Cross-border E-commerce Market Analysis Based on Tmall Global Big Data

HUANGKUN CHEN

1 INTRODUCTION

In the past three years, Tmall Global's cross-border e-commerce industry has demonstrated strong growth momentum, with an average annual growth rate close to 20%, becoming a key driver of the consumer market. Especially in the first half of 2023, the Tmall Global platform welcomed more than 2,000 overseas brands from countries such as Japan, the United States, South Korea, France, and Australia. These brands mainly covered categories including health products, cosmetics, and personal care.

Against this backdrop, an in-depth analysis of product data on Tmall Global is crucial for understanding market dynamics and forecasting future consumption trends. This paper focuses on product data from categories such as cosmetics, personal care, health, luggage, and digital products on Tmall Global. The analysis process includes data cleaning, transformation, and visualization to reveal Tmall Global's sales characteristics, business model, and its position in a highly competitive market. By applying data mining techniques, this paper also explores the future development trends of the cross-border e-commerce market and provides strong data support for decision-making in this field.

The data used in this study includes, but is not limited to, Tmall Global's product information, favorites and ratings fields, and sentiment scores of product reviews. These data not only provide basic information about the products but also incorporate user interaction and feedback, offering rich material for in-depth market analysis.

To fully understand Tmall Global's performance in the cross-border e-commerce field, the tasks of this paper include: First, thoroughly sorting and understanding each field in the dataset, exploring data types and identifying missing values; then, performing necessary preprocessing, including handling missing and abnormal values, and supplementing sentiment scores of products; next, exploring the distribution characteristics of product data and the correlations between variables; and finally, conducting in-depth analysis of various factors affecting sales characteristics through models such as linear regression analysis, principal component analysis (PCA), and principal component regression (PCR). Through these steps, we aim to quantify Tmall Global's performance in the current cross-border e-commerce market competition and provide strong data support to help decision-makers better grasp market trends.

2 DATA DESCRIPTION

The data in this paper is sourced from Tmall Global and covers multiple product categories, aiming to analyze the sales characteristics and consumption trends of the cross-border e-commerce market in depth. The dataset mainly consists of four parts: Tmall Global product information, favorites and ratings fields, sentiment scores of product reviews, and review data for which sentiment scores need to be calculated. The merged dataset covers comprehensive dimensions from basic product information to user interaction and emotional feedback.

First, the file "Tmall Global Product Information.xlsx" provides the basic information of each product, including name, price, and description, which helps us understand product characteristics. Then, the file "Tmall Global Favorites and Ratings.xlsx" records user interaction data such as the number of favorites and ratings, which reflect the popularity and market acceptance of the products. The third file, "Tmall Global Product Review Sentiment Scores.csv," provides consumer sentiment evaluations by analyzing user reviews.

The final file, "comments.csv (to be scored)," contains more unprocessed user review data. After performing sentiment analysis on this data, it can further enrich the user feedback information for the products. The merged dataset includes key fields such as product ID, product name, description, current price, original price, sales volume, number of reviews, shipping address, shipping time, and inventory. In addition, it incorporates multidimensional analysis of product pricing strategies, such as price ranges, average prices, and price binning. Through comprehensive analysis of this data, we can gain deeper insights into product market performance, consumer preferences, and sales trends.

Table 1: Variable Description Table for Tmall Global Product Information

Variable Name	**Variable Description**
Current Price	The current selling price of the product on the Tmall Global platform
Original Price	The original listed price of the product before any discounts
Reference Sales	An estimated sales value reflecting the market acceptance and popularity of the product
Number of Reviews	The total number of reviews received, indicating consumer feedback and attention
Inventory	The quantity of the product available in stock, reflecting supply conditions
Popularity	A composite index of product popularity, typically based on views, favorites, and other engagement metrics.
Rating	The average user rating of the product, reflecting overall customer satisfaction.
Sales Amount	The total revenue generated from product sales.
Discount Ratio	The ratio of the current price to the original price, representing the level of discount offered.

Sentiment Score	A score derived from product reviews using natural language processing, quantifying the emotional tone of customer feedback.

3 DESCRIPTIVE STATISTICS

The objective of this chapter is to deeply analyze the data characteristics of various products on the Tmall Global platform and the correlations between variables using descriptive statistical methods. We will apply data visualization and variable correlation analysis to understand the internal relationships among product characteristics, user behavior, and market trends.

In the data visualization section, various chart types are used to demonstrate the distribution and features of the data. For example, histograms are used to display the distribution of product price ranges, and bar charts are employed to analyze the distribution of product categories. These visualizations provide intuitive support for understanding market dynamics.

The variable correlation analysis section explores the relationships between different data dimensions, including the relationship between current price and sentiment score, sales volume and sentiment score, etc. Additionally, a correlation heatmap is used to analyze the interactions among key variables, including current price, original price, sales volume, number of reviews, inventory, popularity, rating, sales amount, discount ratio, and sentiment score. These analyses help to identify which factors significantly impact product market performance and provide data support for marketing strategies.

In summary, the analysis in this chapter aims to offer a comprehensive understanding of Tmall Global product data through descriptive statistics and visualization methods, revealing the complex relationships between product attributes and market performance.

3.1 DATA VISUALIZATION ANALYSIS

Figure 1 visualizes the quantity distribution of products in different price ranges on the Tmall Global platform. It can be clearly observed that as product prices increase, the number of products drops significantly. Especially in the high-price range exceeding 1000 yuan, the number of products sharply decreases, possibly reflecting the merchants' cautious attitude toward stocking high-priced goods. This distribution indicates the high demand for low-priced products in the market, while high-priced goods, due to their price threshold, have fewer target consumers. This may also relate to consumer purchasing power and habits—low to mid-range products are more readily accepted and purchased by the public.

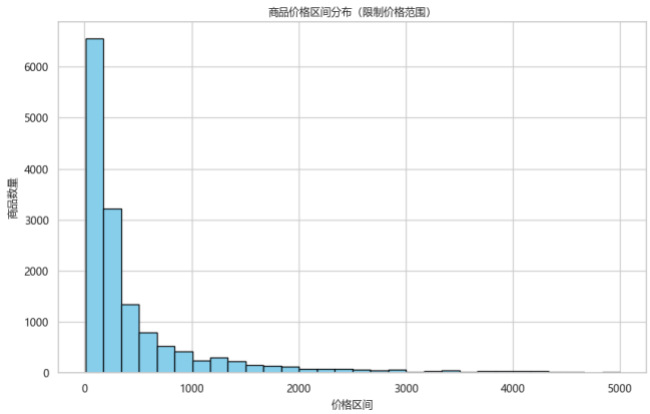

Figure 1. Distribution of product price ranges

Figure 2 shows the quantity distribution of different product categories on Tmall Global. It is evident that the popularity of product categories varies significantly. Categories such as “sports equipment,” “milk powder,” and “health supplements,” which are daily necessities, have higher counts, reflecting their broad consumer appeal due to frequent use. In contrast, categories like “water purifiers” and “beauty devices” have fewer products, possibly because they are not daily essentials or are relatively expensive with lower purchase frequency. Due to high repurchase rates, daily necessities naturally dominate in quantity.

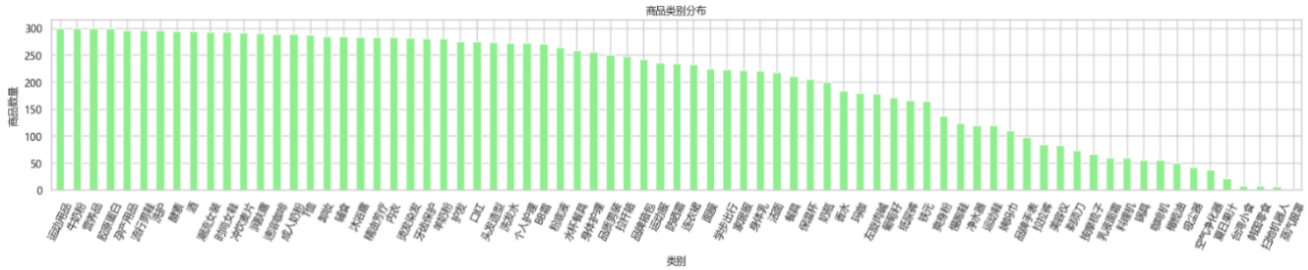

Figure 2. Distribution of product categories

Figure 3 visualizes the distribution of product ratings on Tmall Global. Ratings of 3, 4, and 5 dominate, indicating that most products receive positive consumer reviews. In contrast, ratings of 1 and 2 are fewer, suggesting those products may not meet consumer expectations. A large number of products have a rating of 0, which could be due to serious defects or market rejection. Overall, this rating distribution reflects the general product quality on the platform and consumer satisfaction and acceptance.

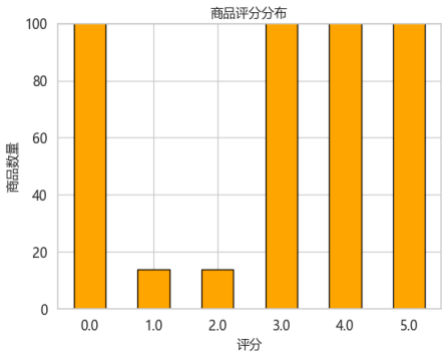

Figure 3. Distribution of product rating intervals

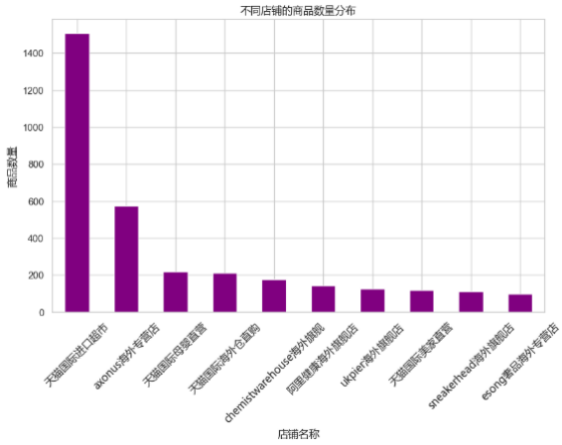

Figure 4. Number of products in top 10 stores

Figure 4 uses a bar chart to present the number of products from the top 10 stores on Tmall Global. The Tmall Global Import Supermarket has significantly more products than other stores, indicating its advantage in product diversity and inventory. The “Axonus Overseas Store” ranks second with relatively high product count, showing its strong market presence on the platform. Other stores, such as Tmall Global Mother & Baby Direct, Overseas Warehouse Direct Purchase, and Chemist Warehouse Flagship Store, have fewer products, possibly due to their focused categories or niche target markets. This distribution provides insights into the market positioning and product strategies of leading stores on Tmall Global.

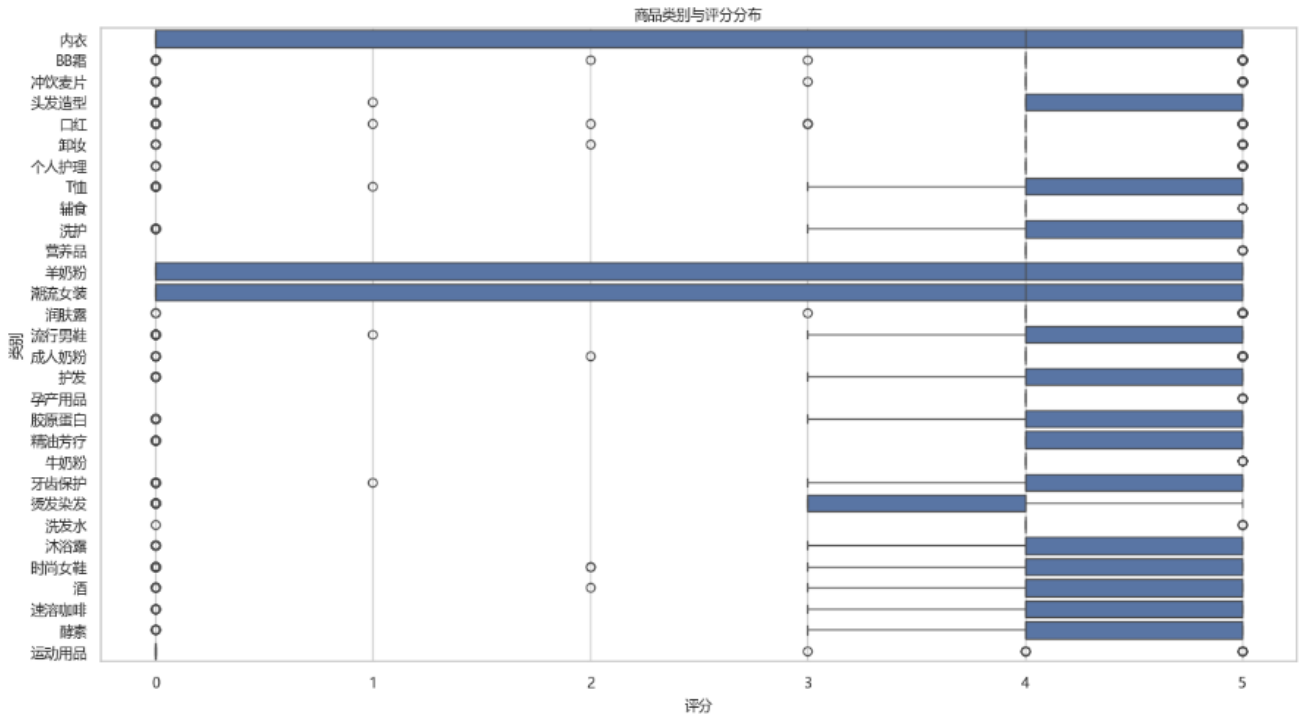

Figure 5. Rating distribution across product categories (partial)

Figure 5 presents a box plot visualizing the distribution of ratings across different product categories on Tmall Global. Categories like branded bags, underwear, branded watches, razors, thermos bottles, goat milk powder, and trendy women's clothing show longer boxes, indicating wider rating ranges and greater variation in consumer feedback. This may reflect inconsistencies in product quality or satisfaction within those categories. Most other categories have shorter boxes, meaning more concentrated ratings and consistent consumer evaluations. The median line shows the central tendency, while the upper and lower bounds of the box indicate rating volatility. Outliers might reflect special cases caused by product issues or unique user experiences. This visualization offers an intuitive way to understand consumer feedback patterns across categories.

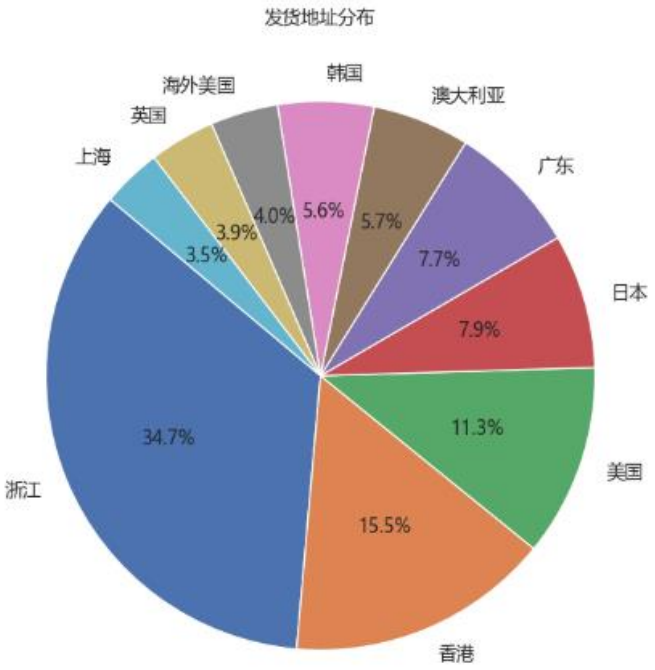

Figure 6. Distribution of the Top 10 Shipping Addresses

Figure 6 visualizes the distribution of the top 10 shipping addresses on the Tmall Global platform in the form of a pie chart. It shows that Zhejiang ranks first with 34.7%, significantly ahead of other regions, reflecting its critical role in cross-border e-commerce logistics. Hong Kong ranks second with 15.5%, the United States is third with 11.3%, and Japan ranks fourth with 7.9%. These data highlight the key positions of these regions in the global e-commerce landscape. The distribution may be influenced by multiple factors. For example, Zhejiang is an important hub for China's e-commerce and logistics, with a developed infrastructure, making it the primary shipping origin on Tmall Global. Hong Kong and the U.S., as key international trade gateways, possess extensive global connectivity and logistical advantages. Japan, a major Asian economy, also enjoys a high global reputation for its products, securing its place in the top shipping origins. These insights not only reveal the geographic characteristics of cross-border logistics but also reflect each region's positioning within the global e-commerce network.

Figure 7 shows a histogram of sentiment scores and their frequencies for products on Tmall Global. The chart clearly reveals that the vast majority of products have sentiment scores above 0.8, indicating that most receive positive consumer feedback. This may reflect either generally high product quality or strong overall satisfaction. However, there is a portion of products with sentiment scores below 0.8, which may suggest defects such as quality issues, poor pricing strategies, inadequate customer service, or logistics problems—leading to customer dissatisfaction. These lower sentiment scores remind merchants to improve on those fronts to enhance user experience and satisfaction. On the whole, this sentiment score distribution provides valuable feedback: high scores can serve as best-practice cases for promotion, while low scores should prompt root cause analysis and corrective actions.

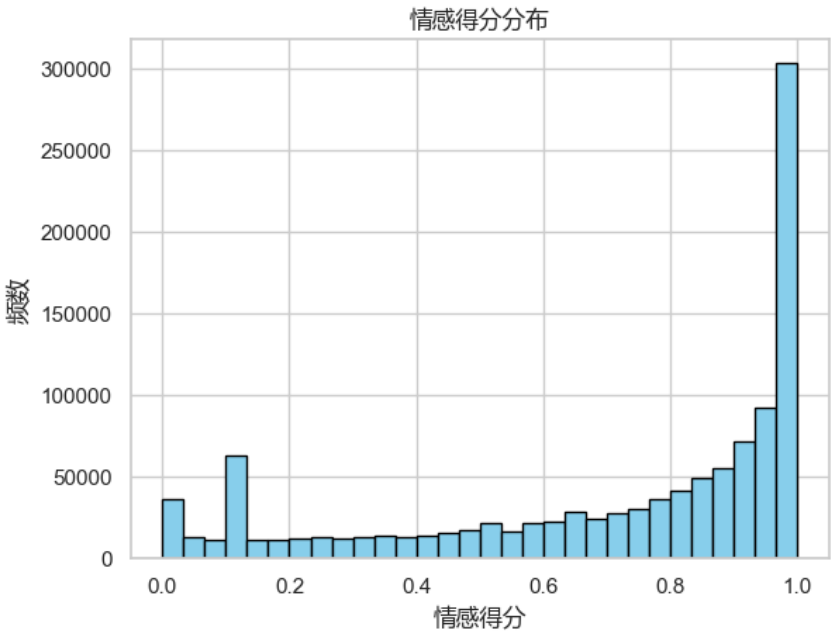

Figure 7. Frequency Distribution of Sentiment Scores

3.2 CORRELATION ANALYSIS BETWEEN VARIABLES

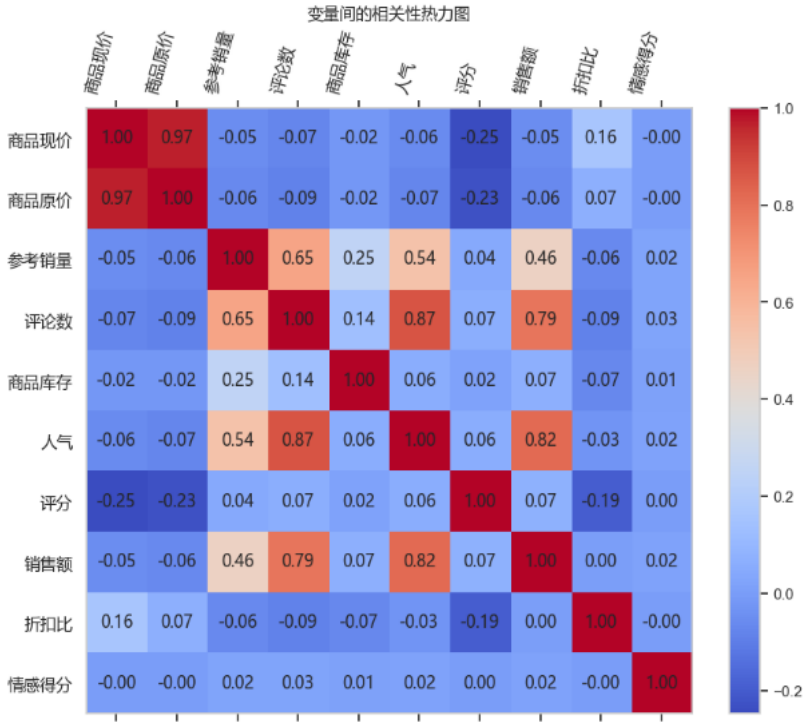

Figure 8: Pearson Correlation Heatmap

When analyzing the cross-border e-commerce market on Tmall Global, we assessed the correlations between various variables based on the Pearson correlation coefficient. The correlation heatmap in Figure 8 presents a range of relationship insights.

First, the correlation between the **current price** and the **original price** of products is very high, reaching **0.97**, indicating a strong linear relationship between these two pricing indicators and reflecting consistency in product pricing strategies.

In addition, there is a positive correlation between **reference sales** and the **number of reviews**, with a correlation coefficient of **0.65**; the correlation with **popularity** is **0.54**, and with **sales amount** is **0.46**. This suggests that products with higher sales volumes are usually accompanied by more reviews and greater popularity, which in turn promotes an increase in overall sales.

Notably, the **number of reviews** and **popularity** are also strongly positively correlated, with a coefficient of **0.87**. Their correlation with **sales amount** is **0.79**, emphasizing the important influence of user feedback on a product's popularity and its sales performance. The correlation between **popularity** and **sales amount** is also high, at **0.82**, further confirming the close relationship between a product's popularity and its market performance. This study speculates that there may be **multicollinearity** among the independent variables in the model, and the number of independent variables is relatively large. Therefore, **Principal Component Analysis (PCA)** can be considered in subsequent modeling to reduce dimensionality and eliminate multicollinearity.

4 REGRESSION ANALYSIS

In this study, we explored the relationships between multiple features of Tmall Global products and their corresponding **sales amount** and **sentiment scores**, aiming to uncover the intrinsic links between product characteristics and market performance through **machine learning modeling**. Based on key features in the dataset—such as current price, number of reviews, popularity, rating, and discount ratio—we built regression models to quantify the relationships between these features and both sales and sentiment outcomes.

4.1 REGRESSION MODELS

In this paper, we selected **three mainstream regression models** to predict the two target variables—**sales amount** and **sentiment score**, respectively.

First, the **linear regression model**, a fundamental and widely used tool in predictive analysis, was applied. It assumes a linear relationship between the target variable and the features. The advantage of this model lies in its simplicity and interpretability, allowing us to intuitively understand how each feature influences the outcome.

Next, we introduced the **random forest model**, an ensemble learning method based on decision trees. Random forest builds multiple decision trees and aggregates their predictions to enhance **accuracy and robustness**. This model is well-suited for handling complex and nonlinear data relationships and offers deeper analytical insights.

Lastly, the **K-Nearest Neighbors (KNN)** model, an instance-based learning method, was employed. It makes predictions based on the values of sample points closest to the target point. KNN excels at capturing **local data patterns** and is particularly useful for analyzing subtle relationships between product features and consumer sentiment scores.

4.2 MODEL EVALUATION

Table 2 lists the relationships between multiple features of Tmall Global products and their corresponding **sales amount**. From the results, we observe that the **standardized coefficients** for **inventory**, **rating**, and **sentiment score** are **270.3371**, **448.9315**, and **37.1742**, respectively, with all **p-values close to 0**. This indicates that these three variables have a statistically significant **positive correlation** with sales, particularly highlighting the strong influence of **ratings and inventory** on sales performance. The **standardized coefficient** of the **current price** is **16.3554**, with an extremely small **p-value** (approximately $\mathbf{4.064 \times 10^{-60}}$), suggesting that current price also has a **significant positive relationship** with sales.

In contrast, the **original price** shows an insignificant correlation with sales. Its **standardized coefficient** is **–0.9692**, and the **p-value** is **0.3325**, indicating that the original price has a **relatively small and statistically insignificant** impact on sales. The **standardized coefficients** for **reference sales** and **popularity** are **4.7296** and **2.7369**, with **p-values** of $\mathbf{2.25 \times 10^{-6}}$ and **0.006202**, respectively. These results demonstrate that both variables have a **significant positive correlation** with sales. Notably, the **number of reviews** has a **standardized coefficient** of **–81.6010** and a **p-value close to 0**, indicating a **significant negative correlation** with sales. This may suggest that a higher number of reviews, possibly containing negative content, could reduce purchasing confidence or that popular products have stabilized, while new items might be trending. In addition, based on the **Variance Inflation Factor (VIF)** analysis, the **VIF** values for **current price** and **original price** are **17.39** and **18.01**, respectively—both exceeding the commonly accepted threshold of 10. This indicates a **strong multicollinearity** between these two variables. All other variables have **VIF values below 5**, suggesting **weak or negligible multicollinearity** among them.

Table 2. Linear Regression Coefficients and Variance Inflation Factors (VIF) for Sales Amount

Variable	Standardized Coefficient	P-Value	VIF
Current Price	16.3554	4.064e-60	17.39
Original Price	-0.9692	0.3325	18.01
Reference Sales	4.7296	2.25e-06	1.93
Number of Reviews	-81.6010	0	5.72
Inventory	270.3371	0	1.09
Popularity	2.7369	0.006202	4.58

Rating	448.9315	0	3.98
Sentiment Score	37.1742	3.042e-302	4.00

From **Table 3**, we can observe that the **current price** of a product shows a **high standardized coefficient** of **682.3480** in relation to **sentiment score**, with a **p-value of 0**. This indicates a clear **positive correlation** between current price and sentiment score. However, the correlation between **original price** and sentiment score is **not statistically significant**, as its **standardized coefficient** is **1.0022** and the **p-value is 0.3162**. This suggests that original price has a **limited impact** on sentiment. The **standardized coefficient** for **reference sales** is **–0.4877**, with a **p-value of 0.6257**, indicating **no significant influence** on sentiment score and suggesting a **negative correlation**.

In contrast, **number of reviews**, **inventory**, and **popularity** all show **significant positive correlations** with sentiment score, with standardized coefficients of **7.6655**, **10.8385**, and **6.7583** respectively. The **p-values** for these variables are all **very small**, confirming their strong and statistically significant relationships with sentiment.

On the other hand, **rating** exhibits a **significant negative correlation** with sentiment score, with a **standardized coefficient** of **–7.4035** and a **p-value close to 0**. This result suggests that, counterintuitively, higher rating scores may not always reflect more favorable sentiment—a pattern that may be attributed to rating inflation or inconsistencies between numerical scores and textual feedback. **Sales amount**, meanwhile, does **not show a statistically significant effect** on sentiment score. Its **standardized coefficient** is **–1.6432** with a **p-value of 0.1003**.

Regarding **multicollinearity**, the **VIF values** for **current price** and **original price** are **17.39** and **17.96**, respectively—well above the common threshold of 10 used to flag significant multicollinearity. This suggests a strong degree of collinearity between these two variables, which can lead to **unstable model estimates** and **interpretation difficulties**. Special caution is needed when interpreting the coefficients of these two variables.

In contrast, other variables such as **reference sales**, **number of reviews**, **inventory**, **popularity**, and **rating** have **VIF values below 5**, indicating **low multicollinearity** and relatively independent relationships.

Table 3. Linear Regression Coefficients and Variance Inflation Factors (VIF) for Sentiment Score

Variable	Standardized Coefficient	P-Value	VIF
Current Price	682.3480	0	17.39
Original Price	1.0022	0.3162	17.96
Reference Sales	-0.4877	0.6257	1.94

Number of Reviews	7.6655	1.783e-14	6.19
Inventory	10.8385	2.271e-27	1.09
Popularity	6.7583	1.397e-11	5.62
Rating	-7.4035	1.328e-13	1.32
Sentiment Score	-1.6432	0.1003	3.74

When conducting regression analysis on **Tmall Global product sales**, we not only relied on **quantitative model evaluation metrics**, such as **Mean Squared Error (MSE)** and **R^2 (coefficient of determination)**, but also used **visualization techniques** to intuitively assess model performance.

Figures 9, 10, and 11 respectively visualize the evaluation results of the **Linear Regression**, **Random Forest**, and **K-Nearest Neighbors (KNN)** models. These visualizations are consistent with the numerical results presented in **Table 4**, providing direct visual evidence of the models' predictive performance.

In **Figure 9**, the visualization of the **Linear Regression** model shows the scatterplot of **actual vs. predicted sales** along with a **residual plot**. Although this model has a relatively **higher MSE**, its **R^2 score is 0.70**, indicating that it is able to capture the general trend of changes in sales to a certain extent.

In contrast, **Figures 10 and 11** illustrate the performance of the **Random Forest** and **KNN** models, both of which demonstrate **near-perfect predictive accuracy**. These models exhibit **low MSE values** and achieve **R^2 scores of 1.00**. In these figures, the predicted values align extremely closely with the actual values, and the points in the residual plots are largely concentrated around the zero line, indicating **high prediction accuracy** on this dataset.

By combining the quantitative analysis in **Table 2** with the visual insights from **Figures 9–11**, we can conclude that although the **Linear Regression** model lags slightly behind in predictive precision compared to Random Forest and KNN, it remains a **valuable analytical tool**, particularly in scenarios that require **strong model interpretability**. Meanwhile, the **Random Forest** and **KNN** models excel at capturing **complex patterns in sales data**, offering powerful tools for understanding the intricate relationships between product characteristics and sales performance.

Table 4. Evaluation of Regression Models for Sales Amount

Model	PCA Applied	MSE	R^2 Score
Linear Regression	No	6278198839060.31	0.70
Random Forest	No	61460209.59	1.00
KNN Regression	No	67277565.47	1.00

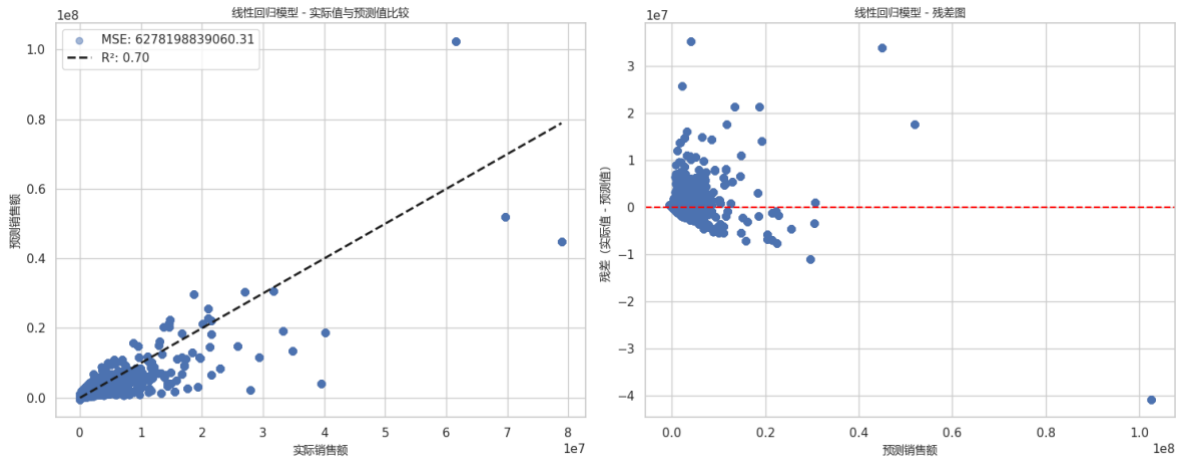

Figure 9. Visualization of Sales Amount Linear Regression Model Evaluation (Without PCA)

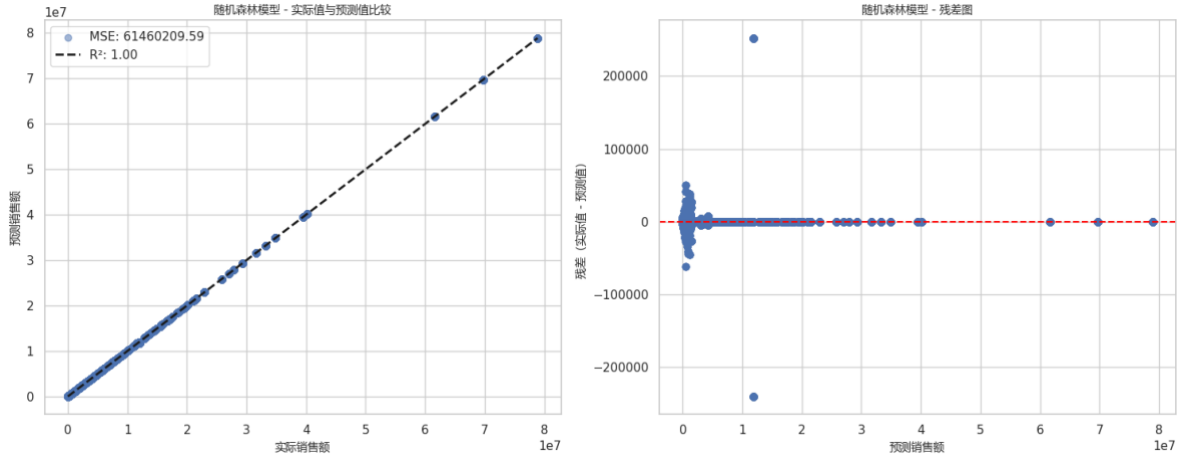

Figure 10. Visualization of Sales Amount Random Forest Model Evaluation (Without PCA)

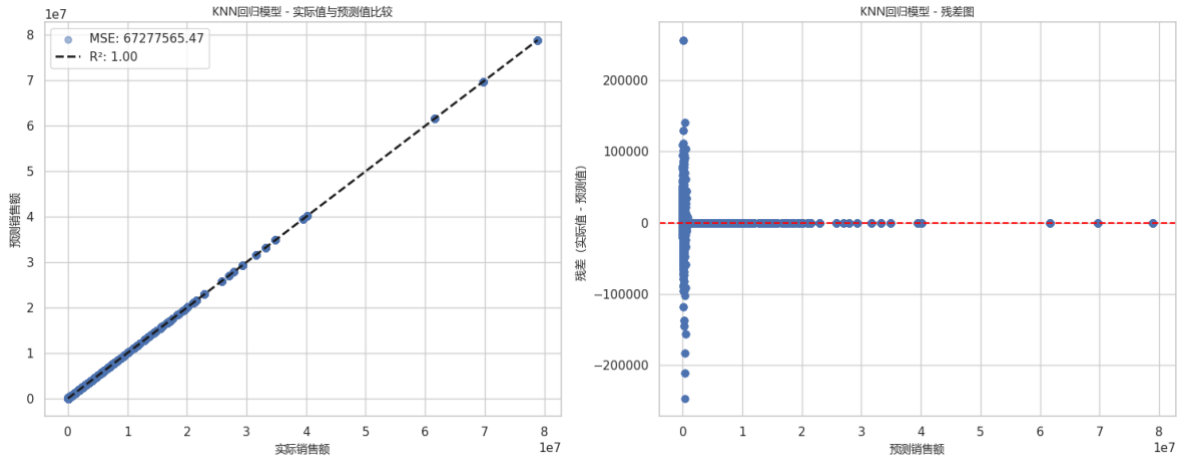

Figure 11. Visualization of Sales Amount KNN Model Evaluation (Without PCA)

Table 5 and its associated visual results reveal that all three regression models—Linear Regression, Random Forest, and K-Nearest Neighbors (KNN)—performed poorly in predicting sentiment scores.

The Linear Regression model yielded a mean squared error (MSE) of 0.10 and an R^2 value near zero, indicating that it was almost entirely ineffective in explaining the variation in sentiment. The Random Forest model showed only a slight improvement, with an R^2 of 0.01, while the KNN model performed even worse, producing a negative R^2 of –0.18, meaning its predictions were less accurate than a simple mean-based approach. The visualizations further support these findings, as they demonstrate a lack of correlation between predicted and actual values, with residuals scattered broadly. These results suggest that the relationship between the selected features and sentiment score is likely nonlinear or more complex than the models can capture. It is also possible that the current set of variables does not adequately reflect the key drivers of consumer sentiment. Factors such as the emotional tone in product descriptions, individual user preferences, or contextual shopping scenarios may significantly influence sentiment but are not represented in the available data. This analysis highlights the limitations of the current approach and suggests that future work should consider more comprehensive data sources and advanced modeling techniques—such as natural language processing or deep learning—to more effectively understand and predict sentiment in the cross-border e-commerce context.

Table 5. Evaluation of Regression Models for Sentiment Score

Model	PCA Applied	MSE	R^2 Score
Linear Regression	No	0.10	0.00
Random Forest	No	0.10	0.01
KNN Regression	No	0.12	-0.18

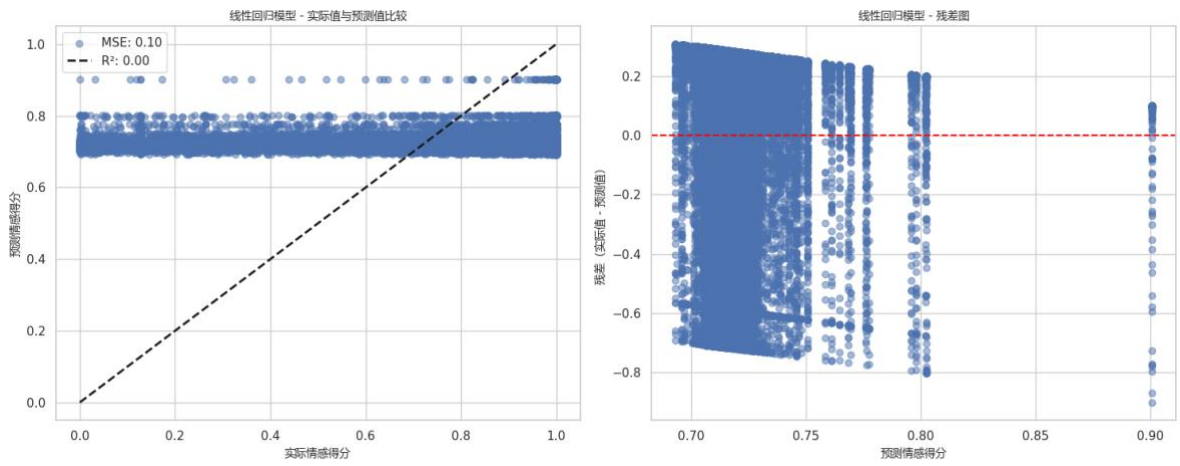

Figure 12. Visualization of Sentiment Score Linear Regression Model Evaluation (Without PCA)

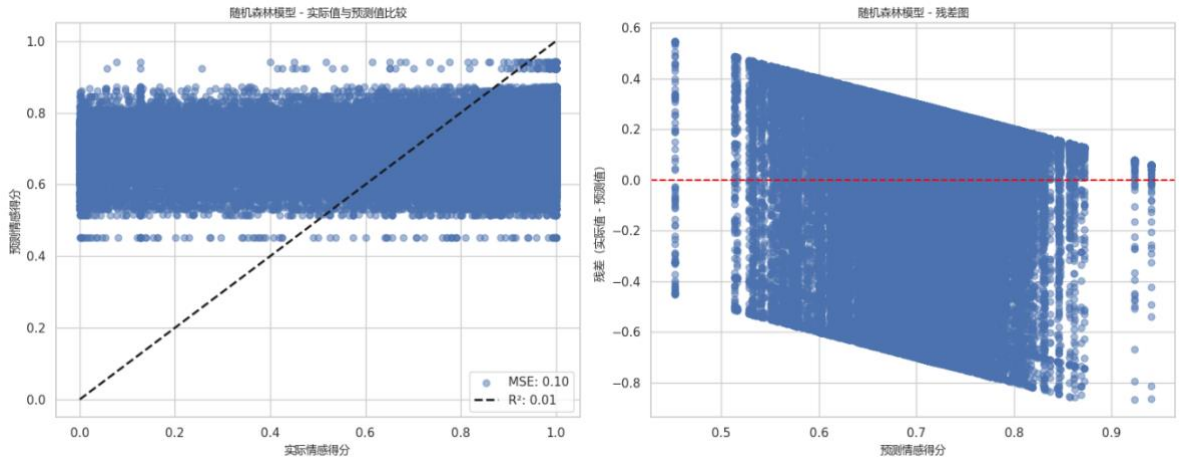

Figure 13. Visualization of Sentiment Score Random Forest Model Evaluation (Without PCA)

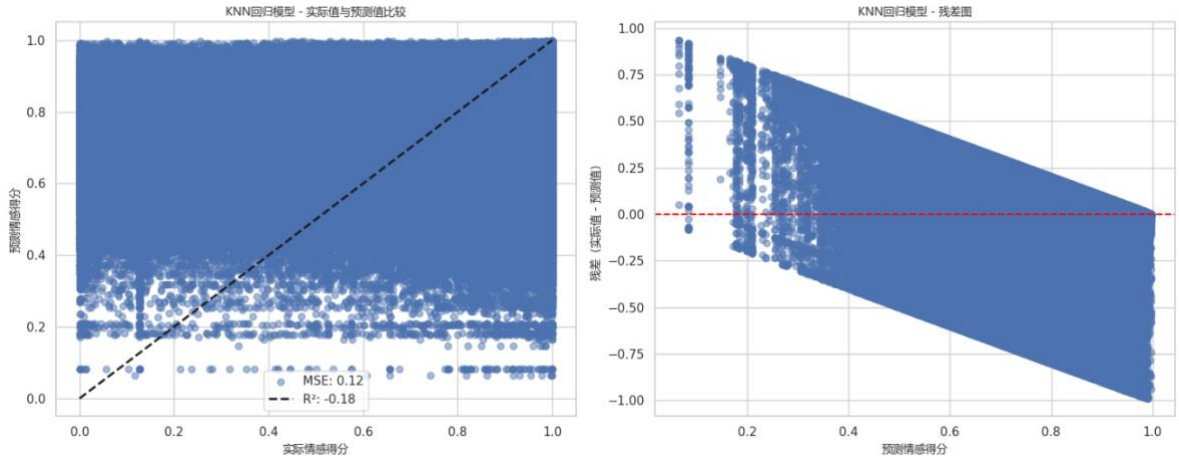

Figure 14. Visualization of Sentiment Score KNN Model Evaluation (Without PCA)

We applied multiple models—Linear Regression, Random Forest, and K-Nearest Neighbors (KNN)—to conduct a comprehensive evaluation of model performance using both quantitative metrics and intuitive visualizations. In predicting sales amount, the Random Forest and KNN models demonstrated strong performance, achieving lower mean squared errors and highly accurate R^2 scores. Although the Linear Regression model was slightly less accurate, it provided greater interpretability, making it a valuable baseline model. These findings suggest that while more complex models offer greater flexibility and precision in handling data, linear regression remains an important analytical tool, especially when model transparency is required. In contrast, all models performed unsatisfactorily in predicting sentiment scores, indicating that the selected features may lack a direct or strong linear relationship with sentiment. This insight suggests that accurately modeling sentiment scores may require deeper data exploration—potentially involving more nuanced factors such as product text descriptions or the emotional tendencies present in user reviews.

4.3 MODEL INTERPRETATION

In this study, **Mean Squared Error (MSE)** was used as a direct metric to compare the prediction accuracy of different models with respect to sales. Additionally, the **coefficient of determination (R^2)**

was used to evaluate each model's ability to explain the variance in the target variable. R^2 values typically range between 0 and 1, where a value closer to 1 indicates that the model effectively captures the variability in the target variable, whereas a value closer to 0 suggests limited explanatory power.

In analyzing these metrics, we also incorporated visualizations to intuitively demonstrate model performance. For instance, in the visualization of sales prediction results, scatter plots were used to show the relationship between actual and predicted values. The closer the points are to the diagonal line, the more accurate the model's predictions. In addition, **residual plots** were employed to show the differences between predicted and actual values. If the residuals are concentrated near the zero line, it indicates that the model's predictions closely match the real-world data.

By combining quantitative indicators with visual interpretation, we were able to conduct a **comprehensive evaluation** of each model's performance in a more interpretable and insightful manner.

4.4 PRINCIPAL COMPONENT ANALYSIS and REGRESSION

Principal Component Analysis (PCA) was first applied to identify the variables that have the greatest impact on product sales and sentiment scores. This step not only helps uncover hidden key factors within the data but also reduces data complexity, making the model easier to interpret and more efficient to compute. After extracting the principal components, we used them as independent variables to reconstruct new regression models. This process aimed to evaluate how the PCA-transformed dataset influences the prediction of sales and sentiment.

To assess the effect of dimensionality reduction, we compared the performance of the PCA-based regression model with that of the regression model built on the original dataset. For evaluation, we continued to use **Mean Squared Error (MSE)** and **R^2** as the primary performance metrics, while scatter plots and residual plots were used to visually present the outcomes of the PCA-based models.

Table 6 presents the **cumulative variance contribution rates** of each principal component in the PCA. These rates reflect how much variance in product sales is explained by each component. Specifically, the **first principal component alone accounts for 55.49%** of the variance, indicating it carries a substantial amount of information and can be considered a major factor influencing sales. The **second component increases the cumulative contribution to 97.03%**, meaning that the first two components together explain the vast majority of the variability in sales. The **third component pushes the cumulative variance explained to 98.83%**, while the **fourth and fifth components** add relatively little additional value, bringing the total to **99.26%** and **99.65%**, respectively.

Therefore, by selecting only the top few principal components as variables, we can **retain most of the original information while significantly reducing model complexity**. This enhances model efficiency without compromising accuracy.

Table 6. Cumulative Variance Contribution of Principal Components (Sales Amount)

Principal Component	Cumulative Variance Contribution (%)
Principal Component 1	55.49%
Principal Component 2	97.03%
Principal Component 3	98.83%
Principal Component 4	99.26%
Principal Component 5	99.65%

In **Table 7**, we applied **Principal Component Analysis (PCA)** for dimensionality reduction and extracted several key components that represent the variability within the sentiment score data. According to the results, the **first principal component accounts for 89.60%** of the total variance in sentiment scores, indicating that the majority of changes in sentiment can be captured by a single comprehensive factor. The **second principal component increases the cumulative variance contribution to 96.69%**, demonstrating that the first two components together explain nearly all of the variability in sentiment scores. As the **third, fourth, and fifth components** are introduced, the cumulative variance contribution further increases to **97.80%**, **98.63%**, and **99.35%**, respectively.

Although each additional component contributes progressively less to the total variance, these subsequent components still help capture **subtle differences** in sentiment scores and provide additional insights into the underlying data structure.

Table 7. Cumulative Variance Contribution of Principal Components (Sentiment Score)

Principal Component	**Cumulative Variance Contribution (%)**
Principal Component 1	89.60%
Principal Component 2	96.69%
Principal Component 3	97.80%
Principal Component 4	98.63%
Principal Component 5	99.35%

4.5 PRINCIPAL COMPONENT REGRESSION MODEL EVALUATION

As shown in **Table 8**, **reference sales** exert a significant influence on sales amount, with a **regression coefficient of 1357.305** and a **p-value of 0.000**, indicating a clear **positive correlation** between the two variables. Additionally, **original price**, **number of reviews**, and **popularity** also have

measurable impacts on sales; however, their regression coefficients are **–127.560**, **–270.871**, and **–275.562**, respectively, suggesting a **negative correlation**. Notably, both original price and review count demonstrate a substantial negative effect on sales.

The coefficients for **current price** and **inventory** are **–36.638** and **25.344**, respectively, indicating relatively minor impacts, with current price showing a **slight negative** correlation and inventory showing a **positive** one. Furthermore, **rating** and **sentiment score** have coefficients of **44.430** and **2.944**, both with **p-values less than 0.003**, reflecting **statistically significant positive correlations** with sales. Among these, **rating** appears to have a more prominent positive influence on sales performance.

Table 8. Principal Component Regression Coefficients (Sales Amount)

Variable	Current Price	Original Price	Reference Sales	Number of Reviews	Inventory	Popularity	Rating	Sentiment Score
Regression Coefficient	-36.638	-127.560	1357.305	-270.871	25.344	-275.562	44.430	2.944
P-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003

As shown in **Table 9**, the regression coefficients for **original price**, **reference sales**, and **popularity** are **23.159**, **13.655**, and **3.701**, respectively, with all corresponding **p-values equal to 0.000**. This indicates a **statistically significant positive correlation** between these variables and sentiment score—particularly for original price and reference sales, which exhibit notably strong positive impacts. The **number of reviews** also shows a positive association with sentiment score, with a coefficient of **3.358** and a **p-value of 0.001**, suggesting that higher review counts are related to greater consumer satisfaction or product appeal.

On the other hand, **rating** displays a **significant negative correlation** with sentiment score, with a **regression coefficient of –9.032** and a **p-value close to 0**, implying that higher ratings may paradoxically be associated with lower sentiment scores. This result may reflect unexpected consumer behavior or be driven by unobserved factors not captured by the model.

The **regression coefficients** for **current price** and **inventory** are **0.234** and **0.844**, respectively, but with **p-values of 0.815** and **0.399**, indicating that their relationships with sentiment score are **not statistically significant**. Similarly, **sales amount** shows a coefficient of **–0.771** with a **p-value of 0.440**, suggesting **no meaningful impact** on sentiment score in this model.

Table 9. Principal Component Regression Coefficients (Sentiment Score)

Variable	Current Price	Original Price	Reference Sales	Number of Reviews	Inventory	Popularity	Rating	Sales Amount
Regression Coefficient	0.234	23.159	13.655	3.358	0.844	3.701	-9.032	-0.771
P-Value	0.815	0.000	0.000	0.001	0.399	0.000	0.000	0.440

Next, we continued to apply **Linear Regression**, **Random Forest**, and **K-Nearest Neighbors (KNN)** regression models, incorporating **Principal Component Analysis (PCA)** for dimensionality reduction. **Table 10** presents the performance evaluation results of the three models after applying PCA.

The **Linear Regression model** exhibited a relatively high **Mean Squared Error (MSE)** of **6,428,915,593,685.24** and a comparatively low **R^2 score of 0.69**, suggesting **limited ability to explain sales variation** and insufficient prediction accuracy. This could be due to the fact that the relationship between sales and the input features is not entirely linear, or that **important information may have been lost** during dimensionality reduction via PCA. **Figure 15** visualizes the evaluation results of the linear model and reveals noticeable discrepancies between predicted and actual values, further supporting the model's performance limitations.

In contrast, the **Random Forest model** demonstrated superior performance with an **MSE of 428,379,060,798.32** and an **R^2 score of 0.98**, indicating that it effectively captures complex relationships between sales and the features, and delivers high predictive accuracy. **Figure 16** clearly visualizes this outcome, showing a near-perfect alignment between predicted and actual values.

Finally, the **KNN regression model** achieved the **best performance**, with an **exceptionally low MSE of 5306.99** and a **perfect R^2 score of 1.00**, signifying **almost zero prediction error** in forecasting sales. This result is reinforced by **Figure 17**, which shows near-identical predicted and actual values.

When evaluating model performance, it is also crucial to consider the relationship between features and the target variable. In this study, the selected features included **current price**, **number of reviews**, **popularity**, **rating**, and **discount ratio**—all of which are key factors influencing product sales. For example, current price directly affects consumers' purchasing decisions; number of reviews and popularity reflect product visibility and acceptance; rating indicates customer satisfaction; and discount ratio can impact perceived value. These factors collectively contribute to the model's predictive power and should be considered when interpreting the results.

Table 10. Evaluation of Regression Models for Sales Amount (with PCA)

Model	PCA Applied	MSE	R^2 Score
Linear Regression	Yes	6428915593685.24	0.69

Random Forest	Yes	428379060798.32	0.98
KNN Regression	Yes	5306.99	1.00

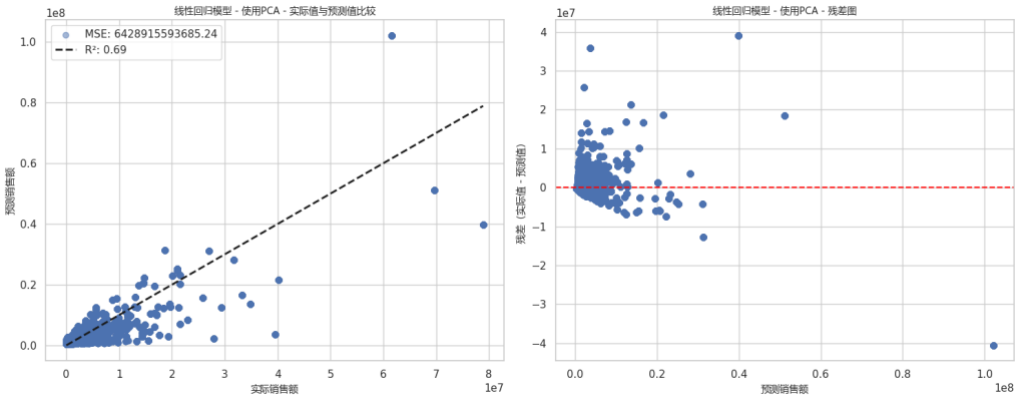

Figure 15. Visualization of Sales Amount Linear Regression Model Evaluation (with PCA)

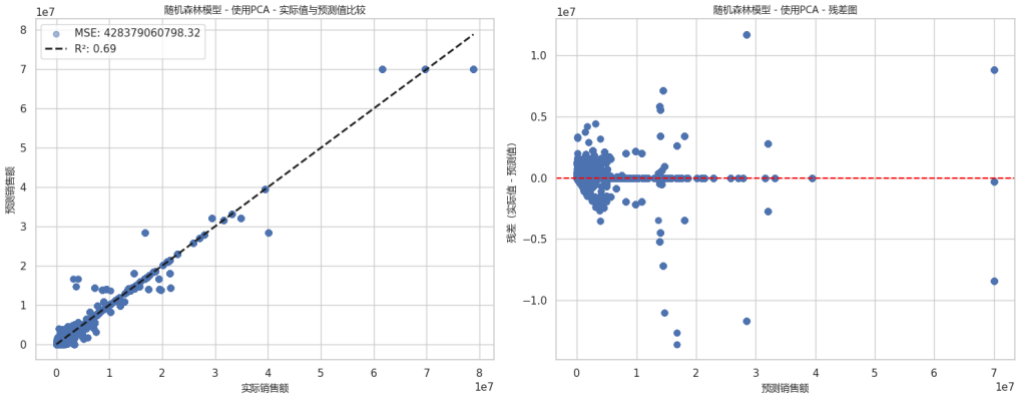

Figure 16. Visualization of Sales Amount Random Forest Model Evaluation (with PCA)

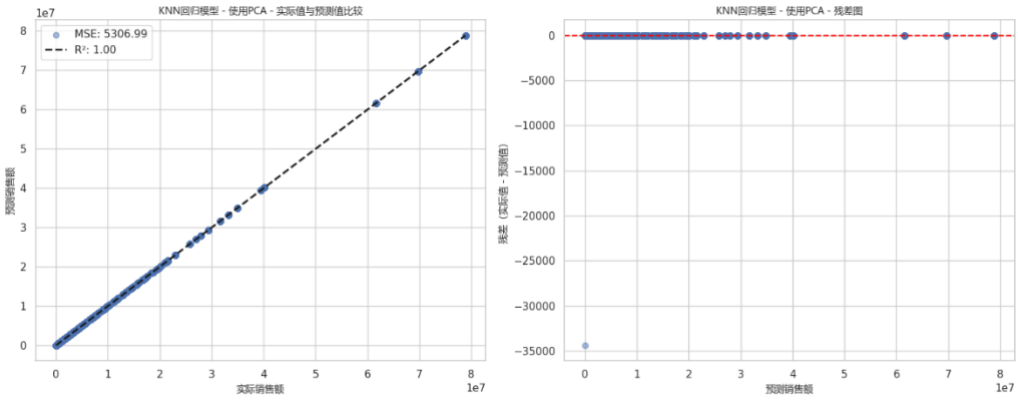

Figure 17. Visualization of Sales Amount KNN Model Evaluation (with PCA)

Table 11 and the corresponding visualizations (**Figures 18, 19, and 20**) provide a detailed overview of the model performance in predicting sentiment scores. The **Linear Regression model** performed poorly, with an **MSE of 0.10** and an **R^2 score of 0.00**, indicating that it was almost completely

ineffective at capturing any meaningful relationship between the features and sentiment score. **Figure 18** further confirms this outcome, showing little to no alignment between predicted and actual values.

The **Random Forest model** demonstrated a slight improvement, but the results remained suboptimal. While its **MSE remained at 0.10**, the **R^2 score increased only marginally to 0.01**, suggesting that even a more complex model like Random Forest struggled to capture underlying patterns in the data. As shown in **Figure 19**, the correlation between predicted and actual sentiment scores was still weak. The **KNN regression model** performed the worst, with an **MSE rising to 0.12** and an **R^2 score of –0.17**, indicating that its predictions were **less accurate than a naive average-based guess**. **Figure 20** illustrates this mismatch, showing a significant discrepancy between predicted and actual values. From the perspective of feature-target relationships, these results suggest that the selected features—**current price, number of reviews, popularity, rating, and discount ratio**—may have a **nonlinear or more complex relationship** with consumer sentiment. The models used may not be sufficient to capture these interactions, highlighting the potential need for more sophisticated modeling techniques or additional feature engineering in future research.

Table 11. Evaluation of Regression Models for Sentiment Score (with PCA)

Model	PCA Applied	MSE	R^2 Score
Linear Regression	Yes	0.10	0.00
Random Forest	Yes	0.10	0.01
KNN Regression	Yes	0.12	-0.17

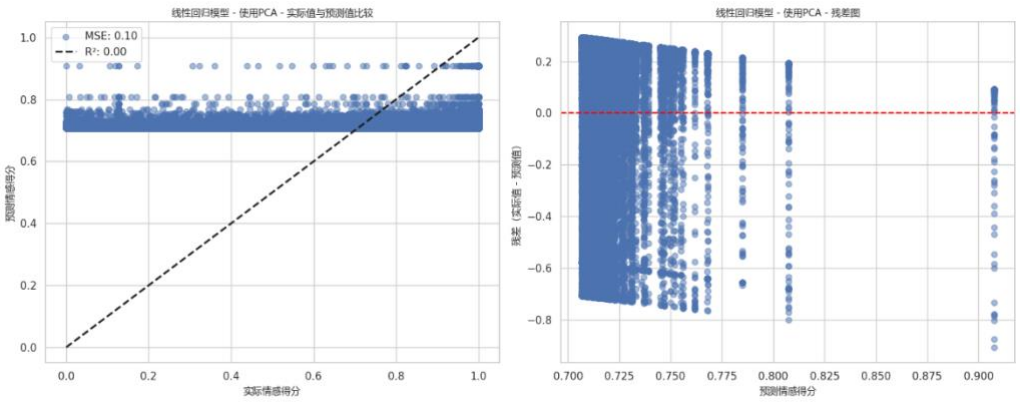

Figure 18. Visualization of Sentiment Score Linear Regression Model Evaluation (with PCA)

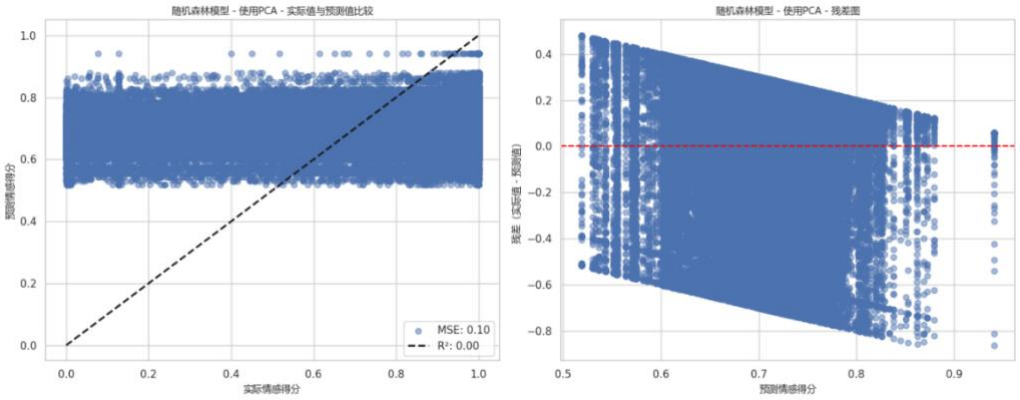

Figure 19. Visualization of Sentiment Score Random Forest Model Evaluation (with PCA)

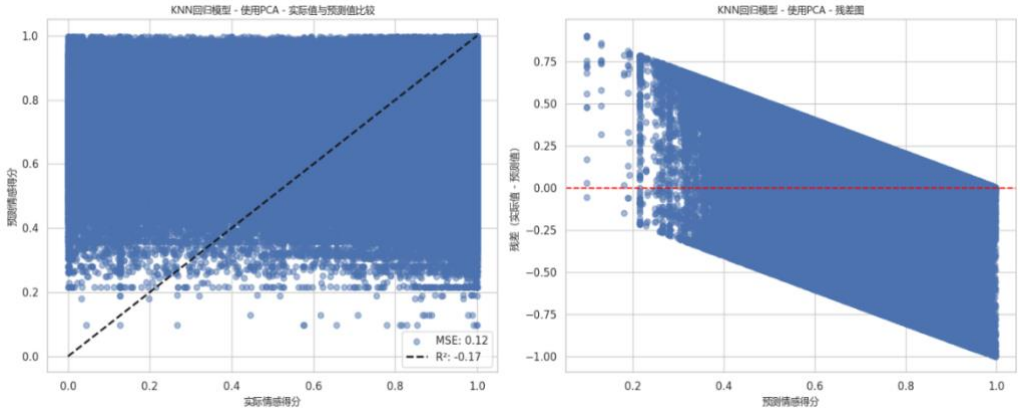

Figure 20. Visualization of Sentiment Score KNN Model Evaluation (with PCA)

5 CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

This study conducted an in-depth analysis of multidimensional data from Tmall Global, utilizing **Linear Regression**, **Random Forest**, and **K-Nearest Neighbors (KNN)** models, combined with **Principal Component Analysis (PCA)**, to predict both product **sales** and **sentiment scores**.

In the prediction of **sales performance**, both the Random Forest and KNN models performed exceptionally well, with the KNN model achieving near-perfect predictive accuracy. This highlights the **robust capabilities of complex models** when dealing with large-scale e-commerce datasets. In contrast, the Linear Regression model demonstrated **clear limitations** due to its theoretical assumptions, particularly its inability to capture **nonlinear relationships** effectively.

However, in the prediction of **sentiment scores**, none of the models yielded satisfactory results. This indicates that the selected features may lack a **direct or strong linear correlation** with consumer sentiment. It also underscores the fact that sentiment prediction is inherently more complex and may

require **richer, multidimensional data** or more **advanced modeling approaches**, such as **text-based sentiment analysis models**, in future research.

5.2 RECOMMENDATION

When working with e-commerce data, selecting the appropriate analytical model is **crucial**. In particular, for complex tasks such as **sales forecasting**, advanced models like **Random Forest** or **K-Nearest Neighbors (KNN)** have demonstrated superior performance and should be prioritized in similar use cases.

However, for more **subjective and multidimensional metrics** such as **sentiment scores**, traditional regression models may fall short in capturing the underlying dynamics. Future research should therefore consider incorporating **advanced analytical techniques**, such as **text-based Natural Language Processing (NLP)**, to enhance the understanding and prediction of consumer sentiment responses. Additionally, this study highlights the **need to expand the dimensionality of data**. Integrating more diverse data types—such as **user behavior data** and **market trend indicators**—can further enrich model inputs, improve predictive accuracy, and enhance generalization capabilities. Given the importance of interpretability in business decision-making, **model transparency** must also be emphasized. Especially when deploying complex machine learning models, it is essential that these models offer **clear, explainable insights** for stakeholders and decision-makers.

Lastly, considering the **ever-evolving nature** of the e-commerce market and consumer behavior, it is recommended that existing models undergo **periodic optimization and validation**. This ensures the **timeliness**, **accuracy**, and **relevance** of predictive outputs in dynamic market environments. By maintaining and iteratively improving models over time, data analytics tools can be more effectively leveraged to provide **robust decision support** for the cross-border e-commerce industry.