Amazon Sale Analysis

CHEN Haonan ; GAN Yuhang ; GAO Qunqing ;

YANG Yichen ; ZHAO Jiayi

Table of Contents

1. Introduction

1.1 Purpose and Overview

This project is an examination of the sales performance of products on the e-commerce platform of Amazon, aiming at how pricing, promotions, and quality signals together have an impact on both sales and top rating outcomes for electronic items. The subject is of great significance since Amazon is situated in a highly competitive environment where the management has to regularly manipulate prices and promotions to lure in customers without sacrificing profits.

The main goal of the analysis is to generate a data-oriented strategy that connects product categories, ratings, and promotion indicators (like coupons, sponsorship, and badges) to suggested managerial actions. More specifically, we tackle the critical business question of finding the right combination of price and promotion tactics to increase monthly sales and revenues while marketing resources are being used more efficiently.

In order to realize these aims, we initially carry out data preprocessing and exploratory analysis on Amazon product data, followed by the implementation of four supporting analytical methods. Firstly, cluster analysis is used to classify products based on the price-rating-sales patterns and thus, to reveal different product segments. Secondly, linear regression is employed to show the relationship between independent variables and sales volume quantitatively. Thirdly, logistic regression is used to predict the probability of a product getting a high rating which, in turn, aids in the creation of so-called "rating-friendly" promotional strategies. Finally, decision tree analysis translates the most influential drivers into intuitive if–then rules, clarifying which combinations of factors are associated with sales for different product categories. Based on all results, we propose integrated recommendations for Amazon's sales and promotion strategy.

1.2 Data Description and Preprocessing

The dataset is made up of Amazon electronics products aggregated on product level with data about prices, discounts, ratings, reviews, sales volume, product categories, and various promotion-related flags. An overview of the original variables along with their definitions is shown in Figure 1.

Attribute	Type	Description
product_title	Nominal	Product title field
product_rating	Interval	Average customer rating
total_reviews	Interval	Number of customer reviews
sales	Interval	Sales
discounted_price	Interval	Unit discounted price of the producted
original_price	Interval	Unit original price of the producted
is_best_seller	Nominal	Whether the product is best seller
is_sponsored	Nominal	Whether the product is sponsored
has_coupon	Nominal	Whether the product has a coupon
buy_box_availability	Nominal	Whether the buy box is available
delivery_date	Datetime	Date/time of the transaction or record
sustainability_tags	Nominal	Whether the product has a sustainability tag
product_image_url	Nominal	Product image url field
product_page_url	Nominal	Product page url field
data_collected_at	Datetime	Data source
product_category	Nominal	Product category
discount_percentage	Interval	Discount or promotion applied

Figure 1 Variables and Explanations

The first step in our analysis was to perform exploratory descriptive statistics (Figure 2) that revealed the presence of missing values in the raw dataset for certain key variables. Thus, we carry out a systematic preprocessing procedure in Excel and SAS:

Variable Name	Type	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
buy_box_availability	C	1	34.33626	.	.	.	.	.	.
data_collected_at	N	.	0	2.0714E9	2.0722E9	2.0718E9	276729.2	-0.38464	946.9017
delivery_date	N	.	28.07967	23980	24013	23988.28	3.470063	1.169784	1354.522
discount_percenta...	N	.	4.831869	0	85.42	6.547151	12.74472	2.372535	6.091001
discounted_price	N	.	4.831869	2.16	5449	243.2273	473.3515	4.426331	24.39259
has_coupon	C	42	0	.	.	.	.	.	.
is_best_seller	C	12	0	.	.	.	.	.	.
is_sponsored	C	2	0	.	.	.	.	.	.
original_price	N	.	4.831869	2.16	5449	257.6111	496.6335	4.259345	22.02058
product_category	C	15	0	.	.	.	.	.	.
product_image_url	C	1	0	.	.	.	.	.	.
product_rating	N	.	2.399531	1	5	4.399431	0.386997	-1.83725	6.978475
sales	N	.	24.63035	50	100000	1293.665	6318.324	10.72364	139.6856
sustainability_tags	C	16	92.01406	.	.	.	.	.	.
total_reviews	N	.	2.399531	1	865598	3087.106	13030.46	24.34288	1082.194

Figure 2 Descriptive Statistical Analysis on the Original Dataset

• Variable removal: Initially, we discard variables such as delivery_date, data_collected_at and product_image_url since they are not of our main concern in the analysis.

• Handling missing values in core variables: The missing values for price, rating, and total_reviews account for only 7.23% of the total data. We prefer to drop rows that lack price, rating, or total_reviews, since imputation would require making strong assumptions and could alter the distribution.

• Sustainability tags: The data documentation states that only products with sustainability tags are recorded; others are shown as missing. Hence, we use 0 ("no sustainability tag") to fill in the missing values of sustainable_tags variables, and 1 to replace those with non-missing values.

• Missing sales: The amount of missing sales values is rather large. By simply deleting all rows with missing sales, we would be greatly reducing our sample size for clustering and subsequent models. In order to keep these products and at the same time not complicate the structure, we take missing sales to mean zero sales during that period.

• Binary indicators: Several nominal variables are converted into binary indicators to simplify model estimation - best_seller1: 1 = has Best Seller badge, 0 = no badge; is_sponsored1: 1 = sponsored, 0 = organic; has_coupon1: 1 = has coupon, 0 = no coupon; buy_box_availability1: 1 = Buy Box available, 0 = not available; sustainability_tags1: 1 = has sustainability tags, 0 = no tags

In addition, we derive a new binary target for later logistic regression: products with rating higher than 4.5 are coded as 1 (high rating), and products with rating less than or equal to 4.5 are coded as 0 (non-high rating).

After preprocessing, the processed dataset (Figure 3) contains no missing values in the variables used for modelling.

Variable Name	Type	Number of Levels	Percent Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
best_seller1	N	2	0	.	.	.	.	.	.
buy_box_availability	C	2	0	.	.	.	.	.	.
buy_box_availabilit...	N	2	0	.	.	.	.	.	.
discount_percenta...	N	.	0	0	85.42	6.614804	12.81265	2.364834	6.042825
discounted_price	N	.	0	2.16	5449	234.3953	461.367	4.63278	27.01291
has_coupon	C	42	0	.	.	.	.	.	.
has_coupon1	N	2	0	.	.	.	.	.	.
is_best_seller	C	12	0	.	.	.	.	.	.
is_sponsored	C	2	0	.	.	.	.	.	.
is_sponsored1	N	2	0	.	.	.	.	.	.
original_price	N	.	0	2.16	5449	248.9191	485.5515	4.430065	24.00624
product_category	C	15	0	.	.	.	.	.	.
product_rating	N	.	0	1	5	4.410364	0.383202	-1.92039	7.696076
product_rating_ov...	N	2	0	.	.	.	.	.	.
sales	N	.	0	0	100000	1034.25	5688.472	11.9277	173.2289
sustainability_tags	C	17	0	.	.	.	.	.	.
sustainability_tags1	N	2	0	.	.	.	.	.	.
total_reviews	N	.	0	1	865598	2999.142	13142.59	24.89101	1100.608

Figure 3 Descriptive Statistical Analysis on the Processed Dataset

A correlation matrix of the interval variables (Figure 4) shows a high correlation between original price and discounted price. Therefore, we keep discounted price as the effective selling price and drop original price from subsequent analyses.

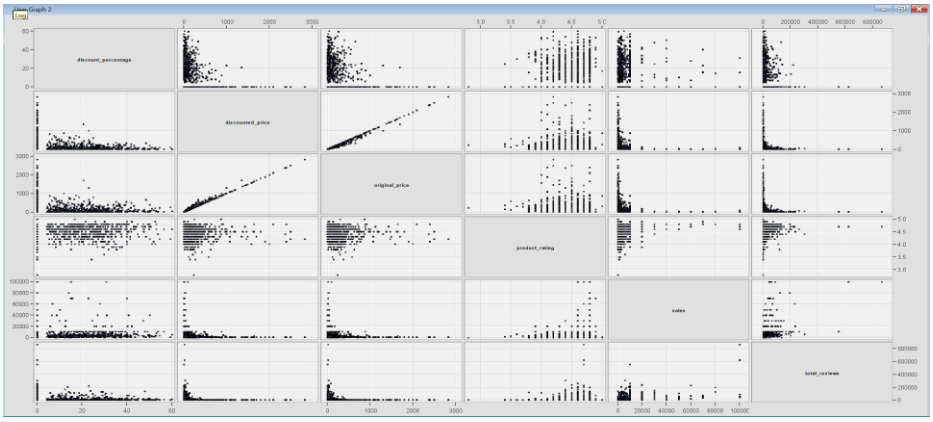


Figure 4 Matrix (discounted_price, original_price, product_rating, sales, total_reviews)

1.3 Research Questions

In this project, we are interested in understanding how Amazon can use pricing and promotional levers to increase product sales and revenue without relying on random or excessive discounting. More specifically, we focus on Amazon electronics products and examine how price, discounts, coupons, badges (such as “Best Seller”), sponsorship, ratings, reviews and product categories jointly

shape monthly sales performance and the likelihood of achieving high ratings. These questions matter because managers on platforms like Amazon have limited marketing budgets and must decide which products to promote, how deeply to discount, and where to invest in quality and reviews in order to maximize ROI and avoid wasting money on ineffective promotions.

To address this business problem, our analysis is guided by the following research questions:

1. Are there distinct types of products that require different promotion strategies?
2. Which factors most strongly drive monthly sales and revenue on Amazon?
3. Which factors interact to influence high ratings and sales outcomes?

These questions link our statistical models directly to a practical goal: designing data-driven promotion and discount strategies that boost Amazon sales and revenue while minimizing unnecessary marketing spend.

2. Cluster Analysis

2.1 Methodology

Cluster analysis is used to categorize products into clear, differentiated market segments and to provide a foundation for targeted pricing, marketing and inventory strategies. We use three interval variables as clustering inputs: discounted_price, product_rating and sales. These capture the core trade-offs between price, perceived quality and demand.

Because our business focus is on how ratings and purchase behaviour differ across product types, we specify four clusters, which offer a good balance between statistical differentiation and managerial interpretability. SAS Enterprise Miner confirms that all three variables have high importance for forming clusters (Figure 5).

Output

Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
discounted_price	discounted_price	3	1	1.00000
product_rating	product_rating	1	2	0.93204
sales	sales	2	1	0.71885

Figure 5 Variable Importance of Cluster Analysis

2.2 Results and Segment Profiles

To compare the commercial contribution of each cluster, we define product-level revenue as the product of discounted price and sales volume. This index allows us to evaluate how much revenue a typical product in each cluster generates.

Table 1 Mean price, rating, sales and revenue by cluster

	discounted_price	product_rating	sales	revenue	Frequency of Cluster	Percentage
Cluster1	25.38710291	4.787024609	19932.88591	506,038.23	894	2.26%
Cluster2	1988.990222	4.149801023	18.36270608	36,523.24	1759	4.44%
Cluster3	16.16938053	4.794690265	89380.53097	1,445,227.82	113	0.29%
Cluster4	156.3260827	4.412486422	352.8581903	55,160.94	36824	93.01%

Table 1 reports the mean statistics by cluster, and the segment statistics plots (Figures 6–8) further illustrate the differences in price, rating, sales and revenue.

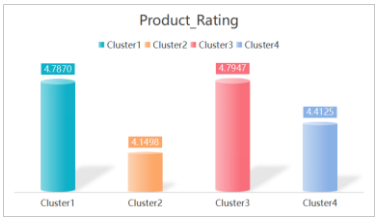
• Cluster 1 – "Value Performers" (894 items): Products with relatively low prices and high ratings, showing relatively high sales and substantial revenue per product.

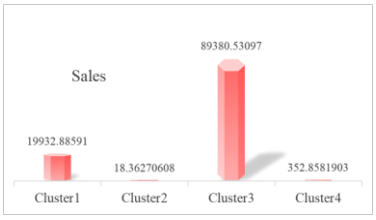
• Cluster 2 – "Premium Under-performers" (1,759 items): Products with the highest prices, lowest ratings and lowest sales, generating the least revenue.

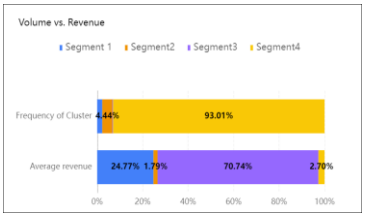
• Cluster 3 – "Star Sellers" (113 items): Products with the lowest prices and highest ratings, achieving the highest sales and the highest revenue per product.

• Cluster 4 – "Mainstream Essentials" (36,824 items): The largest cluster, with moderate ratings, relatively higher prices and relatively low sales and revenue.

The percentage of samples for each group over the total number of samples has been established as shown in the profiles, and even among the various categories of electronic items there are striking differences in performance and the strategic role.







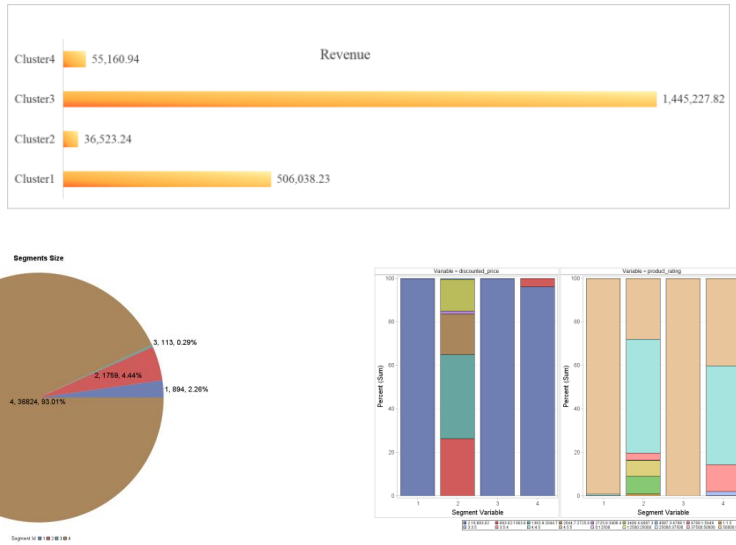


Figure 6 Segment Statistics Diagrams of Different Variables of Cluster

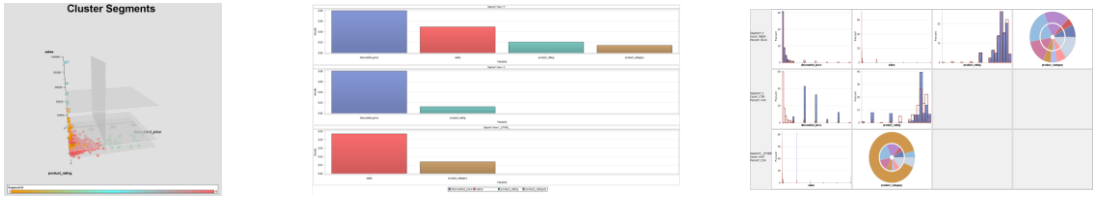


Figure 7 The Results Diagram of Segment Profile

3. Linear Regression Analysis

3.1 Data and Variable Overview

The study utilizes 19,795 product observations derived from the e-commerce platform of Amazon (equally split into training and validation sets). Sales volume is the dependent variable and has the properties of a continuous interval-scaled variable. After going through the data cleaning process (removal of 3,085 records missing price or rating information), the final count of valid records is 19,795.

Table 2 Independent Variables Overview

Variable Type	Variable Name	Nature	Description
Interval	total_reviews	Review Volume	Cumulative product reviews
Interval	discount_percentage	Discount Rate	Product discount percentage (%)
Interval	product_rating	Product Rating	Average rating (0-5 scale)
Binary	best_seller1	Best Seller Badge	0=No, 1=Yes
Binary	is_sponsored1	Sponsored Advertising	0=Organic, 1=Sponsored

Binary	has_coupon1	Coupon Available	0=No, 1=Yes
Binary	sustainability_tags1	Sustainability Label	0=No, 1=Yes
Nominal	product_category	Product Category	15 electronics categories

3.2 Model Specification and Selection

In order to maintain a good balance between the accuracy of prediction and the simplicity of the model, we put side by side three OLS models with sales volume as the dependent variable.

• Full model: Two steps were carried out.

```
The selected model is the model trained in the last step (Step 2). It consists of the following effects:

Intercept  best_seller1  discount_percentage  has_coupon1  is_sponsored1  product_category  product_rating  product_rating_over4_5  sustainability_tags1  total_reviews
```

• Forward stepwise model: The model begins with no predictors (null model) and progressively incorporates variables one at a time with an entry significance level of $\alpha = 0.05$. Nine steps were executed, but the result was the same as that of full model.

```
The selected model is the model trained in the last step (Step 9). It consists of the following effects:

Intercept  best_seller1  discount_percentage  has_coupon1  is_sponsored1  product_category  product_rating  product_rating_over4_5  sustainability_tags1  total_reviews
```

• Backward elimination model: Starts from the full model and removes non-significant variables with retention $\alpha = 0.05$. Discounted_price ($p = 0.6639$, $F = 0.19$) and buy_box_availability1 ($p=0.1113$, $F=2.54$) are dropped. The resulting model has the same Validation VASE as the forward model.

Comparing the three models

Table 3 Comparison of full, forward and backward models

Metric	Full Model	Forward Model	Backward Model
Training R^2	0.2865	0.2864	0.2864
Training RMSE	5,081.49	5,081.58	5,081.58
Validation VASE	21,588,159.63	21,590,797.18	21,590,797.18
Model Complexity	Highest	Moderate	Moderate

Decision Rationale

The three models perform almost identically, with the forward and backward selections achieving virtually the same validation MASE as the full model while using two fewer variables.

To ensure better interpretability and because the forward/backward models are simpler in complexity, we adopt the forward/backward specification as our final model.

Excluded Variables

Step	Effect Removed	DF	Number In	F Value	Pr > F
1	discounted_price	1	9	0.19	0.6639
2	buy_box_availability1	1	8	2.54	0.1113

1. Discounted_Price significance Test Results: F-value=0.19, p-value=0.6639, t-value=(corresponding low t-statistic); Conclusion: Statistically non-significant ($p > 0.05$).

2. Buy_Box_Availability Significance Test Result: F-value=2.54, p-value=0.1113, t-value=1.59; Conclusion: Statistically non-significant ($p > 0.05$).

Availability of Buy Box should in principle increase the ease of purchasing. However, its p-value of 0.1113 does not reach the 0.05 significance level.

Overall, these two variables do not significantly improve the model's predictive power. As a result, both were appropriately removed by the Forward and Backward selection procedures.

3.3 Empirical Results and Statistical Significance

Table 4 Overall Model Fit Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-Statistic	p-value
Model	21	204,918,804,974	9,758,038,332	377.89	<.0001
Error	19,773	510,587,320,922	25,822,451	-	-

Goodness of Fit

R^2=0.2864 (model explains 28.64% of sales variation), Adjusted R^2=0.2856, Model F-statistic=377.89 ($p < .0001$) — extremely significant Prediction Accuracy: Training RMSE=5,081.58, Validation RMSE=4,646.59. The RMSE values for both training and validation are very close to each other, the validation error being a little lower even, because of random sampling. This suggests that strong overfitting is not at all a problem.

Analysis of Parameter Estimates

Category Dummy Variable Coefficients

The category coefficients indicate how far the sales of the categories represented by the coefficients deviate from the baseline (the omitted reference category).

Table 5 Key Category Effects

Category	Coefficient	p-value	Interpretation

Power & Batteries	+6,910.7	<.0001	Highest-performing category, 6,911 units above reference
Printers & Scanners	+56.61	0.8039	No significant difference from reference
Networking	-375.7	0.0995	Slightly below reference
Other Electronics	-104.4	0.2674	Essentially equivalent to reference
Laptops	-99.7	0.2998	Essentially equivalent to reference
Phones	-531.5	<.0001	531 units below reference
Cameras	-550.0	<.0001	550 units below reference
Chargers & Cables	-785.4	<.0001	785 units below reference
Speakers	-723.7	0.0001	724 units below reference
Storage	-1,368.5	<.0001	1,369 units below reference
Headphones	-1,410.4	<.0001	1,410 units below reference
TV & Display	-1,194.0	<.0001	1,194 units below reference
Gaming	-832.2	0.0054	832 units below reference
Smart Home	-454.5	0.1461	455 units below reference

Key Observation

• Power & Batteries' Absolute Leadership: The phenomenal sales potential in this particular category is reflected in the coefficient of +6,910.7 (in comparison with baseline) .

• Premium/Specialty Categories Show Lower Sales: Headphones (-1,410.4), TV & Display (-1,194.0), and Storage (-1,368.5) are the least ranked.

Table 6 Model Coefficient Estimates

	Variable	**p-value**	**Standard Estimate**
1	total_reviews	<.0001	0.1188
2	is_sponsored1	<.0001	-1412.5
4	best_seller1	<.0001	-905.4
5	product_rating	<.0001	822.5
6	discount_percentage	<.0001	21.0339
7	has_coupon1	<.0001	578.6
8	sustainability_tags1	<.0001	375.4
9	Intercept	<.0001	-2216.0

Key Observations

• The highest volume of Reviews is still the major favorable predictor: The total_reviews coefficient is +0.1188 ($p < .0001$), which means that on average one additional review correlates with an increase of about 0.12 units in sales.

• The is_sponsored1 (indicating whether the product is sponsored) has a very strong negative coefficient: This negative coefficient does not imply that sponsored ads are not effective.Because in reality products with lower sales are more likely to purchase sponsored ads, which explains the negative relationship.

• Product rating shows a large and highly significant positive effect on sales. This highlights that rating is a key driver of performance.

Therefore, the next section will dive deeper into strategies for achieving higher product ratings, using logistic regression to analyze the determinants of receiving high ratings.

4. Logistic Regression Analysis

The product rating has been identified as one of the strongest sales drivers, according to the preceding linear regression analysis. However, large discrepancies among different product categories and some marketing tactics, like sponsorship, being linked to decreased sales in the linear model indicate very complicated underlying mechanisms. These findings imply that Amazon will not benefit much from just price reductions or being more visible; rather, it is essential to win and keep top ratings for its sales to perform well on the platform.

Therefore, in this section, we use logistic regression to understand which factors are associated with a product receiving a high rating on Amazon. Our business objective is to identify product and promotion characteristics that significantly increase the probability that an item is highly rated, so that managers can design “rating-friendly” promotion and assortment strategies rather than relying only on price cuts.

4.1 Problem definition and target variable

Customer ratings and reviews are central quality signals in e-commerce. Products with high ratings not only convert better in the short term but also enjoy higher visibility in search results and greater long-term sales momentum. For Amazon and third-party sellers, it is therefore important to understand which controllable levers (such as coupons, discounts, sponsorship, and sustainability

tags) are associated with higher ratings and which practices may harm perceived quality.

To formalize this problem, we define a binary target variable product_rating_over4_5. Products with an average rating strictly higher than 4.5 are coded as 1 ("high rating"), while products with a rating of 4.5 or below are coded as 0 ("non-high rating"). This definition follows the common platform practice of treating 4.5+ stars as a strong quality signal and allows us to use logistic regression to model the probability that a given product will belong to the high-rating group.

Our main research question in this part of the project is: Which product, promotion, and platform attributes significantly affect the probability of achieving a high product rating (> 4.5), and how can these insights guide managerial decisions on promotions and portfolio design?

4.2 Explanatory variables and transformations

We start from the cleaned dataset used in earlier parts of the project, which already includes recoded binary indicators and derived variables.

Name	Use	Report	Role /	Level
is_sponsored1	Default	No	Input	Binary
log_discount_price	Default	No	Input	Interval
has_coupon1	Default	No	Input	Binary
product_category	Default	No	Input	Nominal
sustainability_tags1	Default	No	Input	Binary
log_total_reviews	Default	No	Input	Interval
discount_percentage_DE	Default	No	Input	Interval
buy_box_availability1	Default	No	Input	Binary
best_seller1	Default	No	Input	Binary
product_title	Default	No	Rejected	Nominal
product_rating	Default	No	Rejected	Interval
sustainability_tags	Default	No	Rejected	Nominal
sales	Default	No	Rejected	Interval
is_sponsored	Default	No	Rejected	Nominal
has_coupon	Default	No	Rejected	Nominal
is_best_seller	Default	No	Rejected	Nominal
buy_box_availability	Default	No	Rejected	Nominal
original_price	Default	No	Rejected	Interval
product_rating_over4_5	Yes	No	Target	Binary

Figure 8 The key explanatory variables for the logistic regression analysis

Because several of these variables are highly skewed, especially discounted_price and total_reviews, we apply logarithmic transformations to stabilize variance and make the relationship between the predictors and the log-odds of high rating closer to linear. Specifically, we construct the following transformed variables: log_discount_price = log (1 + discounted_price); log_total_reviews = log (1 + total_reviews); discount_percentage_DE = discount_percentage / 10

These transformations serve three purposes. First, they alleviate the impact of extreme values, such as very expensive products or items with extremely high review counts. Second, they improve the linearity assumption in logistic regression by compressing wide ranges into a more stable scale. Third, they reduce potential collinearity and improve interpretability of the estimated coefficients in terms of proportional rather than absolute changes.

4.3 Model specification and comparison

To evaluate the effect of transformations and variable selection, we estimate four different logistic regression models:

• Reg1: Full model with transformed variables. Includes all candidate predictors, using log_discount_price, log_total_reviews, and discount_percentage_DE instead of their raw counterparts.

• Reg2: Backward selection with transformed variables. Starts from the full transformed model and iteratively removes insignificant variables based on statistical tests until only significant predictors remain.

• Reg3: Forward selection with transformed variables. Starts from an empty model and adds predictors one by one, choosing the variable that most improves model fit at each step.

• Reg4: Baseline model without transformations. Uses the original untransformed variables to provide a benchmark for evaluating the benefit of using transformed predictors.

Fit Statistics

Model Selection based on Valid: Misclassification Rate (_VMISC_)

Selected Model	Model Node	Model Description	Valid: Misclassification Rate	Train: Average Squared Error	Train: Misclassification Rate	Valid: Average Squared Error
Y	Reg	Regression	0.30089	0.20310	0.30537	0.20167
	Reg2	Regression (2)	0.30157	0.20315	0.30605	0.20179
	Reg3	Regression (3)	0.30157	0.20315	0.30605	0.20179
	Reg4	Regression (4)	0.33474	0.20667	0.33521	0.20502

Figure 9 Logistic regression model comparison based on validation misclassification rate

We use the training–validation framework in SAS Enterprise Miner to avoid overfitting. Model performance is evaluated using: Misclassification rate on the training and validation samples. Average Squared Error (ASE), which penalizes large deviations between predicted probabilities and the actual binary outcomes.

Among all configurations, the models with transformed variables (Reg1–Reg3) perform better than the baseline model. The best model reaches a minimum validation misclassification rate of 0.3009, a training misclassification rate of 0.3054, and a minimum validation ASE of 0.2017.

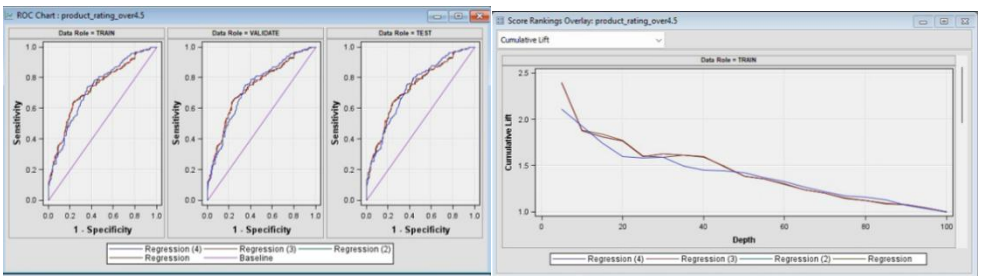


Figure 10 ROC curves and cumulative lift for the logistic regression models

Overall, the three models show very similar performance: their misclassification rates and ASE are close. However, Reg1 achieves the lowest validation misclassification rate and ASE, and a slightly better AIC. We therefore select Reg1 as the final logistic regression model, and use Reg 2 & 3 only as robustness checks to show that our conclusions are not sensitive to the specific stepwise procedure.

4.4 Key findings and coefficient interpretation

Logistic regression analysis has pointed out various strong predictors of high ratings (more than 4.5). The category of SMART_HOME appears to have the most substantial negative impact, to the extent that it is practically impossible for smart home products to receive a rating of 4.5 or more even when other factors are taken into account. This might be due to customer dissatisfaction with the product installation and usage processes, as well as higher expectations from the customers.

On the positive side, three of the variables raise the probability of receiving a high rating significantly:

best_seller1 (Best Seller badge): The products having this badge are likely to be rated very well, which is in line with a feedback loop where the items being popular and well-rated get more visibility and sell more.

has_coupon1 (Coupon availability): Coupons are linked to better ratings, possibly because the customers think they are getting a good deal or because the retailers are selectively discounting products that they are confident in.

log_total_reviews (Number of reviews): The more reviews a product gets, the more likely it is to maintain a high average rating, which is true both for the product quality and for the survivorship (weak products exit before accumulating many reviews).

At the same time, there are also some factors that have negative impacts:

is_sponsored1 (Sponsored listing): Sponsored products are less likely to achieve a 4.5-star rating, implying that sponsorship is mostly used to support weaker or less-established products.

log_discount_price (Discounted price): Premium-priced items are less likely to be given very high ratings, probably because the expectations in the premium segments are stricter.

buy_box_availability1 (Buy Box availability): Negative coefficient indicates that the products winning the Buy Box might be more price-driven or mass-market, where quality and service do not always reach the top tier.

In sum, high ratings indicate a mixture of category effects, quality signs (badges, review volume), and promotion mechanisms (coupons, sponsorship, pricing) that can drive rating.

5. Decision Tree Analysis

5.1 Objective and Model Setup

In the process of creating a regression-type decision tree, we were able to convert the fundamental sales drivers into understandable decision rules applicable to each segment. The target variable is monthly sales for each product.

As input variables we use the significant drivers from the final linear regression model: total_reviews, product_category, is_sponsored1, best_seller1, discount_percentage, product_rating, has_coupon1, and sustainability_tags1.

The splitting criterion is the reduction in average squared error (ASE), so each split aims to minimize the prediction error for sales. The data set is partitioned into the same way as the previous models, and the tree size is chosen based on the validation ASE to avoid overfitting.

Table 7: Decision tree fit statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
sales	sales	ASE	Average square Error	3232399	4043520
sales	sales	RASE	Root Average Square Error	1797.887	2010.85

In the final tree, the training ASE is 3,232,399 and the validation ASE is 4,043,520, with corresponding root ASE (RASE) values of 1,798 and 2,011. Together with the leaf-level sales chart—where total and average sales for each leaf are very similar between the training and validation data, and high-sales leaves keep their relative ranking—this indicates that the segmentation is stable and not strongly overfitted, although the overall predictive accuracy remains moderate.

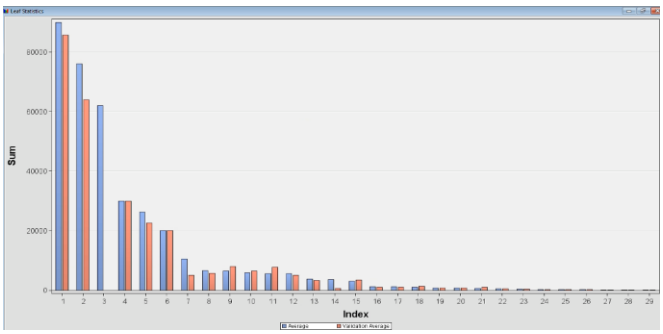


Figure 11 Leaf-level sales for training vs validation data

Compared with the linear and logistic regression models, the decision tree does not aim to maximize

prediction accuracy. Instead, its main value is to link the main drivers together into transparent "if–then" rules for different types of products. Therefore, in the final recommendation section we use the decision-tree segments to propose differentiated strategies for different products rather than to generate precise point forecasts.

5.2 Tree Structure

1. At the top level, the first split is on total_reviews, which divides products into high-review, mid-review and low-review bands. This confirms the earlier regression finding that review volume is the single most powerful driver of sales.

2. Within each review band, the second level mainly splits by product_category. Categories such as Power & Batteries, Printers & Scanners, Phones, Laptops and Other Electronics show clearly different average sales levels even after controlling for review volume.

3. At deeper levels, the tree uses discount_percentage, product_rating, is_sponsored1, best_seller1, has_coupon1 and sustainability_tags1 to refine the predictions inside each segment. These splits show how promotion and quality signals interact with reviews and category to drive sales.

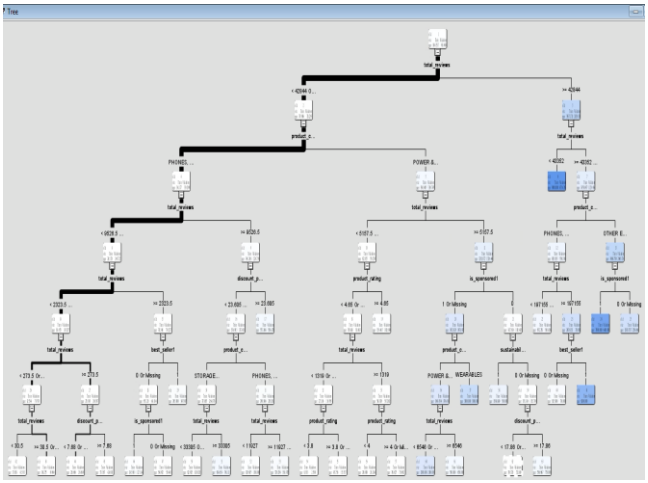


Figure 12 Regression tree for Amazon product sales

5.3 Segmentation Results

To illustrate how the tree works in practice, we summarize three representative decision rules derived from the node rules:

1. "Blockbuster electronics" segment: If a product has very high review volume (around 40,000 or more reviews) and belongs to Power & Batteries or Other Electronics, then the tree predicts extremely high monthly sales (around 70,000–90,000 units).

This pattern suggests that once an electronic product reaches a very large review base in these

categories, further promotion adds little compared to the strong social proof from reviews.

2. “High-potential promoted items” segment: For products with medium review volume (roughly 2,000–40,000 reviews) in categories such as Phones, Laptops, Storage and Cameras, nodes with higher discount_percentage and the presence of coupons have much higher predicted sales than similar nodes without promotion.

This rule highlights that for mid-visibility products, aggressive discounting can effectively convert traffic into sales.

3. “Low-rating long-tail” segment: For products with low review volume (below about 1,300 reviews) and low product_rating (below about 3.8) in Wearables, the tree predicts almost zero sales even if discounts are offered.

This implies that for some long-tail items with weak ratings, price promotions alone cannot compensate for perceived quality problems.

Model Finding

Review volume is the strongest driver of sales; product category shapes the baseline potential; and promotion and quality signals mainly work by lifting sales within each review-category segment.

6. Managerial Recommendations

6.1 Binding strategy

The linear regression results show that Power & Batteries is the strongest-performing category. Therefore, we recommend bundling or cross-promoting these products with weaker categories—such as headphones, storage devices, and TV/display products—to leverage the high conversion power of the leading category. This strategy can increase visibility and sales of underperforming segments and ultimately improve the overall conversion rate and average order value of the electronics portfolio.

6.2 Advertising & Badging Strategy

The linear Regression and logistic regression results suggest that sponsored products tend to have lower ratings and weaker sales, indicating that ads are often used to “rescue” underperforming items. Meanwhile, products with the Best Seller badge are far more likely to achieve ratings above 4.5. We recommend setting minimum quality or rating thresholds for sponsored ads and reallocating advertising resources toward high-potential products. At the same time, Amazon should build a portfolio of Best Seller badge, giving them priority in search ranking, recommendations, and

promotional exposure.

6.3 differentiated management strategy

Both the decision-tree and clustering results indicate that products should follow a differentiated management strategy: allocate stable inventory to high-review blockbusters, use targeted discounts for mid-visibility items, and avoid further promotion for low-rating long-tail products. At the portfolio level, prioritize Star Sellers, adjust or reposition Premium Under-performers, leverage Value Performers for gradual upselling, and manage Mainstream Essentials with cost efficiency. Overall, resources should be allocated based on each segment's performance potential and review quality.

7. Conclusion and Limitations

7.1 Conclusion

Using four complementary methods on Amazon electronics data, we find that sales are driven jointly by three groups of factors: review volume and rating quality, category-specific demand potential, and a small set of promotion and quality signals such as discounts, coupons, badges and sponsorship. Cluster analysis reveals that products fall into distinct strategic segments (such as star sellers and premium under-performers), so a single promotion rule cannot work for all items. Linear regression quantifies how reviews, ratings, price discounts and category membership affect sales volume; logistic regression explains which factors help or hurt the chance of achieving very high ratings. The decision tree then links these drivers into intuitive if–then rules at the segment level, showing how reviews, categories and promotions combine to generate higher or lower monthly sales. Together, these results support a coherent strategy: invest in reviews and ratings, tailor promotions by segment and category, and deploy advertising selectively to amplify already strong products.

7.2 Limitations and directions for future work

This project also has several limitations. First, it uses a single cleaned snapshot of product-level data, so we cannot capture seasonality, product life-cycle dynamics or the long-term impact of promotion decisions. Second, our models are intentionally simple and interpretable; they provide managerial insight but only moderate predictive accuracy for individual products. Third, we treat products independently and do not explicitly model competition between similar items or changes in search ranking and recommendation algorithms.

Future work could extend the analysis by using time-series or panel data, incorporating textual

information from reviews and competitor prices, and experimenting with more advanced ensemble or causal models. Even with more complex techniques, however, it will remain important to translate the results back into transparent rules and segment-level strategies that managers can easily understand and implement.

References:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning: With Applications in R (2nd ed.). Springer.

Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Wadsworth.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. Journal of Marketing Research, 43(3), 345–354.

Filieri, R., McLeay, F., Tsui, B., & Lin, Z. (2018). Consumer perceptions of information helpfulness and determinants of purchase intention in online consumer reviews of services. Information & Management, 55(8), 993–1004.

Li, X., Wu, C., & Mai, F. (2019). The effect of online reviews on product sales: A joint sentiment–topic analysis. Information & Management, 56(2), 172–184.