ISSN 2723-6471

Exploring the Impact of Discount Strategies on Consumer Ratings: An Analytical Study of Amazon Product Reviews

Berlilana[1,*], Arif Mu'amar Wahid [2], Dewi Fortuna[3], Alfin Nur Aziz Saputra[4], Galih Bagaskoro[5]

[1,2,3,4,5] *Magister of Computer Science, Computer Science Faculty, Universitas Amikom Purwokerto, Indonesia*

(Received: November 28, 2023; Revised: December 18, 2023; Accepted: January 14, 2024; Available online: January 29, 2024)

Abstract

This research delves into the influence of discount strategies on consumer ratings within the e-commerce landscape, particularly on Amazon. A logistic regression model assessed how discount percentages and product categories affect consumer ratings. The study followed a rigorous methodology, beginning with comprehensive data collection across diverse product categories on Amazon. This was succeeded by a detailed exploratory data analysis (EDA), data preprocessing, and subsequent model building. The model was then subjected to an extensive evaluation process, encompassing accuracy, precision, recall, F1-score, and ROC-AUC metrics. The evaluation revealed that the model achieved an accuracy of 74.94%, a precision of 72.69%, and a recall of 74.94%. The F1 score was calculated at 69.26%, and the ROC-AUC score was notably 78.24%. These metrics underscore the model's capability to accurately predict consumer ratings influenced by discount strategies. Key findings highlighted the significant predictive power of discount percentages and specific product categories, particularly 'Home & Kitchen', suggesting a complex relationship between discounts, product types, and consumer responses. Theoretically, the study enriches the understanding of consumer behavior in e-commerce, highlighting the nuanced impact of discount strategies on consumer satisfaction, especially in online retail contexts. For e-commerce businesses and marketers, the findings underscore the importance of strategically employing discount strategies and tailoring marketing approaches to specific product categories. This study emphasizes managing customer expectations and maintaining product quality alongside discounts. This research provides valuable insights for optimizing e-commerce strategies and paves the way for future investigations. It opens up avenues for further exploration into factors like product quality, brand reputation, shipping times, and the potential of consumer segmentation and sentiment analysis in enhancing marketing effectiveness. The study marks a significant contribution to the field by linking discount strategies with consumer ratings, using advanced data analytics to inform e-commerce practices in the digital age.

Keywords: Consumer Behavior, Consumer Ratings, Discount Strategies, E-commerce, Logistic Regression, Online Retail

1. Introduction

E-commerce has significantly evolved to become a significant driver in the global economy, impacting both developed and developing nations. In [1] emphasize its potential to foster the growth of SMEs, particularly in regions like China and Asia, propelled by cross-border e-commerce [2], [3], [4]. The pandemic has further accelerated this growth, integrating e-commerce into various industries [5]. The Indian market, for example, is expected to reach remarkable valuations, suggesting a strong link between e-commerce activities and economic growth [6], [7], [8].

Technological adoption within e-commerce, crucial in developing countries, hinges on digital readiness, affecting factors like ease of use and attitudes toward e-commerce [9], [10]. Concurrently, the role of data science in e-commerce has expanded significantly, with big data analytics and AI reshaping the online retail landscape. This is well-articulated by [11], [12], who explore the application of big data from both vendor and customer perspectives. The pandemic highlighted the need for data science in marketing, calling for practical strategies for competitive advantages [13]. The governance and ethical considerations of data management in e-commerce platforms are also crucial aspects of this evolution [14], [15], [16].

Additionally, product reviews have become vital in e-commerce for understanding customer satisfaction. Techniques like text mining and sentiment analysis, as used by [17], [18], extract insights from online reviews, emphasizing user experiences on platforms like Airbnb and Amazon. These methods offer a nuanced understanding of customer preferences, which is crucial for enhancing products and services [17], [18], [19]. The strategic implementation of

***Corresponding author: Berlilana (berli@amikompurwokerto.ac.id)**
DOI: https://doi.org/10.47738/jads.v5i1.163

discounts reveals the balance e-commerce platforms must maintain in consumer loss aversion and repurchase intentions, illustrating the profound impact of discounts on consumer behavior [20], [21].

Analyzing discount strategies in e-commerce is paramount for understanding consumer behavior and optimizing sales strategies. In [22] demonstrate how price discounts can significantly elevate sales volumes in social e-commerce, reflecting the interactive nature of online consumers. This is complemented by [23], who show the positive impact of discounts on customer selection, interest, and satisfaction, especially in sectors like food ordering. In [24] delve into the influence of online deals and discounts on consumer purchase behavior, emphasizing the vital role of discounting in guiding consumer choices and online purchasing trends.

These studies highlight the necessity for e-commerce platforms to comprehend and strategically implement discounts to boost consumer satisfaction and sales. In [25], the influence of demographic factors like age and gender on consumer behavior toward online discounts is investigated, uncovering complex aspects of consumer psychology. Similarly, [26] explores Indian consumers' attitudes towards online discounts, suggesting the potential of such strategies to convert traditional shoppers into online buyers. This underscores the ability of discounts to not only drive sales, expand the consumer base, and foster loyalty.

Furthermore, the role of tailored online promotions in driving purchase intent, as indicated by [26], [27] research on the drivers of online purchase intention, such as promotion, word of mouth, and consumption rituals, showcases the diverse impact of discount strategies. These findings are invaluable for e-commerce businesses in crafting competitive discount policies that attract consumers and enhance their shopping experience. The studies establish a clear connection between discount strategies and consumer behavior, emphasizing the need for e-commerce platforms to leverage data science and analytics to optimize these strategies.

The realm of discount strategies and their influence on consumer behavior and purchasing decisions has been extensively explored in various studies. In [20] acknowledge the effectiveness of discounts in driving consumer buying behavior. In [28] delve into the correlation between consumer types and discount strategies, shedding light on their effect on supplier profitability and e-commerce platform choices. Moreover, [21] investigates the impact of sales promotion strategies on consumers' purchasing decisions, including discounts and gift-giving, revealing the psychological aspects of consumer response to discounting. In [29], [30] further examine the multifaceted influences of discount pricing on consumer behavior, particularly post-purchase perceptions and loss aversion.

These studies are complemented by [31], who explore the effects of quoted discounts and historical promotions on consumer valuations, a crucial aspect for marketers in optimizing discount strategies. In [32] highlights the role of dynamic pricing strategies, especially during economic downturns, and their impact on consumer behavior. In [33] discuss the targeted discounts aimed at specific consumer groups, such as older people and children, and how these strategies can sway purchasing decisions. In [34], it emphasizes the importance of modeling loss-averse consumer behavior in the context of dual-channel supply chain pricing strategies.

This literature review underscores the complex and dynamic relationship between discount strategies and consumer behavior. It highlights the intricate link between these strategies, consumer psychology, and decision-making processes, pivotal in e-commerce success and customer satisfaction. The review also indicates the diverse methodologies employed in these studies, ranging from empirical analyses to theoretical models, providing a comprehensive understanding of the subject matter.

While the existing literature provides extensive insights into the dynamics of discount strategies and consumer behavior, there are notable gaps, particularly in applying advanced data analytics to understand the relationship between discounts and consumer ratings on platforms like Amazon.

Firstly, much of the current research focuses on the immediate impact of discount strategies on sales volumes and consumer purchasing behavior. However, more in-depth studies need to analyze the long-term effects of discounts on consumer loyalty and satisfaction ratings. This includes a need for a comprehensive analysis of how repeated discount exposure might alter consumer perceptions of value and quality over time.

Secondly, while studies like those of [22], [23] touch upon the effects of discounts on consumer behavior, there is a gap in research that systematically integrates and analyzes large-scale consumer data using advanced machine learning techniques, such as deep learning, to predict consumer ratings. Such approaches can offer more nuanced insights into the varying impact of discounts across different product categories and demographic segments.

Thirdly, the existing literature predominantly relies on conventional statistical methods for data analysis. The potential of more sophisticated data science methodologies, including logistic regression and predictive analytics, to unravel complex patterns within consumer ratings in response to discount strategies still needs to be explored. This includes exploring how external factors, like economic trends and social media influence, influence discount strategies in shaping consumer ratings.

Finally, there is a need for more empirical studies that specifically examine the unique dynamics of e-commerce platforms like Amazon. Given Amazon's diverse product range, customer base, and intricate rating system, understanding how its specific discount strategies affect consumer ratings can offer valuable insights for e-commerce businesses. Addressing these gaps requires a blend of advanced data analytics, longitudinal study designs, and platform-specific research to deepen our understanding of how discount strategies influence consumer ratings, particularly in the ever-evolving context of online retail platforms like Amazon.

The central research question for this study is, "How do discount strategies impact consumer ratings on an e-commerce platform, as analyzed through logistic regression?" This question aims to unravel the nuanced relationship between the strategic application of discounts and the resultant consumer ratings on an e-commerce platform like Amazon. It acknowledges the complexity inherent in consumer responses to discounting and seeks to quantify this relationship using logistic regression. This data-driven approach aims to establish a correlation and predict the impact of varying discount strategies on consumer ratings. By focusing on logistic regression, the study intends to leverage its capabilities in handling categorical data and providing probabilistic predictions, particularly suited for understanding consumer rating behaviors.

With this rapid growth, e-commerce has become a crucial sales channel and a battleground for innovative and competitive marketing strategies. One of the most pivotal strategies in the e-commerce landscape is discounts. As a marketing tool, discounts can influence consumer behavior and purchasing decisions. Concurrently, consumer ratings serve as critical indicators reflecting the success or shortcomings of these discount strategies. This study aims to delve deeper into how discount strategies impact consumer ratings in e-commerce, providing new insights into the interplay between marketing tactics and consumer response in the digital era.

2. Literature Review

2.1. Customer Behavior

Consumer behavior refers to the study of individuals and organizations and how they select, buy, use, and dispose of goods and services. It is a crucial aspect of marketing that helps businesses understand the factors influencing consumers' decisions. The first factor influencing consumer behavior is cultural influences. Cultural factors such as values, beliefs, customs, and lifestyles shape consumers' preferences and choices. For example, cultural differences between regions or countries may lead to variations in product preferences and consumption patterns.

Another key aspect of consumer behavior is psychological influences. This involves understanding the mental processes and emotions that drive consumer decisions. Factors like perception, motivation, learning, and attitude play significant roles in shaping how individuals perceive and respond to marketing stimuli. For instance, a positive attitude towards a brand may lead to brand loyalty, while negative perceptions can deter consumers from making a purchase.

Social influences also play a crucial role in shaping consumer behavior. People are social beings, and their choices are often influenced by family, friends, and other reference groups. Social factors include reference groups, family, social roles, and status. For instance, a consumer might choose a particular brand or product because it aligns with the preferences of their social circle or fulfills a specific social role.

Furthermore, personal factors such as age, gender, income, and lifestyle also contribute to consumer behavior. Age, for example, may impact product preferences, with different age groups having distinct needs and preferences. Income levels can influence the affordability of certain products, while lifestyle choices can determine the types of products and brands that align with an individual's values and interests.

In conclusion, consumer behavior is a multifaceted field influenced by cultural, psychological, social, and personal factors. Understanding these factors is essential for businesses to develop effective marketing strategies and cater to the diverse needs and preferences of their target audience. By gaining insights into consumer behavior, businesses can create products, services, and marketing campaigns that resonate with their customers and build long-lasting relationships with them.

2.2. Logistics Regression

Logistic Regression is a statistical method widely used for binary classification problems, where the outcome variable has two possible categories. Unlike linear regression, which predicts a continuous outcome, logistic regression predicts the probability of an observation belonging to a particular category. This method is particularly valuable in situations where the dependent variable is dichotomous, such as predicting whether an email is spam or not, or whether a patient has a certain medical condition.

One key aspect of logistic regression is the logistic function, also known as the sigmoid function, which transforms the output of the linear equation into a range between 0 and 1. This is crucial for interpreting the results as probabilities. The logistic regression model estimates the relationship between the independent variables and the log-odds of the outcome, providing insights into the odds of an event occurring. The model is trained using a process called maximum likelihood estimation, which aims to maximize the likelihood of observing the given set of outcomes based on the estimated probabilities.

Logistic regression offers several advantages, such as simplicity, interpretability, and efficiency in situations with a binary outcome. It is less susceptible to overfitting compared to more complex models and requires relatively small amounts of data for effective training. However, it is important to note that logistic regression assumes a linear relationship between the independent variables and the log-odds of the outcome, which may not always hold true in real-world scenarios.

Despite its strengths, logistic regression has limitations. It may struggle with nonlinear relationships and interactions between variables. Additionally, it is not well-suited for problems with multiple categories in the dependent variable. In such cases, alternative techniques like multinomial logistic regression or other classification algorithms may be more appropriate. Overall, logistic regression is a powerful tool in the data scientist's toolkit, particularly when dealing with binary classification problems and a need for interpretable results.

2.3. Data Science on Online Retail

Data Science plays a crucial role in the realm of online retail, leveraging advanced analytics and algorithms to extract meaningful insights from vast datasets. Firstly, data science is instrumental in understanding customer behavior. By analyzing online shopping patterns, preferences, and purchase histories, retailers can tailor their marketing strategies and personalize the shopping experience. This leads to more effective targeted advertising and recommendations, ultimately enhancing customer satisfaction and loyalty.

Data science helps optimize inventory management. Through predictive analytics, retailers can forecast demand more accurately, ensuring that products are stocked in sufficient quantities and reducing the likelihood of overstock or stockouts. This not only improves operational efficiency but also contributes to cost savings and increased profitability. Moreover, machine learning algorithms can identify trends and patterns in sales data, aiding retailers in making informed decisions about product assortment and pricing strategies.

Fraud detection is a critical aspect of online retail, and data science provides powerful tools for identifying and preventing fraudulent activities. Advanced algorithms analyze transaction data, looking for anomalies and patterns associated with fraudulent behavior. This proactive approach helps protect both customers and retailers, fostering a secure online shopping environment.

Data science contributes to the enhancement of the user experience on online retail platforms. By analyzing customer feedback, website navigation patterns, and user interactions, retailers can make data-driven improvements to their websites or mobile apps. This leads to a more seamless and enjoyable shopping experience, encouraging customers to spend more time on the platform and increasing the likelihood of conversions.

The application of data science in online retail is multifaceted, impacting various aspects of the business. From understanding customer behavior to optimizing inventory management, detecting fraud, and enhancing the user experience, data science is a powerful tool that enables retailers to stay competitive in the dynamic and evolving landscape of e-commerce. As technology continues to advance, the role of data science in online retail is expected to grow, providing retailers with increasingly sophisticated tools to thrive in the digital marketplace.

3. Method

The research adopted a quantitative design, utilizing logistic regression analysis to investigate the impact of discount strategies on consumer ratings on Amazon. Logistic regression was chosen due to its proficiency in analyzing binary outcomes, which aligns with the study's objective of categorizing consumer ratings into distinct classes. This method is particularly effective in examining the relationships and impacts of multiple independent variables, like discount percentages and product categories, on a categorical dependent variable, like consumer ratings. Its ability to handle multiple predictors simultaneously makes logistic regression well-suited for exploring the intricate dynamics of consumer rating patterns in e-commerce. The model-building process encompassed exploratory data analysis for variable selection, data cleaning for missing values and outliers, and data transformation to fit logistic regression requirements. The model's performance was rigorously evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and cross-validation to affirm its robustness and generalizability.

3.1. Data Collection

Data for the study was meticulously collected from Amazon, a leading e-commerce platform, ensuring a diverse and representative dataset. The data encompassed various product categories, with a particular emphasis on those often subject to discounts. Detailed discount information, including original and discounted prices and the discount percentage, was a crucial part of the dataset. This information was central to analyzing the influence of discount levels on consumer ratings.

3.2. Exploratory Data Analysis (EDA)

Following data collection, an exploratory data analysis was conducted. This step thoroughly examined the dataset's structure, including the distribution of key variables like discount percentages and consumer ratings. The EDA aimed to identify initial patterns, anomalies, and relationships in the data, providing a foundation for more in-depth analysis and model building.

The Exploratory Data Analysis (EDA) for this study, which focused on the impact of discount strategies on Amazon's consumer ratings, involved identifying key data trends and potential anomalies. Initially, the dataset's structure and variable types were examined, including the number of rows and columns and the categorization of variables. This phase also entailed a preliminary understanding of critical variables such as product categories and discount information.

The Amazon product reviews dataset was structured with 1,467 rows and 16 columns, encompassing a diverse range of information pertinent to product listings, user interactions, and review details. The data types within the dataset were diverse, reflecting the multifaceted nature of e-commerce data. Specifically, the dataset included the following types: a) Object (String) was predominant in fields capturing textual or categorical data, such as product_id, product_name, category, rating, about_product, user_id, user_name, review_id, review_title, review_content, img_link, and product_link. These fields provide detailed descriptive data about the products, users, and reviews; b) Float64 as Numerical data were represented in this format, including fields like discounted_price, actual_price, discount_percentage, and rating_count. These fields offer quantitative insights into pricing, discounting strategies, and user engagement metrics.

Upon an initial inspection of the dataset, several potential features for modeling were identified and categorized based on their data types and the nature of the information they represent a) Numerical Features, such as discounted_price, actual_price, discount_percentage, and rating_count were identified as numerical features. Although the rating is intrinsically a numerical measure, it was listed as an object type in the dataset, likely due to inconsistent formatting or the inclusion of non-numeric characters; b) Categorical Features, identifiers, and classificatory information, such as product_id, product_name, and category, were classified as categorical features. These features provide a basis for grouping, segmenting, or filtering the data in various analyses; c) Textual Features, descriptive fields containing free-form text, including about_product, review_title, and review_content, were noted as textual features. These fields hold rich qualitative information that could be leveraged for sentiment analysis, topic modeling, or feature extraction; and d) Others, Additional fields, such as user_id, user_name, review_id, img_link, and product_link, were also part of the dataset. While not directly related to product features or user ratings, these fields could provide auxiliary information for user-centric analyses or content linkage.

Addressing missing values was a critical step, with strategies like imputation or removal being implemented based on the nature and extent of missing data. While preparing the dataset for analysis, it was noted that it presented missing values across several columns. An examination of the dataset revealed missing values in various columns, ranging from product identifiers and categories to pricing details and review content. Specifically, columns such as product_id, product_name, category, discounted_price, actual_price, discount_percentage, rating, rating_count, about_product, user_id, user_name, review_id, review_title, review_content, img_link, and product_link had missing values, with most columns missing 2 values and rating_count missing 4.

To address this, a strategic approach was adopted for handling the missing values, tailored to the nature of the data in each column. For categorical and textual data such as product_id, product_name, category, about_product, user_id, user_name, review_id, review_title, review_content, img_link, and product_link, the chosen strategy was to impute with the mode, representing the most frequent value, or to use a placeholder value like "Unknown". This approach aimed to preserve data integrity and ensure consistency in categorical and textual features. The rating column, a mixed-type data field appearing as a numeric variable but listed as an object, required a more nuanced approach. Ultimately, the decision was made to remove all rows containing missing values from the dataset. This led to a cleaned dataset comprising 1,463 rows and 16 columns, indicating the removal of only 4 rows due to missing values. This careful and strategic handling of missing data ensured that the dataset was now suitably prepared, devoid of missing values, and ready for the subsequent in-depth analysis and preprocessing steps. This approach balances data integrity and the practical necessity of a complete dataset for robust analysis.

The following step is the distribution of key variables visualized to understand the underlying trends and patterns, providing valuable insights into the dataset's characteristics. The visualization is shown in Figure 1 below.

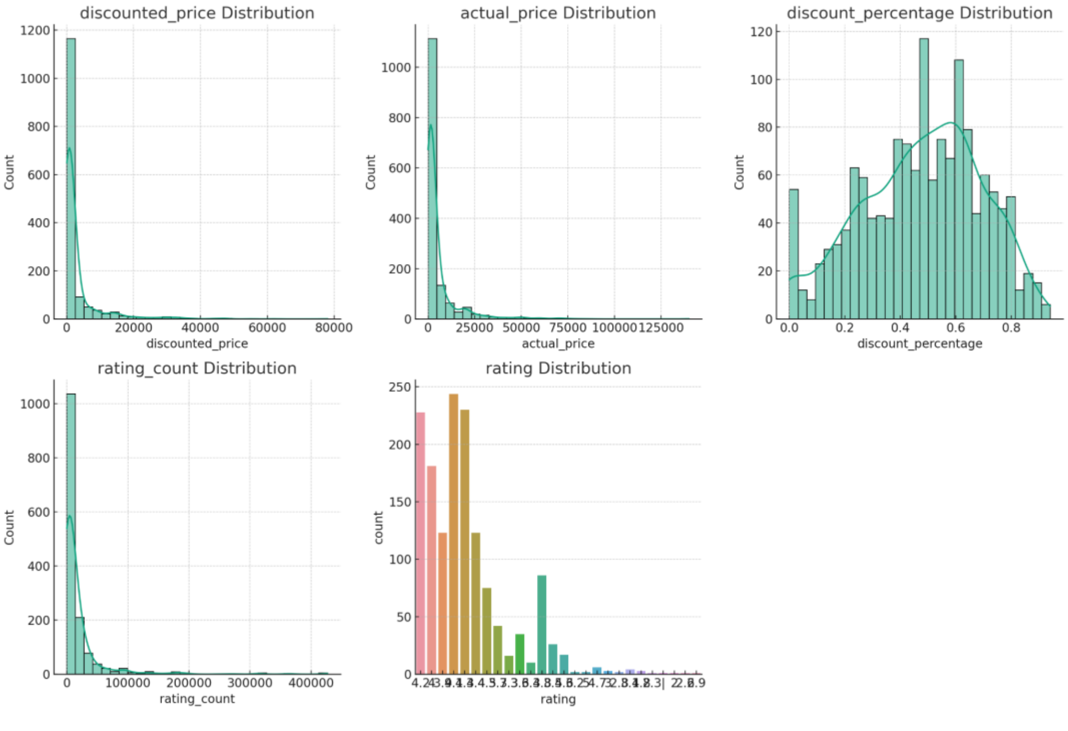

Figure 1. Distribution Plot of Key Variables

The distribution of the Discounted Price showed a right-skewed trend, where most of the values were concentrated in the lower price range. This indicates that many products are available at lower discounted prices, with fewer products having higher discounted prices. Similarly, the Actual Price distribution also exhibited a right-skewed nature, suggesting that products with lower actual prices are more common than those with higher prices. This skewness in pricing data indicates a wider range of lower-priced products on the platform.

The Discount Percentage distribution presented more variability compared to the price distributions. Despite the variety, there was a noticeable concentration of values in specific ranges, hinting at typical discount percentages applied by sellers. This pattern might reflect standard discounting practices or promotional strategies in the e-commerce marketplace. The Rating Count displayed a right-skewed distribution, suggesting that while most products have a relatively low count of ratings, there are a few products with a significantly high count of ratings. This skewness might be attributed to varying levels of customer engagement or product popularity. The distribution of Ratings was particularly intriguing. The concentration of product ratings in specific categories revealed standard ratings given by users, a characteristic often observed in product ratings where certain rating levels (like 4 or 5 stars) are more frequently assigned by customers.

The observations regarding skewness and unusual distributions in these numerical variables – discounted_price, actual_price, discount_percentage, and rating_count – highlight a common trend of right-skewed distributions. Such skewness could potentially impact the performance of predictive algorithms and might necessitate data transformation or normalization to ensure that the data conforms to the assumptions of the analytical models used. In the case of the rating variable, which was treated as categorical for this analysis, the concentration in specific rating values underscores typical consumer rating behavior. Understanding this pattern is crucial for businesses aiming to interpret customer feedback and improve product offerings based on consumer ratings. These visualizations and observations serve as a foundation for further analysis, enabling a more nuanced understanding of the data and guiding the subsequent steps in the modeling process.

In the correlation analysis of the numerical variables within the dataset, intriguing relationships were uncovered, providing valuable insights for model construction and feature selection. The visualization of the correlation analysis is shown in Figure 2 below

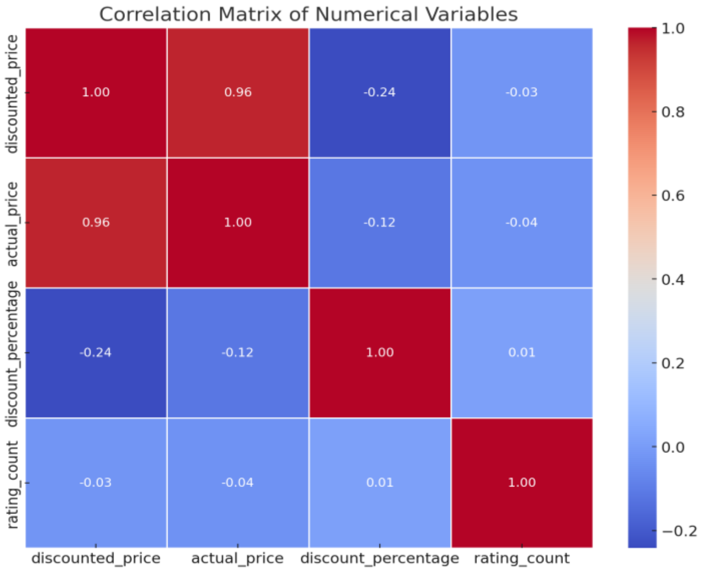

Figure 2. Correlation Matrix of Numerical Variables

A high positive correlation of 0.96 was observed between the Discounted and Actual Prices. This strong correlation suggests a direct relationship where, typically, as the actual price of a product increases, its discounted price tends to increase proportionately. However, the implication of this finding is significant for model building. The high correlation indicates a potential multicollinearity issue if both variables are included in a predictive model, as they provide overlapping information that could affect the model's performance.

A slightly negative correlation of -0.12 was observed between Actual Price and Discount Percentage. This implies that products with higher actual prices tend to have slightly lower discount percentages, though the strength of this correlation is not profound. This suggests that the actual price of a product does not significantly influence the discount percentage applied, indicating that these variables can potentially be considered independently in predictive modeling. When examining the Rating Count's relationship with Discounted and Actual Price, very weak negative correlations of -0.03 and -0.04 were noted. This finding implies a very slight tendency for products with higher prices to have fewer ratings, but the strength of this relationship is not strong enough to suggest a significant influence. This indicates that price and rating count can be independently considered in a model without the concern of strong interdependence.

The analysis of categorical variables within the dataset, mainly their frequency and relationship with the target variable 'rating', unveiled several key insights instrumental for model building and understanding consumer behavior. The visualization is shown in Figure 3 below

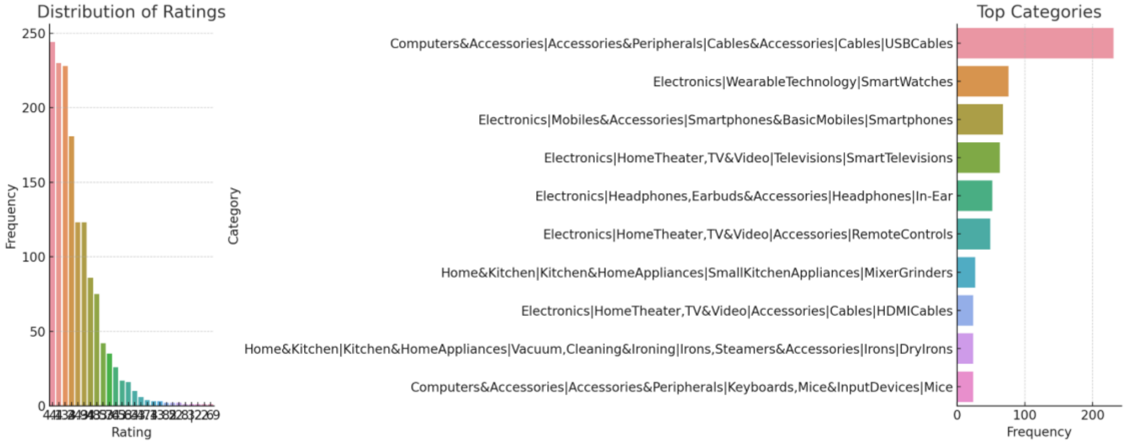

Figure 3. Distribution of Ratings and Categories Plot

In the frequency analysis of different categories, it was observed that certain categories, notably 'Computers & Accessories', 'Electronics', and 'Home & Kitchen' appliances, were more prevalent in the dataset. Specifically, the 'Computers & Accessories | Accessories & Peripherals | Cables & Accessories | Cables | USB Cables' category emerged as the top category with 231 occurrences, highlighting its prominence in the dataset.

Similarly, the Product ID and Product Name variables analysis revealed that certain products and product names appeared more frequently than others. For instance, the product ID 'B07JW9H4J1' was noted to appear 3 times, and the product name 'Fire-Boltt Ninja Call Pro Plus 1.83" Smart Watch' was recorded 5 times. This repetition of certain products and product names indicates their recurring presence in the dataset, which might reflect their popularity or prevalence.

The distribution of the target variable, 'rating', revealed a tendency towards higher ratings, with the most frequent rating being 4.1, followed by 4.3 and 4.2. This distribution showed a concentration in the range of 4.0 to 4.3, suggesting that most products tend to receive good ratings. This trend indicates general consumer satisfaction with the products listed in the dataset.

While the categorical data analysis provided valuable insights into the frequency of categories and the distribution of ratings, it's important to note the limitations of this analysis. The analysis does not directly reveal the relationship between the categorical variables and the rating. Advanced techniques such as Chi-Square tests or logistic regression with categorical predictors could be employed to delve deeper into this relationship. Additionally, due to the large number of unique categories and products, the analysis was focused on the top occurrences, potentially omitting insights from less common but equally significant categories.

3.3. Data Preprocessing

The data preprocessing phase, integral to preparing the dataset for logistic regression analysis, involved several steps to ensure data quality and appropriateness for the model. The initial focus was on data cleaning, beginning with identifying and handling missing values across the dataset. Strategies like imputation for numerical data, where missing values were substituted with mean or median values, and modes or placeholder values for categorical data were employed.

Outlier detection formed a critical part of the data analysis process in this study, with the Interquartile Range (IQR) method employed to identify outliers across various numerical variables in the dataset. Specifically, 217 outliers were detected for the Discounted Price, while the Actual Price had 213 outliers. Interestingly, the Discount Percentage variable had no outliers, and the Rating Count had 141 outliers. The presence of these outliers required careful consideration regarding handling them to ensure the integrity and validity of the subsequent analysis.

Retaining outliers in the dataset was driven by recognizing that these outliers might represent genuine variations in the data rather than errors or anomalies. In e-commerce data, where extreme values can often carry significant meaning, such as luxury or highly discounted products, retaining outliers is a valid approach that can provide a more accurate representation of the real-world scenario.

Duplicate entries in the dataset can compromise the accuracy of the analysis, and hence, a thorough check for duplicates was conducted. Initially, the dataset comprised 1,463 rows. After the check, the number of rows remained unchanged at 1,463, confirming that no duplicates were present in the dataset. This verification ensured that each entry in the dataset was unique and contributed distinct information to the analysis.

The 'rating' column was also carefully examined for non-numeric or unexpected values. Given the quantitative nature of ratings, ensuring that this column contained only numeric values was crucial. Non-numeric values identified in the 'rating' column were converted to NaN (Not a Number), and the corresponding rows were subsequently removed from the dataset. This step was taken to maintain the consistency and reliability of the data, particularly in preparation for numerical analysis and modeling. Initially, the dataset had 1,463 rows, and after this correction, one row was removed, resulting in a total of 1,462 rows.

The feature engineering phase of the study introduced new features into the dataset aimed at enhancing the logistic regression model's performance by capturing additional aspects of the data that might be relevant for predicting the target variable.

One of the significant additions was the creation of the Price Range feature. This new categorical variable was derived from the actual_price and involved categorizing products into distinct price ranges such as 'Very Low', 'Low', 'Moderate', 'High', 'Very High', and 'Luxury'. The introduction of this feature was driven by the notion that different pricing tiers might have varying impacts on consumer ratings. By categorizing products into these price ranges, the model can distinguish between different pricing tiers, potentially offering more nuanced insights into how pricing influences consumer ratings.

Encoded price range variables were created to integrate this categorical variable effectively into the logistic regression model. These dummy variables were formulated as binary variables corresponding to each category in the price range (for instance, price_Very Low, price_Low, etc.). This transformation allows the logistic regression model to systematically incorporate the effect of different price ranges, ensuring that the model can recognize and utilize the nuanced information these price tiers provide.

Another insightful feature that was engineered is the Discount Indicator named discounted. This binary feature indicates whether a product is discounted, with a value of 1 assigned if the discount_percentage is greater than 0, and 0 otherwise. The rationale behind this feature is to capture the presence of a discount explicitly, acknowledging that the availability of a discount could be a significant factor influencing consumer ratings. By converting this information into a binary variable, the model can readily assess the impact of discounts on consumer ratings.

The data transformation process was meticulously carried out, ensuring that the dataset was optimally structured and formatted for effective logistic regression modeling. A pivotal step in this process was the standardization of numerical variables. Key numerical variables, including discounted_price, actual_price, discount_percentage, and rating_count, were standardized using the StandardScaler. This standardization is crucial as it ensures that these variables contribute equally to the model training, preventing any one variable with a larger range or variance from dominating the model's behavior. By bringing these variables onto the same scale, the model can more effectively learn from these features, ensuring that the numerical values are balanced and comparable across the dataset.

Another significant transformation was the one-hot encoding of categorical variables, particularly the category variable. One-hot encoding is a powerful method to convert categorical data into a format suitable for logistic regression and other machine learning models. This process involves creating binary columns for each category within the category variable, where the presence of a particular category is indicated by a 1 and its absence by a 0. This transformation effectively captures the categorical information numerically, allowing the logistic regression model to incorporate it into its analysis.

Finally, the dataset was split into training and testing sets. The dataset was judiciously divided into training and testing sets, marking a significant phase in preparing the data for model training and evaluation. This division was carried out with a 70-30 split, ensuring a balanced distribution between the data used for training the model and the data reserved for testing its performance.

The training set, comprising 70% of the data, includes 1,023 samples of features, denoted as X_train, and an equal number of samples for the target variable, y_train. This substantial portion of the dataset is pivotal for the model's learning process, allowing the logistic regression model to discern patterns, understand relationships between variables, and effectively calibrate its parameters based on the features and the corresponding target values. The remaining 30% of the data from the testing set, which consists of 439 samples of features and an equal number of samples for the target variable, are represented as X_test and y_test, respectively. The testing set plays a crucial role in evaluating the model's performance. It provides an unbiased assessment of how well the logistic regression model generalizes to new, unseen data. This evaluation is fundamental in understanding the model's predictive accuracy, guiding decisions on model adjustments, and ensuring the model is robust and reliable.

3.4. Model Building

The logistic regression model was constructed through a series of critical steps, from variable selection to parameter configuration, all aimed at developing a robust model for analyzing the impact of discount strategies on consumer ratings.

The variable selection process was pivotal in determining the most effective predictors for the model. Insights from exploratory data analysis identified variables with significant correlation to consumer ratings, such as discount percentage and specific product categories within 'Home & Kitchen'. To avoid multicollinearity, variables were carefully evaluated, leading to the inclusion of 'discounted_price' over 'actual_price' due to its direct relevance to the study's focus. Additionally, the 'price_range' variable, derived from 'actual_price', was included to examine the influence of different pricing tiers on consumer ratings.

The logistic regression model's parameters were meticulously configured using a statistical software package. The 'liblinear' solver was chosen for its effectiveness in binary classification and suitability for smaller datasets. Regularization strength, controlling the degree of overfitting, was set to a default value, allowing for a balanced approach that could be fine-tuned during the evaluation phase.

Data preprocessing outputs were integrated into the model, including the results of one-hot encoding for categorical variables and standardized values for numerical variables. This integration was essential for the model to interpret and learn from the data effectively.

Finally, the model was trained on the designated training dataset. The training involved introducing the selected features and corresponding consumer ratings, enabling the model to learn the relationships between variables and the likelihood of different rating outcomes. The training phase was monitored for signs of overfitting or underfitting to ensure the model's reliability and accuracy.

3.5. Model Evaluation

Evaluating the logistic regression model was a critical phase, focusing on assessing the model's performance using various metrics and methods. This evaluation was essential to determine the model's ability to predict consumer ratings based on discount strategies and other relevant factors.

The model's accuracy was initially evaluated as a fundamental measure in classification tasks, indicating the overall proportion of correct predictions. In addition to accuracy, precision and recall were calculated. Precision-measured the proportion of correct positive identifications out of all positive identifications made by the model, while recall assessed the proportion of actual positives correctly identified. The F1-score, the harmonic mean of precision and recall, was also computed.

The ROC (Receiver Operating Characteristic) curve and AUC (Area Under the Curve) analysis were conducted to better understand the model's performance, especially in differentiating between the binary high and low rating classifications. The ROC-AUC score provided a single aggregate performance measure across all possible classification thresholds.

Cross-validation was performed to ensure the model's robustness and generalizability. K-fold cross-validation was used, which involves dividing the dataset into k subsets and training the model iteratively. This method helped assess the model's performance across different subsets of the dataset.

The model's classification capabilities were further explored through a confusion matrix analysis. This matrix provided a detailed breakdown of the model's predictions, showing the number of true positives, false positives, true negatives, and false negatives, offering insight into the types of errors the model was prone to and identifying areas for improvement.

The model evaluation phase employed a comprehensive set of methods, including accuracy, precision, recall, F1-score, ROC-AUC, cross-validation, and confusion matrix analysis. These metrics and methods provided a thorough understanding of the model's performance, strengths, and limitations, enabling an assessment of its ability to predict consumer ratings based on discount strategies.

4. Result and Discussion

The logistic regression analysis on the dataset revealed several important insights regarding the impact of discount strategies on consumer ratings. This section presents these findings, interprets their significance, and discusses their theoretical and practical implications.

The model achieved an accuracy of 74.94%, demonstrating effectiveness in predicting consumer ratings. Precision was noted at 72.69% and recall at 74.94%, suggesting areas for improvement in precision. The F1 score stood at 69.26%. The ROC-AUC score was also 78.24%, indicating proficiency in differentiating between high and low ratings. The analysis identified 'Home & Kitchen' categories, particularly 'Kitchen & Home Appliances' and 'Small Kitchen Appliances,' as significant predictors, along with the 'discount_percentage' feature.

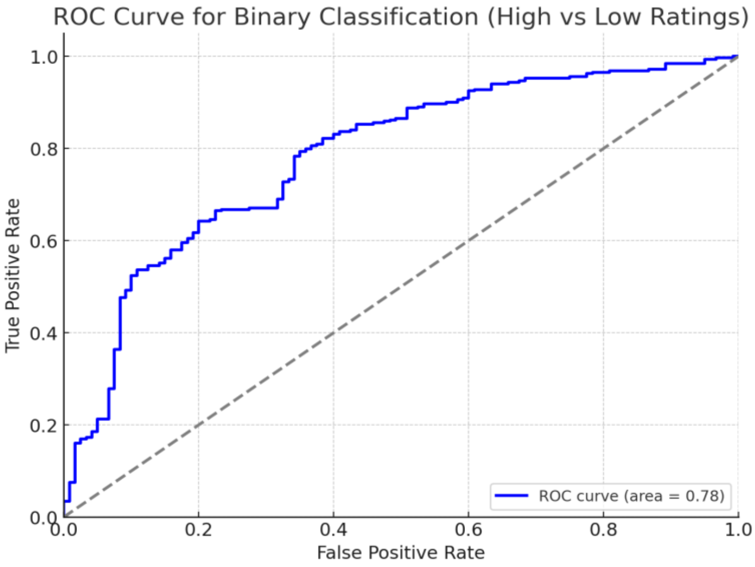

Figure 4. ROC Curve

The Receiver Operating Characteristic (ROC) curve, as shown in Figure 4 above, has been a pivotal analytical tool in distinguishing between high and low ratings in the binary classification scenario. The ROC curve, depicted in blue, illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various threshold settings. This curve represents the model's capability to distinguish between the two rating categories under different classification thresholds.

A particularly noteworthy aspect of the ROC analysis is the Area Under the Curve (AUC), which, for this curve, is approximately 0.78. The AUC is a crucial metric as it quantifies the overall ability of the model to correctly classify the high and low ratings. An AUC of 0.78 indicates a good level of distinction achieved by the logistic regression model. It suggests that the model can differentiate between high and low ratings, effectively discriminating between the two categories across various thresholds.

Interpreting the ROC curve and its AUC provides profound insights into the model's performance. The fact that the curve is significantly above the diagonal gray line, which symbolizes random chance, is a testament to the model's effectiveness. This curve positioning indicates that the model's ability to classify high and low ratings is not merely by chance but is a result of learning from the underlying patterns in the data.

The model's robustness and reliability were further affirmed through 5-fold cross-validation, an essential technique in assessing the model's performance across different subsets of the data. This method provides a more comprehensive understanding of the model's generalizability and stability. The cross-validation results consistently demonstrate the model's steady performance across various folds.

In the first fold, the model achieved an accuracy of 79.02%, setting a promising start to the cross-validation process. The accuracy slightly dipped to 77.56% in the second fold, a common occurrence in cross-validation due to the variability in the data subsets. However, the model's performance peaked in the third fold, reaching an accuracy of 80.00%, showcasing its capability to adapt and learn effectively from different data segments. The fourth and fifth folds exhibited accuracies of 79.90%, indicating a return to the higher performance levels observed in the initial fold. When these individual fold accuracies were aggregated, the mean cross-validation score was 79.28%. This score reflects the model's overall consistent and reliable performance across the different folds. Moreover, the standard deviation of the cross-validation scores was computed to be a mere 0.93%, underscoring the model's stability and consistency in performance across the various data subsets.

The findings from the logistic regression analysis provide critical insights into the nuanced effects of discount strategies on consumer behavior in e-commerce. The accuracy of 74.94% achieved by the model underlines the predictability of consumer response to discounts, indicating that consumers' rating behavior can be systematically influenced by specific discount parameters.

The substantial role of the discount percentage as a predictor suggests that the magnitude of discounts is a significant determinant of consumer satisfaction. This challenges the conventional understanding that merely offering discounts is enough to sway consumer ratings positively. Instead, it highlights the importance of strategically calibrated discounts to elicit favorable consumer responses.

Furthermore, the significant influence of product categories, particularly within 'Home & Kitchen', suggests that the effectiveness of discounts is not uniform across all product types. This finding is crucial for e-commerce platforms and marketers, underscoring the need for a more segmented and targeted approach to discounting strategies. It suggests that a one-size-fits-all discount strategy might not be effective and that understanding consumer segments' specific preferences and price sensitivities is vital.

These results provide a more detailed understanding of the dynamic interplay between discount strategies and consumer ratings. This knowledge is invaluable for e-commerce businesses looking to optimize their discount strategies to enhance customer engagement and satisfaction, ultimately contributing to better business performance in the competitive online marketplace.

5. Conclusion

The exploration into e-commerce and the impact of discount strategies on consumer ratings has led to significant findings. The study's main discoveries include the pronounced influence of discount percentages on consumer ratings and the category-specific nature of these strategies, particularly in 'Home & Kitchen' products. This demonstrates the context-dependent effectiveness of discount strategies. This research observed that discount strategies directly and notably impact consumer satisfaction and perceptions within the e-commerce sector. The logistic regression model, with an accuracy rate of 74.94%, effectively differentiated between high and low ratings. Key predictors identified were the discount percentage and specific product categories, highlighting a nuanced relationship between discounts, product types, and consumer responses.

The findings significantly advance the understanding of consumer behavior in e-commerce. By quantifying the impact of discounts and revealing their variable influence across product categories, the study offers practical insights for businesses. It enriches existing literature on consumer behavior and pricing strategies, providing a framework for optimizing marketing and pricing approaches to improve customer satisfaction.

The study paves the way for future research avenues. Investigating factors like product quality, brand reputation, and shipping times can provide a more comprehensive understanding of consumer satisfaction drivers. Longitudinal studies and cross-platform analyses may uncover evolving consumer trends and platform-specific behaviors. Further exploration into consumer segmentation and sentiment analysis of reviews could deepen the understanding of consumer attitudes and preferences, leading to more tailored and effective marketing strategies. In summary, the study enhances the understanding of consumer behavior in the digital era and opens new research pathways in this dynamic field. It underscores the critical role of discount strategies in influencing consumer ratings in e-commerce.

6. Declarations

6.1. Author Contributions

Conceptualization: B., A.M.W., and D.F.; Methodology: A.M.W.; Software: B.; Validation: B., A.M.W., and D.F.; Formal Analysis: B., A.M.W., and D.F.; Investigation: A.N.A.S.; Resources: G.B.; Data Curation: A.N.A.S.; Writing Original Draft Preparation: A.N.A.S. and G.B.; Writing Review and Editing: A.N.A.S. and G.B.; Visualization: G.B.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Kurnia, J. Choudrie, R. M. Mahbubur, and B. Alzougool, “E-commerce technology adoption: A Malaysian grocery SME retail sector study,” *J. Bus. Res*., vol. 68, no. 9, pp. 1906–1918, Sep. 2015, doi: 10.1016/j.jbusres.2014.12.010.

[2] R. Liu and E. Wang, “Blockchain and mobile client privacy protection in e-commerce consumer shopping tendency identification application,” *Soft Comput. - Fusion Found. Methodol. Appl*., vol. 27, no. 9, pp. 6019–6031, Apr. 2023, doi:

10.1007/s00500-023-08099-8.

[3] T. Kinda, "E-Commerce as a Potential New Engine for Growth in Asia," *Imf Work*. Pap., 2019, doi: 10.5089/9781498317467.001.

[4] S. Ma, X. Guo, and H. Zhang, "New driving force for China's import growth: Assessing the role of cross-border e-commerce," *World Econ.*, vol. 44, no. 12, pp. 3674–3706, 2021, doi: 10.1111/twec.13168.

[5] R. S. Yu and I. E. A, "E-COMMERCE IN CHINA AMID COVID-19 PANDEMIC RESTRICTIONS," *Вестник Российского Университета Дружбы Народов Серия Экономика*, vol. 29, no. 4, Art. no. 4, 2021.

[6] M. K. M, A. P. S, and S. K. R. S, "Open Network for Digital Commerce -ONDC (E-Commerce) Infrastructure: To Promote SME/ MSME Sector for Inclusive and Sustainable Digital Economic growth," *Int. J. Manag. Technol. Soc. Sci. IJMTS*, vol. 7, no. 2, Art. no. 2, Oct. 2022, doi: 10.47992/IJMTS.2581.6012.0223.

[7] Mahmuddin and N. N. Sirait, "E-Commerce Growth and Development, Impact, and Challenges in Indonesia*," Neoclassical Leg. Rev. J. Law Contemp. Issues*, vol. 1, no. 1, Art. no. 1, Oct. 2022, doi: 10.32734/nlr.v1i1.9603.

[8] R. Arisanti, E. R. Utami, A. Muslim, and M. Hayati, "The relationship between economic growth and e-commerce at the beginning of covid-19 pandemic in east Java," *Decis. Sci. Lett.*, vol. 12, no. 1, pp. 149–162, 2023, doi: 10.5267/j.dsl.2022.9.002.

[9] M. Kiselicki, Z. Kirovska, S. Josimovski, and M. Anastasovski, "E-Commerce as a Revenue Generator for Small and Medium Companies in Developing Countries," *Econ. Cult.*, 2022, doi: 10.2478/jec-2022-0015.

[10] A. A. A. P. Andrina, C. J. Kurniadi, I. H. Kenang, and T. F. Sutrisno, "The role of technology acceptance model factors on purchase intention in e-commerce," *BISMA Bisnis Dan Manaj.*, vol. 14, no. 2, Art. no. 2, Apr. 2022, doi: 10.26740/bisma.v14n2.p160-176.

[11] S. Akter and S. F. Wamba, "Big data analytics in E-commerce: a systematic review and agenda for future research," *Electron. Mark.*, vol. 26, no. 2, pp. 173–194, May 2016, doi: 10.1007/s12525-016-0219-0.

[12] S. S. Alrumiah and M. Hadwan, "Implementing Big Data Analytics in E-Commerce: Vendor and Customer View," *IEEE Access*, vol. 9, pp. 37281–37286, 2021, doi: 10.1109/ACCESS.2021.3063615.

[13] O. Fedirko, T. Zatonatska, T. Wolowiec, and S. Skowron, "Data Science and Marketing in E-Commerce Amid COVID-19 Pandemic," *Eur. Res. Stud.*, vol. XXIV, no. Special 2, pp. 3–16, Jun. 2021.

[14] J. Xie, "Discussion on the Mechanism of Irrational Online Shopping Behavior—Based on the Perspective of Mental Accounting Theory," *Open J. Soc. Sci.*, vol. 7, no. 5, Art. no. 5, May 2019, doi: 10.4236/jss.2019.75004.

[15] Gunawan, "Social Commerce from Seller and Region Perspective: A Data Mining for Indonesian E-commerce," *in 2022 International Conference on Data Science and Its Applications (ICoDSA),* Jul. 2022, pp. 268–272. doi: 10.1109/ICoDSA55874.2022.9862835.

[16] Z. Li, "E-commerce Platform Data Governance Environment: Concepts, Elements and Implications," *Front. Bus. Econ. Manag.*, vol. 7, no. 2, pp. 99–104, Feb. 2023, doi: 10.54097/fbem.v7i2.4850.

[17] M. Cheng and X. Jin, "What do Airbnb users care about? An analysis of online review comments*," Int. J. Hosp. Manag.*, vol. 76, pp. 58–70, Jan. 2019, doi: 10.1016/j.ijhm.2018.04.004.

[18] A. Firmanto and R. Sarno, "Aspect-based sentiment analysis using grammatical rules, word similarity and SentiCircle," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 5, pp. 190–201, 2019, doi: 10.22266/ijies2019.1031.19.

[19] C. Wang and M. Hwan Yun, "Cross-Cultural Difference in Product Preference in Consumer Review-Based Text Mining Methods: a Case Study on Smart Band," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 64, no. 1, pp. 1383–1387, Dec. 2020, doi: 10.1177/1071181320641330.

[20] D. Qibtiyah, R. Hurruyati, and H. Hendrayati, "The Influence of Discount on Repurchase Intention," *presented at the 5th Global Conference on Business, Management and Entrepreneurship (GCBME 2020)*, Atlantis Press, Sep. 2021, pp. 385–389. doi: 10.2991/aebmr.k.210831.076.

[21] Y. Gong, W. Hou, Q. Zhang, and S. Tian, "Discounts or gifts? Not just to save money: A study on neural mechanism from the perspective of fuzzy decision," *J. Contemp. Mark. Sci.*, vol. 1, no. 1, pp. 53–75, Jan. 2018, doi: 10.1108/JCMARS-08-2018-0009.

[22] J. Lv, Z. Wang, Y. Huang, T. Wang, and Y. Wang, "How Can E-Commerce Businesses Implement Discount Strategies through Social Media?," *Sustainability*, vol. 12, no. 18, Art. no. 18, Jan. 2020, doi: 10.3390/su12187459.

[23] I. Hasbi, S. Syahputra, S. Syarifuddin, T. I. Wijaksana, and P. Farías, "The impact of discount appeal of food ordering application on consumer satisfaction in Southeast Asia*," J. East. Eur. Cent. Asian Res. JEECAR*, vol. 9, no. 6, Art. no. 6, Dec. 2022, doi: 10.15549/jeecar.v9i6.956.

[24] A. Fagerstrøm, G. Ghinea, and L. Sydnes, "Understanding the Impact of Online Reviews on Customer Choice: A Probability Discounting Approach," *Psychol. Mark.*, vol. 33, no. 2, pp. 125–134, 2016, doi: 10.1002/mar.20859.

[25] S. N. Ahmad and M. Callow, "'Free Shipping' or 'Dollar Off'? The Moderating Effects of List Price and E-Shopping Experience On Consumer Preference For Online Discount," *Int. J. Electron. Commer. Stud.*, vol. 9, no. 1, Art. no. 1, Aug. 2018, doi: 10.7903/ijecs.1542.

[26] S. Vadera, S. K. Suman, and P. Srivastava, "Exploring the behaviour of Indian consumers towards online discounts," *Int. J. Electron. Mark. Retail.*, vol. 10, no. 1, p. 78, 2019, doi: 10.1504/ijemr.2019.10017364.

[27] P. Chatterjee and J. McGinnis, "Customized Online Promotions: Moderating Effect Of Promotion Type On Deal Value, Perceived Fairness, And Purchase Intent," *J. Appl. Bus. Res. JABR*, vol. 26, no. 4, Art. no. 4, Jul. 2010, doi: 10.19030/jabr.v26i4.302.

[28] C. Lan and J. Zhu, "New Product Presale Strategies considering Consumers' Loss Aversion in the E-Commerce Supply Chain," *Discrete Dyn. Nat. Soc.*, vol. 2021, p. e8194879, Jul. 2021, doi: 10.1155/2021/8194879.

[29] L. Efendi and M. Geta, "ANALYSIS OF CONSUMER PSYCHOLOGICAL TOWARDS PRICE DISCOUNTS: A CASE STUDY AT PAKUWON MALL YOGYAKARTA," *J. Int. Conf. Proc.*, vol. 5, no. 5, Art. no. 5, Dec. 2022, doi: 10.32535/jicp.v5i5.2018.

[30] X. Luo and J. Lee, "The Effect of Post-Purchase Discount Format on Consumers' Perception of Loss and Willingness to Return," *J. Asian Finance Econ. Bus.*, vol. 5, pp. 101–105, Oct. 2018, doi: 10.13106/jafeb.2018.vol5.no4.101.

[31] W. Sun, P. Murali, A. Sheopuri, and Y.-M. Chee, "Designing promotions: Consumers' surprise and perception of discounts," *IBM J. Res. Dev.*, vol. 58, no. 5/6, p. 2:1-2:10, Sep. 2014, doi: 10.1147/JRD.2014.2337691.

[32] A. Rohani and M. Nazari, "Impact of Dynamic Pricing Strategies on Consumer Behavior," *J. Manag. Res.*, vol. 4, no. 4, Art. no. 4, Aug. 2012, doi: 10.5296/jmr.v4i4.2009.

[33] D. Edmondson, T. Graeff, L. Matthews, D. Roy, R. Srivastava, and C. Ward, "Consumers honoring veterans and businesses that support them," *J. Consum. Mark.*, vol. 37, no. 1, pp. 77–86, Jan. 2019, doi: 10.1108/JCM-12-2018-2989.

[34] C. Liu, C. K. M. Lee, and K. H. Leung, "Pricing Strategy in Dual-Channel Supply Chains with Loss-Averse Consumers," *Asia-Pac. J. Oper. Res.*, vol. 36, no. 05, p. 1950027, Oct. 2019, doi: 10.1142/S0217595919500271.