CMPE 492
Understanding the Effects of Promotion Mechanisms on Price and Revenue of Sellers in an Online Marketplace

Enver Eren
Mustafa Can Coşkun
Fethi Kahvecioğlu
İbrahim Emre Köksal

Advisor:
Taner Bilgiç
Can Özturan

June 2024

TABLE OF CONTENTS

1. INTRODUCTION . . . 1
   1.1. Broad Impact . . . 1
   1.2. Ethical Considerations . . . 1
2. PROJECT DEFINITION AND PLANNING . . . 2
   2.1. Project Definition . . . 2
   2.2. Project Planning . . . 3
      2.2.1. Project Time and Resource Estimation . . . 3
      2.2.2. Success Criteria . . . 4
      2.2.3. Risk Analysis . . . 4
      2.2.4. Team Work . . . 4
3. RELATED WORK . . . 5
4. DESIGN . . . 8
   4.1. Information Structure . . . 8
   4.2. Information Flow . . . 9
      4.2.1. Descriptive Analytics . . . 10
      4.2.2. Prescriptive Analytics . . . 11
   4.3. Data Summary . . . 12
      4.3.1. Orders . . . 12
      4.3.2. Users . . . 13
      4.3.3. SKUs . . . 14
      4.3.4. Clicks & Delivery . . . 15
      4.3.5. Inventory & Network . . . 15
5. METHODOLOGY . . . 17
6. IMPLEMENTATION AND TESTING . . . 20
   6.1. Implementation . . . 20
   6.2. Testing . . . 22
7. RESULTS . . . 24
8. CONCLUSION . . . 29
REFERENCES . . . 30

1. INTRODUCTION

1.1. Broad Impact

E-commerce's growth has brought new challenges at both strategic and operational levels. In 2023, e-commerce's share of total retail sales worldwide is expected to increase significantly, making strategic promotional campaigns crucial for attracting customers and boosting total revenue and sales. This project explores how different promotional strategies affect seller revenue in an online marketplace, using data from JD.com to develop a data-driven model for optimizing promotion mechanisms.

1.2. Ethical Considerations

This study prioritizes ethical standards in data handling and analysis, ensuring the anonymity and confidentiality of user data through the use of anonymized transactional data from JD.com. Our approach is guided by principles of fairness, non-discrimination, and transparency in developing promotional recommendations for diverse user and SKU groups.

2. PROJECT DEFINITION AND PLANNING

2.1. Project Definition

In this project, effects on promotion strategies such as direct and quantity discounts on pricing and revenue stream of sellers in JD.com, the largest retailer in China, will be examined. The examination will be motivated through econometric and data-driven descriptive and prescriptive analytics. To initiate the analytics of the JD.com promotional strategies, the transactional level data of over 2.5 million customers and 31,868 SKU will be analyzed. The data is offered by the platform of JD.com itself, originally to initiate the MSOM 2020 Data Driven Research Challenge [1].

To provide better customer experience and enhanced revenue stream, e-commerce retailers should be able to build a structured econometric model to further investigate the behavioral actions of the customers by considering certain parameters such as their user classification and price-demand relationship of the products on the marketplace. The necessary strategic and analytic actions that have been taken in the company will have noticeable optimization to the efficiency of their logistic and financial operations [2].

The data that is offered by JD.com contains 7 tables that include their user base, clicks on certain products based on user and SKU id, order transactions, and regional information which all provide critical product and customer-centric information. Also, the transactional order data provides a full customer experience cycle that is initiated by the customer choosing the product on the web and terminated by the delivery of the product at a destination. The time horizon for the order-related transactional data is offered for the March 2018 period which indicates an insufficient time length for time series related analysis such as demand planning and future forecast purposes. However, the scope of this project is building a prescriptive model for the promotional strategies and understanding their impact on certain customer behavior. Hence, the dimensional features will enable to build sufficient in-depth analysis for future endeavors.

E-commerce retailing analytics have been an important aspect of the business organizations in terms of their operations efficiency and financial profitability. And various econometric and behavioral models have been built and implemented in industrial and academic environments. The main aim of this project will be understanding the promotional and discount related strategies' effect on products' demand by considering sell and click actions in an online marketplace with the help of the vector of information that comes within the online marketplace data capturing such as in-depth user-related information and product-based attributes.

2.2. Project Planning

2.2.1. Project Time and Resource Estimation

The project timeline was meticulously planned over several months, incorporating phases of data preprocessing, analysis, model development, and validation. Resources allocated included access to JD.com's transactional data, computational tools for data analysis, and team members' expertise in data science and e-commerce strategies.

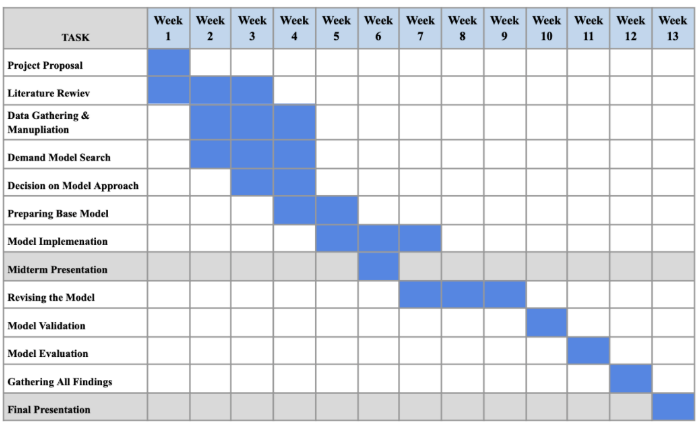

Figure 2.1. Gantt Chart of Project Timeline

2.2.2. Success Criteria

Success is defined by our ability to accurately model the relationship between promotional strategies and revenue outcomes, and to provide actionable recommendations that demonstrably increase seller revenue within the online marketplace.

2.2.3. Risk Analysis

Potential risks include data inaccuracies, assumptions in model development that do not hold in real-world scenarios, and the unpredictability of market responses to recommended promotional strategies.

2.2.4. Team Work

The project was undertaken by a dedicated team of senior students, each contributing specialized skills in data analysis, modeling, and e-commerce strategy. Team collaboration was facilitated through regular meetings, shared documentation, and iterative feedback sessions.

3. RELATED WORK

Along with the advancement of technology and the development of e-commerce, the focus of contemporary sellers have considerably changed. In the dynamic landscape of modern commerce, businesses are continually exploring strategies to enhance their competitive edge and bolster revenue streams. In different countries, research on different datasets was conducted. Among the myriad tactics employed, promotional activities stand out as influential tools that can significantly impact customer behavior, pricing dynamics, and overall business performance. The interplay between promotions, customer engagement, and seller pricing has become a focal point of interest for researchers and practitioners alike.

This literature review aims to provide a comprehensive understanding of existing studies and methodologies related to the analysis of marketplace data. The preliminary focuses on that review are the data analysis methods that have already been used in and how the customer behaviors can be related to promotion strategies.

The most common way implemented in literature is the qualitative approach. This approach is conducted by contacting customers via different channels. Syafeqah Nurul Marzuki [2], searches the factors affecting online shoppers in Malaysia by qualitative approach. Study concluded that even though the promotions are not the most influential factors on customers, it is stated that it is the most effective marketing strategy. Also, Fernando de Oliveira Santini [3] offers a way to handle promotions type in analysis. Monetary and non-monetary types of promotions have different amounts of impact on purchase intentions.

As discussed earlier, promotional activities are one of the most important tools that shapes purchase intentions of customers. However, that fact has not been widely studied empirically. Since the available dataset includes only the transactional data. To overcome this issue, Dennis J. Zhang1 [4] divides eligible customers into two distinct groups, one of which is the treated customers (who received coupons) and the other is the control group (who do not receive coupons). Experiment is conducted to the world's largest railing platform, Alibaba Group, which involves more than 100 million customers and 11 thousand retailers. By implementing this method, five important questions about the effect of promotions and dynamic pricing changes on customers' behaviors.

To develop an optimal price prediction model, Jiawei Wen [5] conducted a study with the data of Airbnb. The aim of the model is to increase marketplace and seller revenue it is essential to optimize and properly price by examining the searching ranks. It is said that it is difficult to use traditional revenue maximization strategies in marketplaces like Airbnb. Thus, he offers a new model that involves two phases. The first phase is concluding a price vector using loss function by using different types of features of booking. The second phase is derived from the price set in the first phase. Calculated price set, along with the probability set of booking, are used to construct an objective function to maximize expected revenue of sellers.

Furthermore, Ghose and Sundararajan's [6] study significantly advances our comprehension of pricing strategies in e-commerce by focusing on demand estimation. They delve into Amazon.com's distinctive approach of utilizing sales ranks to assess demand, demonstrating the conversion of these ranks into demand estimates using Pareto distribution. Through a meticulous combination of purchase experiments and longitudinal sales data, they identify specific parameters critical for precise demand estimation. Additionally, they explore various economic metrics such as price markups and demand elasticities, constructing a comprehensive framework based on sales rank data. Their rigorous application of statistical methods, including OLS regressions, ensures the reliability and accuracy of their findings.

Additionally, research conducted by Huang and Van Mieghem [7] underscores the significance of click tracking, highlighting its potential in unraveling consumer behavior. Their study suggests that analyzing click data holds considerable promise, especially within online marketplaces, as it enables the extraction of valuable insights into demand patterns and price sensitivity. Through the examination of clickstream data, researchers can potentially uncover subtle indicators of consumer interest and purchasing intentions, thereby enhancing strategies for pricing, promotion, and forecasting future demand. Nevertheless, further investigation is needed to fully comprehend the intricate relationship between click data and market dynamics, including its implications for pricing strategies and revenue generation in online marketplaces.

The other study on the click rates is about conversion rates of the click rates, which identifies the metric of the percentage of the visitors that complete their action in the marketplace. Darius Zumstein and Wolfgang Kotowski [8] extend the previous studies of success factors and growth on both retailer and e-commerce. Study investigated the success factors of e-commerce as services (services and sales-promoting instruments), drivers of the conversion rates and drivers of basket value and revenues. In the end, the study states that conversion rates is one of the most important KPI in e-commerce and it has seven different drivers. Also, it is stated that there is a strong positive correlation between conversion rates and online revenues.

4. DESIGN

4.1. Information Structure

Our project employs a robust data model to investigate the impact of promotional strategies on the revenue generation of sellers on JD.com [1]. The data model is visualized through an entity-relationship diagram (ERD), which delineates the structure of the database and the interconnections between its various components.

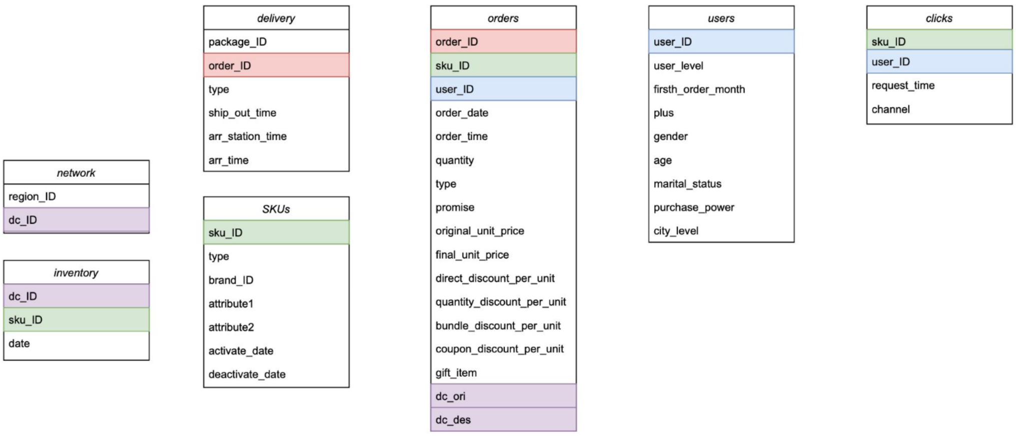

Figure 4.1. Entity-Relationship Diagram (ERD) of the JD.com database.

Users: This entity captures the demographic and behavioral attributes of customers on the platform. Key attributes include the user_ID, which uniquely identifies each user, along with user_level, first_order_month, plus status, gender, age, marital_status, purchase_power, and city_level. This entity is crucial for understanding customer segments and targeting promotions effectively.

SKUs (Stock Keeping Units): Representing the products listed on the marketplace, the SKU entity contains attributes such as sku_ID, type, brand_ID, and other distinguishing features like attribute1 and attribute2. The dates when SKUs are activated or deactivated in the system are also tracked.

Orders: This entity records transactional data and is a central component of our analysis. Each order is identified by an order_ID and contains details about the sku_ID, user_ID, order_date, order_time, the quantity of products ordered, the pricing details before and after promotions, and the delivery origin (dc_ori) and destination (dc_des).

Delivery: Linked to orders, the delivery entity contains information about the shipping process, including the package_ID, associated order_ID, the type of delivery, and timestamps for various stages in the delivery process.

Inventory: This entity tracks the stock levels of SKUs at different distribution centers dc_ID over time.

Network: It provides data on the distribution centers themselves, with each center identified by a dc_ID and associated with a specific region region_ID.

Clicks: This entity captures the online interaction of users with products, recording each click with a timestamp (request_time), the sku_ID, user_ID, and the channel through which the interaction occurred.

The ERD showcases the relationships between these entities, which are essential for the multi-dimensional analysis required in our project. For example, the orders entity is linked to both users and SKUs, reflecting the purchase behavior, which is the focal point of our promotional impact study. Similarly, the clicks entity is tied to users and SKUs, offering insights into customer engagement and product interest.

By leveraging this data model, we can cluster users and products, perform regression analyses to ascertain the effect of promotions on revenue, and simulate various promotional scenarios to prescribe the most effective discount strategies. This structured approach enables a comprehensive understanding of the dynamics between promotions, customer behavior, and revenue outcomes.

4.2. Information Flow

Our project integrates a structured analytics workflow to derive actionable insights and make data-driven recommendations for promotional strategies on an e-commerce platform. This workflow is segmented into two core phases: Descriptive Analytics and Prescriptive Analytics.

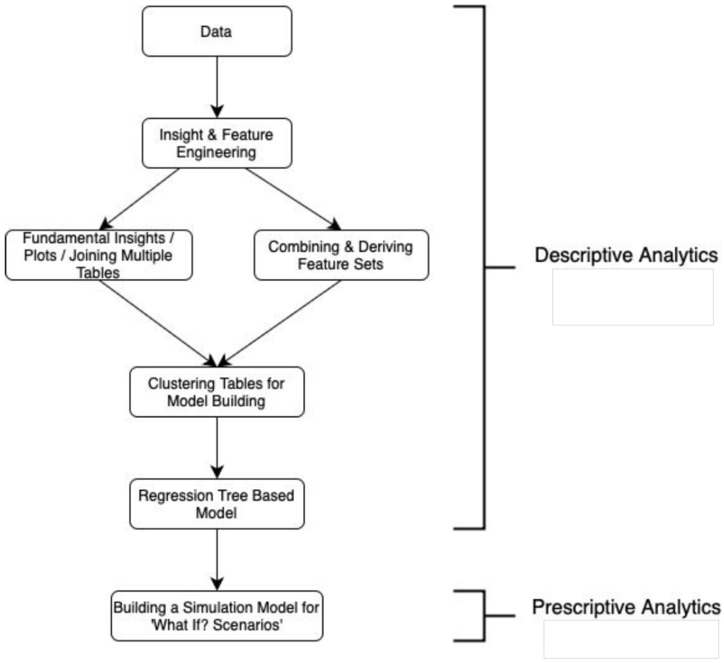

Figure 4.2. Structured Analytics Workflow for Promotional Strategy Optimization.

4.2.1. Descriptive Analytics

The initial phase of our project involves a detailed Descriptive Analytics process where raw data is transformed into a comprehensive understanding of past behaviors and outcomes:

Insights and Feature Engineering: At the outset, the raw data undergoes a feature engineering process to extract relevant attributes and construct informative features that can effectively capture the dynamics of the marketplace.

Fundamental Insights/Plots: Leveraging these features, we generate fundamental insights through various plots and by joining multiple tables. This step provides a visual and quantitative depiction of historical trends and patterns in the data.

Combining and Deriving Feature Sets: By combining and iterating over the derived features, we refine our understanding of the interrelationships within the data, setting the stage for more sophisticated analyses.

Clustering Tables for Model Building: With refined feature sets, we employ clustering techniques to segment SKUs and users into groups with similar characteristics. This segmentation facilitates the tailoring of promotional strategies to specific clusters.

Regression Tree Based Model: Utilizing the clustered data, we construct a regression tree model that enables us to predict the relationship between promotion intensities and revenue across different segments.

Through these steps, Descriptive Analytics provides a solid foundation for understanding the current state of affairs and setting the context for predictive modeling.

4.2.2. Prescriptive Analytics

Building on the descriptive base, we plan to transition into the Prescriptive Analytics phase:

Building a Simulation Model for 'What If?' Scenarios: The insights and models developed from Descriptive Analytics are used to create a simulation environment. This allows us to explore 'What If?' scenarios, where we can control the promotional variables to forecast potential outcomes.

Simulation: In this step, we simulated various promotion strategies using the regression tree model to identify which approaches are likely to maximize revenue. The simulation is instrumental in evaluating the effectiveness of different promotional percentages across user and product segments.

The synergy between Descriptive and Prescriptive Analytics culminates in a sophisticated decision-support tool. This tool provides sellers with data-backed recommendations for promotional discounts, thereby optimizing revenue generation and customer engagement on the JD.com platform.

4.3. Data Summary

4.3.1. Orders

Orders summary is shown in Table 4.1.

	Total	1st Party	3rd Party
Total Orders	486928	N/A	N/A
Total SKUs	9159	328	8831
Total Quantity Sold	669155	338958	330197
Total Revenue	45566801	27462145	18104656
Total Revenue (%)	100.0%	60.3%	39.7%

Table 4.1. Summary of Orders and Revenue

Revenue Distribution: 90% of the total revenue comes from 10.2% of the SKUs shown in Figure 4.3.

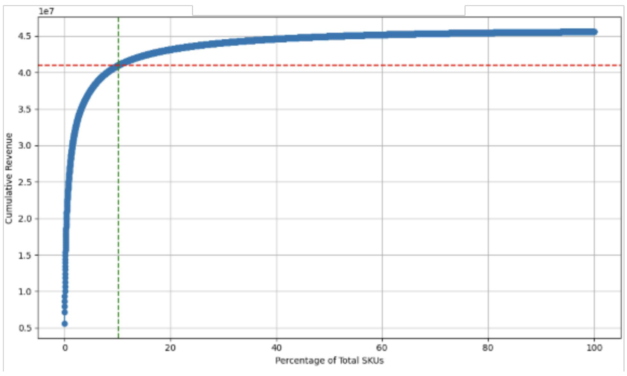

Figure 4.3. Cumulative Revenue vs. Percentage of SKUs

Direct Discount: Discount due to SKU direct discount

Quantity Discount: Discount due to purchase quantity

Coupon Discount: Discount due to customer coupon

Bundle Discount: Discount due to "bundle promotion"

Gift Item: The product can be offered by gift from the seller.

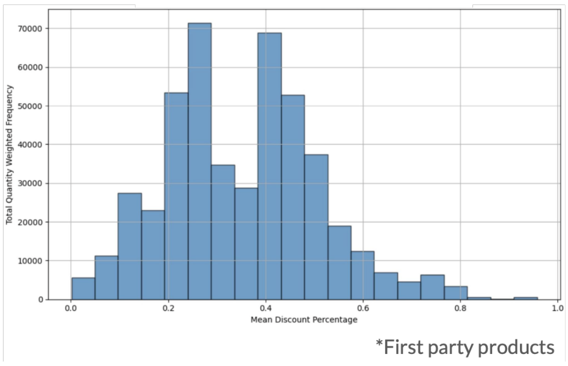

Figure 4.4. Histogram of Mean Discount Percentages Weighted by Total Quantity

4.3.2. Users

Attribute	Values
User Level	-1; 0,1,2,3,4; 10
First Order Month	Year - Month
Plus Membership	0;1
Gender	Female; Male; Unknown
Age	≤15; 16-25; 26-35; 36-45; 46-55; ≥55
Marital Status	Marriage; Single; Unknown
Education	-1; 1,2,3,4 (lowest to highest)
Purchase Power	-1; 1,2,3,4,5 (highest to lowest)
City Level	-1; 1,2,3,4,5 (biggest to smaller)

Table 4.2. User demographics and attributes

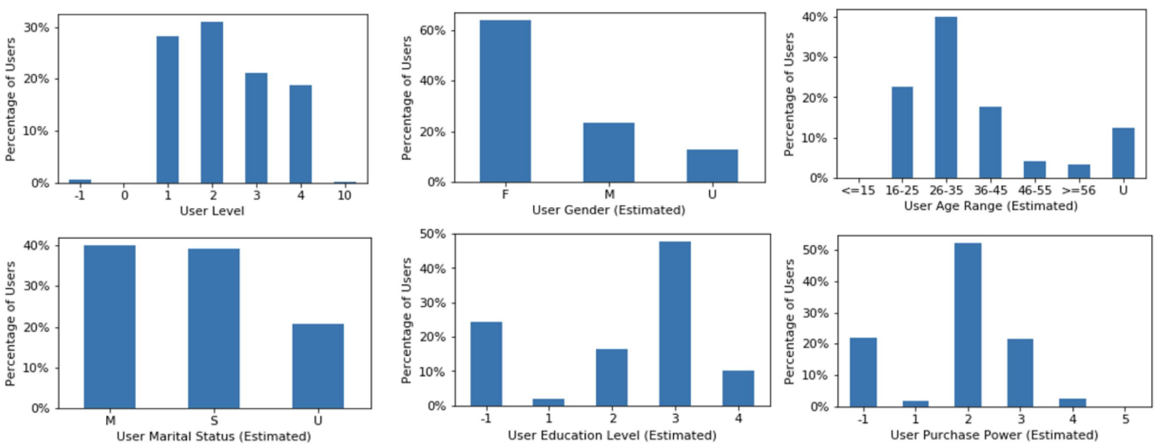

Figure 4.5. Distribution of Users with respect to Attributes (457.298 Users)

4.3.3. SKUs

Distribution of brands and SKUs by party type is shown in Table 4.3.

Party Type	Number of Brands	Number of SKUs
Type 1 (1P)	230	1167
Type 2 (3P)	1828	30701

Table 4.3. Party Types and Counts

Two attribute columns were given in Figure 4.6. Higher value in these attributes indicate better performance.

A1 is scaled 1-4, A2 is 30-100. No specific information related to the sku attributes are given, whether it is tech devices or apparel, etc.

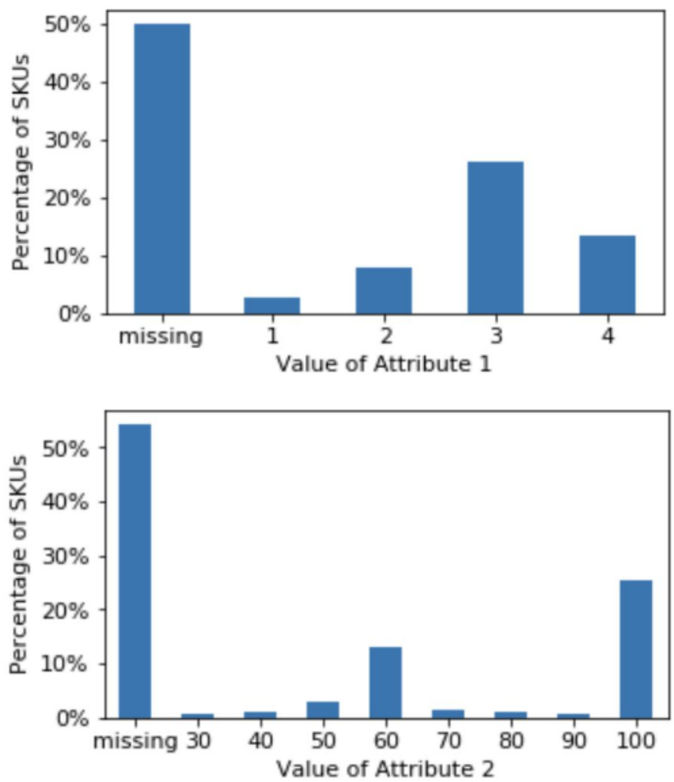

Figure 4.6. Distribution of Attributes A1 and A2

4.3.4. Clicks & Delivery

Clicks: Conversion rates of the SKU's can be obtained in order to eliminate selection bias by calculating the click-order conversion rate.

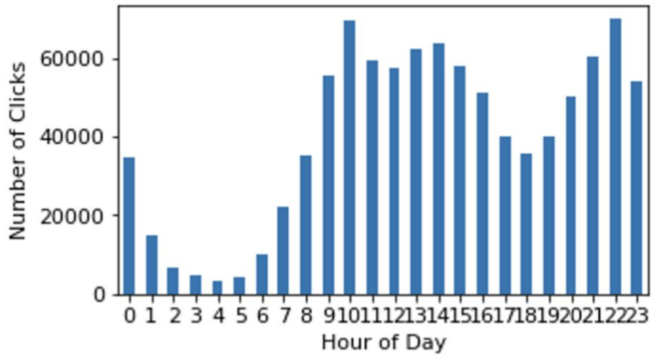

Figure 4.7. Number of Clicks by day time

Party Type	**Number of Delivery**
Type 1 (1P)	244,333
Type 2 (3P)	48,896

Table 4.4. Party Types and Counts for Delivery

4.3.5. Inventory & Network

Inventory: Whether SKU is available or not at the end of the day in the distribution centers(Only available for 1P).

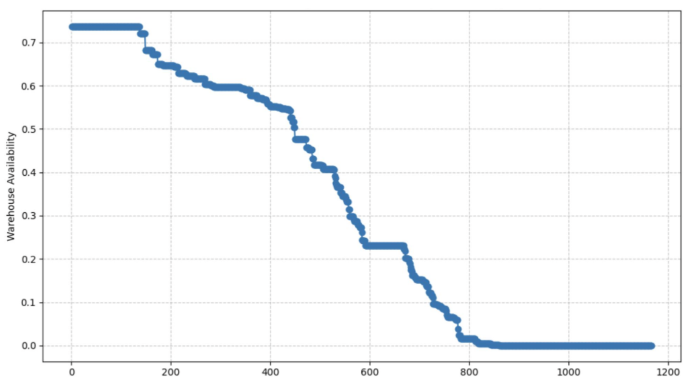

Figure 4.8. Warehouse Availability for the first 1000 SKUs (descending order)

Network: Keeps the hierarchical information of the Regions and the Distribution Centers in these regions.

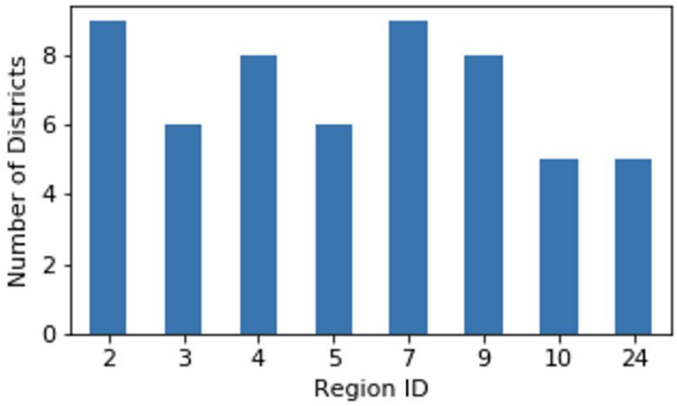

Figure 4.9. Number of Districts in Regions

5. METHODOLOGY

In our project, we aimed to develop a data-driven discount and promotion strategy tailored to different product and user segments to maximize revenue and sales for JD.com. Initially, we explored unsupervised clustering algorithms, but these methods did not yield satisfactory results. Consequently, we shifted our focus to a regression tree methodology to better understand the impact of various promotional strategies on sales and revenue.

Data Analysis: We began by examining transactional data from JD.com, which included detailed user and SKU identifiers. This data allowed us to track specific users and SKUs through order data. The data set provided comprehensive information on users' demographics, behaviors, transaction details, discount types, and pricing mechanisms.

Unsupervised Clustering: Initially, we employed unsupervised clustering algorithms to segment the data. However, these methods did not produce meaningful clusters, leading us to pivot our approach. We then utilized regression trees, which provided a more structured way to analyze the data by segmenting it based on the responses to promotional activities.

Regression Tree Analysis: The core of our methodology involved constructing a regression tree model. This approach allowed us to partition the data into distinct segments, each characterized by unique features. The regression tree model enabled us to:

1.Segment the Data: We divided the dataset into different segments or leaf nodes based on key attributes such as user demographics, product characteristics, and transaction details. This segmentation helped in isolating the effects of various promotional strategies on different groups.

2.Generate Regression Parameters: For each segment, we derived regression parameters to model the relationship between the promotion percentage (discount rate) and total quantity sold. This provided insights into how different levels of promotion affected sales within each segment. The regression parameters were critical in understanding the sensitivity of sales to promotional changes.

3.Model Structure: We set the maximum depth of the regression tree to 6, resulting in 12 leaf nodes. For each leaf node, we trained a linear regression model to quantify the relationship between the discount rate and the total quantity sold. It allowed us to capture the nuances in how different segments responded to promotions.

Simulation: With the regression tree model established, we moved on to the simulation phase. This phase involved:

1.Data Generation: We generated synthetic data points for each discount rate from 0% to 100%. This involved creating data for each leaf node based on the mean discount rate and the error information from the regression summaries. This synthetic data was crucial for running simulations and analyzing various scenarios.

2.Scenario Analysis: We conducted scenario analysis by simulating different promotional strategies. By varying the discount rates, we assessed the potential impact on sales and revenue for each segment. This analysis helped us understand the trade-offs and benefits of different promotion levels.

3.Revenue Optimization: We incorporated a cost function to account for real-life dynamics such as advertising and marketing budgets. This allowed us to optimize the promotional strategies for maximum revenue, taking into consideration the costs associated with higher discount rates. The optimization process ensured that the recommended strategies were not only effective in increasing sales but also cost-efficient.

Model Validation: We validated our model by comparing the generated data with the actual transactional data. This validation ensured the accuracy and reliability of our regression tree model in predicting the impact of promotional strategies on sales and revenue. We used various statistical metrics to assess the goodness-of-fit and predictive power of our model, ensuring its robustness and applicability.

Key Findings: The regression tree analysis revealed several key insights:

- Different user segments and product categories responded uniquely to various levels of promotion. This highlighted the need for tailored promotional strategies.
- Optimal discount rates varied significantly across segments, emphasizing the importance of personalized promotion strategies to maximize revenue.
- Implementing the optimized promotion strategies led to a substantial increase in revenue, demonstrating the effectiveness of the regression tree model in guiding strategic decisions.

By utilizing the regression tree methodology and simulation, we were able to develop a robust and data-driven approach to optimize promotional strategies on JD.com. This methodology provided actionable insights that could be leveraged to enhance revenue and sales on the platform.

6. IMPLEMENTATION AND TESTING

6.1. Implementation

The implementation phase of our project involved developing the regression tree model, running simulations, and optimizing the promotional strategies. The steps taken during this phase are detailed below [9]:

1. Data Preprocessing: The initial step involved preprocessing the transactional data from JD.com. This included cleaning the data, handling missing values, and merging the various tables (users, orders, SKUs, clicks, and deliveries) to create a comprehensive dataset. The order-level data was aggregated to the SKU level to obtain a data set that can be used in regression tree model. Generally, categorical attributes are aggregated as mode and numerical attributes are aggregated as mean. New features were derived from the existing data to capture meaningful patterns. This included categorizing pricing, repeating customers, and rural-urban seperation of users.

2. Click-Weighted Regression Tree Construction: The click-weighted regression tree was constructed using the 'partykit' and 'ggparty' packages in R. The model was designed to segment the data into distinct groups based on key attributes such as user demographics, product characteristics, and transaction level details. The steps included:

- **Data Preparation:** The dataset was prepared by including the click information as a weight parameter.
- **Model Training:** The regression tree model was trained using the *lmtree* function from the *partykit* package, with the click information used as case weights. The model split the data based on key attributes such as user demographics, product characteristics, and transaction details. The maximum depth of the tree was set to 6, resulting in 12 leaf nodes.

- **Generating Regression Equations:** For each leaf node, a linear regression model was trained to establish the relationship between the discount rate and total quantity sold. The regression equations were obtained at the each leaf node to predict the impact of different discount rates on sales.
- **Visualization:** The *ggparty* package was used to visualize the tree structure and the regression models at each node.

3. **Simulation Setup:**

- **Data Generation:** Synthetic data points for each discount rate from 0% to 100% were generated for each leaf node using the regression equation for the leaf and error information from the regression summaries to create the synthetic data.
- **Scenario Analysis:** Various promotional strategies were simulated by varying discount rates. The generated quantities were multiplied with the discounted prices to obtain the revenue for each scenario. The simulation was performed using Python to handle the data generation and scenario testing efficiently. The formula of revenue and cost for each leaf node and each discount rate given below:

$$
\begin{aligned}
&n : \text{Number of SKU's in the leaf}\\
\text{Revenue} &= \sum_{i=1}^{n} \text{Generated Quantity}_i \cdot \text{Original Unit Price}_i \cdot (1 - \text{Discount Rate})\\
\text{Cost} &= \sum_{i=1}^{n} \text{Generated Quantity}_i \cdot \text{Original Unit Price}_i \cdot \text{Discount Rate}
\end{aligned}
\tag{6.1}
$$

4. **Revenue Optimization:** After obtaining the revenue and cost data for each leaf and discount rate, the final data structure is generated. In the mathematical model, exactly one discount rate is assigned to each leaf node. The optimization process was carried out to maximize revenue while considering the cost function. A cost parameter is interpreted as real-life dynamics such as advertisement or marketing budget. Essentially, the cost function acts as a penalty for increasing the discount rate.

$$i : \text{Index of leaves}$$

$$x_{i,dr} = \begin{cases} 1 & \text{if discount rate } dr \text{ is assigned to leaf } i \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Maximize} \quad Z = \sum_{i=1}^{n} \sum_{dr \in D} x_{i,dr} \cdot \text{revenue}_{i,dr} \tag{6.2}$$

$$\text{subject to} \quad \sum_{dr \in D} x_{i,dr} = 1 \qquad \forall i \in \{1, \ldots, n\} \tag{6.3}$$

$$\sum_{i=1}^{n} \sum_{dr \in D} x_{i,dr} \cdot \text{cost}_{i,dr} \leq \text{Budget} \tag{6.4}$$

$$x_{i,dr} \in \{0, 1\} \qquad \forall i \in \{1, \ldots, n\}, dr \in D \tag{6.5}$$

5. Implementation Tools: The implementation was carried out using R, Python, and its libraries such as *ggparty*, *pandas*, and *NumPy* for data processing, model building, simulations. For optimization, the *PuLP* library was utilized.

6.2. Testing

Testing was an integral part of our project to ensure the accuracy and reliability of the regression tree model and the simulation results. The following steps were undertaken:

1.Model Validation:

- **Validation Data:** The regression tree model was validated by comparing the synthetic data with the actual transactional data from JD.com. This validation involved checking the goodness-of-fit for each leaf node's regression models to ensure that the predicted values closely matched the actual values.
- **Statistical Metrics:** P-value of the regressions at each leaf node was examined to assess the significance of the relationships between the discount rates and total quantity sold. Also, standard deviation of the residuals examined and also used in the simulation phase.

2.Scenario Testing:

- Various promotional strategies were tested through simulations. The impact of different discount rates on sales and revenue was analyzed to ensure the model accurately predicted the outcomes.
- **Error Analysis:** Residual errors from the regression models were analyzed to identify any patterns or biases, which helped in refining the model and improving its predictive accuracy.

7. RESULTS

The results section provides a detailed account of the findings from the various stages of our project, from model validation to scenario analysis and revenue optimization. The following subsections present these findings systematically.

1.Click-Weighted Regression Tree: The regression tree construction involved segmenting the data into different groups based on user and product attributes. Each leaf node of the tree represents a segment with a distinct regression equation relating the discount percentage to the total quantity sold. Tree structure is shown in Figure 7.1.

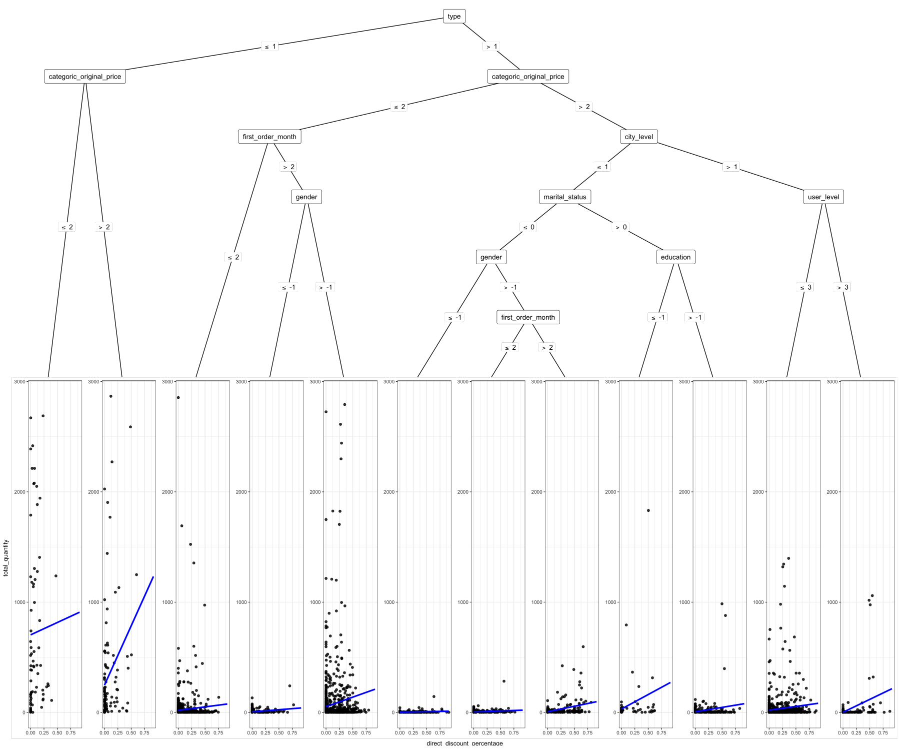

Figure 7.1. Click-Weighted Regression Tree Structure

Regression Equations: The regression equations for each leaf node, along with the p-values, mean residual errors (MRE), and standard deviations of residual errors (SDRE), are summarized in Table 7.1. These equations provide insights into the relationship between the discount percentage and total quantity sold for different segments of users and products.

Leaf Node	Regression Equation(total_quantity)	p-value	MRE	SDRE
1	$703.4 + 224.3 \cdot \text{discount\_percentage}$	0.797	-310.9	780.3
2	$247.6 + 1076.37 \cdot \text{discount\_percentage}$	0.00452	-167.78	523.4
3	$8.941 + 62.723 \cdot \text{discount\_percentage}$	0.0211	-16.94	131.3
4	$611 + 38.212 \cdot \text{discount\_percentage}$	2.38e-06	-3.611	18.51
5	$52.392 + 172.508 \cdot \text{discount\_percentage}$	9.06e-05	-49.39	234.8
6	$2.2508 + 9.7997 \cdot \text{discount\_percentage}$	1.96e-05	-1.251	8.572
7	$5.120 + 16.760 \cdot \text{discount\_percentage}$	0.004909	-3.120	18.5
8	$5.662 + 100.216 \cdot \text{discount\_percentage}$	2.13e-08	-4.66	62.73
9	$30.57 + 263.19 \cdot \text{discount\_percentage}$	0.0559	-27.57	240.9
10	$5.792 + 80.044 \cdot \text{discount\_percentage}$	0.000261	-4.79	74.88
11	$20.361 + 68.033 \cdot \text{discount\_percentage}$	0.00016	-19.36	112.3
12	$1.509 + 233.525 \cdot \text{discount\_percentage}$	5.28e-06	-0.51	130.4

Table 7.1. Regression Equations, p-values, Mean Residual Errors, and Standard Deviations of Residual Errors

2.Simulation Results: By using the regression equations above for each leaf node, we simulated various promotional scenarios to analyze their impact on sales and revenue. The results of the simulations are visualized below.

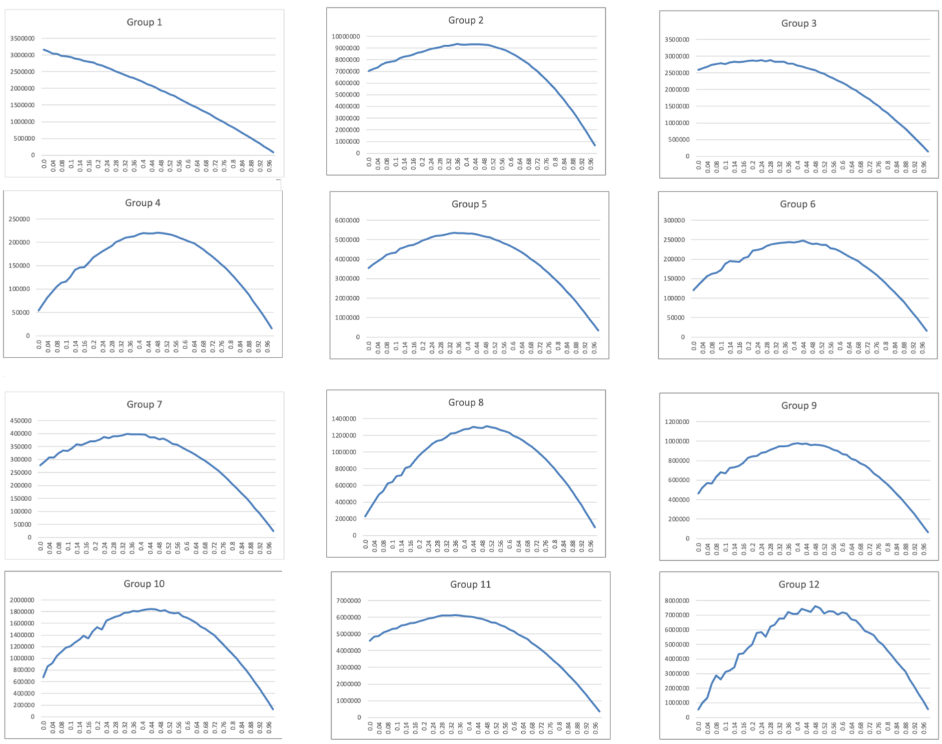

Figure 7.2. Simulated Revenue for Different Discount Rates of Each Leaf Node

3.Revenue Optimization: The revenue optimization process involved maximizing revenue while considering the cost function. The optimization results are given below for different budget constraints.

- Constant Budget(Base Case): 22% Increase in Revenue
- No budget Constraint: 35% Increase in Revenue
- Halved Budget: 12% Increase in Revenue
- Doubled Budget: 32% Increase in Revenue

	Optimized Decision			Current Strategy		
Group	**Discount Rate**	**Expected Revenue**	**Cost**	**Real Discount Rate**	**Real Revenue**	**Real Cost**
1	0,00	3.156.142,41	-	0,08	4.866.527,74	808.387,44
2	0,12	8.142.720,74	1.110.371,01	0,09	11.056.153,95	2.158.638,86
3	0,06	2.738.589,46	174.803,58	0,08	1.594.367,91	345.253,76
4	0,32	204.829,42	96.390,31	0,06	133.687,81	32.528,34
5	0,20	4.958.762,02	1.239.690,50	0,08	4.371.370,26	992.592,74
6	0,22	221.589,03	62.499,47	0,12	170.544,29	61.003,64
7	0,18	370.721,85	81.377,97	0,12	286.772,50	65.632,16
8	0,26	1.102.665,96	387.423,17	0,18	564.489,29	417.442,50
9	0,20	844.650,20	211.162,55	0,15	452.823,02	253.977,76
10	0,24	1.643.352,99	518.953,58	0,12	672.728,43	320.547,72
11	0,12	5.514.404,27	751.964,22	0,14	4.524.994,46	1.885.953,85
12	0,32	6.775.848,23	3.188.634,46	0,12	605.289,65	482.833,34
Total		**35.674.276,6**	**7.823.270,8**		**29.299.749,3**	**7.824.792,1**

Figure 7.3. Constant Budget: 22% Increase in Revenue

Recommendations Based on the Constant Budget Scenario:

- For categories 2, 4, 5, 6, 7, 8, 9, 10, and 12, it is recommended to increase the discount rates to boost revenue. These categories have shown potential for significant revenue growth with higher discount rates.
- For the remaining groups, lowering the discount rates is advised to optimize revenue streams. This approach will help balance the revenue generation across different categories.

	Optimized Decision		Current Strategy	
Group	**Discount Rate**	**Expected Revenue**	**Real Discount Rate**	**Real Revenue**
1	0.00	3.156.142,4	0.08	4.866.527,7
2	0.36	9.343.631,2	0.09	11.056.154,0
3	0.24	2.882.437,8	0.08	1.594.367,9
4	0.48	2.208.839,8	0.06	133.687,8
5	0.34	5.352.073,2	0.08	4.371.370,3
6	0.44	2.476.595,0	0.12	170.544,3
7	0.34	398.517,1	0.12	286.772,5
8	0.5	1.311.760,0	0.18	564.489,3
9	0.4	9.810.400,8	0.15	452.823,0
10	0.44	1.848.072,3	0.12	672.728,4
11	0.34	6.131.854,6	0.14	4.524.994,5
12	0.48	7.608.083,3	0.12	605.289,7
Total		**39.482.155,54**		**29.299.749,31**

Figure 7.4. No Budget Constraint: 35% Increase in Revenue

	Optimized Decision			Current Strategy		
Group	**Discount Rate**	**Expected Revenue**	**Cost**	**Real Discount Rate**	**Real Revenue**	**Real Cost**
1	0,00	3.156.142,4	0	0,08	4.866.527,7	808.387,4
2	0,06	7.611.391,2	485.833	0,09	11.056.154,0	2.158.638,9
3	0,00	2.591.948,7	0	0,08	1.594.367,9	345.253,8
4	0,20	167.289,4	41.822	0,06	133.687,8	32.528,3
5	0,12	4.540.525,5	619.163	0,08	4.371.370,3	992.592,7
6	0,14	195.079,4	31.757	0,12	170.544,3	61.003,6
7	0,14	358.222,6	58.315	0,12	286.772,5	65.632,2
8	0,22	1.010.048,0	284.885	0,18	564.489,3	417.442,5
9	0,18	827.349,3	181.613	0,15	452.823,0	253.977,8
10	0,15	1.386.026,4	244.593	0,12	672.728,4	320.547,7
11	0,06	5.087.515,7	324.735	0,14	4.524.994,5	1.885.953,9
12	0,22	5.797.455,2	1.635.180	0,12	605.289,7	482.833,3
Total		**32.728.993,94**	**3.907.897,05**		**29.299.749,31**	**7.824.792,11**

Figure 7.5. Halved Budget: 12% Increase in Revenue

	Optimized Decision			Current Strategy		
Group	**Discount Rate**	**Expected Revenue**	**Cost**	**Real Discount Rate**	**Real Revenue**	**Real Cost**
1	0,00	3.156.142,4	0	0,08	4.866.527,7	808.387,4
2	0,30	9.177.154,7	3.933.066	0,09	11.056.154,0	2.158.638,9
3	0,14	2.838.651,9	462.106	0,08	1.594.367,9	345.253,8
4	0,40	217.122,4	144.748	0,06	133.687,8	32.528,3
5	0,28	5.226.375,1	2.032.479	0,08	4.371.370,3	992.592,7
6	0,32	240.346,8	113.104	0,12	170.544,3	61.003,6
7	0,34	398.517,1	205.297	0,12	286.772,5	65.632,2
8	0,44	1.303.085,5	1.023.853	0,18	564.489,3	417.442,5
9	0,32	948.281,1	446.250	0,15	452.823,0	253.977,8
10	0,32	1.778.117,5	836.761	0,12	672.728,4	320.547,7
11	0,28	6.116.056,4	2.378.466	0,14	4.524.994,5	1.885.953,9
12	0,36	7.232.514,3	4.068.289	0,12	605.289,7	482.833,3
Total		**38.632.365,22**	**15.644.420,64**		**29.299.749,31**	**7.824.792,11**

Figure 7.6. Doubled Budget: 32% Increase in Revenue

General Findings, Recommendations and Assumptions

- Groups 1 and 2 are identified as Type 1, where JD.com itself acts as the reseller. Since the revenue directly impacts JD.com, the implementation strategy can be applied more easily compared to Third Party Seller categories. This direct control allows for more flexibility in adjusting discount rates and promotional strategies.
- The current budget data does not originate from JD.com. Implementing real budget data that includes actual cost items will provide more accurate results. It is crucial to align the budget with real costs to ensure the effectiveness of the revenue optimization strategies.
- The new strategy is recommended for each leaf node, with the assumption that JD.com does not have fulfillment and network capacity constraints. This assumption is critical for the smooth implementation of the new discount and revenue strategies.

8. CONCLUSION

Throughout this project, we have thoroughly investigated the impact of various promotional strategies on the pricing and revenue streams of sellers on JD.com, the largest retailer in China. By analyzing a comprehensive dataset comprising over 2.5 million customers and 31,868 SKUs, we developed an econometric model to explore the relationship between promotions and consumer behavior.

Our study yielded several critical insights. Different user segments and product categories exhibited distinct responses to promotional activities, emphasizing the need for tailored promotional strategies. Using a regression tree model, we identified optimal discount rates for various segments, leading to significant revenue enhancements. The simulation and scenario analyses demonstrated that strategic discounting could substantially improve seller performance in the online marketplace.

The implementation phase involved detailed data preprocessing, regression model construction, and simulation execution. Our simulations indicated that revenue could be increased by 22% under constant budget constraints, with potential gains of up to 35% in scenarios without budget constraints. This validated our hypothesis that data-driven promotional strategies can optimize revenue.

However, the study also revealed limitations and areas for future research. The limited time horizon of the transactional data restricted our ability to conduct long-term trend analysis. Future research should incorporate extended datasets to further validate and refine the models. Additionally, integrating actual budget data, including real cost items, will provide a more accurate framework for revenue optimization.

In conclusion, this project has provided valuable insights into the strategic management of promotional activities in e-commerce. By leveraging advanced data analytics and econometric modeling, we have offered actionable recommendations to help e-commerce platforms like JD.com optimize their promotional strategies, thereby enhancing customer experience and revenue generation. The methodologies and findings from this study can serve as a foundation for further research and practical applications in e-commerce analytics.

REFERENCES

1. Shen, M., C. S. Tang, D. Wu, R. Yuan and W. Zhou, "Jd.com: Transaction-Level Data for the 2020 MSOM Data Driven Research Challenge", *Manufacturing & Service Operations Management*, Vol. 26, No. 1, pp. 2–10, 2024.

2. Marzuki, S. N. and I. M. Yasin, "Factors Influencing Online Shoppers to Shop at Marketplace or Website: A Qualitative Approach", *Asian Journal of Business and Management*, Vol. 9, No. 4, p. 68, 2021.

3. Santini, F. O., C. H. Sampaio, M. G. Perin, L. B. Espartel and W. J. Ladeira, "Moderating Effects of Sales Promotion Types", , 2015.

4. Zhang, D. J., H. Dai, L. Dong, F. Qi, N. Zhang, X. Liu, Z. Liu and J. Yang, "The Long-term and Spillover Effects of Price Promotions on Retailing Platforms: Evidence from a Large Randomized Experiment on Alibaba", *Management Science*, 2020.

5. Wen, J., P. Vahabi and M. Grbovic, "Revenue Maximization of Airbnb Marketplace Using Search Results", , 2019, pennsylvania State University.

6. Ghose, A. and A. Sundararajan, "Evaluating Pricing Strategy Using e-Commerce Data: Evidence and Estimation Challenges", *Statist. Sci.*, Vol. 21, No. 2, pp. 131–142, 2006.

7. Huang, T. and J. A. Van Mieghem, "The promise of strategic customer behavior: On the value of click tracking", *Production and Operations Management*, Vol. 22, No. 3, p. 489–502, 2013.

8. Zumstein, D. and W. Kotowski, "Success Factors of E-commerce – Drivers of the Conversion Rate and Basket Value", , 2020, zurich University of Applied Sciences, Institute of Marketing Management, 18th International Conference e-Society 2020.

9. Coskun, M. C., E. Eren, F. Kahvecioglu and I. E. Koksal, "Effects of Promotion Mechanisms on Revenue of Sellers in an Online Marketplace", , Jun. 2024, https://github.com/enverern/CMPE492.