2021 2nd International Conference on E-Commerce and Internet Technology (ECIT) | 978-1-6654-3873-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ECIT52743.2021.00093

Sales Prediction based on Machine Learning

Zixuan Huo
International School
Beijing University of Posts and Telecommunications
Beijing, China
huozixuan@bupt.edu.cn

Abstract—**With the increasing influence of the Internet on people's life, the development of e-commerce platforms is more rapid, with users and earnings of these platforms showing a growing trend. In recent years, the strong support of national policies has also provided a good environment for the development of the e-commerce industry. Under the impact of the epidemic this year, the role of the e-commerce industry in the development of the national economy has become more prominent. In such cases, the number and the competitiveness of e-commerce platforms and e-commerce enterprises are increasing. If a platform wants to maintain its advantage in the competition, it must be able to better meet the needs of users, and do a good job in all aspects of coordination and management. At this point, the accurate forecast of the sales volume of e-commerce platforms is particularly important. At present, there are many studies on e-commerce sales prediction, but we are still exploring the prediction model that can be better applied in different scenarios. In this paper, we try and evaluate two linear models, three machine learning models and two deep learning models, finding that machine learning and deep learning models have no advantage in improving the accuracy of sales forecast, but on a predictive basis, models perform better when they include information on calendar and price.**

Keywords-Sales Prediction; Regression; Machine Learning; Deep Learning

I. Introduction

In the past 10 years, the e-commerce industry in China has developed rapidly and is still showing a steady upward trend. Taking data from Alibaba and Amazon, two famous e-commerce companies at home and abroad for example, observing Alibaba’s revenue and revenue growth data as of March 31 from 2015 to 2020, it can be found that its revenue has continued to grow from 76.20 billion in 2015 to 509.7 billion in 2020, and the revenue growth from 2015 to 2019 is also stable and on the rise [1]. In 2020, due to the impact of the COVID-19 epidemic, the growth rate of income has decreased, but the income still continues to grow. The annual financial report data of another major e-commerce giant Amazon also shows a growing trend. Amazon's annual report data shows that its revenue continued to increase steadily from 2015 to 2019 [2]. Although its revenue growth is not as obvious as that of Alibaba, it still continues to rise every year.

At the national level, the implementation of the "E-commerce Law of the People's Republic of China" on January 1, 2019 and the adjustment of cross-border e-commerce retail import policies both show China's emphasis on the e-commerce industry. In addition, at the Third Session of the 13th National People’s Congress, Premier Keqiang Li pointed out that new formats such as e-commerce, online shopping and online services have played an important role in the fight against the epidemic. Support policies should continue to be introduced to comprehensively promote the “Internet Plus” and create digital New economic advantage. The government also clearly pointed out in the report that it will vigorously support e-commerce in the future, and will launch more related policies to promote the Internet economy in the future.

At the same time, due to the large population base and large number of Internet users in China, with the continuous improvement of Internet penetration rate in the future, the further improvement of logistics industry and the convenience brought by mobile payment technology, the development space of domestic e-commerce industry is still very broad. Therefore, good development prospect naturally attracted a number of e-commerce enterprises. According to the data from Tianyancha pro, by November 2020, there are more than 3.78 million e-commerce related enterprises in China, including 570,000 cross-border e-commerce related enterprises. In the first 10 months of 2020, more than 95,000 cross-border e-commerce related enterprises have been added, with a year-on-year growth of 79.22%. Behind the rapid development of the e-commerce industry is the fierce competition of many peer enterprises. If a company want to stand firm in such a fierce competition, data is a core of winning. What kind of data to collect and how to process and apply these data are of great significance for turning potential customers into value customers.

Regardless of whether it is an online or offline company, the purpose is to provide products or services to the society. Therefore, its production and decision-making will be greatly affected by demand forecasts. For merchants, timely meeting customer requirements for delivery dates and other aspects can improve customer satisfaction and enhance competitiveness, reduce corporate inventory and arrange production, and also help merchants make more reasonable pricing and promotion decisions. Transportation management will also be affected by sales predictions. Compared with the traditional retail industry, e-commerce companies respond more quickly to the needs of the market and consumer needs in order to gain a position in the fierce market competition. Therefore, it is very necessary for e-commerce merchants to predict the sales volume in the future.

The indicators that affect the sales forecast of e-commerce products are mainly divided into three categories. The first category is the attribute characteristic index of the product or

978-1-6654-3873-5/21/$31.00 ©2021 IEEE
DOI 10.1109/ECIT52743.2021.00093

business, such as product promotion information, price, daily sales, current business operation time, product customer praise rate, historical transaction volume, current business reputation score and level, product collection popularity. The second category is product reviews and derivative indicators, such as review time, reviewer information, review text, review star ratings, number of review responses. The third category is product online search information, such as keyword search volume. This paper mainly considers the impact of the first category of indicators on sales prediction.

From the perspective of previous studies, linear model [3], machine learning model [4-9] and deep learning model [10, 11] are all common methods to predict e-commerce sales volume.

In summary, we want to compare whether machine learning and deep learning can predict sales more accurately, and whether other information besides historical sales information can help improve prediction accuracy. We used a real data set provided by Walmart and used Walmart's hierarchical sales data to predict daily sales for the next 28 days. In terms of revenue, these data cover stores in three U.S. states (California, Texas, and Wisconsin), including product level, department, product category, and store details. In addition, it has explanatory variables such as price, promotion, day of the week, and special events. We tried 2 linear models, 3 machine learning models, and 2 deep learning models on a data set containing 1941-day sales data. We found that after adding date, price and other information, the performance of the model would be better. However, compared with simple linear regression models, complex machine learning and deep learning models have no advantages in sales prediction.

II. Related Work

Many previous studies have applied different prediction models and obtained good results, especially the machine learning and deep learning models, which have been proven effective in different problems [12-18]. A prediction method applicable to the situation without any historical data of fashion supply chain is proposed in [4], which was based on machine learning. A short-term demand commodity sales prediction model based on LSTM was proposed in [19], which could learn the prediction of future value according to the time series of sales and the emotional rating of comments. It is showed in [20] that the GA-BP algorithm model based on the promotion and historical data of B2C e-commerce platform had a good adaptability to sales prediction. Aiming at the sales forecast problem of cross-border e-commerce enterprises, a three-stage model based on XGBoost was proposed in [5]. ARIMA-NARNN model was proposed in [3] to predict e-commerce sales. In our study, we will further compare the effectiveness of these models in the sales forecasting problem.

In [19], the author considered the short-term sales forecast based on the sentiment of consumer reviews on e-commerce platforms. Based on the historical sales and online review data of an online store selling "multi-flavor chocolate gift boxes" at Taobao's official flagship store, they adopted the LSTM model and achieve the goal of using minimal historical data, manual efforts of data preparation, and computing resources, and making the accuracy of sales forecasting maximal. In [20], the author took promotion activities into account. Through the verification of the actual case of Alibaba, the author adopted the GA-BP algorithm model and the accuracy reached 94%. In [21], taking Taobao (including Tmall) e-commerce platform as the main object, the author captured data from six fields, namely, agriculture and animal husbandry, clothing, personal consumer goods, furniture, second-hand cars and food. They adopted the CNN model and took the AdaBoost model as the comparison model, and obtained that the CNN deep learning model had better prediction accuracy and generalization ability. In [22], considering the low accuracy in the sales prediction of e-commerce products under small sample data, the author, based on the sales data of Lenovo ZUK Z2 mobile phone products in JD Mall, adopted the sales prediction model based on integrated learning XGBoost algorithm and integrates multi-dimensional indicators to build the prediction model based on integrated learning XGBoost. The results showed that the prediction accuracy of this model was better than that of BP, SVM and BP-SVM combination. In different prediction scenarios, different algorithms show different advantages. Therefore, in this paper, we will focus on comparing the prediction effects of linear model, machine learning model and deep learning model.

III. Dataset

A. Dataset Description

The data set used in this paper is the M5 data set provided by WalMart, involving unit sales of 3049 products, divided into 3 product categories (Hobby, Food and Household) and 7 product departments (Hobbies_1, Hobbis_2, Foods_1, Foods_2, Foods_3, Household_1, Household_2), where the above categories are decomposed. These products are sold in ten stores in three states (CA California, TX Texas and WI Wisconsin). The entire data set consists of three data sets, namely, "calendar.csv", which contains information about product sales dates; "sell_prices.csv", which contains information about product prices for each store and date; "sales_train.csv", which contains historical daily unit sales data for each product and store.

B. Dataset Preprocessing

We preprocessed this data set as follows:

(1) We have summarized the sales of different products according to their states and corresponding categories, so as to obtain 9 different time series, corresponding to the following 9 situations: cat_id = FOODS, state_id = CA; cat_id = HOBBIES, state_id = CA; cat_id = HOUSEHOLD, state_id = CA; cat_id = FOODS, state_id = TX; cat_id = HOBBIES, state_id = TX; cat_id = HOUSEHOLD, state_id = TX; cat_id = FOODS, state_id = WI; cat_id = HOBBIES, state_id = WI; cat_id = HOUSEHOLD, state_id = WI. We will predict these nine time series separately in the following parts.

(2) In addition to historical sales, we also use calendar and price-related information. For price information, we took the average of the prices of all commodities in a certain category in a certain state as an input feature that the model may use. For event information, we did not use specific event content, but

converted whether an event occurred or not into a 0-1 variable as an input feature that the model may use.

C. *Dataset Visualization*

We show the sales in CA, TX and WI in Figure 1-3, respectively. In Figure 1-3, the FOODS category has the largest sales volumes in all three states as it has the most number of products.

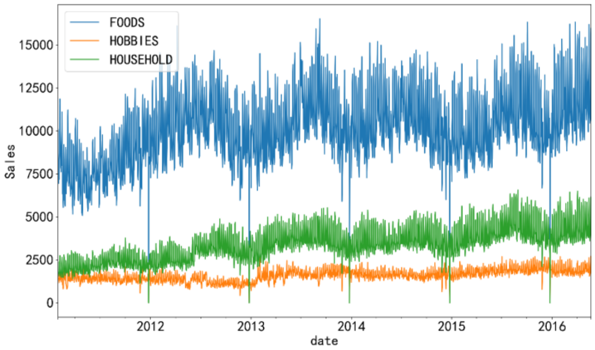

Figure 1. Sales in CA.

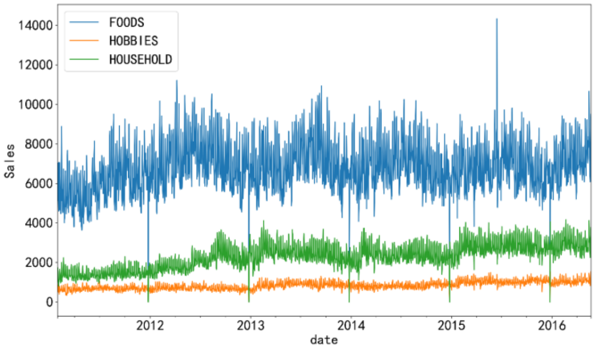

Figure 2. Sales in TX.

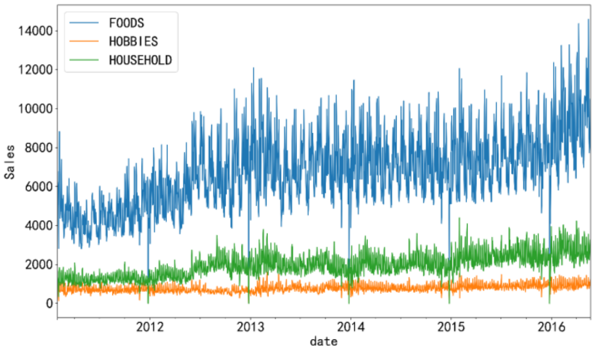

Figure 3. Sales in WI.

IV. Models

In this paper, we compared the performance of two linear models, three machine learning models and two deep learning models.

A. *Linear Models*

The exponential smoothing method is a time series analysis and forecasting method developed on the basis of the moving average method. It predicts the future of the phenomenon by calculating the exponential smoothing value with a certain time series forecasting model. The principle is that the exponential smoothing value of any period is the weighted average of the actual observation value of the current period and the exponential smoothing value of the previous period. The recent data in the sequence is assigned a larger weight, and the forward data is assigned a smaller weight. The reason is that under normal circumstances, the influence of a variable value on its subsequent behavior is gradually attenuated. The first exponential smoothing is for the series without trend and seasonality, the second exponential smoothing is for the series with trend but no seasonality, and the third exponential smoothing considers both the trend and seasonality.

The ARIMA model is the most common statistical model used for time series forecasting. The ARIMA model is very simple, requiring only endogenous variables and no other exogenous variables.

B. *Machine Learning Models*

Regression algorithm is a supervised learning algorithm used to establish the mapping relationship between the independent variable X and the observed variable Y. If the observed variable is discrete, it is called classification; if the observed variable is continuous, it is called regression. The purpose of the regression algorithm is to find a hypothetical function to best fit a given data set. In regression analysis, if only one independent variable and one dependent variable are included, and the relationship between the two can be approximated by a straight line, it is called unary linear regression analysis; if the regression analysis includes two or more independent variables, and the independent variables have a linear relationship, which is called multiple linear regression analysis. For two-dimensional space, linearity is a straight line, for three-dimensional space, linearity is a plane, and for multi-dimensional space, linearity is a hyperplane.

XGBoost is still a Gradient Boosting Decision Tree (GBDT) in essence, and both are boosting methods, but XGBoost is faster and more efficient than GBDT. The core algorithm idea of XGBoost is: continuously add trees, and continuously perform feature splitting to grow a tree. Each time you add a tree, you actually learn a new function f(x) to fit the residual of the last prediction. When we get k trees after training, we need to predict the score of a sample. In fact, according to the characteristics of this sample, each tree will fall to a corresponding leaf node, and each leaf node corresponds to a score. Finally, you only need to add up the scores corresponding to each tree to get the predicted value of the sample.

Random Forest uses the ensemble method of Bagging (bootstrap aggregation). In Bagging, Random Forest will train multiple classifiers independently, and each classifier is trained based on a subset of the training data set. Finally, the prediction results of different classifiers will use the majority rule to derive the final classification results. Compared with a certain classifier, the ensemble model is not easy to make mistakes on a single sample.

C. Deep Learning Models

Deep learning is represented by various neural networks, in which we use Multi-Layer Perceptron (MLP) and Long Short-Term Memory (LSTM). The MLP model we use contains one input layer, one hidden layer and one output layer. We use 256 neurons and ReLU as the activation function in the hidden layer. The LSTM model we use contains one input layer, two hidden layer (i.e., the LSTM layer) and one output layer. We use 100 neurons and ReLU as the activation function in the LSTM layer.

V. Experiments

A. Settings

The models are implemented with Python 3.7, using scikit-learn as the machine learning package and TensorFlow as the deep learning package. We use the model training time and Root Mean Squared Error (RMSE) for evaluation metrics. The data set between 2011-01-29 and 2016-04-24 is used as the training set, and the data set between 2016-04-25 and 2016-05-22 is used as the test set. The target is to predict the sales in the next 28 days.

The Triple Exponential Smoothing model is implemented with the statsmodels package. The ARIMA model is fine-tuned with auto_arima and the optimal model is ARIMA(13, 1, 3), where the 13 is the number of lag observations in the model, 1 is the number of times that the raw observations are differenced, and 3 is the size of the moving average window. The linear models can predict the data in the next 28 days.

Before using the machine learning models, we take the lagged data from the previous 15 days as the input. Then we implement the machine learning models with scikit-learn package. For each type of machine learning model, a total of 28 separate models are built with each model predicting one day. We also compared the different input features with and without calendar and prices information in different runs.

We further preprocessing the sales historical data with standardization, because the deep learning models are sensitive to the input data distribution. We can also predict a sequence for the next 28 days simultaneously with the deep learning models.

B. Results

We first show the results of the comparison of different input features in Table 1. Both the results with and without calendar and prices information are both shown and compared. From Table 1, we find that with more data, the model will perform better after adding information such as date and price. So the following results are using this information as input by default in these models.

TABLE I. A COMPARISON OF DIFFERENT INPUT FEATURES.

cat_id	state_id	Linear Regression	Linear Regression (+Calendar+Price)	Random Forest	Random Forest (+ Calendar +Price)	XGBoost	XGBoost (+Calendar+Price)	MLP	MLP (+Calendar+Price)
FOODS	CA	1014.91	829.57	1171.45	1139.99	1110.56	874.81	1158.41	996.76
HOBBIES	CA	223.16	216.60	229.91	229.25	240.89	217.87	226.12	284.67
HOUSEHOLD	CA	347.42	336.69	357.05	347.71	388.61	367.53	690.05	433.77
FOODS	TX	918.11	881.56	1012.62	967.73	972.21	974.08	774.69	1184.61
HOBBIES	TX	145.56	149.53	157.53	160.14	156.49	151.24	150.98	175.89
HOUSEHOLD	TX	337.67	352.04	363.77	365.22	340.84	390.68	316.77	336.82
FOODS	WI	1184.97	1641.85	1829.74	1860.10	1448.37	1440.27	1850.95	1321.45
HOBBIES	WI	116.63	115.80	119.77	128.20	111.99	138.03	155.41	137.64
HOUSEHOLD	WI	371.75	374.75	325.05	322.82	363.22	351.42	440.20	376.10
Average		517.80	544.27	618.54	613.46	570.35	545.10	640.40	583.08

We then show the performance of different models in Table 2. From Table 2, we found that complex machine learning or deep learning models have no advantages. On the whole, simple linear regression models have achieved the best results. From the perspective of a single time series, Triple Exponential Smoothing achieved the best results on five different series, while the LSTM model achieved the best results on two different series.

TABLE II. A COMPARISON OF DIFFERENT MODELS.

cat_id	state_id	Triple Exponential Smoothing	ARIMA	Linear Regression	Random Forest	XGBoost	MLP	LSTM
FOODS	CA	850.46	873.04	829.57	1139.99	874.81	996.76	1121.08
HOBBIES	CA	170.52	263.72	216.60	229.25	217.87	284.67	198.96
HOUSEHOLD	CA	282.91	353.08	336.69	347.71	367.53	433.77	428.68
FOODS	TX	1184.23	1125.19	881.56	967.73	974.08	1184.61	753.52
HOBBIES	TX	138.61	148.23	149.53	160.14	151.24	175.89	159.52
HOUSEHOLD	TX	334.37	341.08	352.04	365.22	390.68	336.82	322.28
FOODS	WI	1668.76	1983.60	1641.85	1860.10	1440.27	1321.45	1515.48

HOBBIES	WI	107.04	121.39	115.80	128.20	138.03	137.64	162.66
HOUSEHOLD	WI	302.45	343.36	374.75	322.82	351.42	376.10	355.83
Average		559.93	616.97	544.27	613.46	545.10	583.08	557.56

Finally, we show the predicted results from different models for the sales time series in CA in Figure 4-6.

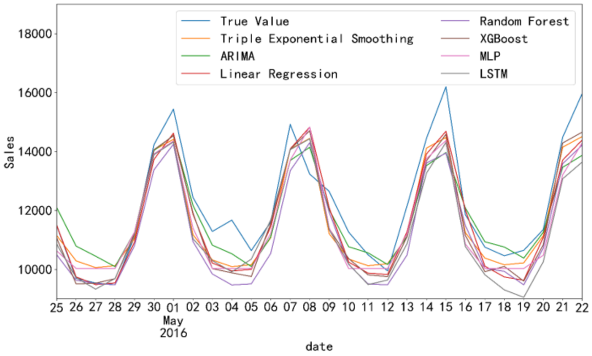

Figure 4. Predictions for CA foods.

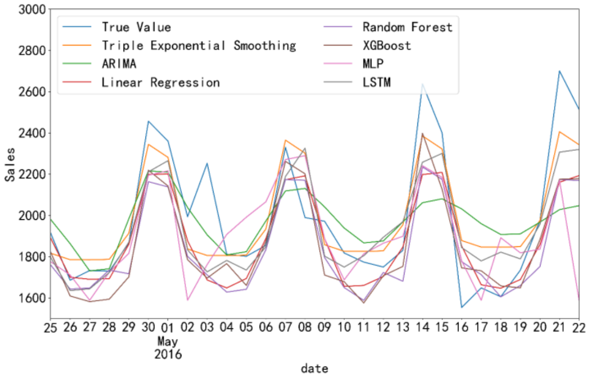

Figure 5. Predictions for CA hobbies.

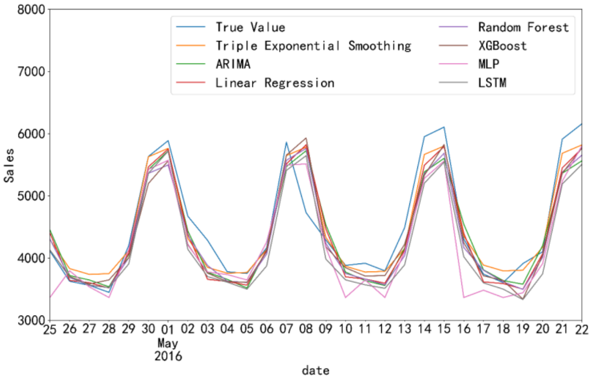

Figure 6. Predictions for CA household.

VI. Conclusion

In this paper, the data set we use covers 3 categories, 7 product departments, and a total of 3049 products, involving ten stores in 3 States. The sales history of these products is 1,941 days or 5.4 years. We first merged the data set, merged the information corresponding to the date and the price-related information into the time series to be predicted, and then tried and compared the prediction models. We focused on trying two linear models, three machine learning models and two deep learning models. By observing the two performance metrics of training time and RMSE, it is found that adding more date and price information is helpful for sales prediction, but machine learning and deep learning models have no obvious advantages in sales prediction. This article only studies some of the predictive indicators in the attribute characteristics of the first type of products or merchants. There are still many influencing factors that have not been taken into consideration. After adding more indicators that may affect e-commerce sales prediction, the performance of the sales prediction model can change. As an indication for further research, the data set used in this article is not very large. Whether there is a significant difference in the performance of machine learning and deep learning on data sets of different sizes can be further explored.

References

[1] Alibaba financial report. Online: http://emweb.eastmoney.com/pc_usf10/FinancialAnalysis/index?color=web&code=BABA.N. Accessed on 2020/12/12.

[2] Amazon's financial report. Online: http://emweb.eastmoney.com/pc_usf10/FinancialAnalysis/index?color=web&code=AMZN.O. Assessed on 2020/12/12.

[3] Li M, Ji S, Liu G. Forecasting of Chinese E-Commerce Sales: An Empirical Comparison of ARIMA, Nonlinear Autoregressive Neural Network, and a Combined ARIMA-NARNN Model[J]. Mathematical Problems in Engineering, 2018, 2018.

[4] Kharfan M, Chan V W K, Efendigil T F. A data-driven forecasting approach for newly launched seasonal products by leveraging machine-learning approaches[J]. Annals of Operations Research, 2020: 1-16.

[5] Ji S, Wang X, Zhao W, et al. An application of a three-stage XGBoost-based model to sales forecasting of a cross-border E-commerce enterprise[J]. Mathematical Problems in Engineering, 2019, 2019.

[6] Sharma S K, Chakraborti S, Jha T. Analysis of book sales prediction at Amazon marketplace in India: a machine learning approach[J]. Information Systems and e-Business Management, 2019, 17(2-4): 261-284.

[7] Zhang B, Tan R, Lin C J. Forecasting of e-commerce transaction volume using a hybrid of extreme learning machine and improved moth-flame optimization algorithm[J]. Applied Intelligence, 2020: 1-14.

[8] Tsai K H, Wang Y S, Kuo H Y, et al. Multi-Source Learning for Sales Prediction[C]//2017 Conference on Technologies and Applications of Artificial Intelligence (TAAI). IEEE, 2017: 148-153.

[9] Li J, Wang Y, Zhao X. Forecast Method of Commodity Sales of E-commerce Enterprises[J]. Statistics & Decision, 2018, 12: 176-179.

[10] Bandara K, Shi P, Bergmeir C, et al. Sales demand forecast in e-commerce using a long short-term memory neural network methodology[C]//International Conference on Neural Information Processing. Springer, Cham, 2019: 462-474.

[11] Pan H, Zhou H. Study on convolutional neural network and its application in data mining and sales forecasting for E-commerce[J]. Electron. Commer. Res., 2020, 20(2): 297-320.

[12] Jiang W, Zhang L. Geospatial data to images: A deep-learning framework for traffic forecasting[J]. Tsinghua Science and Technology, 2018, 24(1): 52-64.

[13] Jiang W. Applications of deep learning in stock market prediction: recent progress[J]. arXiv preprint arXiv:2003.01859, 2020.

[14] Zhao Z, Xu H. Short-term Forecast of Commodity Sales Volume Based on E-commerce Network Data Mining[J]. Logistics Sci-Tech, 2019, 8: 1-7.

[15] Wang J. Online Sales Volume Prediction Based on Items Clustering[J]. Computer Systems & Applications, 2016, 25(10): 162-168.

[16] Jiang W, Zhang L. Edge-siamnet and edge-triplenet: New deep learning models for handwritten numeral recognition[J]. IEICE Transactions on Information and Systems, 2020, 103(3): 720-723.

[17] Jiang W. Time series classification: nearest neighbor versus deep learning models[J]. SN Applied Sciences, 2020, 2(4): 1-17.

[18] Liu S, Li X, Zhao R, et al. Sales forecast of electric business promotion activities based on data mining[J]. Intelligent Computer and Applications, 2019, 9: 338-340.

[19] Shih Y S, Lin M H. A LSTM Approach for Sales Forecasting of Goods with Short-Term Demands in E-Commerce[C]//Asian Conference on Intelligent Information and Database Systems. Springer, Cham, 2019: 244-256.

[20] Zhuang Q, Zhang X, Wang P, et al. A Neural Network Model for China B2C E-Commerce Sales Forecast Based on Promotional Factors and Historical Data[C]//2019 International Conference on Economic Management and Model Engineering (ICEMME). IEEE, 2019: 307-312.

[21] Rong F, Guo M. On Suitability of Online Product Sales Prediction Model Based on Convolutional Neural Networks[J]. J. Northwest Minzu University, 2019, 2: 15-26.

[22] He X, Ma S, Wu Y, et al. E-Commerce Product Sales Forecast with Multi-Dimensional Index Integration Under Small Sample [J]. Computer Engineering and Applications, 2019, 55(15): 177-184.