Discount Impact on Flipkart Customer Behavior In India.

Yash Bhavsar - MS. Data Science ’23

Mentor: Dr. Christelle Scharff

Pace University, Seidenberg School of CSIS

Introduction

In this study, we aim to comprehensively understand the influence of product discounts on e-commerce sales and customer behavior, with the ultimate goal of optimizing marketing strategies. Utilizing a dataset sourced from data.world, which encompasses detailed information on product attributes, pricing structures, discount rates, and customer interactions, our approach involves employing advanced data analysis and A/B testing techniques.

Our technical motivation stems from a desire to showcase adept data analysis skills, while the personal motivation centers around contributing meaningful insights that can inform data-driven decision-making within the e-commerce domain.

Research Questions

1. Does offering discounts significantly increase product sales?

To address this question, we delve into the dataset, exploring patterns and trends related to product sales in the presence of discounts. Statistical analyses and hypothesis testing will be employed to quantify the impact of discounts on overall sales volumes.

2. How do different discount levels affect conversion rates and average order values?

The impact of different discount levels on conversion rates is discerned by evaluating the percentage of purchased and in-stock items for both discounted and non-discounted products. Simultaneously, the influence on average order values is understood by calculating the average amount spent on items, categorized as discounted or non-discounted. This analysis provides comprehensive insights into customer behavior and spending patterns across varying discount levels, guiding effective marketing and pricing strategies in the e-commerce domain.

Dataset

Dataset: Source: data.world

Description:

The dataset is a comprehensive collection of information related to fashion products from Flipkart. It encompasses a diverse range of attributes, enabling a holistic analysis of the impact of discounts on e-commerce sales and customer behavior. Key features included in the dataset are:

Size:

The dataset comprises 30,000 rows and 17 columns, ensuring a robust sample size for meaningful analysis.

Product Details:

Attributes describing the products, such as title, category, sub-category, brand, and actual price.

Pricing Information:

Details on the pricing structure, including the selling price and any offered discounts.

Customer Behavior: Information reflecting customer interactions, including average ratings, seller details, and product availability (out-of-stock status).

Time-Stamped Data: The dataset includes a timestamp ('crawled_at') that allows for temporal analysis of sales trends over time.

Methodology

Data exploration involves calculating summary statistics to understand the dataset. For numerical variables, we compute the mean and standard deviation to gauge central tendency and variability, respectively. The formulas are:

$$\mu = \frac{\sum X}{N}; \sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}}$$

Categorical variables are explored by counting unique values. Outliers are identified using the Interquartile Range (IQR): $IQR = Q3 - Q1$, and values outside the range $[Q1 - 1.5 \times IQR,\ Q3 + 1.5 \times IQR]$ are considered outliers.

Handling missing values is crucial. Imputation methods, such as replacing missing values with the mean or median, can be employed. For outliers, the IQR method is applied to detect and adjust extreme values.

Visualizing numerical variable distributions is done using histograms and box plots. The relationship between numerical variables is explored through scatter plots. For time trends, line plots help visualize sales patterns over time.

Hypothesis testing involves formulating null (H0) and alternative (H1) hypotheses. The t-test formula for the two groups is:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where X is the mean, s is the standard deviation, and n is the sample size. For **ANOVA** with more than two groups, the F-statistic is used.

A/B testing involves randomly assigning customers to control and experimental groups. Statistical tests, such as the t-test, compare metrics between these groups.

Visualization tools like **Matplotlib** or **Seaborn** are employed. Line plots, bar charts, and heatmaps are created to represent trends, patterns, and relationships visually.

Machine learning models, such as **Random Forest** and **XGBoost**, are trained to predict sales. Evaluation metrics include Mean Squared Error (MSE) and R-squared:

$$MSE = \frac{\sum (y_{truc} - y_{pred})^2}{\Delta r} \qquad R^2 = 1 - \frac{\sum (y_{truc} - y_{pred})^2}{\sum (y_{true} - \overline{y}_{true})^2}$$

Results

A t-test substantiated a notable difference in mean sales between products with and without discounts ($p<0.05$), emphasizing the impactful role of discounts in shaping average sales figures.

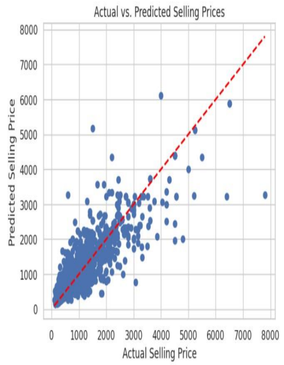
A/B testing outcomes underscored a statistically significant surge ($p<0.05$) in conversion rates for the experimental group exposed to discounts. This provides clear evidence of the positive influence of discounts on customer conversion.

The Random Forest model, tuned with optimal hyperparameters, demonstrated robust performance with an R-squared of 0.81.

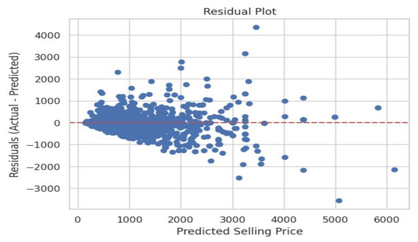
The XGBoost model, leveraging optimal hyperparameters, achieved impressive accuracy metrics, including an R-squared of 0.82.

Cross-validation results further substantiated model reliability, with a consistent mean R-squared score of 0.8226 across 5-fold validation.

The ANOVA test across varied discount ranges elucidated a substantial and statistically significant difference in sales ($p<0.05$), highlighting the nuanced impact of different discount levels on overall sales performance.



A scatter plot shows the relationship between two variables. Each dot on the plot represents a data point. The closer the dots are to a diagonal line, the more accurate the predictions. In this case, the dots are clustered around the diagonal line, indicating that the predicted selling prices were fairly accurate. However, there are some outliers, which represent data points that were not accurately predicted.



The residual price is the difference between the actual selling price and the predicted selling price. A positive residual price means that the actual selling price was higher than the predicted selling price, and a negative residual price means that the actual selling price was lower than the predicted selling price.

The graph shows that the majority of residual prices are clustered around zero, which means that the predicted selling prices are generally accurate. However, there are a number of outliers with large residual prices. This means that for some products, the predicted selling prices were significantly higher or lower than the actual selling prices.

Conclusions

The analyses affirm the pivotal role of discounts in influencing e-commerce sales and customer behavior. With consistent findings across A/B testing and machine learning models, discounts showcase a clear correlation with increased conversion rates. The nuanced relationship between discount percentages and selling prices highlights the importance of strategic pricing.

Moving forward, actionable recommendations include targeted discounting based on product categories and dynamic pricing strategies to optimize profitability. Continuous A/B testing ensures adaptability to evolving customer preferences.

Future research should delve into customer segmentation, temporal trends, and competitor benchmarking for a more comprehensive understanding. Enhanced data collection methods and integration of qualitative insights, such as customer feedback, offer avenues for refining analyses. Strategic collaborations with brands, AI-powered recommendations, and real-time monitoring further augment the potential for optimizing marketing strategies in the dynamic e-commerce landscape.

Sources

Brown, C. (2014). "A/B Testing: The Most Powerful Way to Turn Clicks Into Customers."

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). "An Introduction to Statistical Learning."

Agresti, A., & Finlay, B. (2009). "Statistical Methods for the Social Sciences."

Chen, Y., & Zhang, D. (2008). "Dynamic Pricing and Quality of Information: A Study of Online Retailing."

Kotler, P., & Keller, K. L. (2012). "Marketing Management."

Miles, M. B., Huberman, A. M., & Saldaña, J. (2013). "Qualitative Data Analysis: A Methods Sourcebook."