2024 6th International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) | 979-8-3315-4179-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/MLBDBI63974.2024.10823944

Sales Volume Forecast of Sports Fashion Clothing Based on Machine Learning Algorithm
A LGBM-SHAP Approach

Ben Niu
Beijing Institute of Fashion Technology
Beijing, China
Niuben@bift.edu.cn

Yuechen Gao
Beijing Economic Manacement School
Beijing, China
1078510566@qq.com

Ruiliang Guo
Beijing Institute of Fashion Technology
Beijing, China
20210010@bift.edu.cn

Xinyan Shao
Beijing Institute of Fashion Technology
Beijing, China
1188919@qq.com

Weijuan Zhang*
Beijing Institute of Fashion Technology
Beijing, China
* 19045402@qq.com

***Abstract*—The forecast of sales volume is of great significance to the development of e-commerce. In this paper, a sales forecasting model based on LGBT and SHAPE algorithm is proposed. The LGBT model is used to train the forecasting model, and the SHAPE model is used to quantitatively analyze the contribution of each feature, which not only retains the good forecasting effect of machine learning, but also improves the interpretability of the results. Based on the daily frequency data of a sports brand, this paper finds that the prediction effect of the model is very good, and the model also finds out that the transaction conversion rate, the number of visiting users, the page views and the average price of a single product are the main factors affecting sales, which provides some reference for improving the explanatory power of the machine learning model in finance.**

Keywords- sales volume; forecast; LGBT; SHAP; interpretability

I. INTRODUCTION

In today's competitive e-commerce environment, accurately predicting sales volume (especially fashionable products) has become one of the key factors for the success of enterprises [1]. With the rapid changes of consumer behavior and market trends, enterprises must make a general forecast of sales volume to cope with the changing demand. Therefore, it is of great significance to master the appropriate methods and tools for sales forecasting to reduce the inventory risk of enterprises and meet customer needs [2,3].

Light Gradient Boosting Machine (LGBM) is a gradient lifting framework based on decision tree, which is famous for its efficiency and performance, especially on large data sets [4]. Its advantage lies in its optimized tree growth strategy and efficient feature selection, which makes the training speed faster and the memory occupation lower [5], and it is especially excellent when dealing with high-dimensional sparse data [6]. It has been widely used in finance [7], medicine [8], autonomous driving [9] and other fields.

However, decision tree-based ensemble learning algorithms such as LGBM are black-box models [10]. Although the prediction accuracy is high, it is difficult to directly see the specific contribution of each feature from the model. SHAP can quantify the contribution of each feature to the prediction results and provide the influence of each feature at the individual or global level [11]. This transparency enhances the trust in the model, especially in high-risk areas, such as finance, medical care or law, and the logic behind the decision must be clear [12].

In this paper, a prediction model using LGBM model and SHAP model to explain the results is proposed, which gives consideration to the model effect and interpretability. Using the data of a sportswear brand to verify, it can better predict the sales volume and analyze the main factors affecting the sales volume, which provides some reference for the development of e-commerce.

II. METHODOLOGY

A. LGBM Model

In principle, LGBM model uses the negative gradient of loss function as the residual approximation of the current decision tree to fit the new decision tree. In addition, LGBM adopts the histogram algorithm, the core of which is to discretize the continuous floating-point features into k discrete values and construct a Histogram with a width of k. Then traverse the training data and count the cumulative statistics of each discrete value in the histogram. In feature selection, we only need to traverse to find the optimal segmentation point according to the discrete values of histogram.

Suppose the training set has n instances $x_1, \ldots, x_n$; The characteristic dimension is s; At each gradient iteration, the negative gradient direction of the loss function of the model data variables is expressed as $g_1, \ldots, g_n$. Decision tree divides the

data into nodes through the optimal segmentation point. The gradient descent tree measures the information gain by the variance after segmentation, for example, O represents the training set of a fixed node, and d represents the segmentation of feature j, which is defined as shown in the following formula (1).

$$V_{j|O}(d) = \frac{1}{n_0}\left[\frac{\left(\sum_{x_i\in O:x_{ij}\le d} g_i\right)^2}{n^j_{l|O}(d)} + \frac{\left(\sum_{x_i\in O:x_{ij}>d} g_i\right)^2}{n^j_{r|O}(d)}\right] \quad (1)$$

Where $n_0 = \sum I\left[x_i \in O\right]$, $n^j_{l|O} = \sum I\left[x_i \in O: x_{ij} \le d\right]$ and $n^j_{r|O}(d) = [\sum x_i \in O: x_{ij} > d]$. Then traverse every split point of each feature, find $d^*_j = \mathrm{argmax}_d V_j(d)$, and calculate the maximum information system gain, and then divide the data into left and right sub-nodes according to the split points of different features.

B. SHAP Model

SHAP model is based on the concept of Shapley value in game theory, which assigns a contribution value to each feature, indicating the average contribution of the feature in all possible combinations [13]. The basic idea is to calculate the different marginal contributions of the feature in all feature sequences through the marginal contributions of the feature added to the model, and finally calculate the SHAP value of the feature, that is, the average value of all marginal contributions of the feature.

Suppose the i[th] sample is x_i, the first feature of the ith sample is x_{ij}, the marginal contribution of the feature is mc_{ij}, and the weight of the edge is w_i, where $f(x_{ij})$ is the SHAP value of x_{ij}, then the SHAP value of the first feature of the i[th] sample is calculated as shown in the following formula (2).

$$f\left(x_{ij}\right) = mc_{i1}w_1 + \cdots + mc_{i1}w_n \quad (2)$$

The predicted value of the model for this sample is y_i, and the baseline of the whole model (usually the average value of the target variables of all samples) is y_{base}, then SHAP value obeys the following equation (3).

$$y_i = y_{base} + f(x_{i1}) + f(x_{i2}) + \cdots + f(x_{is}) \quad (3)$$

Where $f(x_{i1})$ is the contribution of the first feature in the ith sample to the final prediction value y_i, and the SHAP value of each feature represents the change of model prediction under the condition of this feature. When the value is greater than 0, it means that the feature improves the predicted value, otherwise, it will reduce the predicted value.

C. Evaluation Criteria of the Model

After the model is established, an evaluation standard must be set to test the effect of the model, evaluate the predicted results and show the practical value of the model. The standard commonly used to evaluate regression models is mean absolute error (MAE). In this paper, the MAE and R^2 are selected to test the model effect. The calculation method is shown in the following formulas (4) and (5).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y} - y| \quad (4)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

Where $y = (y_1, y_2, \ldots, y_m)$ represents the true value of the data, and $\hat{y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m)$ represents the predicted value of the model.

D. Summary of Model Steps

1) In order to prevent over-fitting, this paper divides the processed data into training set, test set and verification set, and inputs the training set and verification set into AutoML for model training to find the best model.

2) Using grid search to find the optimal parameters of the best model.

3) Input that data of the optimal parameters, the train set and the verification set into the optimal model for training, and obtaining a regression model.

4) The evaluation index is constructed to test the effect of the optimal regression model.

5) Through the quantitative analysis of the characteristics of the best model by SHAP-LGBM model, the goal of this problem is obtained.

III. DATA

A. Data Sources and Descriptive Statistics

In this paper, the sales data of a sportswear brand from June 1, 2016 to November 5, 2021 were obtained from Alibaba Cloud Tianchi platform, and the data frequency was daily, with a total of 1953 days. Specifically, it includes 11 indicators, and Table 1 shows the descriptive statistics of these 11 indicators during the whole sample period.

Table 1 Descriptive statistics table

Indicator	mean	std	min	max
Page views	441324	403037	52322	10165460
Visitors	98051	133747	20293	3266915
Per capita page views	4.82	1.07	1.28	8.45
Average stay time	84.96	22.73	6.45	153.67
Jump rate %	48.61	8.28	15.33	93.54
Transaction customers	9123.1	9988.9	950	173656
Transaction volume	10662.25	11866	1032	205006
Transaction amount	960969	1489201	105305	33158260
customer unit price	96.89	25.46	21.40	311.15
Sold goods	59610	67655	4420	958449
Transaction conversion rate %	10.56	3.88	0.49	33.62

Transaction Volume in the table is the dependent variable that needs to be predicted and analyzed in this paper, while other indicators are the independent variables.

B. Data Preprocessing

First of all, the data needs to be processed with missing values. Figure 1 counts the remaining quantity of each indicator after missing values are removed, and it can be found that the quality of the data is good, and all indicators have no missing values, and the data of 1953 days are retained.

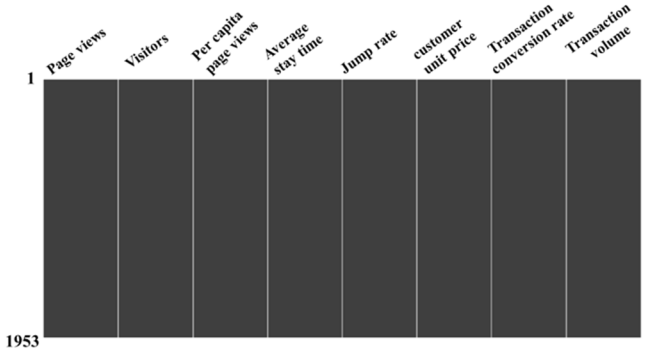

Figure 1. Feature importance diagram.

Then, it can be seen from the table that the orders of magnitude of each index are quite different. At the same time, in order to remove dimensions, all indexes are normalized in this paper, and the calculation formula is shown in the following formula (6).

$$y_{new,i} = \frac{y_i - y_{min}}{y_{max} - y_{min}} \tag{6}$$

In addition, we found that the data of Transaction customers, Transaction amount and Sold goods are highly correlated with Transaction volume, and they are all related to sales volume, so these indicators should be removed when forecasting and analyzing.

Finally, according to the length of time, this paper takes the data from June 1, 2016 to December 31, 2019 as the training set, and takes the data from January 1, 2020 to November 5, 2021 as the test set, in which the data in 2021 is the verification set.

IV. EMPIRICAL RESULTS

After substituting the relevant indicators into the model analysis, this paper first analyzes and explains the importance of each indicator affecting the sales volume, then reports the forecast results of this model, and finally shows the forecast chart.

A. Indicator Importance Analysis

Firstly, the importance of each index is analyzed, and the model results are shown in Figure 2 below. The larger the value, the more important the impact of this indicator on sales volume, and the importance in the figure decreases from top to bottom.

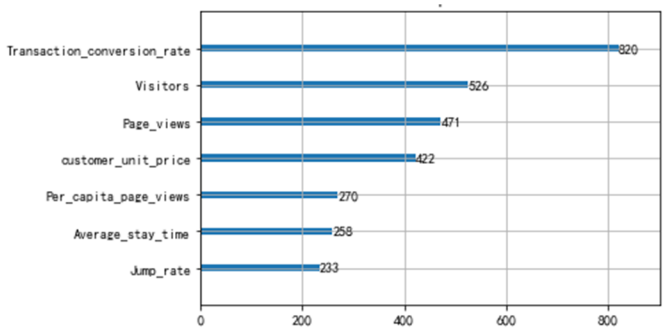

Figure 2. Feature importance diagram.

From the results, we can see that Transaction conversion rate, number of visitors, page views and unit price are the most important factors affecting sales, among which the first three are mainly affected by the degree of platform promotion and the attractiveness of product pages, while the unit price is only affected by the cost and profit of the merchants themselves. The following is an in-depth explanation of the important reasons for each indicator.

The indicator of Transaction conversion rate reflects the proportion of users visiting websites/apps who actually complete transactions. The higher the conversion rate, the better the website can attract users to trade, which is very important for increasing sales.

The Number of visitors directly affects the size of potential buyers. The more visitors, the more users are exposed to the products, thus increasing the sales base. This is a basic indicator.

Page views reflect the browsing behavior and interest of users on the website /App. The higher the number of page views, the more interested users are in products or services, and the higher the conversion potential.

Unit price directly affects whether customers are willing to buy the product. If the price is much higher than the customer's expected price, it is almost impossible to make a deal, and if the price is too low, it will lead to too little profit or even a loss, so this is a particularly important indicator.

Finally, appropriately increasing investment in platform promotion, carefully designing product pages and attracting consumers as much as possible are the most important factors affecting sales. On the other hand, merchants should analyze the sales price with the maximum profit according to the relative relationship between sales volume and unit price, and also need to pay attention to the sales prices of competitors of related products and make comprehensive pricing.

B. Prediction Eesult Analysis

For model evaluation, we calculated the mainstream evaluation indicators, and the results are shown in Table 2. As can be seen from the table, the performance of the model is very good, each prediction error is very small, and the R-square is relatively large, which means that the model explains the numerical change of sales volume to a great extent.

Table 2 Predictive performance evaluation table

Evaluating indicator	Value
Mean Squared Error	0.00030
Root Mean Squared Error	0.01722
Mean Absolute Error	0.00345
R^2	0.91235

The prediction result of this model is shown in ure 3 below, where the blue line represents the real data and the orange line represents the prediction data.

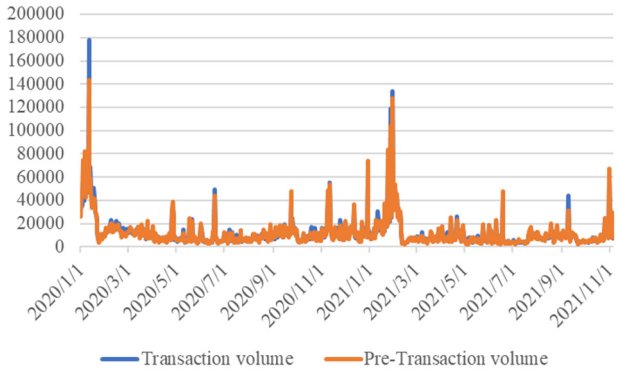

Figure 3. Prediction result chart.

It can be seen that the trading volume fluctuates at the daily level, and there is a certain periodicity, which are difficult to predict. The model in this paper has a good forecast effect on sales volume, and even captures the jumping change of sales volume at the daily level, so it is a reliable model as a whole.

V. CONCLUSIONS

In this paper, a sales forecasting model of sportswear based on LGBT algorithm and SHAPE explanation model is proposed. By collecting the daily frequency sales data of a sports brand in recent six years, this paper investigates the influence of 11 influencing factors on sales. Firstly, the LGBT algorithm is used to establish the sales forecast model, and then the contribution degree of each feature is quantitatively analyzed by the SHAPE model. The evaluation results show that the model has good accuracy and explanatory ability in forecasting sales volume.

SHAPE model analysis results show that the transaction conversion rate, the number of visiting users, the number of page views and the average price of a single product are the main factors affecting sales. This provides a basis for enterprises to deeply analyze the key points that affect sales. In particular, increasing promotional activities can effectively improve the conversion rate; Optimizing product pages helps to improve user stickiness and conversion rate; Reasonable pricing is accepted by customers while ensuring profits.

In addition, LGBT algorithm performs well in forecasting sales, and the predicted value is highly matched with the actual sales trend, which shows its application prospect in the field of sales forecasting. However, this study is only based on a single brand data, and the prediction effect needs further verification on a wider sample.

Generally, the model maintains the prediction accuracy and further improves the explanatory ability of the model through SHAPE. It provides a reference for e-commerce enterprises to improve their sales level and decision-making, and helps enterprises to actively regulate the key factors affecting sales, improve product attractiveness and customer stickiness, thus reducing inventory risk.

REFERENCES

[1] Zhang, L., Bian, W., Qu, W., Tuo, L. and Wang, Y., 2021, April. Time series forecast of sales volume based on XGBoost. In Journal of Physics: Conference Series (Vol. 1873, No. 1, p. 012067). IOP Publishing.

[2] Lyu, F. and Choi, J., 2020. The forecasting sales volume and satisfaction of organic products through text mining on web customer reviews. Sustainability, 12(11), p.4383.

[3] Sakurai, S., 2011. Prediction of sales volume based on the RFID data collected from apparel shops. International Journal of Space-Based and Situated Computing, 1(2-3), pp.174-182.

[4] Aziz, R.M., Baluch, M.F., Patel, S. and Ganie, A.H., 2022. LGBM: a machine learning approach for Ethereum fraud detection. International Journal of Information Technology, 14(7), pp.3321-3331.

[5] Osman, M., He, J., Mokbal, F.M.M., Zhu, N. and Qureshi, S., 2021. Ml-lgbm: A machine learning model based on light gradient boosting machine for the detection of version number attacks in rpl-based networks. IEEE Access, 9, pp.83654-83665.

[6] Xi, B., Li, E., Fissha, Y., Zhou, J. and Segarra, P., 2024. LGBM-based modeling scenarios to compressive strength of recycled aggregate concrete with SHAP analysis. Mechanics of Advanced Materials and Structures, 31(23), pp.5999-6014.

[7] Mousavi Anzahaei, S.M. and Nikoomaram, H., 2022. A comparative study of the performance of Stock trading strategies based on LGBM and CatBoost algorithms. International Journal of Finance & Managerial Accounting, 7(26), pp.63-75.

[8] Kanber, B.M., Smadi, A.A., Noaman, N.F., Liu, B., Gou, S. and Alsmadi, M.K., 2024. LightGBM: A Leading Force in Breast Cancer Diagnosis Through Machine Learning and Image Processing. IEEE Access.

[9] Shangguan, Q., Fu, T., Wang, J. and Fu, L., 2022. A proactive lane-changing risk prediction framework considering driving intention recognition and different lane-changing patterns. Accident Analysis & Prevention, 164, p.106500.

[10] Ekanayake, I.U., Meddage, D.P.P. and Rathnayake, U., 2022. A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP). Case Studies in Construction Materials, 16, p.e01059.

[11] Van den Broeck, G., Lykov, A., Schleich, M. and Suciu, D., 2022. On the tractability of SHAP explanations. Journal of Artificial Intelligence Research, 74, pp.851-886.

[12] Mokhtari, K.E., Higdon, B.P. and Başar, A., 2019, November. Interpreting financial time series with SHAP values. In Proceedings of the 29th annual international conference on computer science and software engineering (pp. 166-172).

[13] Sundararajan, M. and Najmi, A., 2020, November. The many Shapley values for model explanation. In International conference on machine learning (pp. 9269-9278). PMLR.