Research on sales forecasting of online products based on machine learning methods--Taking Meituan photo studio as an example

Yiran You

School of Management Science and Engineering, Tianjin University of Finance and Economics, Tianjin, China

YiranYou@stu.tjufe.edu.cn

Abstract. With the rapid development of the digital economy, the way consumers purchase goods and enjoy services has gradually shifted to various online platforms. As a leading comprehensive platform, Meituan has accumulated massive market data in the field of photo services. In this paper, based on the Colab environment, eXtreme Gradient Boosting (XGBoost) and Random forest (RF) in the field of machine learning are used to introduce SHapley Additive exPlanations (SHAP) analytics to model and analyze its sales data and predict future sales. The model was optimized using grid search and stochastic search, combined with several metrics such as R-Square (R^2), Median Absolute Error (MdAE), and Log Mean Absolute Percentage Error (Log MAPE) for a comprehensive assessment of the model effectiveness. The results show that RF outperforms XGBoost in both initial and optimized models. In particular, with the introduction of interaction features, RF can effectively capture complex nonlinear relationships and significantly improve the accuracy of sales volume prediction, while XGBoost performs poorly in the face of data imbalance and extreme values, with large prediction errors. This study provides an important reference for merchants to optimize their marketing strategies and improve user experience, which has important theoretical value and practical significance.

Keywords: Online platform, sales prediction, random forest regression, SHAP analysis, interaction features

1 Introduction

With the rapid development of the digital economy, online platforms have become the main way for consumers to purchase products and services. As one of the leading life service platforms in China, Meituan's online sales data contains a wealth of market information [1]. For example, in the field of photographic services, users can select merchants, compare service prices, and conduct transactions through the platform. How to effectively predict product sales based on these data, and then help merchants optimize their marketing strategies, is an important current research topic.

In recent years, the application of machine learning algorithms in data analysis and prediction has received much attention. Identifying key factors and accurately predicting them by analyzing historical data can support companies in formulating efficient strategies and improving the efficiency of resource allocation [2].

CONF-CIAP 2025, January 17-23, Eskisehir, Turkey

DOI 10.4108/eai.17-1-2025.2355320

Early studies mainly used traditional statistical methods such as regression analysis and time series analysis. However, it has limitations in high-dimensional and massive data scenarios. In recent years, the application of machine learning in sales volume prediction has gradually increased researchers have realized more accurate prediction of product sales volume through methods such as support vector machines, random forests (RFs), and decision trees [3, 4]. Especially on high-traffic e-commerce platforms, sales prediction models not only need to consider each influencing factor but also take into account the complexity of the market and the diversity of consumers [5]. In addition, deep learning methods are gradually being applied to the field of sales forecasting. Scholar Chen combines particle swarm optimization and Long Short-Term Memory Network (LSTM) to propose a merchandise sales prediction model, which makes the difference between the prediction results and the actual sales fluctuate in a small range (-2.4% to 1.82%), which is significantly better than the error range of the standard LSTM model, indicating that shows that deep learning prediction accuracy and stability in big data scenarios are better than traditional methods [6]. At the same time, while machine learning models often provide highly accurate predictions, their black-box nature makes the results difficult for the general public to intuitively understand. Thus, interpretable machine learning has become an important research area in recent years [7]. For this reason, the SHapley Additive exPlanations (SHAP) additive interpretation method was proposed by Mangalathu to quantify the feature importance and reveal the model decision mechanism, effectively quantify the contribution of features to the prediction results, and provide a new way of thinking about model interpretation and practical application [8].

In this study, based on the sales data of Meituan Photo Studio, a series of data processing work was carried out to introduce interaction features, feature importance, SHAP analysis, apply different machine learning algorithms to construct a product sales prediction model, reveal the main factors affecting sales for prediction, and optimize the model using grid search and random search. Ultimately, the model performance is evaluated through multi-dimensional indicators. Provide effective marketing strategies and more comprehensive and accurate guidance for various online platforms.

2 Methodology

2.1 Data source and description

The data for this study comes from the Meituan platform, focusing on its photo studio online product information. Focusing primarily on merchants in the Beijing area, it contains 4,298 pieces of data and 18 features across the following fields: dianpu name, such as Korean Bride STUDIO Travel Wedding Photography (Beijing Store), dianpu star, ranges roughly from 2.0 to 5.0. Type, like wedding photography, etc. Price-related fields, including dianpu price avg, product prices online, etc. Shooting related fields, such as pic num, and jingxiu num. Shooting related fields, such as pic num, jingxiu num. Additional services, such as negative gift situation. The study selects numerical features that have some correlation with sales and that directly reflect product pricing, service quality, and shooting results. The features are selected to accomplish the task of sales forecasting to ensure the representativeness of the data and the usefulness of the analysis results.

Data processing includes the steps of data Cleaning: removing features that are not relevant to sales volume prediction to reduce noise and improve model efficiency. Missing values were

filled in using the mean, median, or plurality to ensure data completeness. Text processing: clean up product names with regular expressions, retaining key numerical information for analysis. Data consolidation: Integrate multiple data sources and harmonize data formats and feature names to ensure consistency and accuracy. The final dataset covers multidimensional features such as product price, store rating, online price, and number of shots, and the distribution of key features is explored through descriptive statistical analysis to lay the foundation for subsequent research.

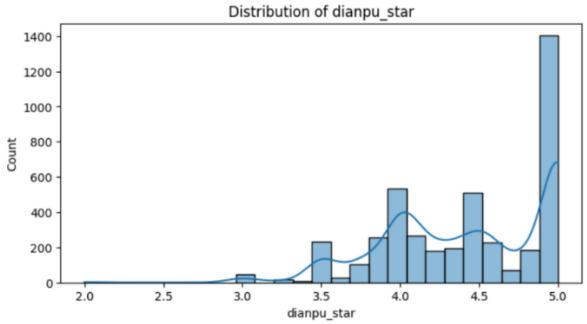

Fig. 1. The Distribution of dianpu star (Picture credit: Original).

As shown in Figure 1, the horizontal coordinate shows the store's star rating, which ranges from a minimum of about 2.0 to a maximum of 5.0. The vertical coordinate indicates the number of stores in each star range. The distribution of store ratings is concentrated in the higher score bands, especially above 4.0, with a significant proportion of 5.0 ratings. This suggests that the majority of stores in the dataset have higher ratings, possibly reflecting users' tendency to choose highly rated stores or stores with better service quality overall. However, an unbalanced distribution of ratings may cause the model to be more skewed toward highly rated samples. To this end, this study enhances the scoring discrimination through feature interactions, such as using the product of dianpu star and product prices as a new feature star price interaction to optimize the prediction performance.

2.2 Methodology introduction

In feature selection, three scenarios were designed in this study to train the model and validate the conclusions: Case 1 (no interaction 1): contains dianpu star, dianpu price avg, product sales, product prices, product prices online, pic num, and jingxiu num. Case II (Interaction): based on case I the interaction feature star price interaction is introduced. Case 3 (no interaction 2): the top three features obtained by feature importance ranking, i.e., product prices online, product prices and dianpu price avg, are chosen.

The main models used in this study include: (1) eXtreme Gradient Boosting (XGBoost) regression model:

XGBoost is a decision tree model based on gradient boosting, which aims to gradually improve the prediction accuracy by constructing a series of weak learners (decision trees) [9]. Its optimization goal is to minimize the objective function with the following formula:

$$L = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (1)$$

Where L is the loss function (e.g., mean square error (MSE)), $l = (y_i, \widehat{y_i})$ is the loss function, and $\Omega(f_k)$ is the regularization term to control model complexity and prevent overfitting.

(2) RF regression model:

An RF consists of multiple decision trees, each trained based on independent randomly sampled data, and the final prediction is the average of the predictions of all the trees, denoted as:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{2}$$

Where $\boldsymbol{T}$ is the number of trees, $h_t(x)$ denotes the predicted value of the tth tree, and $\hat{\boldsymbol{y}}$ is the final predicted value. RFs are used for node splitting by randomly selecting features to effectively cope with noise improve robustness, and help identify key features through feature importance analysis.

2.3 Degree of influence of features and parameter optimization

To understand the impact of features on sales volume prediction, this study calculates the correlation matrix of the features and plots a heat map to reveal the linear relationship between the features. Highly correlated features may cause redundancy and affect model performance. The contribution of each feature to the model's prediction results was quantified through SHAP analysis with the following formula:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N|-|S|-1)!}{|N|!} (\nu(S \cup \{i\} - \nu(S)) \tag{3}$$

Where N is the set of all features, $\boldsymbol{\nu(S)}$ is the output of the model on the feature set S, and $\boldsymbol{\phi_I}$ denotes the SHAP value of feature $\boldsymbol{i}$. This approach helps to identify the most critical features in sales volume forecasting. For parameter optimization of the model, this study uses a combination of grid search and cross-validation. Grid search selects the optimal parameters by combining them under multiple parameter combinations, while cross-validation further ensures the generalization of the model and reduces the risk of overfitting. The best parameters were finally selected for model training by cross-validation with 3 and 5 folds.

2.4 Evaluation indicators and synthesis analysis

In the evaluation of the model, R-Square (R^2) was first chosen to measure the ability of the model to explain the fluctuations in the target variable, ranging from 0 to 1. The closer the value is to 1, the better the model fit. However, when the data are unevenly distributed or have extreme values, R^2 may not be sufficient to accurately reflect the performance of the model, and therefore a comprehensive assessment in combination with other metrics is required. Secondly Median Absolute Error (MdAE) is insensitive to extreme values and provides a more robust reflection of the typical error magnitude of the model and is suitable for data where extreme values exist. Once again Log Mean Absolute Percentage Error (Log MAPE) is more suitable for measuring relative prediction error than the traditional Mean Absolute Percentage Error (MAPE) as the logarithmic treatment reduces the asymmetric effect of large error values. Through the comprehensive analysis of the above indicators, we comprehensively assess the model's fitting ability and resistance to extreme values, and deeply analyze the characteristics of the prediction error distribution to verify the reliability and robustness of the model's performance.

3 Statistical analysis

3.1 Relevance of each feature

In machine learning, the level of correlation between features and target variables affects the performance of the model. The nonlinear models used in this study, RF and XGBoost, are capable of capturing complex nonlinear relationships. Thus even if the correlation between individual features and the target variable is low, the model can still improve the prediction performance through feature interaction. Low-correlation features can also sometimes reduce redundancy, avoid multicollinearity problems, and enhance the generalization ability of the model.

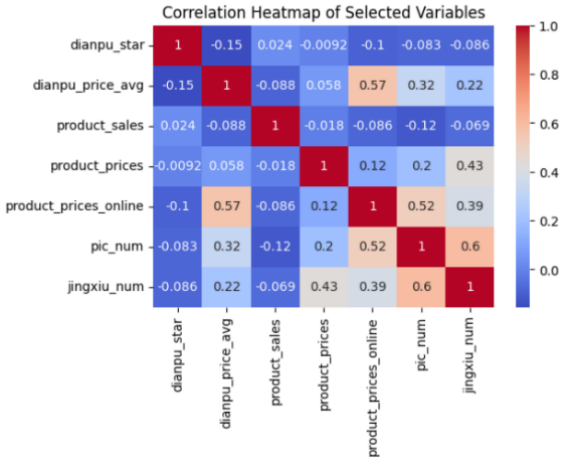

Fig. 2. Feature Correlation Heatmap (Picture credit: Original).

As in Figure 2, further explanations of the variables are derived: The target variable product sales has a low correlation with the other variables, indicating that these variables have a low impact on sales. Interaction features were added to the study to better capture sales changes. The correlation between pic num and product prices online is 0.52, showing a moderate positive correlation between the number of pictures and online prices, i.e. more pictures may lead to slightly higher prices. The correlation between jingxiu num and pic num is 0.6, indicating that the number of shots is significantly and positively correlated with the number of refined images.

3.2 Characteristic importance

As in Table 1, the importance of each feature of the RF model is extracted and ranked, and the top three to four features with the highest importance are shown.

Table 1. Feature Importance of the RF Model (Table credit: Original).

	Feature	RF Feature Importances
1	Product prices online	0.479490
2	Product prices	0.221858
3	Dianpu price avg	0.185461
4	Pic num	0.068901

Based on the results presented in Table 1, the following ranking of the importance of the characteristics can be observed. product prices online: has the highest significance (0.479490) and is more than twice as high as the other higher variables, suggesting that online prices are a key factor in sales volume, directly influencing consumer purchasing decisions. Online prices usually have a direct impact on consumers' purchasing decisions. Product prices also of significant importance, influencing consumer choice and distribution channels. However, product prices may cover a wider range of sales channels than online prices. pic num indicates that the content of the service appeals to the consumer and may increase willingness to buy. The small effect of dianpu price avg indicates that the average store price has a limited impact on sales. Based on these analyses, it can be hypothesized that price sensitivity is the main driver and marketing strategies should focus on pricing. On service marketing, increasing the number of shots may improve the conversion rate.

3.3 SHAP analysis

Firstly, it is explained that it is normal for the results of the feature importance analysis of XGBoost and RF to be different from the results of the SHAP analysis due to the different computational approaches and focuses. RF feature importance is statistically derived from the tree structure, measured based on the information gain (e.g., Gini coefficient or MSE) when splitting nodes. The SHAP analysis, however, is based on the Shapley value in game theory, which quantifies the marginal contribution of features to the model output, overcomes the problem of multiple covariance among features, and provides a more objective assessment.

As can be seen in Figure 3, each point in the graph represents the SHAP value of a sample, with red points indicating high values and blue points indicating low values. SHAP values for product prices and product prices online stand out in the negative direction, indicating that high values of these features usually result in lower model output, e.g., fewer sales. A high value of dianpu star tends to increase sales and supports the findings of the feature importance analysis.

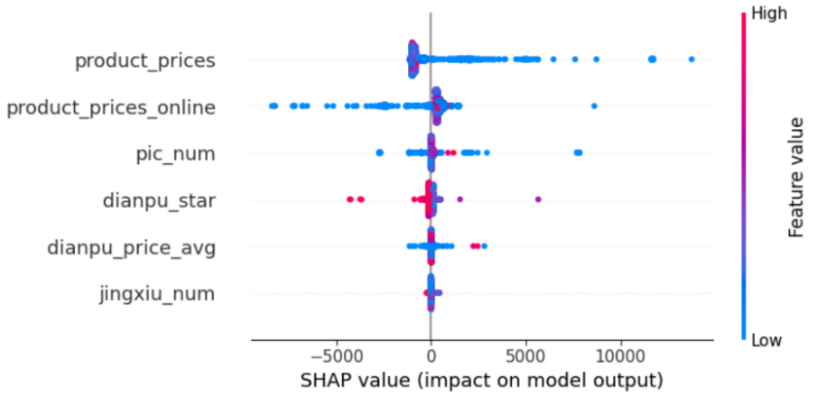

Fig. 3. SHAP Value Scatter Plot (Picture credit: Original).

3.4 Model optimization and evaluation

As shown in Figures 4, 5, the performance comparison of XGBoost and RF in the sales prediction task before and after optimization in three cases contains three sub-figures, the horizontal coordinates are the names of the different models, which are Initial XGBoost,

Optimized XGBoost, Initial RF and Optimized RF. The left subplot has a vertical coordinate of R^2, unitless, and ranges from 0 to 1. The center subplot has a vertical coordinate of MdAE, and the unit is the number of products sold. The vertical coordinate of the right subplot is Log MAPE in percent (%). It is used to compare and evaluate the prediction effect before and after model optimization.

Case II, interaction, hyperparameter tuning for random search, is visualized in Fig. 4. Instead, the optimized R^2 decreases, and MdAE and Log MAPE deteriorate significantly, showing that the optimization does not lead to a positive effect. Suggests that this scenario may be overfitting or underfitting.

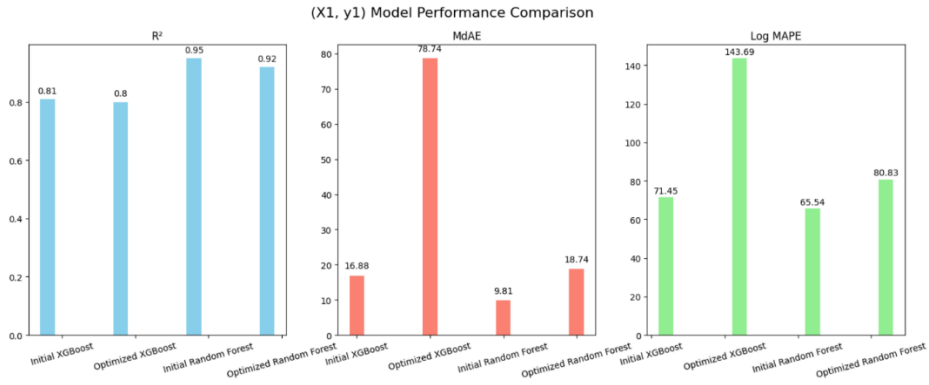

Fig. 4. Bar Chart Comparing Model Performance Before and After Optimization in Interactive Contexts (Picture credit: Original).

Case 3, no interaction2, is optimized by grid search, as in Figure 5. The resulting R^2 remains stable or slightly elevated, indicating a more robust fit of the model to the overall trend. However, both MdAE and Log MAPE increased, indicating a decrease in prediction accuracy for individual values, which may affect the model's ability to predict detail.

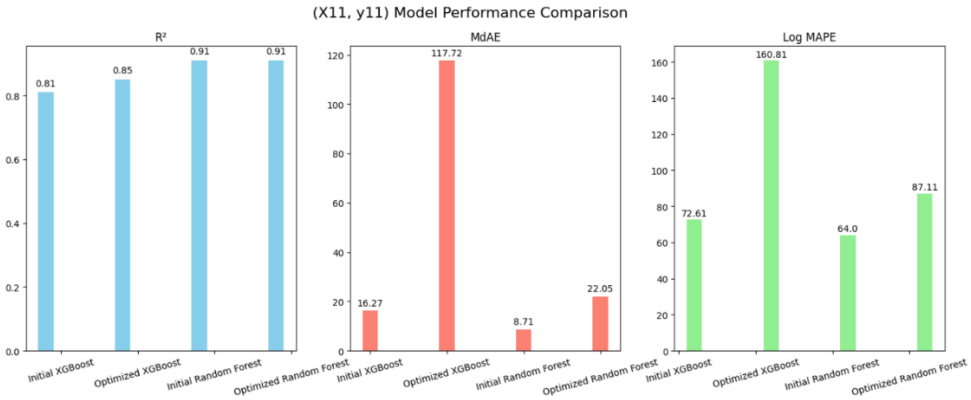

Fig. 5. Bar Chart Comparing Model Performance Before and After Optimization in Non-Interactive Scenario 2 (Picture credit: Original).

It is found that RF outperforms XGBoost's R^2 in all cases, especially in the interaction context where it reaches 0.95, indicating that the model is able to fit the sales data to a great extent, higher than XGBoost's 0.81. Its MdAE and Log MAPE metrics also performed well, especially in the no-interaction2 and interaction contexts, with a MdAE of 8.71 and high predictive accuracy. Under the interaction feature, the Log MAPE of RF reaches 65.54%, which is still lower than Initial XGBoost's 71.45%, further verifying the advantage of RF in sales prediction.

3.5 Visualization of the final result

To show the prediction results of the models more intuitively, in this study, a comparative visualization of sales prediction is carried out, especially for the Initial RF model under the combination of interactive features and the Optimized XGBoost model under the combination of no-interactive features1 to present the difference between the actual sales and the predicted sales. In Fig. 6, the X-axis represents the randomized serial number of the data points and the Y-axis represents the product sales. The red lines/dots, indicate True Values, i.e. actual sales. The blue lines/dots, indicate Initial RF's predicted values for sales. Orange lines/dots indicate the predicted value of sales by the Optimized XGBoost model. Fluctuations in actual sales can have significant peaks and troughs, reflecting the seasonality of sales or the impact of unexpected events. The predicted values of the two models may exhibit deviations from actual sales at some data points, showing the predictive power of the models under specific conditions. The predicted values from Initial RF may be smoother overall, showing the strength of the model in capturing trends in sales volume. The overall data is visualized in Figure 6.

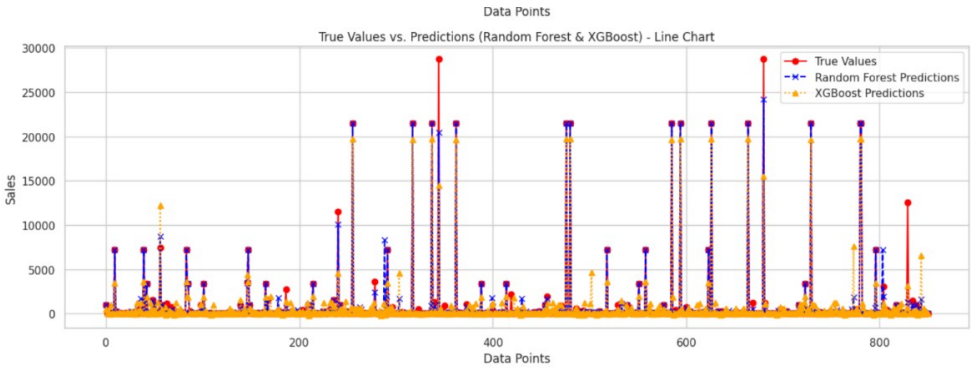

Fig. 6. Comparison Chart of Predictions from Two Models (Overall Data) (Picture credit: Original).

Due to the huge amount of data, to avoid the chart is too complex and unclear, the strategy of dividing the data sequentially into 12 segments and drawing charts separately was adopted for group visualization. Further, a randomized grouping strategy was used to ensure the representativeness of the data and to avoid the influence of order effects. This method ensures that each segment of data contains samples from different categories and locations to fully reflect model performance. For example, sequential segmentation may be subject to bias due to seasonal or cyclical fluctuations, while randomized grouping helps to reduce this effect. This is illustrated in Figure 7.

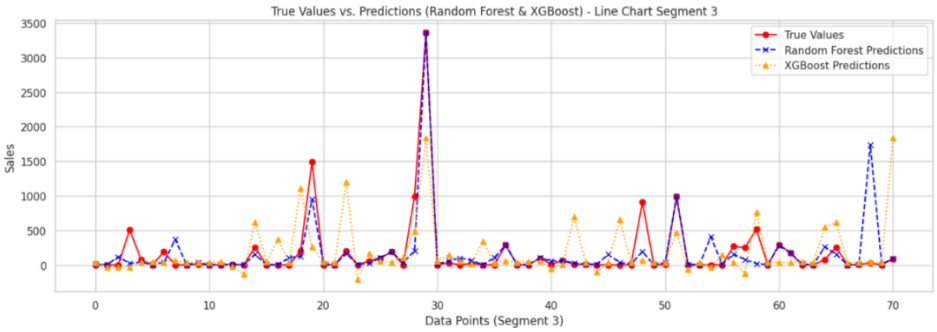

Fig. 7. Comparison Chart of Predictions from Two Models (Randomly Divided into 12 Groups) (Picture credit: Original).

The predictive effects of the model are presented in a more representative way through the presentation of segmented graphs after randomized grouping. The line graph visualization of an arbitrarily selected section of data shows that the Initial RF model is closer to the true value in terms of predicted values, showing better stability and lower error. In contrast, Optimized XGBoost's predictions are relatively more volatile, but it remains competitive in terms of explanatory power, R^2, after model tuning.

4 Discussion

This study provides a comprehensive assessment of the performance of multiple machine learning models in sales volume forecasting and discusses the strengths and weaknesses of the models and their practical application value in conjunction with feature importance analysis. Through SHAP analysis, the study demonstrates the consistency and fairness of feature importance interpretation in high-dimensional data scenarios, which complements the traditional feature importance analysis approach of RFs and improves the understanding of the model decision-making process. This is supported in the literature, with Ana and Lundberg (2024) noting that SHAP has significant advantages in terms of model interpretability and transparency [10]. The model optimization results show that the optimized RF excels in several performance indicators, especially in the case of interacting features, its R^2 value reaches 0.95, and the prediction accuracy is significantly improved. Meanwhile, the model outperforms other models in MdAE and Log MAPE metrics, demonstrating strong adaptability and robustness to complex data relationships. The optimized XGBoost, on the other hand, although improved in R^2, is weaker in handling unbalanced data and extreme values, resulting in higher MdAE and Log MAPE. The study also found that XGBoost's predictions fluctuated significantly, especially in the prediction of extreme cases, through random group visualization analysis. To address this issue, the study recommended a combination of training and testing using Augmented Data Augmented Data (AD) techniques to cope with the lack of sufficient data to achieve the desired accuracy in many real AI data processing cases [11].

In order to enhance the practical application of the sales forecasting model, merchants can combine the results of the study to develop more accurate marketing strategies. For example, in response to the significant impact of the number of shots on sales revealed in this study, merchants should invest in high-quality product images and video production, and can use short

video ads for store promotion. Related studies have shown that credibility, expertise, and attractiveness of video advertisements are positively correlated with consumer purchases, while authenticity and brand heritage influence consumer purchase behavior in a U-shaped manner [12]. In addition, merchants can regularly analyze the impact of their pricing strategies on sales and adjust their strategies in light of sales forecasts, especially during holidays and promotional seasons, by targeting potential best-selling products and implementing targeted promotional strategies. At the same time, to cope with the rapid changes in the market, merchants are advised to utilize big data for strategic marketing and dynamic capabilities needed to improve market responsiveness [13]. By dynamically monitoring market trends, merchants can more flexibly adjust their product lines and pricing strategies in response to changes in consumer demand and challenges in the competitive environment.

In summary, this study not only reveals the application potential of machine learning models in sales forecasting but also provides practice-oriented improvement suggestions for merchants, providing theoretical basis and technical support for optimizing marketing decisions and improving sales performance.

5 Conclusion

The results of this study show that the RF model has stronger predictive power and stability in sales volume forecasting, especially when faced with complex features and interactions. It outperforms XGBoost in error control and prediction accuracy and significantly outperforms XGBoost in optimized performance. In practice, merchants can utilize RF's sales prediction results to develop more accurate pricing and promotional strategies, especially during holiday or promotional seasons. By adding high-quality images to display, adding short videos to promote your store and optimizing your pricing strategy, you can effectively boost sales. In addition, merchants should leverage the dynamic capabilities of big data, combining market feedback and external data to continuously optimize their marketing strategies to adapt to market changes and enhance competitiveness.

References

[1] Zhao, W., & Zhao, Q. (2022). Profitability analysis of Meituan. *Old Brand Marketing, (02)*, 169-171.

[2] Zhao, Q., Bai, L., Xu, W., et al. (2019). Monthly sales prediction of unmanned supermarket under multiple linear regression model. *Times Economy and Trade, (13)*, 35-37.

[3] Martins, E., & Galegale, N. V. (2023). Sales forecasting using machine learning algorithms. *Revista de Gestão e Secretariado (Management and Administrative Professional Review)*.

[4] Pavlyshenko, B. (2019). Machine learning models for sales time series forecasting. *Data, 4*(15).

[5] Rajasree, T., & Ramyadevi, R. (2024). Time series forecasting of sales data using hybrid analysis. In *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)* (pp. 732-735).

[6] Chen, I., & Zhang, S. (2023). Research on merchandise sales prediction based on deep learning. *Information and Computer, 35*(12), 111-113.

[7] Yang, J., & Cao, J. (2021). Tree-based interpretable machine learning of the thermodynamic phases. *Physics Letters A, 412*, 127589.

[8] Mangalathu, S., Hwang, S.-H., & Jeon, J.-S. (2020). Failure mode and effects analysis of RC members based on machine-learning-based Shapley Additive exPlanations (SHAP) approach. *Engineering Structures, 219*, 110927.

[9] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16) (pp. 785–794). Association for Computing Machinery.

[10] Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and Translational Science, 17*(11), e70056.

[11] (2021). Augmented data and XGBoost improvement for sales forecasting in the large-scale retail sector. *Applied Sciences, 11*(17), 7793.

[12] Meng, L. (Monroe), Kou, S., Duan, S., & Bie, Y. (2024). The impact of content characteristics of short-form video ads on consumer purchase behavior: Evidence from TikTok. *Journal of Business Research, 183*, 114874.

[13] Brewis, C., Dibb, S., & Meadows, M. (2023). Leveraging big data for strategic marketing: A dynamic capabilities model for incumbent firms. *Technological Forecasting and Social Change, 190*, 122402.