# Research on Product Sales Forecast Based on Test Sales Comment Data

Li Li*, Hao Li, Jiamei Yang

School of economics and management,Nanjing University of Science and Technology Nanjing, Jiangsu, China

* Corresponding Author Email: levi98h@foxmail.com

**Abstract.** This study predicts product sales based on product sales during the trial period and text mining of online reviews of products. This study uses the BERT model to judge the sentiment tendency of reviews and identify product attributes, build a sales forecast model based on the extracted review features and the sales and product features during the trial sales period, and evaluate the effect of the forecast model. Among the prediction models constructed in the study, the multiple linear regression model has the worst effect, the prediction effect of random forest and BP neural network is close, and the prediction effect of XGBoost model is the best.

**Keywords:** Test Sales, Online Reviews, Sales Forecast.

## 1. Introduction

Test marketing (also known as market testing) is a process that many new products go through before they reach commercialization. Through product test marketing, on the one hand, the market potential can be judged through the sales volume of the product, so as to make actual production arrangements. On the other hand, in the product evaluation given by consumers, the enterprise can learn the market feedback of the product in detail. At present, there are few studies on trial sales, the data during the trial sales period is not fully mined, and there are few studies on how the data during the trial sales period can provide ideas for enterprises to improve new products, marketing strategies.

## 2. Literature Review

### 2.1 Research on sales forecast methods

The existing quantitative data prediction methods are based on the relevant theories, according to the object to be predicted and the existing data, select a suitable mathematical model, reveal the relationship between different factors. Analysis and forecast of the future development trend and status. He Xijun [1] and others built a prediction model based on ensemble learning XGboost (eXtreme Gradient Boosting) to predict mobile phone sales. The results show that the prediction accuracy of this model is significantly improved compared with BP (Back Propagation) neural network, SVM (Support Vector Machine) and BP-SVM combined prediction model.

Through previous research, it can be found that quantitative forecasting is often used to predict future sales based on historical data. At the same time, considering a variety of influencing factors and the high learning ability of machine learning methods, many scholars use machine learning methods to make predictions.

### 2.2 The impact of online reviews on sales

Existing studies have shown that there is a close relationship between product online review data and product sales. Based on Amazon's camera online review data, Wang Wei [2] and others combined sentiment analysis technology and econometric model, and the research shows that product reviews have a positive impact on product sales. Scholars such as Zhang Xinyue [3] collected the top 50 sales samples of Dangdang.com, established a regression model, and concluded that product sales are largely affected by the number of online reviews.

When studying the impact of online product reviews on product sales, most current studies choose to quantify online product reviews from the following two types of indicators. The first category is to describe the intuitive features of reviews, including the number of reviews, the length of reviews, and the responses to reviews; the second category of indicators is to describe the content features of online product reviews, and to study the sentiment tendency of reviews through text mining of actual product reviews.

# 3 PREDICTION MODEL PROPOSED

## 3.1 Determination of influencing factors

There are many influencing factors of product sales, and different scholars choose different prediction indicators when forecasting product sales.

**Table.1.** Selection of factors affecting sales

| Author | Influencing factors | Dependent variable |
|---|---|---|
| Chitragupta [4], et al. | Comment valence, number of comments, variance of comments | Movie Box Office |
| Pillow [5], et al. | Product promotion information, price, daily sales | Product sales |
| Jimenez [6], et al. | Historical sales data, promotional information, e-commerce product features | Product sales |

## 3.2 Comment features

As can be seen from Table 1, in the existing research, the influencing factors of product sales mainly include two categories, one is the characteristics of product reviews and related derivative indicators, including the number of reviews, the length of reviews and other intuitive basic features of reviews. As well as the content characteristics of comments such as comment sentiment, comment differences, etc., which involve many aspects of comment characteristics; the other type is the basic attribute characteristics of products, including sales price, product type, etc., and some studies will also consider historical sales or history. Factors such as sales over the same period. Therefore, this study will determine the influencing factors of product sales from the above aspects, so as to build a product sales forecast model.

(1) Analysis of product characteristics

The price of the product and the type of product are important factors that need to be taken into account in the forecasting model when making all product sales forecasts. In the field of economics, product price is related to sales volume, and user demand is greatly affected by product price. In addition, the sales of different types of products will also have significant differences, and the type of products is also one of the factors that will affect the sales of products. In the process of constructing the prediction model, since the product type is a categorical variable, in this study, the categorical variable of the product type is converted into a dummy variable, and then added to the prediction model for research.

The main work of this study is to forecast sales based on product test sales, so sales during the test sales period are also an important factor that needs to be considered. Therefore, when building a forecast model for a product, it is necessary to consider the impact of product price and product type, and the sales volume during the trial sales period of the product will also be added to the forecast model.

(2) Sentiment feature analysis

For the unstructured text corpus of online product reviews, the use of machine learning methods can better analyze product reviews, and the newly proposed BERT model (Bidirectional Encoder Representations from Transformers) performs well on many natural language processing tasks, and its practical application is relatively low. The BERT model pre-trains the network on a large number

of corpora, initializes the parameters, and then uses a small amount of professional domain corpus to fine-tune the model based on the pre-training results, and finally achieves better results.

### 3.3 Prediction model construction

According to the summary of the previous section, the relevant variables and specific measurement methods selected for sales forecast in this study are shown in the table. The independent variables include the product's review characteristics, the product attribute characteristics included in the review, the characteristics of the product itself, and the past sales of the product, a total of 15 variables, and the dependent variable is the future sales of the product. According to the variables and their measures in Table 2, a sales forecast model of the product is constructed to predict the future sales of the product.

**Table.2.** Description of relevant variables for sales forecast

| Variable category | | Variable names | Variable measure |
|---|---|---|---|
| Dependent variable | | Product sales | Product sales |
| Argument | Comment Basic Characteristic | Comment emotion | Comment sentiment value |
| | | Comment length | Overall number of comments |
| | | Number of comments | Whether the comment was replied to |
| | | Comment reply | Average review text length |
| | | Number of comments | Overall number of comments |
| | Comment on product features | Exterior | Sentiment value of product appearance |
| | | Comfort | Emotional Inclination Value of Product Comfort |
| | | Design | Emotional Inclination Value of Product Design |
| | | Material | Emotional tendency value of product material |
| | | Cost-effective | The emotional preference value of price-performance ratio |
| | | Size | Sentiment value of product size |
| | | Work | Emotional inclination value of product workmanship |
| | Product features | Product price | Product selling price |
| | | Product type | Type of product |
| | Test sales | Product test sales | Product trial period sales |

## 4 Empirical research

The data of this study comes from an e-commerce website, which provides consumers with a large number of clothing products, and users' purchases and feedback form a large number of orders and comments.

### 4.1 Product feature extraction and comment sentiment analysis

The programming language used in this experiment is python, and the deep learning framework is Tensorflow. The pre-training model used in this experiment is the BERT-Base, Chinese model provided by Google. The network structure of the pre-training model has a total of 12 layers, the hidden layer has 768 dimensions, adopts a 12-head mode, and has a total of 110M parameters. The Chinese model used in this experiment is pre-trained on Chinese Wikipedia articles, and this experiment uses the curated review data to fine-tune this language model.The optimizer used during the training of the model is Adam, the learning rate is 5e-5, and the sequence length is set to 128.

The evaluation indicators of the prediction results of the final BERT model are shown in Table 3. In the three-category process of review sentiment tendency, the model prediction result has a precision rate of 0.913, a recall rate of 0.912, and an F1 value of 0.912; while in the classification process of product attribute features, the average precision rate is 0.946, and the average recall rate is 0.916, with an average F1 value of 0.931. It can be seen that the recall rate of the two classification problems is not much different, and the precision rate and F1 value of product attribute features are slightly higher than the three classifications of emotional tendencies, but the evaluation index values of the two classification problems are all above 0.9, it can be seen that the effect of the BERT model is relatively good in both emotional tendency judgment and identification of product attributes contained in reviews.

**Table.3.** BERT model result evaluation index table

| Evaluation indicators | Emotional tendencies | Product attributes |
|---|---|---|
| Accuracy | 0.913 | 0.946 |
| Recall | 0.912 | 0.916 |
| F1 | 0.912 | 0.931 |

### 4.2 Research on product sales forecast

(1) Selection of test sales time

In this study, the weekly sales changes of six product types in the early stage of sales were calculated, and the results are shown in Figure 1. As can be seen from Figure 1, the sales of the six types of products will have a large increase in the fourth week of sales, and the changes in sales in the first three weeks are relatively stable. Therefore, in this study, the product sales the first 3 weeks are used as the trial sales period of the product, and the product reviews and sales during this period are studied, so as to provide a reference for the company's sales forecast and product optimization.

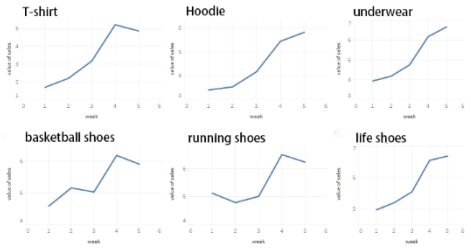



**Figure 1.** Product sales trends

(2) Selection of test sales time

In order to match the actual operation situation and provide a better reference for the company's product strategy after trial marketing, in this study, we use the new product data of an e-commerce website from January 2017 to December 2019, and we use the new product data from January 2017 to June 2019 as the training set, and use the new product from July 2019 to December 2019 as the test set. Use the training set to train the model, use the obtained training model to predict the product sales of the test set, and compare the prediction result with the actual sales data to obtain the prediction accuracy of each prediction model. For enterprises, the increase and decrease ratio of product sales are relatively concerned, so the MAPE value of the model prediction result is used as one of the reference indicators to measure the error rate of the prediction model. In addition, in addition to measuring the error rate in addition to the indicators, the RMSE value of the model prediction results is also used to evaluate the overall error of the model.

The multiple linear regression model, BP neural network model, random forest model and XGBoost model are constructed, and the model is trained on the training set, and the optimal parameter values of the model are obtained. After that, the trained model is used to predict the product sales in the test set, and the prediction results of each model are evaluated. The prediction results of each model are shown in Table 4.

**Table.4.** Comparison of prediction results of each model

| Predictive model | RMSE | MAPE |
|---|---|---|
| Multiple Linear Regression | 1.099 | 19.957% |
| BP neural network | 0.851 | 15.259% |
| Random forest | 0.827 | 13.357% |
| XGBoost | 0.593 | 8.399% |

As can be seen from Table 4, on the whole, the errors of the prediction results of the four prediction models are all below 20%, among which the multiple linear regression model has the worst prediction effect, the MAPE value is close to 20%, and the BP neural network model and random. The prediction results of the forest model are not much different, and the prediction effect of random forest is slightly better. Among the four models, the XGBoost model has the best prediction effect on this test set, and it is also the only prediction model with an error of less than 10%. The comparison between the prediction results of each model and the actual sales volume of the product is shown in Figure 2.

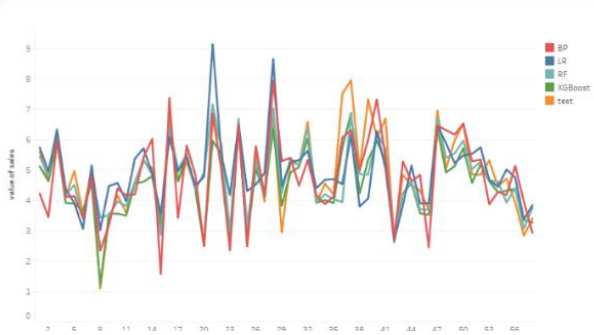



**Figure 2.** Prediction results of each model5 conclusion

## 5 Conclusions

In this study, the BERT model was used to identify the sentiment tendency of reviews and the product attribute features included in the reviews. The BERT model performed well in the three evaluation indicators of precision, recall and F value.

When constructing the sales forecast model, this study considered the influence of sales during the trial period, product attributes, review features, and product features included in the reviews, and constructed a multiple linear regression forecast model and a BP neural network forecast model and XGBoost prediction model. The results show that the prediction performance of the XGBoost model is the best.

## References

[1] He Xijun, Ma Shan, Wu Yuying, Jiang Guorui. E-commerce product sales forecast based on multi-dimensional index fusion under small sample [J]. Computer Engineering and Applications, 2019, 55(15): 177-184.

[2] Wang Wei,Wang Hongwei. Influence of Featured Views on Purchase Intention: A Sentiment Analysis Approach for Online Reviews [J]. Systems Engineering Theory and Practice, 2016, 36(1): 1-14.

[3] Zhang Xinyue, Sun Kelin, Lin Boxia, et al. An Empirical Study of the Impact of Online Reviews on Merchandise Sales [J]. E-commerce, 2017(7): 51-52.

[4] Chintagunta P K, Gopinath S, Venkataraman S. The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets [J]. Marketing Science, 2010, 29(5): 944-957.

[5] Pillo G D, Latorre V, Lucidi S, et al. An Application of Support Vector Machines to Sales Forecasting Under Promotions[J].4OR, 2016, 14 (3) :309-325.

[6] Jimenez F, Snchez G, Garca J M, et al. Multi-objective Evolutionary Feature Selection for Online Sale Forecasting[J]. Neurocomputing, 2017, 234 (C) :75-92.