International Journal of Computer Applications Technology and Research
Volume 8–Issue 04, 116-121, 2019, ISSN:-2319–8656

An Approach of Data Processing and Sales Prediction Model for E-commerce Platform

Guangpu Chen
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, Sichuan

Zhan Wen
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, Sichuan

Yahui Chen
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, Sichuan

Yuwen Pan
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, Sichuan

Xia Zu
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, Sichuan

Wenzao Li
School of Communication Engineering
Chengdu University of Information Technology
Chengdu, Sichuan
School of Computing Science,
Simon Fraser University,
Burnaby BC, V5A 1S6, Canada

Abstract: For the e-commerce platform, obtaining sales status of the stores is important for formulating sale strategies, risk assessment plans and loan appraisal. The traditional way to obtain sales status is mainly based on the subjective judgment of relevant practitioners and the analysis of mathematical statistical models composed of historical data. These methods are inaccurate and too dependent on people's judgment. Therefore, using data mining and machine learning technology to predict the sales amount came into being. In this paper, we propose a method to process a great deal of data from China's famous e-commerce platform called Jingdong. This method can make the messy data become uniform data sets which are more suitable for machine learning. Based on the uniform data sets, two sales prediction models are used to predict the sales amount of the stores in Jingdong. In experiment, 9-month historical sales and behavior data of 10,000 stores in Jingdong platform are processed by the proposed method. Furthermore, two prediction models including GBDT(Gradient Boosting Decision Tree)+DNN(Deep neural network) and GA(genetic algorithm) are used to predict the sales amount of stores in following 3 months. To verify the accuracy of the prediction, we import WMAE（Weighted Mean Average Error）score. In experimental results, the best WMAE is 0.39, which means accuracy is 61%. It shows the method of data processing and prediction models are effective compare with other models. This indicates the proposed method and model can be used for sales prediction in e-commerce platform.

Keywords: e-commerce; sales prediction; big data; data mining; machine learning

1. INTRODUCTION

Recent years, the rise of e-commerce platforms in China has led to online shopping becoming part of people's daily lives. The online shopping economy has become an inseparable part of China's economic development. Forecasting and analyzing plays an important role when the platforms try to make the right business decisions. Along with the explosion data, the traditional statistical analysis predict method has been unable to adapt to the ever-changing market. The sales forecast is also more favored for emerging Technologies such as big data and machine learning.

At present, the forecast of sales mainly has the following major categories, one is based on user research, the other is based on commodity research, and the third is hybrid research. These three types of research methods are based on the research of traditional store sales models. In the face of the rise of e-commerce platforms in the current environment, traditional forecasting methods are gradually unable to apply. There is still a lot of research space in the sales forecasting research of e-commerce platform in the new era of information.

Currently,Jingdong Financial Supply Chain has more than 100,000 companies and provided 250 billion yuan in loans. The majority of these enterprises are small, medium and small enterprises.Supply chain finance is a financial activity based on the industrial supply chain. Its purpose is to support financial activities through the industrial supply chain.In the past few years, based on the accumulation of Jingdong Big Data, Jingdong Finance has successively launched three core products of Jing Baobei, Jing Small Loan and Movable Finance, which greatly improved the financing difficulties and high financing costs of small and micro enterprises.In order to achieve accurate subsidies for each store, regular measurement and tracking of the operation status of each store has become an important evaluation standard for loans.Only accurate estimates of the future sales of the store can accurately assess their funding needs and set a reasonable loan amount. This paper will establish a forecasting model through the past sales records, product information, product evaluation, advertising costs and other information of each store on the platform, and predict the sales of each store in the next three months.

This study is based on the sales records, product information, product evaluation, and advertising costs of 10,000 stores on the Jingdong platform for the past 9 months. In order to predict sale amount in the following three months, we need to dig deeper into these data and build a machine learning prediction model. In the research, firstly, we will analyze and optimize the original data to make it more relevant to the model. Secondly, we will fit and predict the data through a machine learning model. Finally, we selected the best model' parameters that can accurately predict future sales by comparing the weighted mean absolute error (WMAE) of the predicted results.

This paper mainly has two contributions: First, we propose a method of data processing for e-commerce platform. This method can make these data become uniform data sets and it's more suitable for machine learning. Second, based on the processed data sets, a sale prediction model is used to predict the sale amount of the stores in e-commerce platform. It can provide a reference for e-commerce platform to provide loan for stores.

The rest of the paper is organized as follows: Sect.2 Introduces research on topics similar to this paper and detail the theoretical knowledge of the research methods. Section 3 presents the detailed process of the experiment. Sect.4 presents the analysis and conclusion of the result of the experiment. The five section presents the shortcomings of the experiment and propose some suggestions for the improvement.

2. RELATED WORKS

2.1 Research Status of Sales Forecast of E-commerce Platform

At present, the forecast of sales mainly has the following major categories, one is based on user research, the other is based on commodity research, and the third is hybrid research. These three types of research methods are based on the research of traditional store sales models. In the face of the rise of e-commerce platforms in the Internet environment, traditional forecasting methods are gradually unsuitable[1]. There is still a lot of research space in the sales forecasting research of e-commerce platform in the new era of information.

2.2 Feature Engineering

Feature engineering is an engineering activity that maximizes the extraction of features from raw data for use by algorithms and models[2]. Feature processing is the core part of feature engineering, including data preprocessing, feature selection, dimensionality reduction, and so on.In this project, we will use the sklearn library in Python to implement a series of operations on feature engineering.

2.3 Principal Component Analysis(PCA)

Principal component analysis(PCA) is a method for dimension reduction of features. It replaces the original features with a concise number of new features[3]. These new features are linear combinations of original features. It combinations maximize sample variance and try to make the new features uncorrelated. This makes it easy to study and analyze the influence of each feature on the model, and effectively reduce the complexity of models and increase the speed of the training model.

2.4 Gradient Boosting Decision Tree

GBDT(Gradient Boosting Decision Tree) is an iterative decision tree algorithm consisting of multiple decision trees[4]. All trees vote for the final result.It is recognized as one of algorithms with strong generalization.The idea of GBDT can be explained by a simple example. If a person is 30 years old, we first use 20 years old to fit and find that the loss is 10 years old. Then, we use 6 years old to fit the remaining losses and find the gap is 4 years old. In the third round ,we used 3 years old to fit the remaining gap and find the gap is only one year old. If the number of iterations is not yet complete, we can continue to iterate below. For each iteration, the age error of the fit will decrease.The specific steps of the algorithm are as follows.

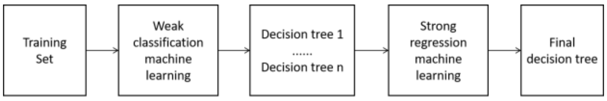
Training Set
Weak classification machine learning
Decision tree 1
......
Decision tree n
Strong regression machine learning
Final decision tree

Fig. 1:Process of Gradient Boosting Decision Tree

In this project, on the one hand, the algorithm is used to calculate the ranking of feature importance when crossing features, and on the other hand, the algorithm is used to transform features when building regression prediction models.

2.5 Deep Neural Network

DNN(Deep neural network) is a multi-layered neural network, also known as Multi-Layer perceptron(MLP).DNN is similar to the hierarchical structure of traditional neural networks. The system consists of a multi-layer network consisting of an input layer, a hidden layer (multilayer), and an output layer[5]. Only adjacent nodes have connections, the same layer, and cross-layer nodes. There is no connection between each other, and each layer can be regarded as a logistic regression model.Different from traditional neural networks, DNN adopts a forward-propagating training method.This is more suitable for processing high dimensional data.

In this project, we use DNN to process features transformed by GBDT and build prediction model.This can give full play to its advantages.

2.6 Genetic Algorithm

Genetic algorithm is a computational model that simulates the natural evolution of Darwin's biological evolution theory and the biological evolution process of genetic mechanism[6]. It is a method to search for optimal solutions by simulating natural evolutionary processes.In machine learning applications, the algorithm automatically selects the most appropriate model and parameters based on the input training data and tags, just like evolution.The steps of the algorithm are as follows.

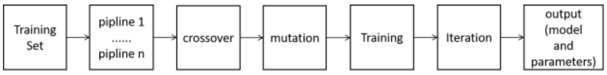

Fig. 2:Process of Genetic Algorithm

In this project, we use genetic algorithms to automatically select the algorithm that best fits this training set.We want to prove that model training can be finished and get good scores even without any machine learning foundation.We hope that we can promote the algorithm to make more effort on data analysis processing.

3. RESEARCH AND DESIGN OF SALES FORECAST MODEL FOR E-COMMERCE PLATFORM

3.1 Introduction to The Experimental Process

Based on the general process of machine learning training data, the experimental process of this paper is as follows.

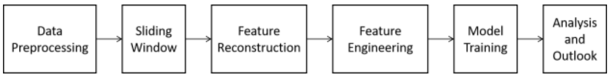

Fig. 3:Experimental Process

1)Data preprocessing:Perform preliminary inspection of origin data to handle missing, error, and abnormal data.

2)Sliding window:Divide origin data to fit model training and expand training set size to increase training coverage.

3)Feature reconstruction:Due to a mismatch between the original data and the predicted target,rebuild the training set based on the predicted target.

4)Feature engineering:Further optimize the training set generated by feature reconstruction and deeply mine more features that are beneficial to the training model.

5)Model training:The training set generated by the first few steps is input into a preset machine learning model, and the weighted mean absolute error (WMAE) of the output result is used as a measurement standard for selecting the optimal parameter.

6)Analysis and outlook:Analyze the results obtained from model training and draw conclusions.

3.2 Introduction of The Origin Data

This project uses the order quantity, sales, number of customers, evaluation number, advertising cost and other data of several stores within 270 days before 2017-04-30 provided by Jingdong Finance as training data, and sales of 90 days after the end of each month as labels.This project will use this data to build a predictive model to predict the total sales for the platform within 90 days after 2017-04-30. The original data structure used in this project is shown in Table 1 to Table 5.

Table 1. Orders data

Name	Type	Description	Example
shop_id	int	Store id	1631
Figures	Good	Similar	Very well
pid	int	Product id	41
ord_dt	date	Order date	2016/9/4
ord_cnt	int	Order count	1
sale_amt	float	Sales Amount	19.82
user_cnt	int	User count	1
rtn_amt	float	Return amount	0
rtn_cnt	int	Return count	0
offer_amt	float	Discount Amount	0
offer_cnt	int	Discount count	0

Table 2. Products data

Name	Type	Description	Example
shop_id	int	Store id	1014
pid	int	Product id	8141588
brand	int	Brand id	785
cate	int	Cate id	243
on_dt	date	put on shelves date	2017/2/9
off_dt	date	pull off shelves date	2017/5/1

Table 3. Comments data

Name	Type	Description	Example
shop_id	int	Store id	1494
good_num	int	Positive feedback	2
mid_num	int	Neutral feedback	0
bad_num	int	Negative feedback	0
dis_num	int	Number of share order	0
cmmt_num	int	Number of comments	2
create_dt	date	Create time	2016/8/7

Table 4. Advertisements data

Name	Type	Description	Example
shop_id	int	Store id	1036
charge	float	Advertising recharge	65298.6
consume	float	Advertising spending	25096.86
create_dt	date	Create time	2016/9/30

Table 5. Total sales data

Name	Type	Description	Example
shop_id	int	Store id	2143
sale_amt_3m	float	Sales within 90 days after the end of the month	72983.09
dt	date	Last day of the month	2016/12/31

As those tables shows,the data in Tables 1 to 4 will be used as training data,Table 5 is the target data.Among them, about 400W data in Table 1 account for 2G.,Table 2 has about 14W data, accounting for 0.8G,Table 3 is about 80W data, accounting for 0.5G,Table 4 is about 30W data, accounting for 0.3G.

3.3 Data Preprocessing

Data processing is mainly divided into three parts.The first is the processing of abnormal data.There are some obvious erroneous data in the original data used in this project,for example, the date on which the item is placed is later than the date of the removal, the number of return orders is more than the number of orders, etc.Since the overall proportion of such data is not high, this time it will be deleted directly.Secondly,there are still some missing data in the original data,mainly in Table 1, so we used the average of the last 6 days of missing data to fill in missing values.Finally, the purpose of this modeling is to predict sales for the three months after April 30, 2017, but the training data includes special festivals such as November 11, December 12(China Online Shopping Festival),for example, we plot the sales of the stores numbered 1630 in October and November as follows. Comparing the sales of these two months, it can be seen that there will be a significant abnormality in the sales of special event days (November 11).

To deal with this situation, the project adopted a smoothing treatment for these special sales days, using the average of the two days before and after the day instead of the day.

3.4 Sliding Window

The data used in this experiment belongs to time series data and contains a lot of time information.A time series is a sequence in which the values of the same statistical indicator are arranged in chronological order in which they occur.By processing such information through sliding window division, on the one hand, the training data set can be expanded, and on the other hand, the influence of time on the prediction effect of the model can be weakened as much as possible.The specific division method is as follows.

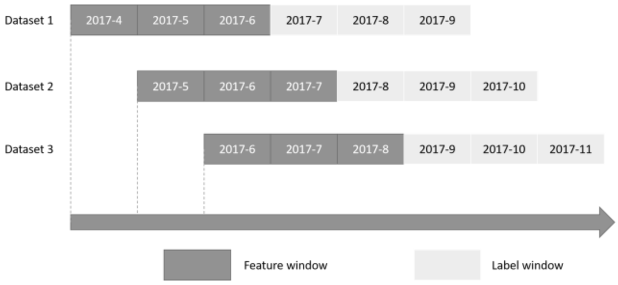

Fig. 4:Schematic Diagram of Sliding Window Division

3.5 Feature Reconstruction

As can be seen from the above tables, the original data of this experiment is a one-to-many situation, and the goal of our model is the sales of each store in the next 3 months. These multi-dimensional features cannot be used as input features of the prediction model. Therefore, it is necessary to perform a dimension reduction and reconstruction process on the original data through statistics, sampling, crossover, etc., so that the prediction model can be successfully constructed.

Financial data often accompany with multiple collinearity and financial leverage effects, discount, income, and cash flow are related to each other and have many combined features. Therefore, during the feature reconstruction, not only need comprehensive characteristics, but also need some sampling and crossing characteristics .The process of feature reconstruction is as follows.

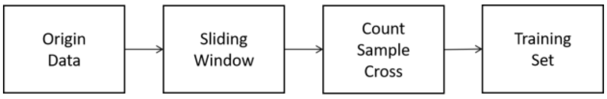

Fig. 5:Process of Feature Reconstruction

3.6 Feature Engineering

After feature reconstruction,we initially formed the training set required for the prediction model, and features that can use to build the model are 80 dimensions. But as in the field of data mining, the data and features determine the upper limit of machine learning, and the models and algorithms just approximate this upper limit.Therefore,we still need to mine more and better features,so the Feature engineering is proposed.

3.6.1 Data Cleaning

The first thing we need to do is to clean up the training set.The missing rate of training set is as follows.

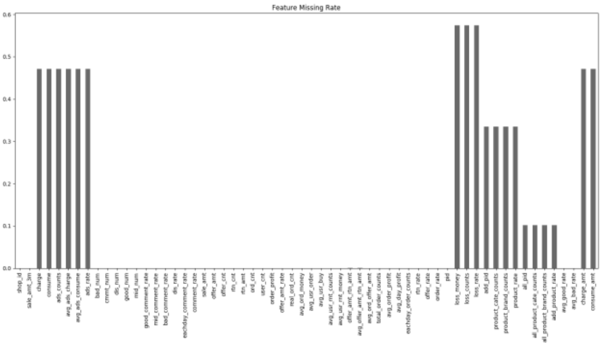

Fig. 6:Missing Rate of Training Set

As can be seen from the figure, about 20% of the features have missing data, including advertising features, order features, and product features.

For advertising features,the reasons for the missing are mainly the preference of the merchant and the size of the sliding window is smaller than the advertising investment period.In order to weaken the impact of this situation on the model, we set missing data as a special attribute value, which is different from any other attribute value and is filled with "-999" in the actual programming process.

Then about the order and product features,the main reason for the lack of data is due to the solution adopted by feature reconstruction,like there is no loss of orders or put on new goods in the sliding window period.Because the lack of these data is due to the absence of this situation,we fill all missing data with "0".This will make it work in the model.

3.6.2 Feature Crossing

Feature crossing is a mathematical combination of two or more category attributes into one feature.This project takes the Grid Search method on the feature crossing,make the reconstructed 60 features with no missing values are multiplied by each other to obtain a total of 870 features.Then, put the 870 features into the GBDT algorithm model as a training set for sales forecasting model.The parameters of GBDT are as follows.

After the training, use the method of ”get feature importance” to get the ranking of the importance of those features, and the top 100 are as follows.

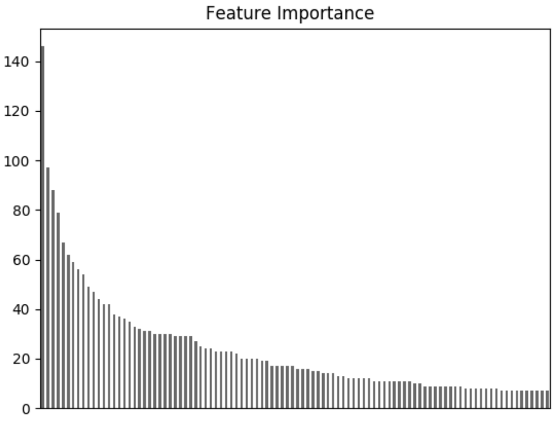

Fig. 7:Feature Importance of Cross-Features

3.6.3 *Feature Selection*

After the above processing, we have a total of about 200 features. Although it enhances the persuasiveness and prediction accuracy of the regression prediction model, it also increases the time for model training, and too many features make the model too complex and generalized, which is easy to cause dimensionality disaster. Therefore, in this project, we use PCA（principal component analysis）to select the most effective features from the original features to reduce the dimension of data set. The parameters of PCA are as follows.

After numerous experiments, the project got the highest score when reserve 90% features.So far, we have obtained a total of 161 features for model training.

3.7 Model Training

3.7.1 *GBDT+DNN*

The first model is constructed by GBDT combine DNN,the specific combination is as follows.

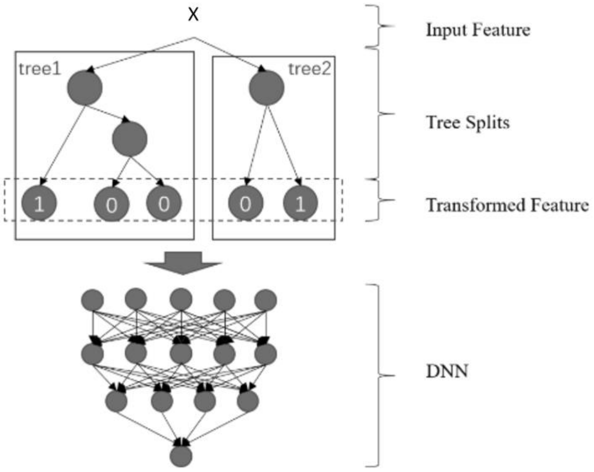

Fig. 8:Combination of GBDT and DNN

The purpose of this is to firstly exploit the advantages of GBDT algorithm mining feature combination. Second, DNN deep learning is more suitable for processing massive data and multi-dimensional input features. Through the combination of the two, the advantages of the respective algorithms can be fully utilized.The parameters are selected as follows.

3.7.2 *Genetic Algorithm*

The second model is built in two steps. First, the training data is input into the genetic algorithm to automatically select and iterate out the best regression prediction model suitable for the project. Secondly, the training set is input to the model of the first step output for training.The specific process is as follows.

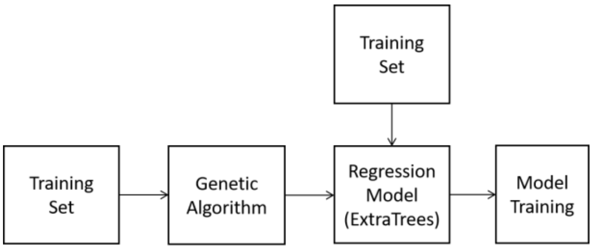

Fig. 9:Process of Genetic Algorithm

The purpose of this is to try genetic algorithm, a method of streamlined machine learning.This algorithm does not require an in-depth understanding of complex algorithm principles, and can automatically select the best predictive model based on input data.It will provide new ideas for establishing machine learning prediction models.The parameters of Genetic algorithm and the model output by it as follows.

4. ANALYSIS AND CONCLUSION

4.1 Result of Prediction

After a series of processing in Chapter 3, we generated the optimal training set used by the training model and selected the appropriate training model based on the characteristics of the training data.Through experiments,we found that feature engineering has greatly improved the model prediction results compared with different algorithms choices.We compared the scores and feature importance of the training sets that with or without feature engineering under each algorithm as follows.Also we used K-fold cross-validation in the experiment,and took WMAE as the evaluation criteria.The predicted scores for training sets used under different algorithms and with or without feature engineering are shown as follows.

Table 6. Scores of the training sets that with or without feature engineering under each algorithm

Models	WMAE (without feature engineering)	WMAE (with feature engineering)
LR	0.7853	0.7125
RF	0.5132	0.4589
SVR	0.4982	0.4464
GBDT	0.5044	0.4478
GBDT+DNN	0.4522	0.3982
GA	0.4500	0.3901

As can be seen from the table, firstly, the result has greatly improved under each algorithm when the training set is processed by feature engineering. Secondly, the two sets of algorithms finally adopted in this experiment have a certain improvement on the project compared with other regression prediction algorithms. These can prove the quality of the prediction.

4.2 Conclusion

As can be seen from the above table, both feature engineering and algorithm model selection have a certain improvement on regression prediction results.This proves that the project can effectively process data and predict future sales.After summarizing, this paper mainly has the following contributions:

1)Adding analysis of financial data in the process of data preprocessing and feature reconstruction.

2)Further mining of data through feature engineering and get a great improvement in scores.

3)After deep mining the characteristics of financial data, the method of inputting DNN after GBDT extended feature number was used and some improvements were successfully achieved.

4)Automatically select regression prediction models by genetic algorithm and obtain good scores, which provides new ideas for model selection.

5. PROSPECT

The desensitization data used in this study was provided by JD,therefore, we are unable to know the specific products sold in each store.I think it is very important for mining potential information.Second, we are unable to quantify the direct or indirect impact of unscheduled promotions and government policies on the platform, so these factors are not included in the training model, it may be benefit to the model.In addition, in terms of algorithms, the prediction effect can be further improved by means of model fusion.It will take a lot of time and resources to find better algorithms or fusion solutions.These outlooks are all issues that need to be considered in the future to explore such topics.

6. ACKONWLEDGEMENTS

The data used in this research is from JingDong Finance, and we are very grateful to the agency for supporting this experiment.

This research is supported by Science and Technology Department of Sichuan Province, Fund of Science and Technology Planning (No. 2018JY0290), Science and Technology Project of Sichuan Province(No. 2019YFS0493), Humanity and Social Science Research Project of Ministry of Education, Youth Foundation(17YJC190035) and Meteorological Information and Signal Processing Key Laboratory of Sichuan Higher Education Institutes (No. QXXCSYS201606)

7. REFERENCES

[1] Hui-Chih Hung, Yu-Chih Chiu, Huang-Chen Huang, et al. An enhanced application of Lotka–Volterra model to forecast the sales of two competing retail formats[J]. Computers & Industrial Engineering, 2017, 109:325-334.

[2] Goodness C. Aye, Mehmet Balcilar, Rangan Gupta, et al. Forecasting aggregate retail sales: The case of South Africa[J]. International Journal of Production Economics, 2015, 160:66-79.

[3] Andrés Martínez, Claudia Schmuck, Sergiy Pereverzyev, et al. A machine learning framework for customer purchase prediction in the non-contractual setting[J]. European Journal of Operational Research, 2018.

[4] Giuseppe Piras, Fabrizio Pini, Davide Astiaso Garcia. Correlations of PM10 concentrations in urban areas with vehicle fleet development, rain precipitation and diesel fuel sales[J]. Atmospheric Pollution Research, 2019.

[5] Patrícia Ramos, Nicolau Santos, Rui Rebelo. Performance of state space and ARIMA models for consumer retail sales forecasting[J]. Robotics and Computer-Integrated Manufacturing, 2015, 34:151-163.

[6] Bin Zhang, Dongxia Duan, Yurui Ma. Multi-product expedited ordering with demand forecast updates[J]. International Journal of Production Economics, 2018, 206:196-208.

[7] Maobin Li, Shouwen Ji, Gang Liu, et al. 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.

[8] Eric W. K. See-To, Eric W. T. Ngai. Customer reviews for demand distribution and sales nowcasting: a big data approach[J]. Annals of Operations Research, 2018, 270(1-2):415-431.

[9] A.L.D. Loureiro, V.L. Miguéis, Lucas F.M. da Silva. Exploring the use of deep neural networks for sales forecasting in fashion retail[J]. Decision Support Systems, 2018.

[10] G. Dellino, T. Laudadio, R. Mari, et al. Microforecasting methods for fresh food supply chain management: A computational study[J]. Mathematics and Computers in Simulation, 2018, 147:100-120.