Amazon Sales Analysis

Jiaren Li, Yizhou Chen, Camilla Ren, Aaron Bai, Pengwei Xu

Abstract* — *In recent years, the application of analytics in e-commerce has increasingly become a key factor in competitive success. This report explored the combination of machine learning and statistical analytics to enhance Amazon's sales strategy. Focusing on product category, consumer spending, and sentiment analysis, our analyses incorporated advanced analytical methods such as the LightGBM model, random forest model, and NLP to identify relationships between product discounts, consumer ratings, and sentiment-driven purchase patterns.

I. Background

Advanced machine learning techniques in e-commerce had transformed the way companies like Amazon used large datasets to improve sales strategies and interact with consumers. As e-commerce became more complex, the need for sophisticated analytics to optimize product performance increased; a seminal study by Ghose and Ipeirotis (2010) had found that detailed analysis of pricing strategies and customer feedback could significantly impact sales and satisfaction.

Further research had highlighted the important role of machine learning in deciphering complex consumer data to predict market trends and customer preferences, including work by Thobani (1970) and recent insights by Li and Karahanna (2015). These approaches provided frameworks for organizations to proactively adapt strategies for a predictive advantage in competitive markets. In our work, we aimed to extend this foundational research through the application of various machine learning techniques to the analysis of massive product data from Amazon, synchronizing discounts, consumer ratings, and sentiment analysis to improve sales strategies and pricing models. Our goal with this study was to provide actionable insights that would help Amazon develop precise marketing strategies and pricing models that appeal to a diverse customer base.

The dataset used to build the models in this report had included more than 1,000 Amazon products, each with detailed information taken directly from Amazon's official website. This comprehensive dataset included 16 columns: Product ID, Product Name, Category, Discount Price, Actual Price, Discount Percent, Rating, Number of Ratings, Product Information, User ID, Username, Review ID, Review Title, Review Content, Image Link, and Product Link. For our analysis, we had specifically selected the most relevant columns, focusing on those that provided significant insight into product performance and consumer preferences: Product ID, Category, Discount Price, Actual Price, Discount Percentage, Rating, and Rating Count.

II. Data Prepocessing

For this analysis, we focused on ensuring data integrity and usability through various cleaning and transformation steps. First, we examined the dataset for zero values and missing data in selected columns. We chose to impute the column mean instead of removing missing values, thus preserving the completeness of the dataset. In addition, we addressed formatting inconsistencies by removing currency symbols (e.g. ₹) and commas from the Discount Price and Actual Price columns and converting these values to float types for accurate calculation.

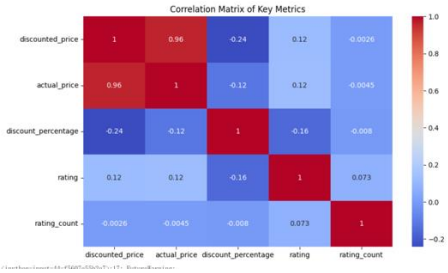
Other transformations included converting the Discount Percentage from a percentage string format (e.g., "10%") to a decimal float format (e.g., 0.1). We also restructured the category data by using a function to split the category column into a number of columns based on the '|' separator to allow for a more granular level of analysis. For instance, "Home and kitchen | heating, cooling, and Air Quality" was divided into separate category columns like category 1, category 2, etc. Missing values in the new category columns were filled in with empty strings if they were less than the maximum value of the data set.

To analyze the impact of these categories on Product Ratings and Discount Percentage, we aggregated the data by each new category column and summed the average ratings associated with each. This approach provided clear insights into how different categories influence consumer perceptions and pricing strategies, and allowed for detailed analysis and visualization.

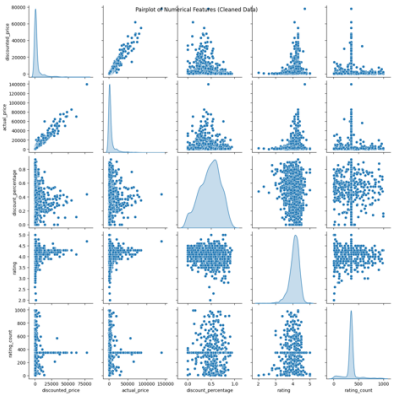
Finally, we identified a significant correlation between Actual Price and Discount Price, highlighting the dependence of discount practices on original product pricing. The specific relationship between pricing strategies and consumer feedback within this dataset was highlighted by the fact that no other significant correlations were found.

III. Visulizations

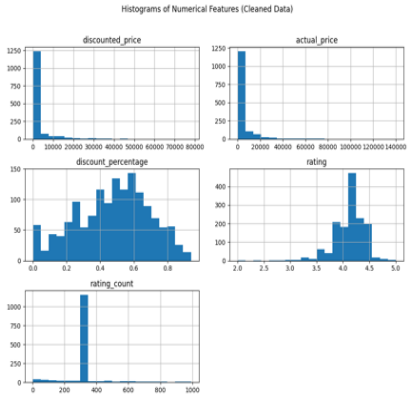
The correlation matrix showed a strong correlation between actual price and discounted price, providing a clear visualization of the relationship between key variables. However, there were no significant correlations between ratings and other variables, suggesting that pricing strategies may not have a direct impact on customer ratings.



The scatterplot matrix, which provides a overview of the pairwise relationships between numerical attributes, further supported these findings. In particular, a linear relationship was confirmed between actual and discounted prices. This suggests that discounts are typically proportional to actual prices. Conversely, there was no apparent relationship between prices and ratings, meaning that other factors have an impact on customer satisfaction.



Histograms of the numerical features, plotted on the diagonal of the scatterplot matrix, highlighted the distribution of each variable after cleaning. These visualizations showed that most of the ratings clustered around four, reflecting the generally positive feedback from customers. Understanding consumer behavior and tailoring marketing strategies accordingly requires insights such as these.



IV. Modeling and Analysis

A. *Product Category Analysis*

In our analysis of the product categories, the main objective was to determine whether there was a relationship between the rating and the product category. We created two types of visualization for each category column: bar charts to show the average rating associated with each category, and scatter plots to visualize the relationship between the percentage of discount and the rating, using different colors for the categories. The bar graphs used the x-axis for the different category values and the y-axis for the average ratings to visualize how the average ratings varied between categories.

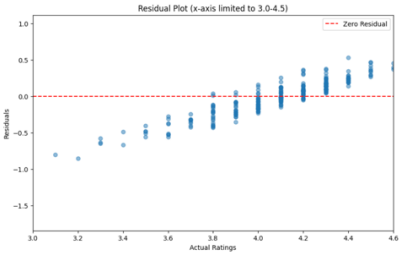
When analyzing the scatter plots illustrating the relationship between discount percentage and rating for different category columns (categories 1 to 5), a key observation emerged: The patterns across these scatter plots appeared remarkably similar, indicating a lack of significant variation or relationship between the different categories and the observed ratings.

To explore this further, we built two models: one containing both categorical and numerical variables, and another containing only numerical variables (actual price, discount price and percentage). If the full model had a lower RMSE, it would indicate a better fit than the reduced model, potentially demonstrating the importance of the categorical feature in the prediction of ratings. We chose to use LightGBM for our analysis because of its efficiency in dealing with a large number of categories and its ability to automatically manage the categorical features. This

approach allowed us to take full advantage of the complexity of the categorical data, while still benefiting from the speed and performance advantages of the model.

We then split the data into eighty per cent for training and twenty per cent for testing to build a LightGBM model for predicting scores. The first model did not include the categorical features and achieved an RMSE of 0.27423 on the test data, while the second model did include the categorical feature and achieved an RMSE of 0.2655. The categorical feature did not have a significant relationship with rating, as indicated by the small difference between the two results. The residuals from both models were similar, supporting this conclusion. Only actual price, discount price and rating significantly influenced the prediction of ratings.

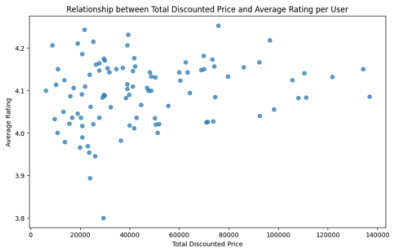
A limitation was observed in the residuals where data points consistently fell around the line y=0. Transforming the rating (y-value) could potentially improve the RMSE result. In addition, most of the ratings in the dataset were concentrated between 3.5 and 4.5, with very few ratings in other intervals, although the ratings ranged from 0 to 5. This distribution reduced the ability of the model to learn effectively due to the lack of data at the extremes, which significantly increased the difficulty of the predictive model. Therefore, the selection of a more balanced data set could lead to different and potentially better results.



B. *Consumer Segment Analysis*

We analyzed the relationship between user ratings and prices (actual and discounted) grouped by user ID to determine correlations and patterns. Key metrics included only total discounts per user, total actual prices per user, and average ratings per user.

As shown in the scatter plot, a weak positive correlation was found between total discounted price and average ratings per user. Although the trend was not entirely linear, it was observed that users who spent more or received bigger discounts often gave higher ratings. Most of the scores were concentrated in the 20,000 to 40,000 price range but were scattered, suggesting that merchants' discounting activities alone did not guarantee higher satisfaction. Some outliers were also found, like the lowest outlier score at a discounted price of 28,000, which indicated that factors other than product prices such as product quality, expectations, or delivery experience also influenced customer satisfaction. While deep discounting could improve ratings, it was not the only factor that could enhance user satisfaction. It was also important to address other factors such as product quality and delivery times.



Then, we built a random forest model to predict future user ratings. After cross-validation and Randomized SearchCV optimization, the mean square error (MSE) was 0.0097. Through the improvement of the residual graph, it was shown that the postmortem analysis using Cook distance improved the stability of the model. The actual vs. forecast graph showed that the forecast was very close to the actual score.

In conclusion, offering discounts could improve user satisfaction, but optimizing factors such as product quality, delivery, and user experience was also critical. In the future, if further analysis of product categories, user demographics, and purchase frequency is added, this will provide greater insight into the main factors that improve customer satisfaction.

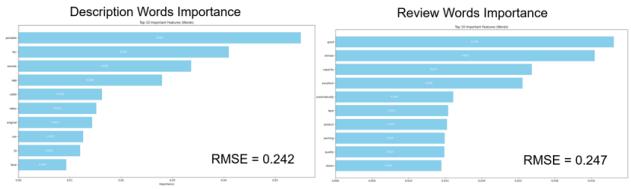
C. *Customer Sentiment Analysis*

This chapter focused on applying Natural Language Processing techniques to analyze the product description and customer review datasets. By using clustering algorithms, dimensionality reduction, and feature importance evaluation, we aimed to uncover patterns and assess the impact of text data on product ratings.

The text data from product descriptions and customer reviews were cleaned by removing null values and irrelevant words. This step ensured higher-quality input for subsequent analyses. We applied Principal Component Analysis for dimensionality reduction and performed K-means clustering to categorize the data into several clusters. We used the elbow method and silhouette score, and we observed:

For the description of the product, there was no clear elbow point in the elbow method. Silhouette scores increased with more clusters, indicating that K-means might not have been the optimal approach for this dataset. However, the results showed that reviews should have been grouped into 11 clusters based on K-means analysis. In future studies, the existing results could be used to quickly classify and predict the types of commodities according to the descriptions of commodities.

Also, to evaluate the influence of text features on ratings, a Random Forest model was used. The top 500 words were selected as features. For product description, the model achieved a root mean squared error (RMSE) of 0.242 on the test set. Similarly, for customer review, processing yielded an RMSE of 0.247. The importance of words is shown as follows. While the random forest model achieved good results, other advanced models such as XGBoost or LightGBM could be tested to further reduce RMSE.



V. Conclusion and Future Works

Our analysis successfully demonstrated how advanced machine learning and statistical analysis could be used to optimize Amazon's sales strategies by increasing the efficiency of product sales and improving profit margins. By integrating LightGBM and Random Forest models, and using NLP for sentiment analysis, we analyzed the complex interplay between product discounts, consumer reviews, and buying patterns. Our findings showed that while discounts directly correlated with consumer ratings, factors like product price and quality also played a crucial role in influencing customer satisfaction.

The LightGBM model revealed that categorical features, such as product category, had minimal impact on predicting consumer ratings. However, numerical variables such as actual price and discount rates were more predictive of customer satisfaction and likelihood to purchase. The use of Random Forest models in our analysis of consumer segments showed that deeper discounts were correlated with higher user ratings, suggesting that strategic discounting could not only improve user satisfaction but also enhance overall sales performance. However, our sentiment analysis using NLP techniques highlighted that while discounting increased reviews, optimizing product quality and delivery experiences was equally important to maintain high customer satisfaction.

Further research could explore the impact of real-time analytics on sales strategies. This could improve the ability to predict seasonal trends and shifts in consumer behavior. Additionally, a more balanced data set across a wider range of reviews could refine the predictive accuracy of our models and provide a richer insight into the full range of customer preferences and behaviors.

In summary, our study confirmed the powerful role of sophisticated analytics in helping Amazon refine its sales strategy, and underscored the importance of targeted pricing, strategic discounts, and continuous improvement in product and service quality in maximizing profitability and customer satisfaction.

References

[1] *Cheung, C. M. K., Limayem, M., Chan, G. W. W., Kwong, T., & Zhu, L. (2015). Online Consumer Behavior: A Review and Agenda for Future Research. Management Science, 32(6), 1238-1263.*

[2] *Ghose, A., & Ipeirotis, P. (2010). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics | IEEE Journals & Magazine | IEEE Xplore. https://ieeexplore.ieee.org/document/5590249/*

[3] *Thobani, S. (1970). Improving e-commerce sales using machine learning.*
https://dspace.mit.edu/handle/1721.1/118511

[4] *TJ, K. (2023, January 17). Amazon Sales Dataset. Kaggle.*
https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset/data

Appendix

Jiaren Li for Customer Sentiment Analysis; Yizhou Chen for Consumer Segment Analysis; Aron Bai for Product Category Analysis; Camila Ren for Background and Conclusion; Pengwei Xu for Dataset analysis.