Understanding the Relationship Between Consumer Purchasing Behavior and Sales

Group 9

Haritha Rajendra Rao Savithri (dal136929)

Naveena Paleti(nxp230055)

Sai Shripad Achutuni (Axs240134)

Salwa Niaz Preet (sxp240086)

Contents

Introduction

The rapid expansion of e-commerce has revolutionized consumer shopping habits, making it a cornerstone of the global economy. According to a recent report by Boston Consulting Group (BCG), e-commerce is projected to grow by 39% and reach an estimated $8 trillion by 2027. In this landscape, understanding customer behavior is more critical than ever, as businesses strive to stay competitive and maximize revenue. This report leverages advanced machine learning algorithms—such as linear regression, neural networks, CART, and random forests—to analyze the impact of customer behavior on sales performance on online platforms. By uncovering actionable insights, this analysis serves as a valuable resource for e-commerce business owners, marketers, and decision-makers aiming to optimize strategies, enhance customer engagement, and capitalize on the immense growth potential of digital commerce.

Problem Statement

Understanding the Relationship Between Product Ratings, Reviews, and Consumer Purchasing Behavior

This study aims to analyze how product ratings and reviews influence purchasing behavior, specifically in terms of sales and revenue. It is important to note that this study focuses on consumer purchasing behavior in the context of online retail platforms. By examining the correlation between these factors, the goal is to identify the extent to which customer feedback affects consumer decisions in this specific context.

Data

The data set we will be using has been taken from the software, Helium 10. This software is an Amazon seller software that gives information on keywords, their search volumes, search trend, clicks and conversion rates, trending product niches, product ratings, reviews, sales, revenue, competing products and more. Our data set focuses on the American Amazon marketplace. Our data set consists of information on 30 variables for 8411 products for the month of August, 2024.

The data set has the following variables:

1. Keyword Phrase: The product/ the phrase customers type in to look for a product.
2. Search Volume: The number of times the phrase has been searched for on Amazon in a particular tie period.
3. Search Volume Trend (%): Change in average search volume of the keyword phrase in one month compared to the average search volume of the preceding one month.
4. Search Frequency Rank: A metric that ranks keyword phrases based on how frequently customers search for them in a month. Shows popularity of the keyword phrase compared to the others during a time period. Higher the rank (smaller the number) more the popularity. (500th is more popular than 1000th).
5. Search Frequency Rank Trend (%): Change in search frequency rank in one month compared to the search frequency rank of the preceding one month.
6. Top 3 ASINs Total Click Share: Proportion of clicks the top 3 products with the same keyword phrase receive relative to other products having the same keyword phrase.
    - ASINs (Amazon Standard Identification Numbers)
7. Top 3 ASINs Total Click Share Trend: Change in total click share of top 3 ASINs in one month compared to the total click share of the preceding one month.

8. Top 3 ASINs Total Conv. Share: The proportion of total conversions (sales) top 3 ASINs with the same keyword receive relative to all other products having the same keyword phrase.
9. Top 3 ASINs Total Conv. Share Trend: Change in total conversion share of top 3 ASINs in one month compared to the total conversion share of the preceding one month.
10. Top 1 ASIN: The unique ID of the product that is top-performing.
11. Top 1 ASIN Click Share: Proportion of clicks the top product with a particular keyword phrase receives relative to other products having the same keyword phrase.
12. Top 1 ASIN Click Share Trend: Change in total click share of top product in one month compared to the total click share of the preceding one month.
13. Top 1 ASIN Conv. Share: The proportion of total conversions (sales) top product with a particular keyword receives relative to all other products having the same keyword phrase.
14. Top 1 ASIN Conv. Share Trend: Change in total conversion share of top product with a particular keyword phrase in one month compared to the total conversion share of the preceding one month.
15. Top 2 ASIN
16. Top 2 ASIN Click Share
17. Top 2 ASIN Click Share Trend
18. Top 2 ASIN Conv. Share
19. Top 2 ASIN Conv. Share Trend
20. Top 3 ASIN
21. Top 3 ASIN Click Share
22. Top 3 ASIN Click Share Trend
23. Top 3 ASIN Conv. Share
24. Top 3 ASIN Conv. Share Trend
25. Competing Products: Total number of products with the same keyword sold by competitors.
26. Top 3 Clicked ASINs Monthly Average Age: The average time, in months, the top 3 ASINs have been made available on Amazon. It shows how long they have been on Amazon.
27. Top 3 ASINs Total Monthly Sales: The total number of units of the top 3 ASINs sold on a monthly basis.
28. Top 3 ASINs Total Monthly Revenue: The total monthly revenue of the top 3 ASINs.
29. Top 3 ASINs Total Review Count: The total number of reviews of the top 3 ASINs.
30. Top 3 ASIN Total Average Rating: The total average rating of the top 3 ASINs.

Descriptive Statistics

The following tables show the minimum and maximum values along with the three quartiles and mean of each quantitative variable in our data set. Overall, the data shows a broad range of values across all variables, indicating significant variation in product performance and search trends. We have separated the table for variables that have missing values to evaluate what variables are missing and what percentage of the entire data set do the missing values constitute. This formed the basis of data-cleaning and pre-processing. Given that some variables have more than 15% data missing, it was not logical to remove all such observations. We had to go deeper to understand the importance of the variables, assess the relationships between variables and recognize whether the pattern was random or systematic.

Variable	Min	Q1	Median	Mean	Q3	Max
Search Volume	36,119	44,726	61,896	102,461	98,286	4,066,797
Search Frequency Rank	1	2,501	5,000	5,589	7,500	556,591
Top 3 ASINs Total Click Share (%)	2.3	20.5	32.9	38.75	51.52	100

Click Share Trend (%)	-73.8	-1.6	0.9	1.762	3.6	99.9
Conversion Share (%)	0	5.9	17.2	21.78	32.5	100
Conversion Share Trend (%)	-100	-0.7	0.4	1.136	3.1	100
Top 1 ASIN Click Share (%)	0.9	9	15.3	22.11	26.9	100
Top 1 ASIN Conv. Share (%)	0	1.3	7.3	11.47	15.9	99.2
Top 2 ASIN Click Share (%)	0	5.6	8.7	10.06	12.9	47.6
Top 2 ASIN Conv. Share (%)	0	1.1	4.5	6.096	9	73.7
Top 3 ASIN Click Share (%)	0	4.1	6	6.589	8.5	32.6
Top 3 ASIN Conv. Share (%)	0	0.9	3.2	4.221	5.9	100
Competing Products	1	583	1,000	9,792	7,000	400,000

Variable	**Min**	**Q1**	**Median**	**Mean**	**Q2**	**Max**	**NA's**
Top 3 Clicked ASINs Monthly Avg. Age	1	27	50	59.78	83	262	3475 (35%)
Top 3 ASINs Total Monthly Sales	0	9,672	30,403	52,354	71,330	456,345	1570 (16%)
Top 3 ASINs Total Monthly Revenue	0	215,312	598,729	1,225,966	1,426,428	27,357,655	1468 (15%)
Top 3 ASINs Total Review Count	0	7,132	24,499	54,438	65,013	1,766,871	1320 (13%)
Top 3 ASIN Total Average Rating	0.3	4.3	4.5	4.249	4.6	5	1316 (13%)
Search Frequency Rank Trend	-18,975	-11	8	-11.1	25	100	64 (0.64%)
Search Volume Trend	-94.3	-10.7	6.5	133.9	27.5	72,507	68 (0.68%)

Data Cleaning

Our dataset includes various products, each with own unique characteristics. These products can have a wide range of values for each variable (e.g., conversion rate, monthly age, sales, etc.). Each observation (product) is different and these differences are crucial for understanding trends and making predictions. For instance, a product might have a high conversion rate but a missing value for monthly age. The product can be an old FMCG (Fast-Moving Consumer Goods) product with consistent sales over time, which is likely to have a high value for its monthly age. However, the product is also likely to be a new, trending item that has seen a surge in sales for the current month.

These two scenarios are fundamentally different, and simply imputing the missing monthly age with a mean value would distort the data, as it ignores the distinct characteristics of each product. A mean imputation would likely represent neither of the products accurately and could lead to incorrect assumptions. Additionally, simply removing the rows would have meant deleting 3475 entries that could have valuable insights to give. Therefore, omitting was also not a viable option.

We observed that missing data was not random. Thus, data cleaning was an iterative task. We identified patterns, and formed new rules, tailoring to the data's nuances.

Step 1

We observed that when all conversion rates were zero, there were missing values for sales and revenue. Thus, we made a rule that if all conversion rates are zero, then sales and revenue should also be zero because no conversion means no sales and revenue. The table shows the difference in missing values before and after we applied the rule.

Variable	Missing Values	
	Before	**After**
Sales	1570	565
Revenue	1468	462

Step 2

The variables monthly age, review count, rating, sales and revenue are important variables for analysis. We understood we could not assume data points for all these variable when they are all missing at once for a given product. Here, omitting the entries made more sense.

From this point onward, we systematically observed the missing values for these five variables, first identifying instances where all five were missing, then where four were missing, followed by three, and so on, until we ensured that no more than one of these variables was missing at a time. By the end of this step, our data set was reduced to 8412 entries.

During the process, we made the rule that if age, rating, and reviews are still missing, the rows should be removed. Upon reviewing the data, it was observed that many keywords consisted of common words like "from," "to," and "if". These entries were deemed irrelevant for the analysis and we understood it was better to exclude them to maintain data quality and relevance.

Variable	Missing Values	
	Before	**After**
Sales	565	3
Revenue	462	4
Monthly Age	3475	2171
Review Count	1320	18
Rating	1316	14

Step 3

The missing values for monthly age, which amounted to 2,171, were predicted using linear regression. We preferred this approach over mean imputation because linear regression allows for more accurate predictions based on existing data patterns and by leveraging the correlations with other variables. To maintain the integrity of the dataset while also reducing the bias that could arise from assuming a uniform value for missing data, this was the best next step for data cleaning.

Step 4

For the remaining missing data, we chose to go for the clustering technique because it identifies patterns and relationships within the data by grouping similar observations, giving relevant predictions.

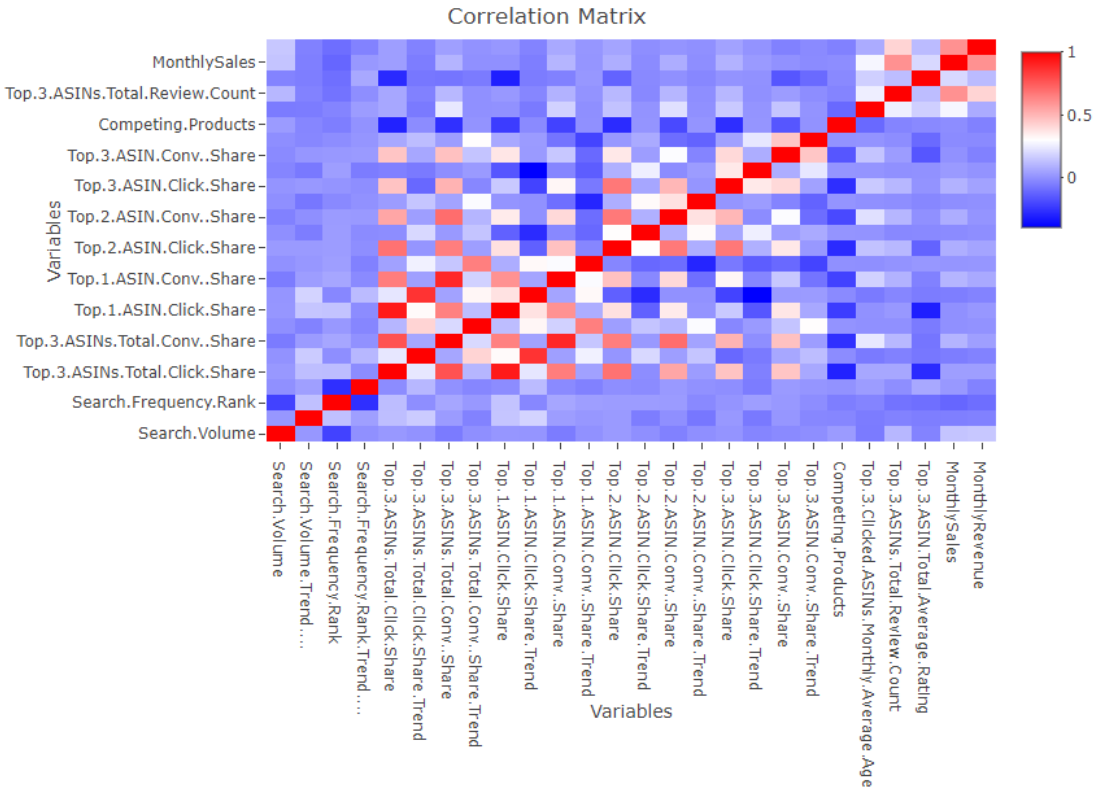
Step 5

In the end, we removed unwanted columns and got our final dataset of 8412 observations, with no missing values.

Data Visualization

To understand the relationships between variables and to check distributions of each variable, we visualized the correlation matrix as a heatmap, plotted histograms, and also observed scatter plots to assess the relationship between our target variable, sales, and all other variables.

Correlation Matrix – Heatmap



The heatmap shows us that most variables have a weak negative relatonsip between them. Our target variable is Monthly Sales. Revenue and review count are positively related to sales and seem to have a strong and moderate impact, respectively. When sales increase, revenue should increase so the relationship is natural and not so much of a cause and effect. However, more review count seems to be a good predictor for sales. There seems to be a weaker correlation between competing products and sales, suggesting that while the number of competing products might have some effect, it is not as influential as other factors such as review count. Similarly, search frequency rank seems to be moderately and negatively correlated with sales, which makes sense. A product with a lower search rank may not be very popular or trending, thus its sales are expected to be low.

Key Takeaways for Sales:

- Strong positive relationship with revenue and review count
- Moderate negative relationship with search frequency rank
- Rest seem to have weak positive relationships with sales

Side-by-side Histograms

To understand the distributions of all variables, side-by-side histograms are immensely useful. Not only can we identify normal or skewed variables, we can also set the basis for data pre-processing. Additionally, we can see the range of variables, the central tendency and also identify gaps.

Positively Skewed Variable(s)

Competing products, monthly revenue and sales, search frequency rank and trend, search volume rank and trend, all click shares and their conversion shares, and finally the average monthly age of listings. These are likely to be scaled and transformed before conducting machine learning algorithms that are sensitive to scale changes, like regression.

Negatively Skewed Variable(s)

Total average rating and search frequency trend.

Symmetric

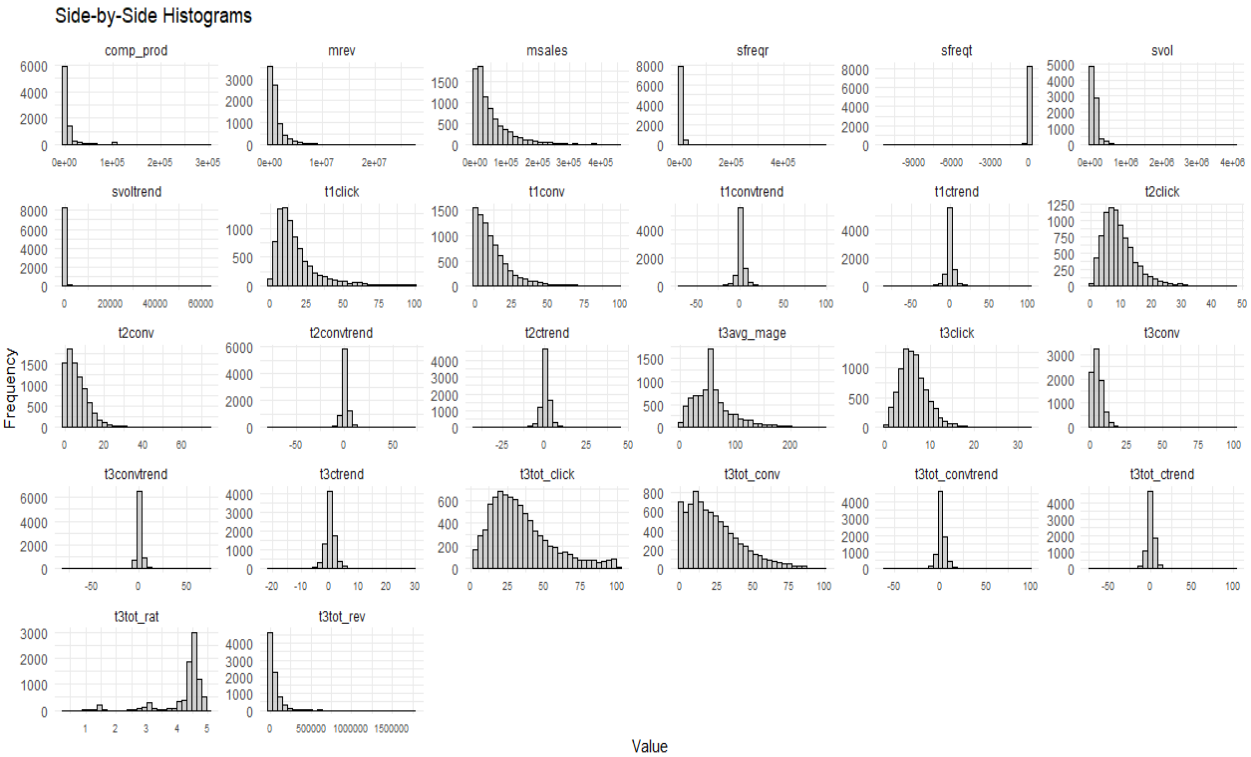
All click and conversion trends

Outliers

Several variables, especially those with positive skew, show extreme values. This is particularly true for competing products, revenue, search frequency rank and trend, review count, and search volume.

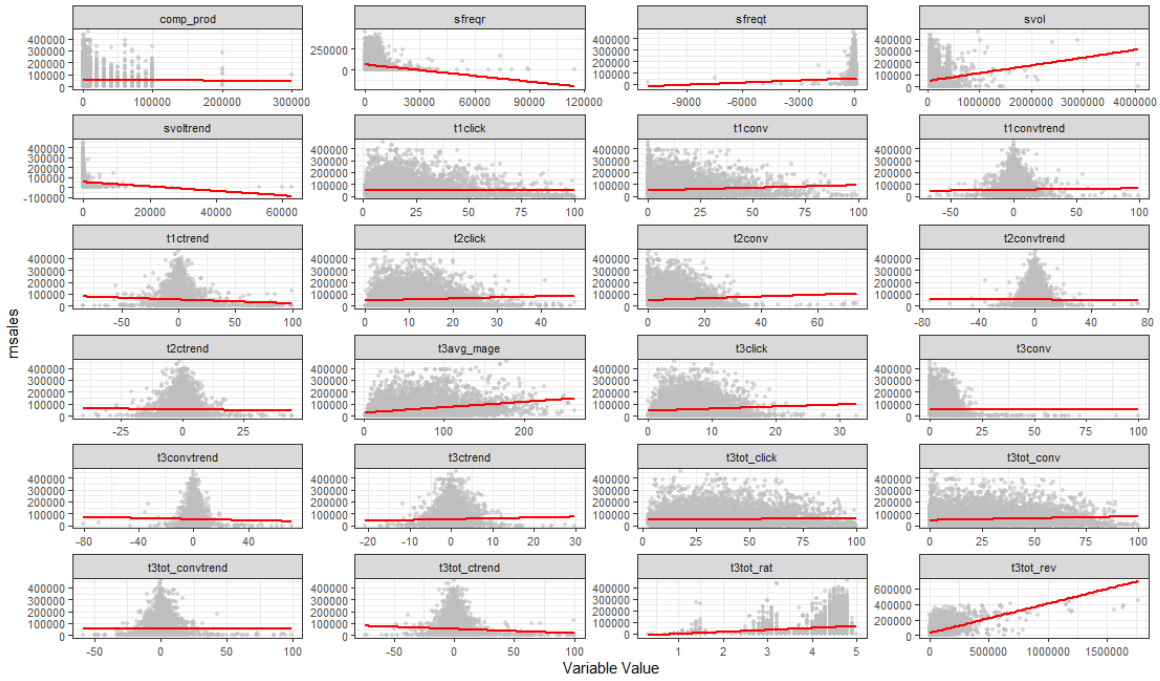
Concentrated Data

The trend variables show some variance but have concentrated high-frequency peaks mostly at 0 values, indicating that the products were equally popular in the previous month.



Side-by-side Scatter Plots

- Variables such as search volume, monthly age of listings, rating and review count show a positive correlation with sales, meaning increasing values in these variables are likely to drive higher sales.
- Search frequency exhibits a negative correlation with sales, suggesting that as the rank rises, sales tend to decline.
- Some variables, like conversion trends do not seem to have a significant relationship with sales.

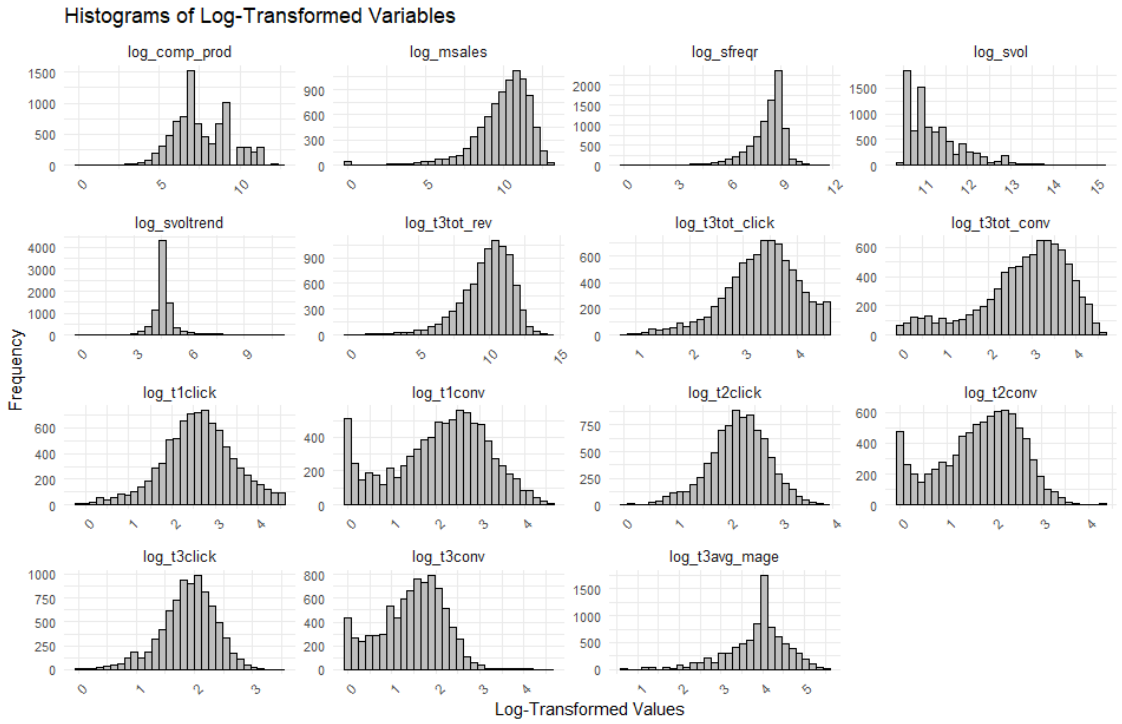


Data Pre-processing

Log-Transformation

As shown and discussed above, some variables were positively skewed and had a wide range of values. Using these variables without transforming would likely give unreliable estimates especially in regression outputs. Therefore, we chose to log-transform some variables. The resulting distributions are shown in the histogram below.

While some variables now show a roughly symmetric distribution (competing products, individual listing clicks and conversions, search volume trend, age), sales, search frequency rank, review count and total clicks and conversion have become negatively skewed. In order to understand the change in skewness better, we also found out exact values of the before and after log transformation skewness metric, given in the table below. Log transformation reduced skewness for all variables, bringing their distributions closer to symmetry.



Variable	Before Log	After Log	Variable	Before Log	After Log
msales	2.14001	-1.79258	**t3click**	1.13345	-0.54348
sfreqr	6.18412	-1.67588	**t2conv**	2.3736	-0.36359
t3tot_rev	6.58447	-1.10613	**t1conv**	2.15772	-0.35362
t3avg_mage	1.37552	-1.03365	**t2click**	1.55464	-0.32352
t3tot_conv	1.04177	-0.88591	**t1click**	2.15439	-0.1803
t3tot_click	1.03464	-0.60336	**t3conv**	5.70782	-0.1569
comp_prod	4.06377	0.29715	**svol**	11.0606	1.39758
svoltrend	33.7143	2.61844			

Data Partitioning

For the analysis, our final data has been partitioned into training and validation sets so that we can see how well our models can fit new data. A random 60% of data is selected as the training set and remaining 40% as validation set.

Each of the models will be trained first and then run on validation data to check performance of the model. Depending on the model, we will give error analysis and precision and accuracy metrics to decide which model works best for our data.

Data Modeling

Linear Regression

This model aims to find which qualitative variables are good predictors for sales. We first run a simple model, selecting a small subset of predictors that we think should have an impact on sales. Then, we conduct stepwise regression to find the best predictors of sales. Once the best model is chosen, we will interpret the results of that regression output.

Simple Model

```
Call:
lm(formula = log_msales ~ log_svol + log_svoltrend + sfreqr +
    log_t1click + t1ctrend + t1convtrend + log_t2click + t2ctrend +
    t2convtrend + log_t3click + t3ctrend + t3convtrend + log_comp_prod +
    t3avg_mage + log_t3tot_rev + t3tot_rat, data = train.df)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3252 -0.2846  0.1153  0.4569  3.0460

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.0034221  0.0113800  -0.301 0.763644
log_svol       0.0343174  0.0149930   2.289 0.022126 *
log_svoltrend  0.0404177  0.0136066   2.970 0.002988 **
sfreqr        -0.0286605  0.0163146  -1.757 0.079023 .
log_t1click    0.0588818  0.0193788   3.038 0.002390 **
t1ctrend      -0.0846542  0.0158719  -5.334 1.01e-07 ***
t1convtrend    0.0545724  0.0136020   4.012 6.11e-05 ***
log_t2click   -0.0303896  0.0271356  -1.120 0.262804
t2ctrend      -0.0034590  0.0146372  -0.236 0.813196
t2convtrend    0.0506677  0.0130479   3.883 0.000104 ***
log_t3click    0.0063877  0.0224514   0.285 0.776030
t3ctrend      -0.0045588  0.0142225  -0.321 0.748577
t3convtrend    0.0252470  0.0126977   1.988 0.046831 *
log_comp_prod  0.0780418  0.0147877   5.277 1.36e-07 ***
t3avg_mage    -0.0001312  0.0126516  -0.010 0.991728
log_t3tot_rev  0.5502673  0.0139916  39.328  < 2e-16 ***
t3tot_rat      0.0931542  0.0127100   7.329 2.68e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.808 on 5029 degrees of freedom
Multiple R-squared:  0.3703,    Adjusted R-squared:  0.3683
F-statistic: 184.8 on 16 and 5029 DF,  p-value: < 2.2e-16
```

Full model

We first run the full model, choosing all variables as predictors. Here, we have not taken the total cumulative click share, trend, conversion share and trend of the 3 top-performing stores. Instead, we have taken them individually for the three stores in order to avoid collinearity.

```
Call:
lm(formula = log_msales ~ log_svol + log_svoltrend + sfreqr +
    log_t1click + t1ctrend + log_t1conv + t1convtrend + log_t2click +
    t2ctrend + log_t2conv + t2convtrend + log_t3click + t3ctrend +
    log_t3conv + t3convtrend + log_comp_prod + t3avg_mage + log_t3tot_rev +
    t3tot_rat, data = train.df)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2066 -0.2830  0.1119  0.4526  3.2533

Coefficients:
               Estimate Std. Error t value             Pr(>|t|)
(Intercept)   -0.006355   0.011283  -0.563             0.573303
log_svol       0.056085   0.015071   3.721             0.000200 ***
log_svoltrend  0.029734   0.013528   2.198             0.027999 *
sfreqr        -0.009997   0.016405  -0.609             0.542320
log_t1click   -0.012908   0.021489  -0.601             0.548071
t1ctrend      -0.057763   0.016254  -3.554             0.000383 ***
log_t1conv     0.153853   0.019605   7.847  0.00000000000000515 ***
t1convtrend    0.012569   0.014469   0.869             0.385078
log_t2click   -0.064130   0.028504  -2.250             0.024499 *
t2ctrend       0.005806   0.014821   0.392             0.695283
log_t2conv     0.047404   0.021524   2.202             0.027688 *
t2convtrend    0.034268   0.014384   2.382             0.017237 *
log_t3click   -0.011729   0.023941  -0.490             0.624201
t3ctrend      -0.003386   0.014283  -0.237             0.812620
log_t3conv    -0.042356   0.018135  -2.336             0.019553 *
t3convtrend    0.045153   0.014161   3.189             0.001439 **
log_comp_prod  0.080507   0.014676   5.486  0.00000004322113330 ***
t3avg_mage    -0.010715   0.012665  -0.846             0.397563
log_t3tot_rev  0.537725   0.013939  38.578 < 0.00000000000000002 ***
t3tot_rat      0.066279   0.013007   5.096  0.00000036047092286 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8009 on 5026 degrees of freedom
Multiple R-squared:  0.3818,    Adjusted R-squared:  0.3794
F-statistic: 163.3 on 19 and 5026 DF,  p-value: < 0.00000000000000022
```

Forward Step Model

```
Call:
lm(formula = log_msales ~ log_t3tot_rev + t3tot_rat + log_comp_prod +
    log_t1conv + log_svol + t2convtrend + log_t1click + log_t2click +
    t1ctrend + t3convtrend + log_t3conv + log_svoltrend + log_t2conv,
    data = train.df)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1981 -0.2793  0.1098  0.4568  3.2379

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  -0.006394   0.011278  -0.567             0.570775
log_t3tot_rev  0.533985   0.012988  41.114 < 0.0000000000000002 ***
t3tot_rat      0.065862   0.012894   5.108        0.00000033775 ***
log_comp_prod  0.083654   0.014481   5.777        0.00000000806 ***
log_t1conv     0.158819   0.018015   8.816 < 0.0000000000000002 ***
log_svol       0.062672   0.011652   5.379        0.00000007843 ***
t2convtrend    0.033872   0.012772   2.652             0.008023 **
log_t1click   -0.015757   0.019937  -0.790             0.429381
log_t2click   -0.065149   0.019877  -3.278             0.001054 **
t1ctrend      -0.048973   0.013357  -3.666             0.000248 ***
t3convtrend    0.043033   0.013130   3.278             0.001054 **
log_t3conv    -0.046878   0.017005  -2.757             0.005861 **
log_svoltrend  0.024729   0.012293   2.012             0.044306 *
log_t2conv     0.041711   0.020906   1.995             0.046077 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8006 on 5032 degrees of freedom
Multiple R-squared:  0.3815,    Adjusted R-squared:  0.3799
F-statistic: 238.7 on 13 and 5032 DF,  p-value: < 0.00000000000000022
```

Backward Step Model

```
Call:
lm(formula = log_msales ~ log_svol + log_svoltrend + t1ctrend +
    log_t1conv + log_t2click + log_t2conv + t2convtrend + log_t3conv +
    t3convtrend + log_comp_prod + log_t3tot_rev + t3tot_rat,
    data = train.df)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1963 -0.2790  0.1105  0.4549  3.2379

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  -0.006387   0.011278  -0.566             0.571175
log_svol       0.062179   0.011635   5.344        0.000000094812 ***
log_svoltrend  0.023826   0.012239   1.947             0.051625 .
t1ctrend      -0.053374   0.012141  -4.396        0.000011242923 ***
log_t1conv     0.154304   0.017085   9.032 < 0.0000000000000002 ***
log_t2click   -0.070687   0.018601  -3.800             0.000146 ***
log_t2conv     0.042070   0.020900   2.013             0.044182 *
t2convtrend    0.033689   0.012769   2.638             0.008358 **
log_t3conv    -0.048544   0.016874  -2.877             0.004033 **
t3convtrend    0.043125   0.013129   3.285             0.001028 **
log_comp_prod  0.086977   0.013856   6.277        0.000000000374 ***
log_t3tot_rev  0.533706   0.012982  41.110 < 0.0000000000000002 ***
t3tot_rat      0.068625   0.012411   5.530        0.000000033720 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8006 on 5033 degrees of freedom
Multiple R-squared:  0.3814,    Adjusted R-squared:  0.3799
F-statistic: 258.6 on 12 and 5033 DF,  p-value: < 0.00000000000000022
```

Both Step Model

```
Call:
lm(formula = log_msales ~ log_svol + log_svoltrend + t1ctrend +
    log_t1conv + log_t2click + log_t2conv + t2convtrend + log_t3conv +
    t3convtrend + log_comp_prod + log_t3tot_rev + t3tot_rat,
    data = train.df)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1963 -0.2790  0.1105  0.4549  3.2379

Coefficients:
               Estimate Std. Error t value             Pr(>|t|)
(Intercept)   -0.006387   0.011278  -0.566             0.571175
log_svol       0.062179   0.011635   5.344        0.000000094812 ***
log_svoltrend  0.023826   0.012239   1.947             0.051625 .
t1ctrend      -0.053374   0.012141  -4.396        0.000011242923 ***
log_t1conv     0.154304   0.017085   9.032 < 0.0000000000000002 ***
log_t2click   -0.070687   0.018601  -3.800             0.000146 ***
log_t2conv     0.042070   0.020900   2.013             0.044182 *
t2convtrend    0.033689   0.012769   2.638             0.008358 **
log_t3conv    -0.048544   0.016874  -2.877             0.004033 **
t3convtrend    0.043125   0.013129   3.285             0.001028 **
log_comp_prod  0.086977   0.013856   6.277        0.000000000374 ***
log_t3tot_rev  0.533706   0.012982  41.110 < 0.0000000000000002 ***
t3tot_rat      0.068625   0.012411   5.530        0.000000033720 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8006 on 5033 degrees of freedom
Multiple R-squared:  0.3814,    Adjusted R-squared:  0.3799
F-statistic: 258.6 on 12 and 5033 DF,  p-value: < 0.00000000000000022
```

Model Evaluation

Model	ME	RMSE	MAE	Adj R sq	AIC
Simple	0.0085538	0.7802414	0.5197582	0.3683	12187.88
Full	0.015885	0.778174	0.522741	0.3794	12100.54
Forward	0.015982	0.7786	0.522768	0.3799	-2230.57
Backward	0.015965	0.778983	0.522914	0.3799	-2231.95
Both	0.015965	0.778983	0.522914	0.3799	-2231.95

ME (Mean Error): Measures bias; closer to 0 indicates less bias.

Simple model with ME of 0.0085538 appears to be the least biased model with minimal systematic error. Full model has a slightly higher bias but still close to 0. Forward model's ME is comparable to the full model. Backward and both have the same ME, showing no improvement.

RMSE (Root Mean Squared Error): Measures overall prediction error; lower values indicate better accuracy.

Simple model has higher prediction error compared to other models. Full has a slightly better accuracy than the simple model. Forward model has RMSE of 0.7786 which is very similar to the full model, showing marginal improvement. Backward and both are identical and have a slightly higher error than forward.

MAE (Mean Absolute Error): Measures average magnitude of errors; lower values indicate better performance.

Simple model has the lowest average error among models. Full has a slightly higher average error and forward only shows minor improvement. Backward and both are Marginally worse than Forward with a negligible difference.

Adjusted R-squared (Adj R^2): Indicates the proportion of variance explained, adjusted for predictors; higher values are better.

The simple model has the lowest adjusted R^2 (0.3683), explaining the least variance among all models. The full model shows a moderate improvement. The forward model achieves the highest adjusted R^2 of 0.3799, indicating slightly better performance. Both the backward and both models match the forward model's adjusted R^2, showing similar explanatory power.

Variable	Simple	Forward	Backward	Both
log_comp_prod	O	O	O	O
log_svol	O	O	O	O
log_svoltrend	O	O	O	O
log_t1click	O	O	X	X
log_t1conv	X	O	O	O
log_t2click	O	O	O	O
log_t2conv	X	O	O	O
log_t3click	O	X	X	X
log_t3conv	X	O	O	O
log_t3tot_rev	O	O	O	O
sfreqr	O	X	X	X
t1convtrend	O	X	X	X
t1ctrend	O	O	O	O
t2convtrend	O	O	O	O
t2ctrend	O	X	X	X
t3avg_mage	O	X	X	X
t3convtrend	O	O	O	O
t3ctrend	O	X	X	X
t3tot_rat	O	O	O	O

AIC (Akaike Information Criterion): Measures model quality relative to complexity; lower values are better.

The simple model has the highest AIC, making it the least optimal model. The full model shows significant improvement with a lower AIC. The forward model drastically improves with an AIC of -2230.57, reflecting a better fit with fewer predictors. The Backward model slightly outperforms Forward with an AIC of -2231.95, indicating even better performance.

Final Model

Based on these metrics, it appears that the backward and both step regression is the most optimal one. They have the lowest AIC of -2231.95, indicating the best model fit, and match in performance with the highest Adjusted R^2. Additionally, they achieve this with fewer predictors compared to the forward model, making them more parsimonious while maintaining optimal performance.

Coefficients Interpretation

log_svol: A 1% increase in svol (search volume) leads to an increase of approximately 0.062% in sales.
log_svoltrend: A 1% increase in the trend of search volume results in an increase of about 0.024% in sales.
t1ctrend: A 1% increase in the trend of click-through rate for the top ASIN results in a decrease of 0.053% in sales.
log_t1conv: A 1% increase in the conversion rate for the top ASIN results in a 0.154% increase in sales.
log_t2click: A 1% increase in the click share of the second ASIN leads to a 0.071% decrease in sales.
log_t2conv: A 1% increase in the conversion share of the second ASIN results in a 0.042% increase in sales.
t2convtrend: A 1% increase in the trend of conversion share for the second ASIN leads to a 0.037% increase in sales.
log_t3conv: A 1% increase in the conversion share of the third ASIN results in a 0.049% decrease in sales.
t3convtrend: A 1% increase in the trend of the third ASIN's conversion rate increases sales by 4.3%.
log_comp_prod: A 1% increase in the number of competing products increases sales by 0.09%.
log_t3tot_rev: A 1% increase in the total review count of the top 3 ASINs increases sales by 0.534%
t3tot_rat: A 1-unit increase in the total rating of the top 3 ASINS increases sales by 0.069%.

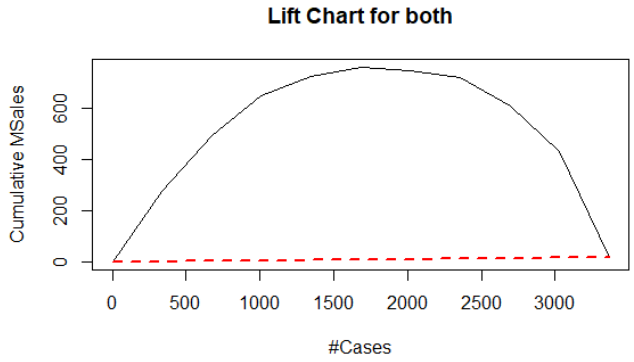
The adjusted R-squared indicates that the model explains about 38% of the variation in the log-transformed sales. While not extremely high, it might be acceptable for a complex system with multiple factors. Additonally, the F-statistic (258.6, p-value < 0.001) shows that the model overall is statistically significant.

Recommendations based on model

- Review count for the top 3 products has the most significant positive impact on overall sales.
    - Encourage customers to leave reviews by offering incentives (e.g., discounts, loyalty points).
- Conversion rates for the top product are critical, with a strong positive impact. A 1% increase in the conversion rate for the top product can lead to significant sales growth.
    - Business owners should improve product descriptions, images, and offer targeted discounts or promotions.
- Search volume positively influences sales. For every 1% increase in search volume, sales grow by 0.062%.
    - Businesses can invest in SEO and keyword optimization to drive traffic. They can also use this as an indicator to stock their inventories.
- The declining click-through trend for the top product negatively affects sales. This might indicate reduced customer interest or ineffective advertising.
    - Revamping ad campaigns for the top product, focusing on fresh creatives and better targeting can help businesses. It must be ensured that product remains relevant through pricing and feature updates.

Performance Metrics

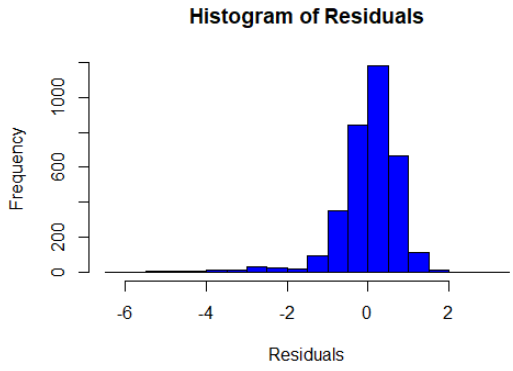
Lift Chart Analysis

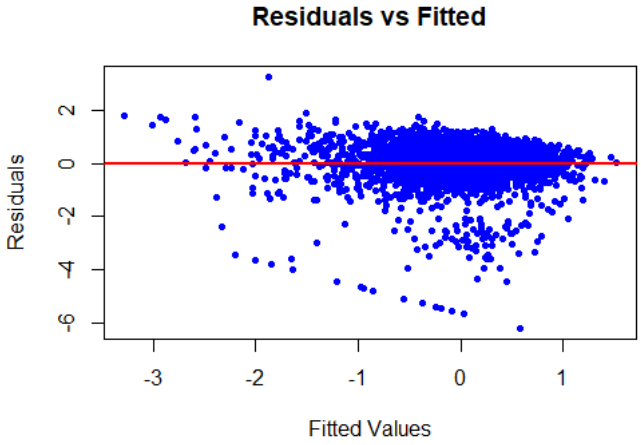


The black line in the lift chart being significantly higher than the red dashed line demonstrates that the model is effective at identifying and prioritizing cases with high predictive power, outperforming a random baseline. This separation between the two lines indicates the model's strong ability to differentiate valuable cases early in the process.

The curve's shape, which rises sharply at first and then levels off, shows that the model captures the majority of its predictive capability in the initial portion of the cases. This suggests that as more cases are considered, the model's incremental ability to add predictive value diminishes. Essentially, the model is most efficient at prioritizing high-value cases but provides diminishing returns when applied to larger datasets. This tapering effect is typical for predictive models that leverage ranked outcomes or probabilities.

Residual Analysis





These residual plots are of our final model, the backward/both regression output. The residuals show some heteroscedasticity (non-constant variance), as the spread of residuals decreases with increasing fitted values. This suggests that the model might not perfectly capture all the variability in the data,

particularly for larger predicted values. There are residuals particularly below -4, indicating outliers in the data. The presence of a funnel-shaped pattern indicates that the model might predict better for certain ranges of the response variable but less accurately for others.

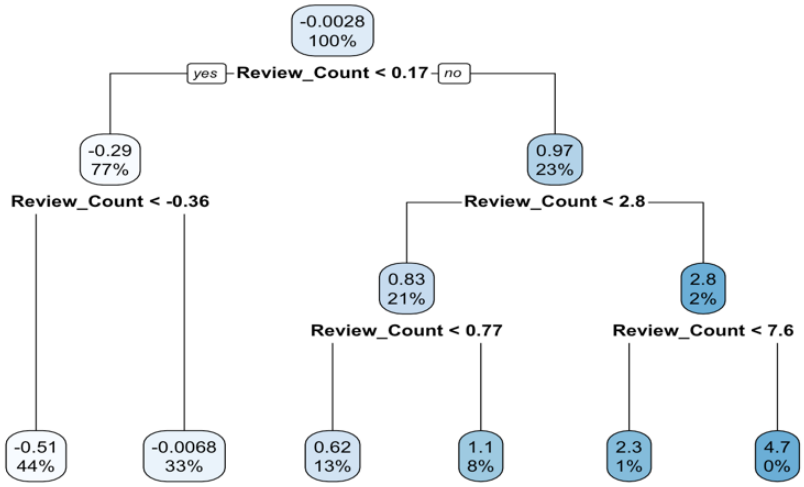
Decision Tree

We are using both regression and classification decision trees here, because:

- Classification simplifies decision-making, while regression provides detailed predictions.
- Improved Business Strategies: Classification identifies key product groups, and regression quantifies their impact

Regression Decision Tree

- Purpose: To predict the sales value (numerical).
- Use Case: Useful for quantitative insights (e.g., forecasting sales/revenue).



Recommendations based on model

1) Key Decision Points:

- The most important splits are based on Review Count < 0.17 and Review Count < 2.8.
- Focus on increasing reviews for products below these thresholds to improve performance.

2) Predictions Increase with Higher Review Count:

- At the root node, the prediction starts low (-0.0028).
- As the Review Count increases, the predicted value climbs:
- At Review Count >= 2.8, the prediction reaches 2.8
- At Review Count >= 7.6, the prediction is 4.7
- Incentivize customers to leave reviews for products near these critical thresholds.

3) Most Data Falls on the Left Side:

- 77% of the data has a Review Count < 0.17, suggesting that most products in your dataset have relatively few reviews
- Use strategies like discounts, loyalty programs, or post-purchase reminders to encourage more reviews.
- Highlight these products in email marketing or promotions to boost visibility and sales

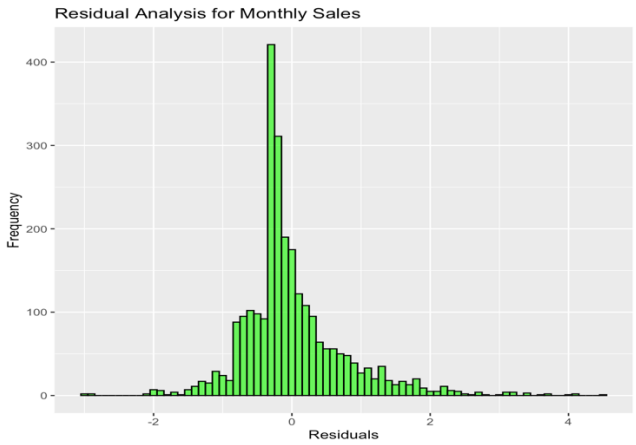
Performance Metrics

Residual Analysis

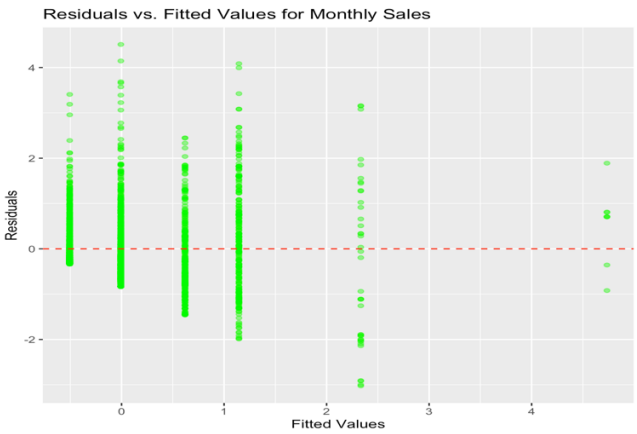
Residual analysis histogram shows a well-behaved model:

- Residuals are centered around 0, symmetrical, and have a narrow spread.

- This indicates that model is performing well and doesn't show signs of bias or systematic prediction errors.

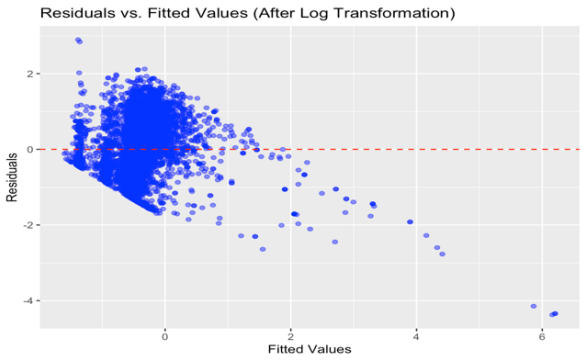


- Residuals Cluster Around Zero-Most of the residuals are close to the red dashed line, meaning the model's predictions are generally accurate.
- For smaller predicted sales (left side of the plot), the residuals are more spread out. This means the model struggles to predict low sales accurately.
- For higher predicted sales (right side of the plot), the residuals are tightly clustered around the red line. This suggests the model is more accurate at predicting higher sales.
- Some points are far away from the red line, indicating outliers or cases where the model's prediction is far off from the actual value.
- The model performs moderately well but is less accurate for lower sales predictions.



Next Steps

- Address heteroscedasticity (e.g., apply a log or square root transformation to the target variable).



- The log transformation has reduced the variance of residuals across the fitted values
- Funnel-shaped pattern - Residuals for low fitted values (left side) are more spread out. Residuals for higher fitted values (right side) are more tightly clustered.
- This suggests that the model still struggles to predict accurately for lower sales values, even after the transformation.
- The transformation has helped reduce heteroscedasticity and made the residual spread more uniform, but it hasn't fully eliminated the issue
- The model is less accurate for lower sales values, which could affect business decisions if low sales predictions are important

Errors

Target Variable	**RMSE**	**MAE**	**R-Squared**
Monthly Sales	0.75419	0.5265366	0.434131

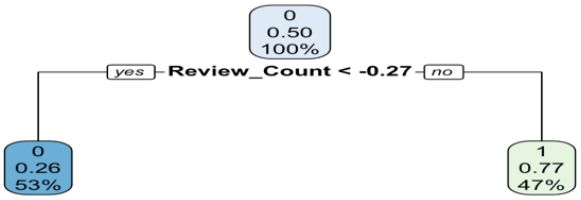
- Root Mean Squared Error (RMSE): 0.75419
    1. Measures the average magnitude of errors, penalizing large errors more.
    2. Indicates moderate error levels in predictions.

- Mean Absolute Error (MAE): 0.5265366
    1. Measures the average absolute error, showing predictions are off by ~0.53 on average.

- R-Squared: 0.434131
    1. Explains ~43.41% of the variance in Monthly Sales, indicating a moderate fit.
    2. Suggest the model is missing key factors or relationships.

Classification Decision Tree

- Purpose: To predict whether sales fall into High or Low categories.
- Use Case: Helps in binary decisions (e.g., targeting products for promotions).



Converting the continuous target variable (Monthly Sales) into a binary classification problem:

- 1 (High Sales): Sales above the median.
- 0 (Low Sales): Sales at or below the median.

Recommendations based on model

1)Increase Review Count:

- Products with low Review_Count (< -0.27) are more likely to have low sales. Efforts should focus on increasing reviews for these products, such as:
    1. Post-purchase emails requesting reviews.
    2. Offering incentives for leaving feedback.

2) Prioritize High-Review Products: Products with high Review Count (>= -0.27) are likely to achieve high sales.

- Marketing campaigns and promotions should target this segment.

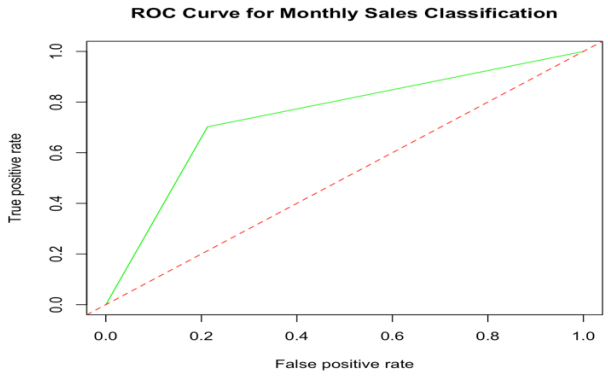
3)Set Review Goals:

- Aim for a Review Count above -0.27 as a threshold for better sales performance.
- For new products, implement strategies to quickly gather reviews to cross this critical threshold.

Performance Metrics

Lift Chart & ROC Analysis

The ROC curve for our model indicates good classification performance since it stays well above the random guessing line. Calculating AUC to quantify its performance further and optimize the classification threshold for this specific use case.



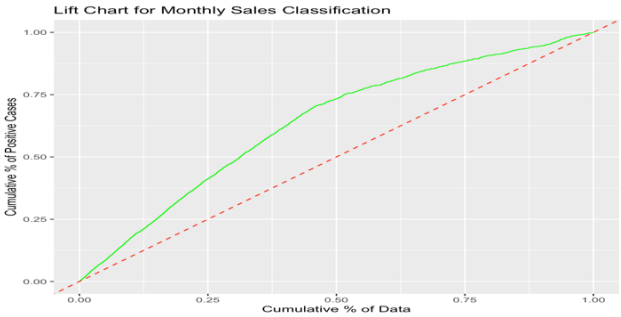
AUC for Classification Model: 0.7446565

Better Than Random: Our model is significantly better than random guessing (AUC > 0.5).

Balanced Predictions: With AUC > 0.7, the model is effective at predicting both classes without extreme bias toward one.

While AUC = 0.74 is good, it suggests room for improvement:

- The model might be missing important predictors or relationships in the data.
- There may be some overlap between the two classes (High Sales and Low Sales), which makes classification harder.



The Lift Chart indicates that your model performs well, especially in the top percentages of predictions, capturing a significant number of positive cases efficiently. The consistent gap between the green curve and the red line highlights the model's predictive power compared to random guessing.

- The green curve is higher than the red dashed line, meaning our model is much better than random guessing.
- A steeper green curve means your model captures High Sales cases early (with fewer products), making it more efficient.

Accuracy, Precision and Recall

Target Variable	ACCURACY	PRECISION	RECALL	F1-Score	AUC
Monthly Sales	0.7446565	0.7255454	0.7870229	0.7550348	0.7446565

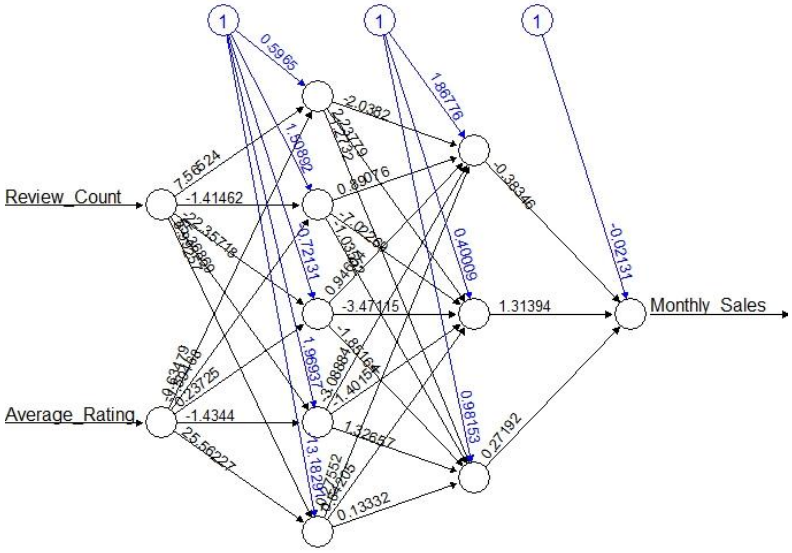
- Accuracy: 74.47% of predictions were correct, indicating good overall performance.
- Precision: 72.55% of predicted High Sales were accurate, showing room for improvement in avoiding false positives.

- Recall: 78.70% of actual High Sales were correctly identified, meaning the model captures most positive cases.
- F1-Score: 75.50%, highlighting a balance between precision and recall.
- AUC: 74.47%, reflecting strong overall classification ability.

Performance Summary

- The model is better at identifying High Sales (high recall) but slightly struggles with avoiding false positives (lower precision).
- The F1-Score indicates a balanced performance, suitable for imbalanced datasets.
- The model performs well overall and can be used effectively to identify High Sales.
- For business decisions where avoiding false positives is critical, consider fine-tuning the model or adjusting thresholds to improve precision.
- Decision Tree: Offers a slight improvement over Linear Regression by capturing non-linear relationships, but it may suffer from overfitting, resulting in suboptimal performance.

Neural Networks

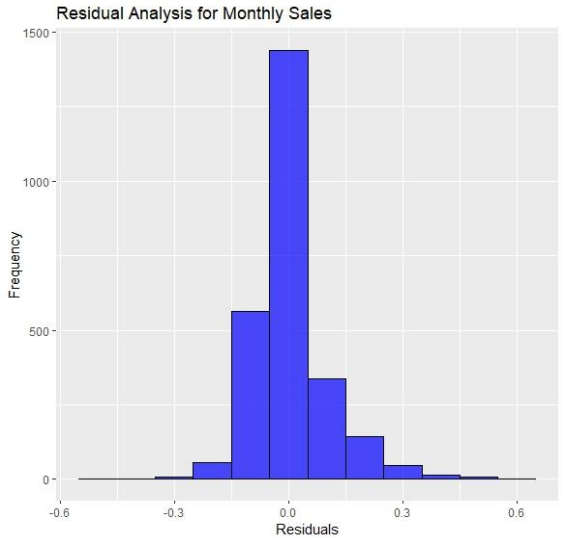


Error: 4.605665 Steps: 1227

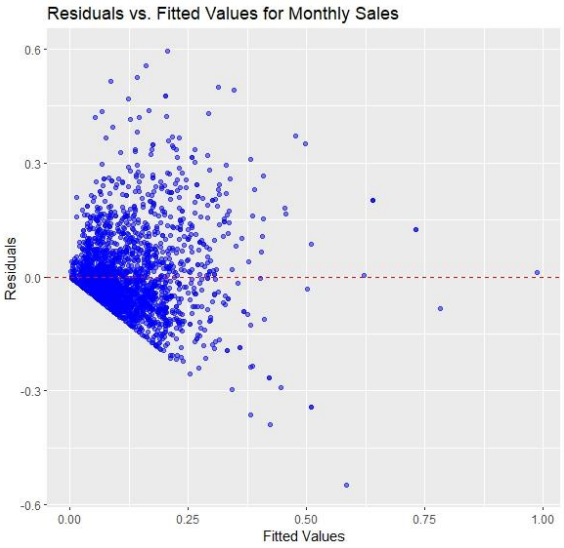
When compared to possible benchmarks, the neural network model's RMSE (Root Mean Squared Error) of 0.0981 indicates good predictive accuracy for the target variable. The anticipated and actual values differ significantly, as noted in the AUC (Area Under Curve) of 0.8314. This suggests that patterns in consumer buying behaviour affected by ratings and reviews are well captured by the model.

Performance Metrics

Residual Analysis



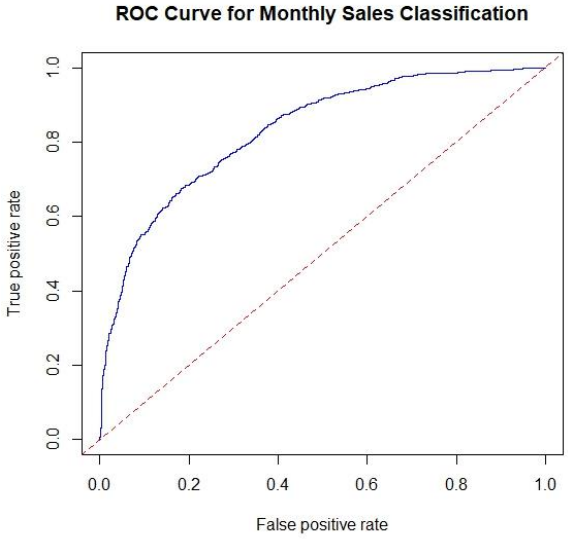
The distribution of residuals for monthly sales appears to be approximately normal with a slight left skew. This suggests that the model may slightly overestimate sales for some months.



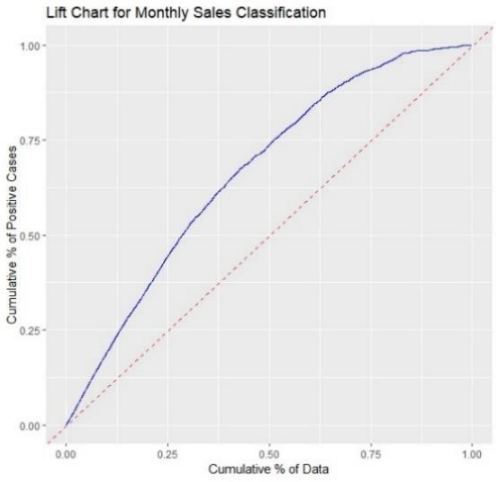
The residuals vs. fitted values plot shows no pattern, meeting model assumptions, but a few outliers need review.

Lift Chart & ROC Analysis

The ROC curve shows a high true positive rate and low false positive rate, indicating strong model performance in classifying sales months.



The lift chart shows improved positive case identification over random selection, with the curve above the diagonal, highlighting effective prioritization of high sales months.



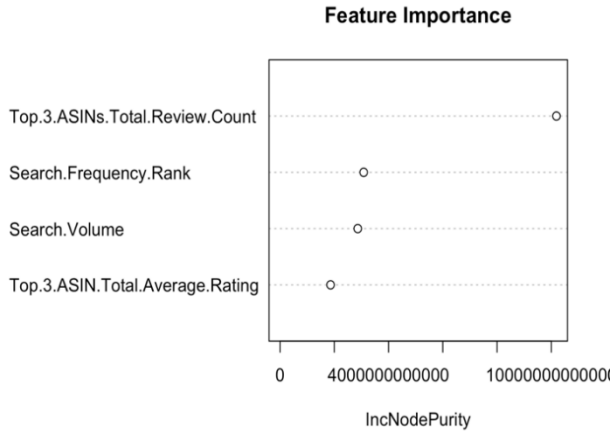
Errors, Accuracy, Precision and Recall

TARGET VARIABLE	MODEL	R-squared	ACCURACY	PRECISION	RECALL	AUC
Monthly Sales	Neural Network	0.4587335	0.5041985	0.5021097	0.9992366	0.8313746

TARGET VARIABLE	MODEL	RMSE	MAE
Monthly Sales	Neural Network	0.09814219	0.0665477

Random Forest

Feature Importance



The provided graph visualizes the feature importance based on the metric IncNodePurity from a model (likely a decision tree or random forest). IncNodePurity measures how much a particular variable reduces impurity (e.g., Gini impurity or mean squared error) in the nodes where it is used as a split criterion.

Observations

1. Top.3.ASINs.Total.Review.Count:
    - This feature has the highest importance, indicating it significantly contributes to the model's predictive performance.
    - Its IncNodePurity value is the largest among all features, suggesting that total review count for the top 3 ASINs is highly predictive.
2. Search.Frequency.Rank:
    - This feature is the second most important.
    - Its importance implies that the rank of search frequency also plays a critical role in predictions.
3. Search.Volume:
    - Slightly lower in importance than Search.Frequency.Rank, but still a relevant predictor.
    - It signifies that the volume of searches has a substantial impact on the outcome.
4. Top.3.ASIN.Total.Average.Rating:
    - This feature has the least importance among the four.
    - Although less influential, the average rating for the top 3 ASINs still contributes meaningfully to the model.

Conclusion

The analysis highlights that customer engagement metrics such as total review count and search frequency rank are critical for the model's performance. Understanding these metrics could aid in

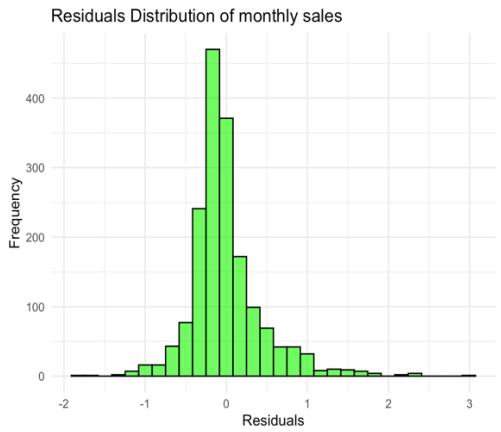
strategic decision-making for product optimization, customer targeting, or market analysis. Lower-ranked features, like average rating, might still provide additional insight but are less impactful compared to the leading predictors.

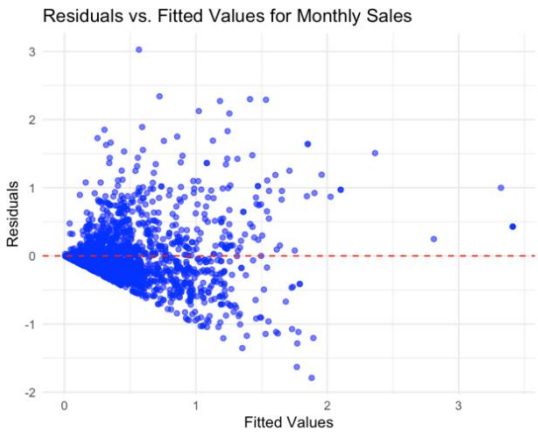
Recommendations based on model

- The Top 3 ASINs Total Review Count is the most influential factor affecting sales.
    - Increase the total number of reviews for the top-performing products.
    - Aim for at least a 10% increase in review count per product quarterly to remain competitive.
    - Monitor and manage review quality, ensuring at least 70%-80% positive reviews, as negative reviews can deter potential buyers.
- This measures how often customers search for a product. A lower rank (closer to 1) means higher customer interest. Products with a high search frequency rank likely attract significant traffic.
    - Focus on optimizing search rankings through targeted SEO for product listings.
    - Identify keywords that drive the most searches and integrate them into product titles, descriptions, and tags.
- Search volume reflects the number of searches a product receives. A high search volume indicates potential demand.
    - Target high-search-volume keywords in paid advertising campaigns to boost visibility.
    - Maintain a weekly review of search trends to adjust inventory and promotions accordingly.

Performance Metrics

Residual Analysis





1. Residual Distribution

- Description: The first histogram displays the distribution of residuals from the predictive model for monthly sales.
- Observations:

- The residuals are approximately centered around zero, indicating that the model has low bias in predictions.
- The distribution appears symmetric, resembling a normal distribution, which suggests that the errors are randomly distributed.
- The majority of the residuals lie between -1 and 1, with only a few outliers beyond this range.

- Implication: A roughly normal distribution of residuals validates the assumption of normality, essential for linear regression models and other statistical models.

2. Residuals vs. Fitted Values

- Description: The second scatter plot compares residuals against fitted values of monthly sales.
- Observations:
  - Residuals are scattered around the horizontal line at zero, without any clear pattern, which indicates homoscedasticity (constant variance of errors).
  - The spread of residuals increases slightly with larger fitted values, suggesting potential heteroscedasticity at higher sales volumes.
  - A few residuals lie significantly far from the majority, indicating the presence of outliers or regions where the model might underperform.
- Implication:
  - The lack of a strong pattern in the residual plot supports the assumption that the model is well-specified.
  - However, the slight increase in variability at higher fitted values may require further investigation, possibly through transformation or weighting.
- The residual analysis confirms that the model performs reasonably well, with errors largely distributed normally and without systematic bias.
- Potential issues to address include slight heteroscedasticity at higher sales values and the influence of outliers.
- Next steps might include examining outliers for data quality or incorporating transformations to stabilize variance if needed.

Errors, Accuracy, Precision and Recall

TARGET VARIABLE	MODEL	RMSE	MAE	R-sq
Monthly Sales	Random Forest	0.06456	0.0323475	0.4090108

The table summarizes the key performance metrics for the predictive model used to forecast monthly sales:

- RMSE (Root Mean Squared Error): 0.06456

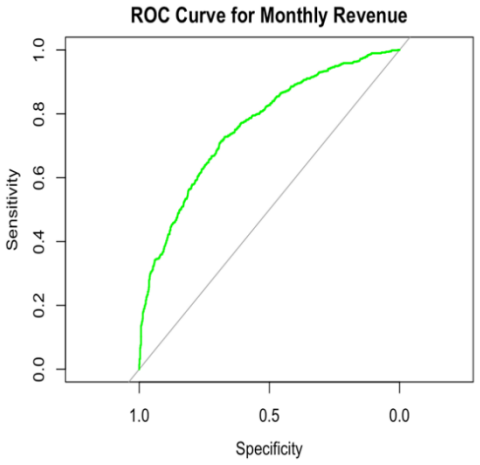
- Indicates the average magnitude of prediction errors, with a lower value signifying better accuracy. The model's RMSE suggests that on average, predictions deviate from actual values by approximately 6.5%.

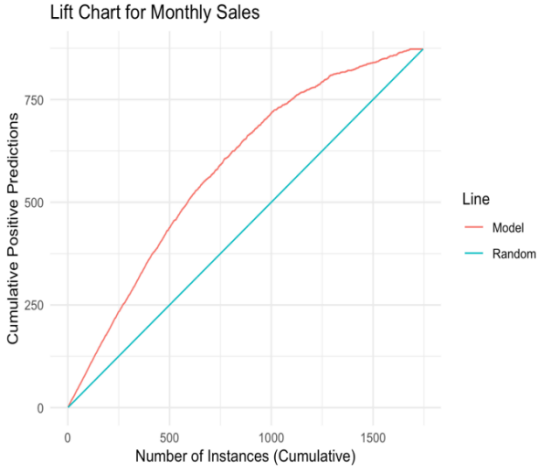
- MAE (Mean Absolute Error): 0.0323475
  - Represents the average absolute difference between predicted and actual sales. The model's MAE implies that predictions are off by about 3.2% on average, showcasing relatively small errors.

- R-Squared: 0.4090108
  - Explains the proportion of variance in the target variable (monthly sales) accounted for by the model. An R-squared of ~41% indicates that the model captures a moderate level of variance, leaving significant room for improvement.

- The model demonstrates reasonable predictive accuracy as reflected by the low RMSE and MAE values.
- However, the R-squared value suggests that the model only partially explains the variability in monthly sales. This indicates potential for model refinement through feature engineering, incorporating additional predictors, or exploring more complex modeling techniques.
- Next steps might involve:
  1. Investigating unexplained variance to identify potential new features.
  2. Addressing the observed heteroscedasticity in residuals.
  3. Experimenting with non-linear models or ensemble methods for improved performance.

Lift Chart & ROC Analysis





1. ROC Curve

- Description: The ROC curve plots sensitivity (true positive rate) against 1-specificity (false positive rate) for the predictive model.
- Observations:

- The ROC curve deviates significantly from the diagonal, indicating good model performance.
- The shape of the curve suggests that the model achieves a balance between sensitivity and specificity for most thresholds.
- This implies that the model effectively discriminates between different outcomes.

2. Lift Chart

- Description: The lift chart compares the cumulative positive predictions of the model against a random model.
- Observations:
  - The model's cumulative positive predictions (red line) outperform the baseline random predictions (blue line) across the number of instances.
  - The separation between the two lines indicates that the model has predictive power and performs significantly better than random guessing.
- The ROC curve and lift chart demonstrate that the model is effective in predicting monthly sales, with strong classification performance.
- To further improve the model, consider fine-tuning hyperparameters or incorporating additional features to increase predictive accuracy.

Models Evaluation & Comparison – Selecting the best model

MODEL	RMSE	MAE	R Squared
Linear Regression	0.77898	0.522914	0.3799
Decision Tree	0.75419	0.5265366	0.434131
Neural Networks	0.09814	0.0665477	0.4587335
Random Forest	0.09674	0.0645698	0.4090108

Linear Regression: Performs poorly as it assumes a linear relationship between features and the target variable. The high RMSE and low R-Squared indicate it is insufficient for capturing the complexity of the data.

Decision Tree: Offers a slight improvement over Linear Regression by capturing non-linear relationships, but it may suffer from overfitting, resulting in suboptimal performance.

Neural Networks: Achieves a notable reduction in error metrics and the highest R-Squared value, indicating it captures complex patterns and explains the maximum variance in the target variable.

Random Forest: Delivers the best RMSE and MAE values, highlighting its superior accuracy and robustness. While its R-Squared is slightly lower than Neural Networks, it is less prone to overfitting and more interpretable, making it the most reliable model overall.

Business Insights and Final Recommendations

In all models, regardless of predictive accuracy, review count stood out as the most important predictor for sales. Other than that, product rating and search frequency rank also appeared to have an impact on sales. Based on our entire analysis, our findings suggest the following business insights:

- **Impact of Review Count on Monthly Sales.**
    - Encourage customers to leave reviews through post-purchase email campaigns or incentives like discounts on future purchases.
- **Impact of Average Rating on Monthly Sales**
    - Focus on maintaining high product quality to sustain favorable ratings.
    - Address negative reviews promptly and improve low-rated products to rebuild consumer confidence.
- **Impact of Search Frequency Rank and Search Volume**
    - Focus on improving the rank of products with lower sales by using SEO techniques
    - Identify products with moderate ranks that show potential for improvement.

Conclusion

In conclusion, our analysis highlights the critical factors influencing sales performance on online platforms, providing actionable insights for e-commerce businesses. The total review count of top products emerges as the most significant driver, emphasizing the need to amplify customer feedback and maintain a positive review landscape. Search frequency rank and search volume underline the importance of discoverability, urging businesses to optimize product visibility through strategic SEO and marketing efforts. Additionally, maintaining high product ratings strengthens consumer trust and boosts conversions. By leveraging these insights, e-commerce businesses can strategically enhance their product performance, align with customer preferences, and capitalize on the projected growth of the e-commerce market, which is expected to reach $8 trillion by 2027. These findings offer a roadmap for informed decision-making, ensuring sustained competitiveness in a rapidly evolving digital marketplace.

Group Members and Tasks Division

Group Member Name	Tasks Done
Haritha Rajendra Rao Savithri (dal136929)	Scheduling Teams Meetings, Data Cleaning, Classification & Regression Decision Trees
Naveena Paleti(nxp230055)	Data Visualization, Random Forest Model
Sai Shripad Achutuni (Axs240134)	Data Visualization, Neural Networks Model
Salwa Niaz Preet (sxp240086)	Data Cleaning, Data Visualization, Linear Regression Model