Research on Precise Demand Forecast of Retail Commodities Based on LSTM-ARIMA Joint Model

Haoran Song¹ *, Dengyu Du², Yao Li³

¹College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, Jiangsu, China
²College of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, Jiangsu, China
³School of Economics, Nanjing University of Posts and Telecommunications, Nanjing 210023, Jiangsu, China

*Corresponding author: haoransong@njupt.edu.cn

Abstract. Aiming at the retail forecasting problem, this article uses software such as MATLAB, SPSS, and Python as tools, and uses dual correlation analysis, LSTM neural network-ARIMA joint model, and BP neural network-based subclass-skc fusion model, etc., to get an impact on skc sales. The importance of the factors and accurate forecasts of sales.

1. Introduction
My country's economic development has entered a new normal period, and consumption will become the most important "engine" for my country's economic growth. The retail industry has a special position in promoting consumption [1]. Under the trend of retail digitalization, new retail will emerge. Different from the traditional retail industry, due to the improvement of people’s living standards and the gradual emphasis on the pursuit of spiritual life, consumers are gradually no longer limited to buying goods at cheap prices, but paying more attention to the quality, characteristics, styles, etc.

It can be seen that the new retail industry pays more attention to the digitalization of stores and pays attention to the analysis and prediction of existing data. In the context of the "Internet +" development strategy proposed by the country, many opportunities are provided for the new retail industry, but there are also corresponding challenges. The new retail industry needs to formulate corresponding strategies to deal with it, and the production model will transition to a variety of types and small batches. The significant increase in the difficulty of managing inventory is the number one issue facing small businesses and retail stores.

Whether the historical sales data can be processed and analyzed, and as accurately as possible to predict the future sub-category level, and even the sales volume of the skc (single single color) level, is related to whether new retail companies can seize the era in this new retail trend. Opportunities to achieve sustainable development and rank among the forefront of the industry.

2. Model
We are required to perform statistical analysis on the data first, that is, to filter all the data in the attachment, calculate the total sales in a given time range, and then sort them in descending order
according to the sales volume to obtain the top 50 sales target skc. Further, the relevant data whose sales date is in the holiday period are screened out to obtain a data set with no outliers. Finally, you can use the skc data set to perform double correlation analysis and make a heat map.

2.1. Data preprocessing

Due to the large amount of data given, it is necessary to preprocess the data with Python and Excel first to better analyze the factors affecting sales. The steps for data screening and elimination are as follows:

Step1 Filter the top 50 skc: Filter the top 50 skc sales data in the attachment 1: sales info from July 1, 2018 to October 1, 2018. That is, calculate the sales of each skc according to formula (2), and then arrange them in descending order.

\[ \text{Sum}_j = \sum_{i=1}^{n} c_i \]  

In the formula, represents the number of skc types, and represents the total number of days of data contained from July 1, 2018 to October 1, 2018.

Then sort them in descending order according to the obtained sales of each type of skc, and filter the top 50 skc with sales from July 1, 2018 to October 1, 2018.

Step2 Eliminate unavailable data: Because from May to July 2018, some sub-category target data corresponding to skc, inventory information on holidays and other data are missing, so 50 skc must be screened again before analyzing influencing factors, Get the data set. A total of 17 skc containing complete data was finally obtained.

Analyze the outliers of the data left and analyze whether there is any abnormality in the single-day price of each skc, that is, first calculate the actual daily price of each skc, and divide the actual single-day cost by the single-day sales, that is

\[ \text{unitc}_i = \frac{c_i}{s_i} \]  

The discount can be calculated by dividing the actual selling price by the listed price:

\[ d_i = \frac{\text{unitc}_i}{\text{tag}_i} \]  

Draw a box plot and find that a set of skc contains outliers. After excluding unavailable data, 16 sets of skc are finally obtained. There is no abnormal data. Due to space limitations, this article only gives some of the actual daily selling prices of the 16 skc Box plot.
Figure 1. Box plot of part of skc after removing unavailable data.

Note: The upper and lower black lines represent the upper and lower edges, the upper and lower blue lines represent the upper and lower quartiles, and the red line represents the median.

Observing the box plot, there are no outliers outside the upper and lower edges, and the data set has passed the outlier test, and the data set can be used for further analysis.

Each skc and its influencing factors and sales related data table are shown in the attached Table 1.2 target test complete skc and influencing factor data, only part of the factor data corresponding to skc and the daily average sales of National Day are given in the article (as an average, so it contains decimals):

Table 1. Part of the factor data corresponding to skc and the average daily sales of National Day.

| tiny_class_code | tag_price | ie     | National Day holiday average sales | skc         |
|-----------------|-----------|--------|------------------------------------|-------------|
| 27050401        | 223.75    | 253.4286 | 0.8715                             | 690572118833 |
| 27050401        | 173.75    | 276.1429 | 0.8476                             | 790572118833 |
| 27050401        | 123.75    | 432.4286 | 0.8894                             | 896572118954 |
| 27071209        | 123.75    | 324.2857 | 0.8693                             | 496572333586 |
| 27112849        | 73.75     | 622.4286 | 0.8920                             | 302572775577 |
| 27164944        | 123.75    | 251.7143 | 0.9211                             | 496573321265 |
| 27196225        | 173.75    | 393.5714 | 0.8621                             | 102573650748 |

Note: tiny_class_code is the small class number, tag_price is the marked price, ie is the inventory quantity.

Step3: Descriptive statistics: Perform descriptive statistics on the selected skc data, and give descriptive statistics on the data during the 16 skc National Day, including the minimum, maximum, mean variance and skewness:

In terms of National Day, the overall sales of each single-color skc on the National Day are around 6, and the variance is not large, indicating that generally the sales of each skc are not much different. In terms of price, from 73 to 249 yuan, the average selling price is about 142 yuan. There is a large gap in the inventory of different skc, ranging from 81 to 622, with a large variance. The strength of the discount is generally stable, fluctuating around 0.88.

2.2. Pearson correlation coefficient model

The Pearson correlation coefficient was proposed by the British statistician Pearson in the 20th century. It measures the degree of linear correlation between two data sets [2], with a value range of [-1, 1], and a positive value means positive Correlation, a negative value indicates a negative correlation. If there are two variables X and Y, the Pearson correlation coefficient between the two is:
Under the condition that the null hypothesis is established, that is, under the premise that there is no correlation between the two variables, calculate the probability value (p value) that the two variables are not correlated. If the p value is small, it means that the two variables are not correlated. If the probability of is very small, we can reject the null hypothesis and accept the alternative hypothesis. If the significance level is 0.05, it means that the null hypothesis is accepted and rejected under the 95% confidence level. If p<0.05, then the original hypothesis can be rejected Hypothesis, accept the alternative hypothesis that there is a significant linear relationship between the two variables.

Therefore, when the p-value is much greater than 0.05, even if the correlation coefficient is large, we cannot say that there is an obvious correlation between the two variables, and generally we must first discuss the size of the correlation coefficient when the p-value meets the requirements.

2.3. Use Pearson correlation coefficient model to solve factor correlation
Because the Pearson correlation coefficient needs to meet the normal distribution in the population, we perform the Jarque-bera test (JB test) on the sample. After performing the JB test on the data, we found that basically all passed the test and accepted compliance. For the assumption of normal distribution, see Appendix 1.2 for the Matlab program. For the sake of intuition, given the Q-Q chart of the price, it can be seen that the sample points are basically close to a straight line, indicating that the sample basically obeys the normal distribution, as shown in Figure 2.3:

![Q-Q chart of price.](image)

After the normal distribution test is basically passed, the Spss software is used to longitudinally analyze the correlation coefficient between each skc and other factors except for the sub-category target and obtain the significance.

The correlation coefficients between sales volume and factors during different holidays are very different, which is caused by the different consumer psychology of consumers during different holidays.

1) Marked price: No matter what holiday, the marked price and the overall sales volume are negatively correlated. This is obvious. Consumers are more willing to buy high-quality and inexpensive goods. It is nothing more than double twelfth, double eleven and other festivals as shopping festivals. Value price-performance ratio.
2) Inventory: Under normal circumstances, we think that if the inventory is large, the sales of skc will be high. This is a characteristic of non-shopping festivals. For example, National Day, a festival that is not mainly shopping, people choose to buy goods and are more inclined to buy goods Plenty of it. On the Double Twelfth Shopping Festival, whether it is offline physical stores or online retail online stores, there are measures to prepay and restock. This is a commonly used "hunger marketing" strategy for merchants, so there will be less inventory and sales during Double Twelve. Instead, it is higher.

3) Discount: When the discount is negatively correlated, the sales volume is high when the discount is strong. Chinese people pay more attention to face. Most of the consumption during the National Day comes from shopping in physical stores, and discounts are not high during this holiday, so the impact of discounts on sales is not significant. However, Double Twelve is a shopping festival, and merchants have a lot of discounts. People also hope to consume by grabbing shopping coupons. Sales are significantly related to discounts.

Next, horizontally analyze the impact of sub-categories on the sales of skc. The so-called horizontal comparison is to compare the correlation coefficients of various factors and sales in different sub-categories on the same holiday to compare whether different sub-categories have a significant impact on sales.

Also, during the National Day period, the correlation coefficient between the sub-category 27050401 and the sub-category 27196225 and the sales volume is very different. This is caused by the attributes of the product. The latter has a large correlation coefficient with inventory and only a weak correlation with discounts. This shows that the problem we studied is likely to be seasonal clothing products, the latter is likely to be seasonal autumn clothing, while the former has a large correlation coefficient with discounts, indicating that this small category may be out-of-season discounted products.

For the sake of rigor, we also conducted partial correlation analysis. High-order partial correlation analysis is to analyze the correlation between two sets of variables under the condition of controlling the linear influence of other variables, so that the influence of other variables can be eliminated [3]. The results show that the partial correlation coefficients are basically consistent with the results of correlation analysis.

To sum up, combined with dual correlation analysis: Pearson correlation analysis and partial correlation analysis, the factors that affect skc sales during the four holidays have different rankings. Generally speaking, the vertical influence is important. Price>discount>inventory, but the specifics are the ranking of different festivals has changed, and different sub-categories determine the overall ranking of other influencing factors from a horizontal perspective. This result is in line with consumer psychology and is also of practical significance.

3. Model evaluation and promotion
This article mainly solves the problem of determining the influence factors of skc sales volume and forecasting future sales volume. A number of models have been established to analyze and solve the problem in detail. The specific advantages and disadvantages are as follows:

3.1. Advantages and disadvantages of the model:

3.1.1. advantage: 1. In the data screening, we use Python to process the data format, and use MATLAB to draw the data outlier box plot, which greatly improves the efficiency of only using Excel to clean the data.
               2. In the whole process of relevant analysis and forecasting, a variety of visualizations are used for intuitive display. The amount of data is very large, and the results obtained are tables with thousands of rows at every turn. We use them as time series diagrams and thermal Figures and so on are displayed, giving the results very well.
               3. Introducing the LSTM neural network into the traditional ARIMA algorithm can greatly enhance the accuracy and robustness of the prediction model, which allows our prediction model to be more widely used in the prediction of enterprise sales.
4. Combine the BP neural network with the weight coefficients we defined to realize the prediction that considers the factors of small classes and holidays. The introduction of the weight coefficient is in line with the actual trend of skc sales.

3.1.2. Disadvantages: 1. Due to the large amount of data, not all the data are complete enough, so the correlation coefficient of some skc is not given because there are many missing data for related factors.
2. When using the time series ARIMA model for forecasting, we consider June-September 2019 as the original data. In fact, if the sales data for 2018 are also taken into account, the seasonal characteristics of holidays can be more clearly reflected.

3.2. Promotion of the model:
1. The factors considered in the correlation analysis are mainly price, discount and inventory, which are based on the digital results of retail stores. Horizontal promotion: You can consider the upstream data of the supply chain, such as the cost of ingredients, transportation costs, etc., to conduct a combination correlation analysis of multiple factors. Longitudinal view: You can add more customer factors, such as buying behavior, return rate, click-through rate of online e-commerce, search engine search data, etc. In the era of big data, the richer the source and type of data, the more you can find the hidden relationship between sales volume.
2. The prediction based on the LSTM neural network-ARIMA joint model has been able to predict future sales more accurately, but adding the various types of data mentioned above as input to retrain and improve the neural network, and select the appropriate activation function and hidden layer, can get more accurate retail demand forecast, and then inject new vitality into the retail industry.

The small class-sk fusion model based on BP neural network is also not a perfect model. The form of weight coefficients in the model is relatively simple. If you can find more suitable weight coefficients through continuous attempts, then the result of the fusion model Will be better.

References
[1] Wang Haibo. Research on the evolution of retail formats in my country [D]. Beijing Jiaotong University, 2016.
[2] Manfred Mudelsee. Estimating Pearson's Correlation Coefficient with Bootstrap Confidence Interval from Serially Dependent Time Series[J]. Kluwer Academic Publishers-Plenum Publishers, 2003, 35(6).
[3] Yan Likun. Application of correlation coefficient and partial correlation coefficient in correlation analysis [J]. Yunnan Institute of Finance and Trade, 2003(03): 78-80.
[4] Zhai Jing, Cao Jun. Combination forecasting model based on time series ARIMA and BP neural network [J]. Statistics and Decision, 2016(04): 29-32.
[5] Cui Tian. Research on Brand Automobile Sales Forecast Based on Internet Search Data [D]. Xi'an University of Technology, 2019.
[6] Xu Weize. Forecasting economic and financial time series: Comparison of ARIMA and LSTM models [J]. Shanxi Agricultural Economics, 2020(03): 36-37.
[7] Cheng Zhiyi. Research on the transformation and upgrading of my country's traditional retail industry to the O2O model [D]. Harbin University of Commerce, 2015.
[8] Zhang Qian. Research on simulation and optimization of customer experience under the O2O model of traditional retail industry [D]. Zhengzhou University, 2015.