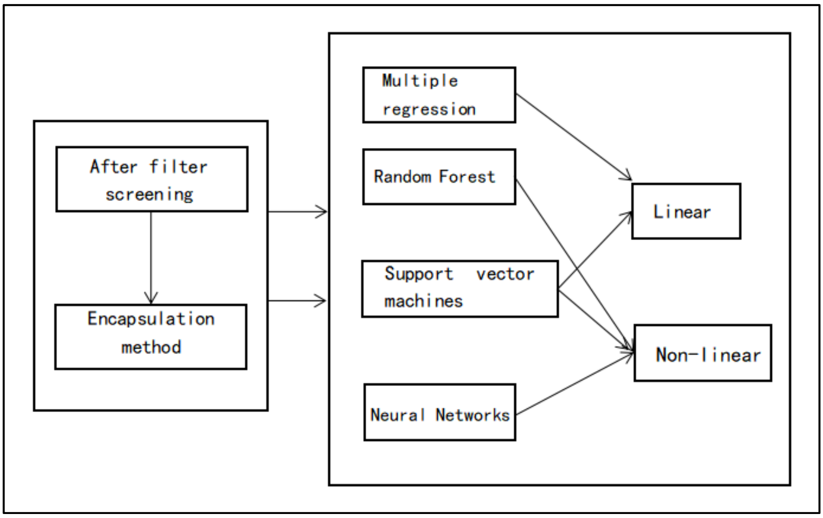


Fig. 9. Combined structure of the model.

Based on whether the evaluation criteria are independent of subsequent learning algorithms, feature selection methods can be divided into filter methods and encapsulation methods. The filter method does not rely on a specific modeling technique, it selects features based on the statistical characteristics of the data . In contrast, the encapsulation approach uses a portion of the data to model, looks at the results, and then repeats the process, selecting variables in each iteration. These methods have important application value in feature selection and can improve the performance and efficiency of the model.

3.6 Result analysis

3.6.1 Comparison of results

In the case of ensuring that the error rate is within a certain range, MSE is used as the model evaluation index. The R^2 parameter is a commonly used indicator to measure the goodness of fit of a linear model. In order to avoid sample bias caused by a single random sampling, this experiment uses the mean as the prediction evaluation index of the model. Among them, the "features" in the table are the characteristic attributes selected from the data. This is the smallest subset of features that can effectively explain the target.

The following are the prediction results of the training set, as shown in Table 4. The first and second columns are the selected models and features respectively, and the 3rd/4th/5th columns are the test result data R^2, RMSE, and MAPE respectively.

Table 4. Prediction results of training set.

Regression model	**Feature**	**Parameter settings**	**R^2**	**RMSE**	**MAPE**
Llinear regression	all	——	0.8694	5093	0.0676
Linear regression	123478	——	0.8615	5244	0.0668
Random forest	all	ntree=300	-	3024	0.0335
Random Forest 2	1234578	ntree=70	-	3215	0.0382
Support vector regression linear	All	kernel="linear",cost=10,gamma=0.0001	0.8558	5329	0.0648
Support vector machine linear 2	123478	kemel="linear",cost=10,gamma=0.0001	0.8514	5442	0.0671
Support vector machine nonlinearity	All	kemel="radial",cost=100,gamma=0.001	-	4859	0.0589
Support vector machine nonlinearity 2	123478	kernel="radial",cost=100,gamma=0.001	-	5046	0.0641
Neural network nonlinearity	All	size=18,maxit =1000,linout=F	-	310	0.0049
Neural network nonlinearity 2	12345678	size=40,maxit =1000,linout=F	-	4045	0.058

When evaluating the performance of the model, RMSE and MAPE show consistent trends. The nonlinear regression model shows advantages over the linear regression model in terms of overall performance. It is particularly worth mentioning that the random forest model showed the best fitting effect among all models.

From the numerical observation of the R^2 parameter, it can be known that multivariate linear programming and support vector machines have shown good results in model fitting, and the performance of the two is quite close. However, the performance of the R^2 parameter of the neural network is not ideal, showing poor fitting effect.

Table 5. Test set prediction results.

Regression model	**Feature**	**R^2**	**RMSE**	**MAPE**
Linear regression	all	0.8694	5088	0.0666
Linear regression	123478	0.8621	5251	0.0654
Random forest	all	**-**	3018	0.0328

Random Forest 2	1234578	-	3264	0.0377
Support vector regression linear	All	0.8563	5352	0.0655
Support vector machine linear 2	123478	0.8522	533	0.0658
Support vector machine nonlinearity	All	-	4854	0.0585
Support vector machine nonlinearity 2	123478	-	5045	0.0646
Neural network nonlinearity	All	-	299	0.0039
Neural network nonlinearity 2	12345678	-	4054	0.0585

The above are the test set prediction results, as shown in Table 5. After in-depth analysis and modeling of the collected quantitative data on influencing factors, the constructed prediction model demonstrated a high level of accuracy. This result shows that based on the currently available data, we have been able to build a predictive model with good performance.

3.6.2 Summary of results

By comparing the prediction results, we observed that, except for the BP neural network model, other models showed higher prediction accuracy in daily sales prediction. It is worth noting that nonlinear models do not significantly outperform linear models in performance. This implies that in the current situation, the impact model of various factors on sales tends to be more linear. However, nonlinear models have the ability to capture more nonlinear relationships. Therefore, after selecting the best model, we randomly selected 30 consecutive days of data to predict sales, and compared it with the actual 30-day sales data to verify the accuracy of the model. As shown in Figure 10.

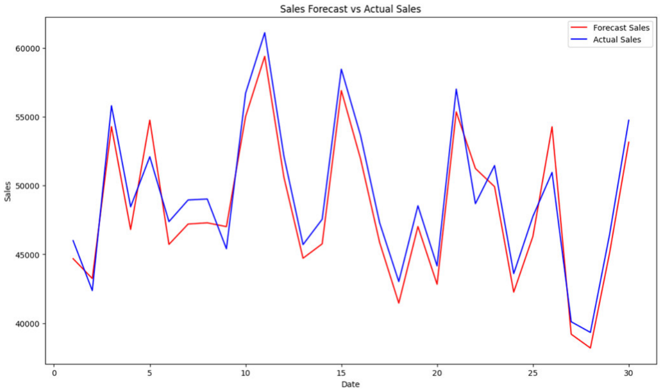


Fig. 10. Comparison chart of forecasted sales and actual sales.

The red line in the figure is the predicted sales volume, and the blue line is the actual sales volume. It has been calculated that the margin of error between forecast sales and actual sales is approximately 14%. It already has relatively reliable and practical use value.

4 Conclusion

This article compared different prediction models and found that the support vector machine using radial basis kernel has the best prediction effect and is the most stable. Although the prediction accuracy of multiple linear regression is close, support vector machine has stronger nonlinear fitting ability and self-processing ability, and is more robust. Therefore, support vector machines have higher application value in e-commerce sales forecasting. It should be noted that the performance of different algorithms can also reflect data characteristics, but compared with multiple linear regression, it does not significantly improve the prediction accuracy, which indicates that there is currently a strong linear relationship between explanatory variables and sales.

In academic research, we focus on building prediction processes under the guidance of CRISD-DM and collating related knowledge and technologies. From an application point of view, we try to ensure the comprehensiveness of knowledge as much as possible, but there may be insufficient depth in the optimization of predictive models. First, there is still room for optimization in feature selection and parameter selection in the modeling process. Although the research focus of this article is on methodology and prediction process, which compromises the optimization of the model, research in this direction is of great significance and can be integrated into the prediction process proposed in this article in the form of modules. Secondly, the current forecasting process considers only a single scenario in the modeling process, and does not conduct detailed analysis of forecasting needs under special scenarios. For example, sales forecasts under big promotions, brand day promotions and other modes, as well as in the period before and after promotions, require more detailed research. Such research will require more dimensions of data and deeper customer analysis, combined with expert opinion. Such prediction tasks are more demanding, and according to literature research, data mining algorithms during major promotions may have significant advantages over traditional prediction methods.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No.075-15-2022-1121).

References

1. D. Chen, S.L Sain, Journal of Statistical Computing and Simulation **89(5)**, 861-879 (2016)
2. Y. Ma, J. Zhang, Information Fusion **73**, 110-121 (2021)
3. L. Xu, C. Jiang, J. Wang, J. Yuan, Y. Ren, IEEE Access **2**, 1149-1176 (2021)
4. Y. Chen, Tsinghua science and technology **6(1)**, 75-79 (2001)
5. L. Huang, M. Benyoucef, Electronic Commerce Research and Applications **12(4)**, 246-259 (2013)
6. M.H. Huang, D.Q. Chen, Y.W. Fan, Information Systems Boundaries **19(4)**, 739-751 (2017)
7. Su Wei, Application and standardisation of data mining tools J. Computer Engineering (2004)
8. Dong Shishi, J. Integration Technology **2(1)** (2013)
9. J. Brownlee, Mastering machine learning algorithms: understanding how they work and implementing them from scratch. Machine Learning Savvy (2016)
10. D. Lahat, T. Adali, C. Jutten, IEEE **103(9)**, 1449- 1477 (2015)
11. Pan Wuming, Pan Yunhe, J. Computer Application Research (2004)
12. Wang Jacai, Chen Qi, Zhao Jieyu., J. Computer Engineering (2003)
13. W.H. Inmon, *Building the Data Warehouse* (Beijing, Machinery Industry Press, 2000)

14. Jiawei Micheline Kamber, *Data Mining Concepts and* Techniques (Beijing, Machinery Industry Press, 2001)
15. J.M. Chen, *Data Warehousing and Data Mining Techniques* (Electronic Industry Press, 2002)
16. S.J. Pan, Q. Yang, IEEE Knowledge and Data Engineering Repertoire **22(10)**, 1345-1359 (2010)
17. H. Wang, W. Chen, Z. Xu, Journal of Physics: conference series **1593(1)**, 012062 (2020)
18. Y. Wang, J. Ma, Q. Zhang, Information Fusion **73**, 110-121 (2021)
19. G. Zhang, X. Li, Y. Luo, Energy **162**, 164-173 (2018)
20. S.J. Pan, Q. Yang, IEEE Transactions on Knowledge and Data Engineering **22**(10), 1345-1359 (2010)
21. G. Shmueli, N.R. Patel, P.C. Bruce, *Business intelligence data mining: concepts, techniques and applications in Microsoft Office Excel and XLMiner* (John Wiley & Sons, 2010)
22. R.S. Sutton, A.G. Barto, *Reinforcement learning: an introduction* (MIT Press, 2018)
23. H. Li, J. Zhang, Chinese Journal of Management Science **10(4)**, 82-88 (2002)
24. Z.H. Zhou, Integration methods: foundations and algorithms. crc Press. (2012)
25. X. Luo, Journal of Consumer Psychology **15(4)**, 288-294 (2005)
26. F. Tingli, C. Niu, Y. Song, ASSEHR **687**, 1020-1030 (2022)
27. Y. Siming, Engineering and Technology **47** (2023)