Decoding Sales Drivers In E-Commerce

(Causal Analysis)

TEAM E-COMMERCE

Amna Gul
Bhakti Barve
Jay Maniyar
Jagriti Sharma
Kashish Kalra
Lebogang Mooketsi
Pranali Rahangdale

USC Marshall
School of Business

INTRODUCTION

In the dynamic landscape of e-commerce, understanding what truly drives sales is crucial for platform success. Every day, Wish faces critical questions about the effectiveness of its platform features and rating systems.

Like many organizations in the e-commerce space, Wish struggles to effectively leverage its historical sales data, trends, and other influential factors to make data-driven decisions that improve profitability. There is a pressing need for a comprehensive strategy and analysis that can accurately forecast sales performance and provide actionable insights, enabling businesses to stay ahead in dynamic market conditions. This challenge is particularly acute in optimizing sales performance and identifying the factors that most significantly impact total units sold.

The impact of understanding these factors is crucial for several reasons:

- Revenue Growth: By pinpointing the most impactful drivers, Wish can boost sales through targeted strategies and resource allocation.
- Market Strategies: Tailoring approaches like pricing models, promotional campaigns, and product features can strengthen market position and drive customer engagement.
- Competitive Advantage: In a crowded e-commerce market, understanding key sales drivers helps Wish to stay competitive by aligning strategies with customer preferences.

Within this context, Three key challenges consistently emerge in their operations:

First, the impact of product ratings on customer purchasing decisions remains unclear. While conventional wisdom suggests that higher-rated products should sell better, the relationship isn't always straightforward. Some highly-rated products struggle to gain traction, while others with moderate ratings sometimes unexpectedly outperform. This creates uncertainty in predicting sales performance and challenges in guiding merchants about the importance of maintaining high product ratings.

Second, Wish employs urgency messaging as a marketing tactic to drive sales, displaying banners like "Only 3 left!" or "Limited time offer!" across various products. However, the platform lacks concrete evidence about whether these messages genuinely influence customer buying behavior or if they've become just another element of background noise in the increasingly cluttered e-commerce space.

Third, the platform offers ad boost features that allow products to gain greater visibility, but questions persist about their true effectiveness in driving sales. With merchants investing significant resources in these boosts, understanding their actual impact on customer buying behavior is crucial for both the platform and its sellers.

This is where our research project steps in. Using a dataset of 1,573 observations from the Wish platform, we aim to uncover the causal relationships between these three key factors and sales performance. Our research questions are precisely targeted:

1. What is the causal effect of product ratings on sales?
2. Does urgency messaging directly affect customer buying behavior?
3. Does ad boost directly affect customer buying behavior?

These questions aren't just academic exercises - they address core business needs. Understanding the causal impact of product ratings can help Wish better predict sales performance and guide their rating system development. Knowing the true effect of urgency messaging can inform marketing strategies and interface design. Understanding the impact of ad boosts can help optimize pricing and recommendations for this feature.

Our approach employs sophisticated causal inference techniques to extract reliable insights despite working with a relatively small dataset. We'll utilize methods such as propensity score matching, Double machine learning, Inverse Propensity Weighting, Causal Forest, etc. to establish causal relationships rather than mere correlations.

The practical implications of this research are significant. If urgency messaging proves highly effective, Wish might invest in more sophisticated implementations of this feature. If ad boosts show a strong causal impact on sales, the platform could develop more nuanced boost options.

Understanding the true impact of product ratings could lead to improved rating systems and better sales predictions. Ultimately, this project aims to transform data into actionable insights that can drive Wish's decision-making process. By uncovering the causal relationships between these key factors and sales, we'll provide Wish with the evidence needed to optimize its platform features, improve the shopping experience, and drive growth in the competitive e-commerce marketplace.

This focused approach will allow us to dive deep into each research question, ensuring robust and reliable findings that can directly inform Wish's strategic decisions. The insights gained from this analysis will not only benefit Wish's current operations but also contribute to the broader understanding of e-commerce dynamics and customer behavior.

To align our analysis with the business challenges faced by the Wish platform, we define the following hypotheses that reflect both the strategic importance of key business levers and the methodological rigor needed for causal inference. These hypotheses are rooted in real-world challenges faced by e-commerce platforms, such as low-rated products struggling to gain traction, the mixed impact of urgency tactics, and uncertainty around discount strategies.

Rather than relying on simple correlations, we pursue causal insights using a layered approach that incorporates confounder detection, matching methods, and machine learning techniques to isolate the true drivers of sales.

HYPOTHESIS

Hypothesis 1: Impact of Ratings on Sales

Business Motivation:

Sellers with lower ratings may find it difficult to gain trust, resulting in poor sales. The platform needs to determine whether improving ratings (or compensating with discounts) could effectively boost sales for underperforming merchants.

Hypotheses:

- **Null (H0)**: Product ratings do not causally influence total units sold.
- **Alternative (H1):** Higher Product ratings lead to an increase in total units sold.

Methodological Design:

- **Confounder Detection:** Applied regression-based confounder identification
- **Causal Methods:** Double Machine Learning (DML) for continuous treatment variables

Hypothesis 2: Effect of Urgency Messaging and Ad Boosts

Business Motivation:

Urgency cues (like limited time offers) and paid ad placements are used to influence buyer psychology. It’s important to understand if these features truly convert into higher sales or merely add noise or cost.

Hypotheses:

- **Null (H0):** Urgency texts and ad boosts have no causal effect on units sold.
- **Alternative (H1):** These interventions lead to a statistically significant increase in units sold.

Methodological Design:

- **Confounder Detection:** Regression-based checks for urgency and ad boost groups
- **Causal Method:** Propensity Score Matching, using urgency/ad_boost as binary treatment and matched controls to estimate ATT (Average Treatment Effect on the Treated)

Hypothesis 3: Impact of Discount Disparity on Sales

Business Motivation:

Products often show a large disparity between the listed retail price and the actual sale price. It's critical to assess whether these discounts influence consumer behavior and at what threshold they become effective.

Hypotheses:

- **Null (H0):** Price discount disparity has no causal effect on units sold.
- **Alternative (H1):** Larger perceived discounts causally increase units sold.

Methodological Design:

- **Confounder Detection:** Regressed pricing disparity and units sold on other variables to identify shared confounders
- **Causal Method:** Applied PSM across discount tiers and considered modeling heterogeneous treatment effects if needed

UNDERSTANDING OUR DATA

Data Overview

The dataset was sourced from Kaggle and originates from the Wish.com e-commerce platform. It includes products that appear in the search results when the keyword **"summer"** is entered and collected during August 2020.

- **Observations:** 1,573 rows
- **Features:** 43 columns
- **Unit of Analysis:** Each row is a unique product listing on the Wish platform.

Variable Categories

- **Product Information:** product_id, title, price, retail_price, units_sold, rating, color, product_variation_size_id, product_picture, etc.
- **Merchant Information:** merchant_id, merchant_title, merchant_rating, merchant_rating_count, merchant_profile_picture, etc.
- **Meta Data:** crawl_month, theme, etc.

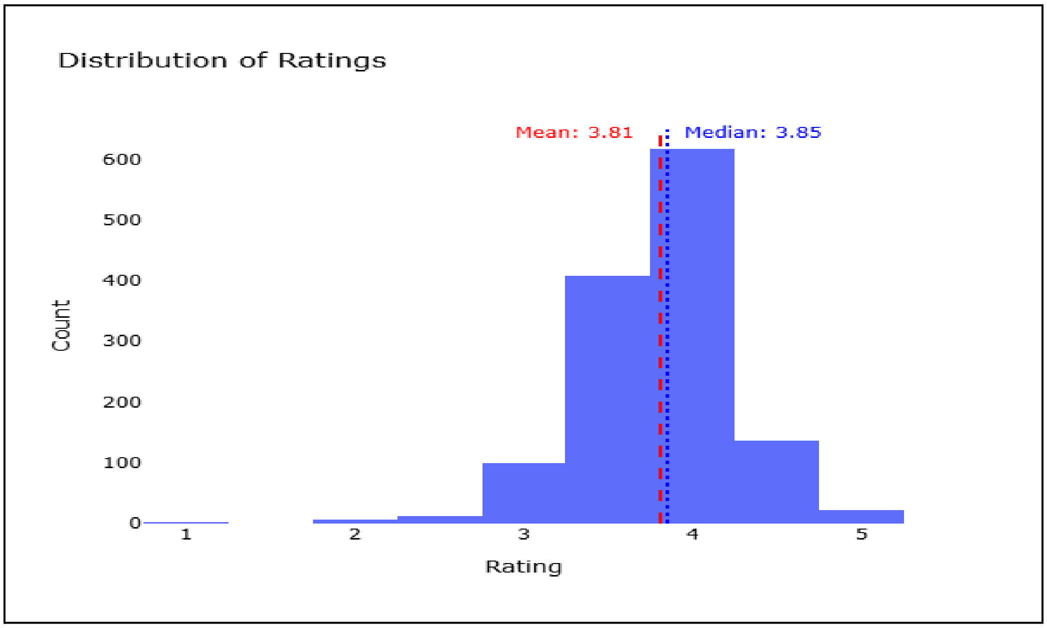
Data Quality Issues

A detailed audit of the raw dataset highlighted the following quality concerns:

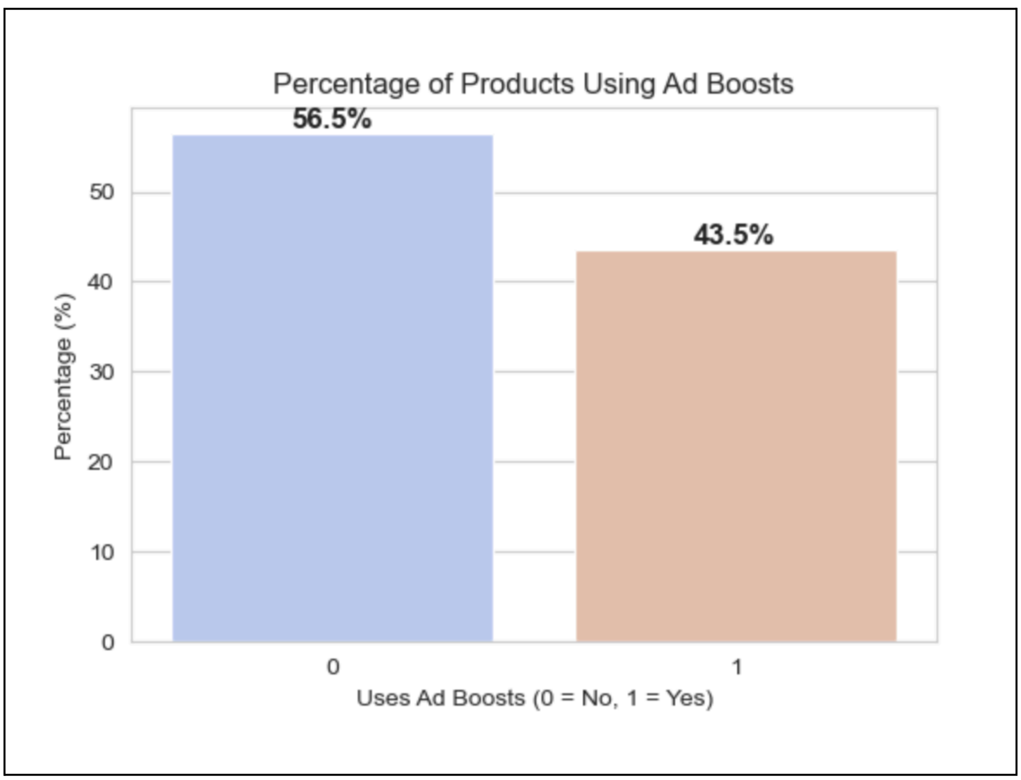
- **Missing Data:**
  - has_urgency_banner: Missing in 1,071 rows (~68% of dataset)

    - Other fields such as product_color, merchant_profile_picture also contain missing values.
- **Duplicate Entries:**
    - Identified and removed 34 duplicate rows.
- **Class Imbalance:**
    - Severe imbalance in binary fields like badge_local_product (~98% False, ~2% True).
    - Categorical variables such as product_color contain many rare levels (e.g., 83 color values appearing fewer than 10 times).
- **Data Inconsistency:**
    - merchant_rating values present despite merchant_rating_count = 0, suggesting spurious entries.
    - Inconsistent formats in product_variation_size_id, e.g., 'S', 's.', 'SIZE-S', 'small', 'S Pink'.

For a more comprehensive understanding, deep diving through visuals:



Histogram of Rating V/S Count of Products



We can see that the majority of products sold were not being promoted.

METHODOLOGY

DATA PREPROCESSING

The initial data preparation focused on streamlining the dataset by removing unnecessary features and handling missing values. We eliminated 17 non-essential columns such as Title, Merchant Name, and Product URL to concentrate on relevant features for analysis. A systematic approach was taken to address missing values, with specific handling for the "Has Urgency Banner" column while maintaining data integrity across other variables. The preprocessing phase then progressed through several critical steps to enhance data quality and model effectiveness:

- Missing Value Treatment:
  - "Has Urgency Banner": Replaced missing values with 0
  - Other columns: Applied complete case deletion

Missing values can significantly impact model performance and lead to biased results. For the "Has Urgency Banner" column, replacing missing values with 0 was logical as it indicated the absence of an urgency banner. For other columns, complete case deletion was chosen over imputation methods to ensure data reliability, as imputing values could have introduced artificial patterns in the dataset.

- Class Imbalance Handling:
  - Applied SMOTE for "Has Urgency Banner" (70:30 ratio)
  - Removed binary columns with severe imbalance (≥90:10)

Class imbalance can lead to biased model predictions towards the majority class. SMOTE (Synthetic Minority Over-sampling Technique) was specifically applied to the "Has Urgency Banner" feature to achieve a more balanced 70:30 ratio, creating synthetic examples of the minority class while avoiding overfitting. Binary columns with extreme imbalances (90:10 or greater) were removed as they provided limited predictive value and could potentially introduce noise into the model.

- Categorical Variable Treatment:
    - Consolidated rare categories in the "Color" column into "Less Frequent Item"
    - Size standardization: Mapped 91 original sizes into 4 distinct groups

Categorical variables with numerous rare categories can lead to sparse data and unreliable model predictions. For the "Color" column, infrequent categories were grouped into a "Less Frequent Item" category to reduce dimensionality while preserving important categorical information. The size standardization was particularly crucial, as it simplified the 91 different size variations into 4 meaningful groups, making the data more manageable while maintaining practical significance.

- Quality Control:
    - Addressed spurious ratings
    - Validated categorical groupings

Quality control measures were implemented to ensure data reliability. Spurious ratings were identified and handled to prevent them from skewing the analysis. The categorical groupings were validated to ensure they maintained business relevance while effectively reducing dimensionality. This step was crucial for maintaining the balance between data simplification and preservation of meaningful information.

These preprocessing steps were carefully implemented to ensure robust data quality while preserving meaningful patterns and relationships within the dataset, ultimately preparing it for effective modeling and analysis. Each step was chosen based on its ability to address specific data challenges while maintaining the integrity of the information needed for accurate modeling.

DETECTING THE CONFOUNDERS

In observational studies, confounders are variables that influence both the treatment assignment and the outcome, potentially biasing the estimated causal effects. Identifying and adjusting for these confounders is critical to ensure valid inference.

To detect potential confounders in our dataset, we followed a two-step statistical approach:

1. **Association with Outcome**:

   For each candidate variable, we regressed the outcome variable on that feature to check whether it significantly affects the outcome. If the p-value of the regression coefficient was less than 0.05, the variable was flagged as potentially outcome-relevant.

2. **Association with Treatment**:

   Next, we regressed the treatment variable on the same candidate features to assess whether they also influence treatment assignment. Again, a p-value below 0.05 was used as the threshold for statistical significance.

A variable was considered a confounder **only if** it was significantly associated with both the treatment and the outcome (i.e., $p\text{-value} < 0.05$ in both regressions). This process helped ensure we only controlled for variables that could bias the treatment effect estimation if ignored.

By systematically applying this criterion, we curated a set of confounding variables to include in our causal inference models, thereby improving the robustness of our ATE estimates.

Treatment= Rating	**Treatment= uses_ad_boost**	**Treatment= has_urgency_banner**
rating_count	product_variation_inventory	inventory_total
product_variation_inventory	-	uses_ad_boosts
merchant_rating	-	merchant_has_profile_picture
merchant_has_profile_picture	-	origin_country_CN
product_color_yellow	-	product_color_black
variation_size_label_encoded	-	-

Identified Confounders for Each Treatment Variable Based on Statistical Association with Both Treatment and Outcome

MODELS USED

To estimate the causal effect of treatment variables on the outcome, I employed a diverse set of causal inference techniques. Each method brings its own assumptions, strengths, and limitations, allowing for a more robust and comprehensive analysis when used in combination. Below is a summary of the models used:

Linear Regression (with Covariates)

A baseline model to estimate treatment effects while controlling for observed confounders. It assumes linearity and no unobserved confounding.

Propensity Score Matching (PSM)

Matches treated and control units with similar propensity scores to simulate a randomized experiment. This helps reduce selection bias but may discard unmatched units.

Inverse Probability Weighting (IPW)

Weights observations by the inverse of their propensity to receive the treatment, creating a pseudo-population where treatment is independent of observed confounders. Used for both ATT and ATE estimation.

Double Machine Learning (DML)

Combines machine learning models to estimate nuisance parameters (outcome and treatment models), then uses orthogonalization to isolate the treatment effect. This method handles high-dimensional confounders and reduces overfitting bias.

Causal Forest

A non-parametric ensemble method that estimates heterogeneous treatment effects (CATE) and can be averaged to compute ATE. It captures non-linear relationships and treatment effects heterogeneity across subpopulations.

RESULTS

Causal Effect of Product Ratings on Sales

Method	Estimated ATE	95% Confidence Interval	p-value	Significant?
OLS Regression	Coeff: 1022.12	[45.81, 1998.43]	0.04	Yes
Linear Reg DML	985.99	[448.52, 1521.46]	0.000	Yes
Lasso DML	1421.95	[305.07, 2538.81]	0.012	Yes
Causal Forest	1966.24	[-2260.67, 6431.76]	0.68	No

All methods estimate a strong and statistically significant positive effect of product ratings on units sold. ATEs range from +986 to +1966, with tight confidence intervals in most models.

We selected Linear DML because it produced a precise, statistically significant estimate with a narrower confidence interval than Lasso and better model stability than Causal Forest.

The result strongly aligns with business intuition - highly rated products attract more buyers. Customers rely on social proof, and positive reviews increase trust and conversion.

Causal Effect of Adboost on Sales

Method	Estimated ATE	95% Confidence Interval	p-value	Significant?
OLS Regression	Coeff: -104.75	[-1037.706, 828.203]	0.826	No
IPW	-127.01	[-1024.250, 770.228]	0.781	No
PSM	928.10	[220.87, 1621.01]	0.0136	Yes
DML	-88.52	[-969.87, 792.83]	0.8440	No
Causal Forest	128.79	[–5138.39, 5395.97]	0.9618	No

Only PSM finds a statistically significant positive effect (+928 units sold). Other methods show small, uncertain effects with wide confidence intervals, lacking significance.

We selected PSM due to its statistically significant ATE, tight confidence interval, and excellent covariate balance, making it most reliable for this small dataset.

The positive effect of ad boosts aligns with expectations - promoted products should get more visibility and sales. However, the lack of consistent results across methods suggests that boosting is effective only in specific contexts, reinforcing the need for targeted campaigns.

Causal Effect of Urgency Banners on Sales

Method	Estimated ATE	95% Confidence Interval	p-value	Significant?
OLS Regression	Coeff: -466	[-1239, 306]	0.237	No
IPW	-503.14	[-1253.95, 247.67]	0.189	No
PSM	-1192.36	[-1599.99, -784.73]	0.000	Yes
DML	-710.06	[-1861.77, 441.65]	0.226	No
Causal Forest	-446.49	[-607.07, -285.91]	0.000	Yes

Most models estimate a negative effect, but only PSM and Causal Forest show statistically significant results. ATEs range from –446 to –1192, suggesting urgency banners may reduce sales.

We chose Causal Forest for its statistically significant estimate with a narrower confidence interval and its ability to capture complex interactions in consumer behavior.

The finding challenges common marketing practice - urgency cues are expected to boost conversions. Here's why this counterintuitive result might make sense:

Why Didn't the Urgency Banner Drive Sales as Expected?

The negative correlation between urgency banners and unit sales presents an interesting paradox and is indeed counterintuitive. There may be several reasons for this:

- **Banner fatigue** plays a significant role, as customers become increasingly desensitized to urgency messages, potentially viewing them as manipulative or triggering a "cry wolf" effect that diminishes credibility.
- **Psychological reactance** is another crucial factor, where customers actively resist pressure tactics to preserve their sense of autonomy in decision-making.
- **Trust issues** emerge when perpetual urgency messaging appears deceptive, damaging brand credibility and customer confidence.
- **Price sensitivity** also comes into play, as urgency banners often accompany full-price items, causing value-conscious customers to wait for genuine sales. Furthermore, the pressure created by urgency can lead to decision paralysis, where customers feel overwhelmed and abandon their purchase intentions rather than making rushed decisions.
- **Alternative effects** may include customers bookmarking items for later purchase or engaging in comparison shopping, effectively counteracting the intended immediate purchase behavior. Understanding these complex psychological responses is crucial for optimizing urgency messaging strategies and improving conversion rates in e-commerce platforms.

BUSINESS APPROACH

Solving Urgency banner paradox

Upon observing a concerning negative Average Treatment Effect (ATE) for urgency banners, we conducted a deeper analysis to understand the true impact on sales performance. Our investigation revealed a striking contrast in treatment effects. The Average Treatment Effect (ATE) of -446.49 ($p<0.001$) indicates that implementing urgency banners across all products would, on average, decrease sales by approximately 446 units. However, **the Average Treatment Effect on the Treated (ATT) of 1096.66 ($p<0.001$**) demonstrates that products currently displaying urgency banners experienced a substantial increase of about 1,097 units in sales.

This substantial divergence between ATE and ATT suggests strong treatment effect heterogeneity and effective targeting mechanisms in current urgency banner deployment. The positive ATT indicates that marketers have successfully identified products where urgency messaging resonates with customers, likely those with genuine scarcity or time-sensitive appeal. However, the negative ATE implies that universal application of urgency banners would be counterproductive, potentially because overuse could diminish their credibility or effectiveness. This pattern might also reflect sophisticated consumer behavior - urgency messaging appears to work well for products where scarcity or time pressure is authentic and relevant but could backfire when perceived as artificial or manipulative.

These findings have important implications for marketing strategy. Rather than expanding urgency banner usage, efforts should focus on refining the selection criteria for banner placement, understanding what product characteristics make urgency messaging effective, and maintaining the credibility of these promotional tools. The significant difference between ATE and ATT also suggests that current deployment strategies are well-calibrated, successfully identifying products where urgency messaging can drive sales while avoiding potential negative effects on products where such messaging might deter purchases.

Tailored rating insights at product level

Based on our analysis using Double Machine Learning, we estimate that a 1-point increase in a product's average rating causally leads to an increase of approximately 985 units sold, holding all other factors constant. To better inform investment decisions, we scale this estimate: a 0.1-point increase in mean rating is therefore expected to result in an increase of approximately 98.5 additional units sold. Merchants can pursue several strategies to boost their product ratings—such as improving delivery times, enhancing product quality, upgrading packaging, or proactively resolving customer complaints. While these strategies can be effective, they typically involve a cost.

To support merchants in making cost-effective decisions, we provide tailored recommendations based on a simple principle: only invest in improving ratings when the cost of the strategy is less than the revenue expected from the increased units sold. Since the additional revenue depends on the product's unit price, the decision to invest is highly product-specific. For example, a 0.1-point increase in rating may be worth the investment for a \$30 product (leading to ~\$2,955 in additional revenue), but not for a \$2 product (only ~\$197).

In summary, we recommend rating investment strategies only when the cost per product is lower than the expected revenue gain from the projected lift in units sold, ensuring a positive return on investment.

Optimizing AdBoost use based on product-level insights

Based on our causal analysis, using AdBoost can significantly increase product visibility and drive additional sales. From our model, we estimate that activating AdBoost leads to a meaningful lift in units sold (ATE= 928). However, AdBoost incurs a cost of approximately 15% of the total revenue generated during the promotion.

To determine whether AdBoost is worth it, we compare the additional revenue it generates against its cost. For example, consider a product whose revenue increases from $1,600 to $16,448 after AdBoost. This results in a revenue gain of $14,848, but also incurs an AdBoost cost of $2,467.20 (15% of post-boost revenue). After accounting for the cost, the net revenue gain is $12,380.80, yielding a return on investment (ROI) of over 5x, which strongly supports using AdBoost.

However, the benefit is not uniform across all products. For instance, in another case, revenue increases only marginally from $160,000 to $167,424, while the AdBoost cost amounts to $25,113.60, leading to a net loss of $17,689.60—an ROI of −0.70, indicating that AdBoost is not worth it in this scenario.

This variation in performance suggests that AdBoost should not be applied uniformly. Instead, we recommend that merchants use the following rule of thumb: Only use AdBoost if it increases your net revenue by more than ~17.6%, the break-even point based on a 15% ad cost.

Ultimately, the decision depends on both the effectiveness of AdBoost for the specific product and the product's price point and margin. We help merchants assess this tradeoff by calculating the net gain and ROI from actual performance data, ensuring that advertising spend leads to meaningful, profitable growth.

RECOMMENDATIONS

Based on the results of the causal inference analysis and the specific challenges faced by Wish.com, we propose several targeted recommendations aimed at improving platform-level and merchant-level sales performance. These recommendations are designed to be actionable, scalable, and grounded in both statistical evidence and business relevance.

First, Wish should invest in improving product and merchant ratings. The analysis revealed that ratings have a statistically significant and positive causal effect on units sold, as confirmed by Double Machine Learning (DML) and Lasso DML models. Higher-rated products consistently outperform lower-rated ones in terms of sales volume. To capitalize on this, the platform should prioritize quality control mechanisms, incentivize customers to leave reviews, and enhance the visibility of highly rated products in search results and curated listings. These measures not only drive short-term revenue but also build long-term brand credibility and consumer trust.

Second, a more selective and data-driven ad boost strategy should be adopted. Although the effectiveness of ad boosts varied across modeling techniques, Propensity Score Matching (PSM) demonstrated a significant positive Average Treatment Effect (ATE). However, this effect was not universal. Therefore, merchants should be encouraged to use paid promotions only when the projected revenue uplift exceeds the associated ad spend. Wish could consider deploying a decision-support tool or ROI calculator to help merchants determine when boosting is financially advantageous, thereby optimizing marketing budgets.

Third, we recommend that Wish enhance its data collection practices. The dataset used in this study was limited to a single month, which restricted our ability to apply time-based causal methods such as Difference-in-Differences (DiD). Expanding the dataset to include monthly or quarterly snapshots would allow for a more robust understanding of treatment effects over time, control for seasonality, and better assess changes in customer behavior. Furthermore, integrating behavioral data—such as session durations, click-through rates, and cart abandonment metrics—could significantly enrich future causal models.

Fourth, while urgency banners showed mixed results, with only PSM indicating significant sales uplift, this variability suggests the need for a more nuanced strategy. We recommend that urgency banners be tested using A/B experiments across different product categories, price ranges, and customer segments. Rather than applying urgency messages platform-wide, which may lead to user fatigue or skepticism, they should be used sparingly and strategically to preserve their psychological impact.

Finally, we advise Wish to move toward multivariate treatment effect analysis. Current modeling isolates the effect of each treatment independently. However, customer decision-making is influenced by a combination of factors—such as ratings, pricing, and urgency messaging—simultaneously. Employing multivariate causal models or uplift modeling techniques would allow Wish to assess the interaction effects of these levers, thereby enabling more sophisticated, personalized marketing interventions.

REFERENCES

https://medium.com/@shivangi.choudhary_97000/a-comprehensive-guide-on-causal-inference-in-retail-bf641f5a506a

https://roundtable.datascience.salon/data-driven-decision-making-using-causal-inference-for-e-commerce-growth

https://mathco.com/article/causal-inference-and-statistical-tests-for-business-analytics/

https://www.onlinescientificresearch.com/articles/mitigating-bias-in-ecommerce-recommendation-systems-a-causal-inference-approach.pdf

https://thesai.org/Downloads/Volume7No12/Paper_38-Novel_Causality_in_Consumers_Online_Behavior_Ecommerce_Success_Model.pdf