TEAM 2025104

INDEX

1. INTRODUCTION

Problem Statement Description

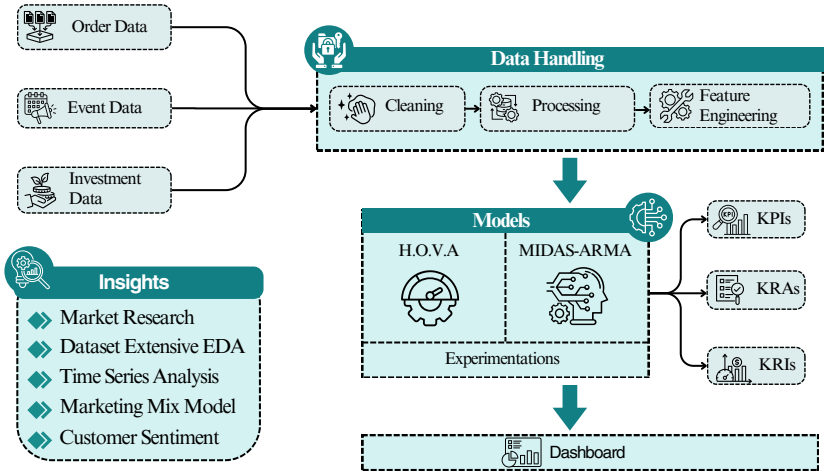
An e-commerce firm specializing in electronics based out of Ontario, Canada, named ElectroMart has raised concerns regarding the impact of marketing expenditures from July 2023 to June 2024, questioning their effectiveness in driving revenue growth. As a data analyst, the goal is to support the marketing team in optimizing the budget allocation for the upcoming 12 Months. This requires a data-driven approach to assess the actual impact of various marketing levers on revenue and recommend an optimal spending strategy.

Business Context

After a year of significant marketing spending, ElectroMart aims to analyze the effect of various strategies, measure ROI, identify key revenue drivers, and optimize budget allocation to maximize impact and improve marketing efficiency for the next year.

Expected Outcomes

The goal is to optimize marketing strategy by identifying key revenue drivers, measuring marketing ROI, reallocating budget for maximum impact, targeting high-performing products, selecting the most effective marketing channels, and leveraging data-driven insights to enhance decision-making and improve overall marketing efficiency.



Dataset Description

Customers_Orders_Data.csv contains transactional details of customer order , including payment methods, delivery timelines, product pricing, and location-specific customer information**,** ***Canada Holiday List.xlsx*** contains a comprehensive list of Canadian public holidays. ***Media data-Sale Calendar-NPS Scores_Data.xlsx*** contains data on marketing expenditures, dates of special promotional events and customer satisfaction (NPS scores) and the company's stock index over the period of 12 months. **Hierarchy Details.xlsx,** ***SKU_details.xlsx*** provides detailed product segmentation with SKU identifiers, super categories, categories, subcategories, and product verticals.

2. DATA VISUALIZATION

Product Based Analysis

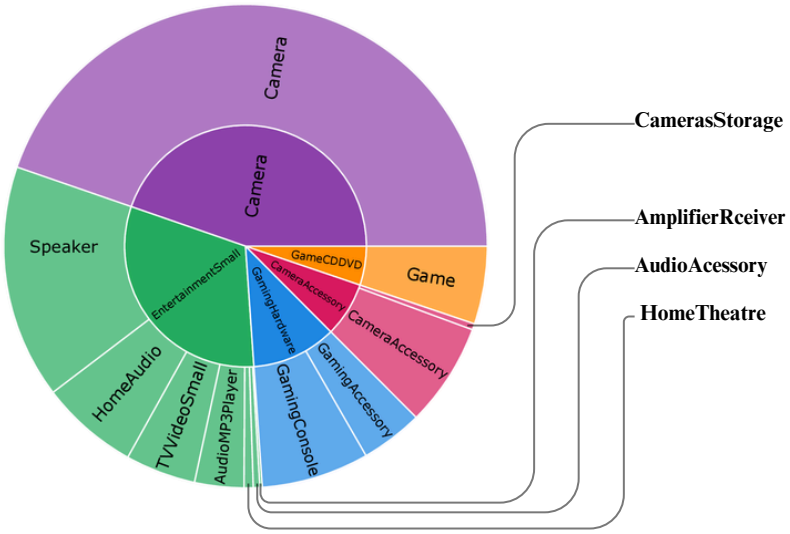


Figure 1: Total GMV by Product Analytic Category

Camera and EntertainmentSmall lead in GMV, warranting focused marketing, while GamingHardware, CameraAccessory, and GameCDDVD need better promotions. Budgeting should prioritize high-revenue categories while using targeted discounts or bundling for weaker segments.

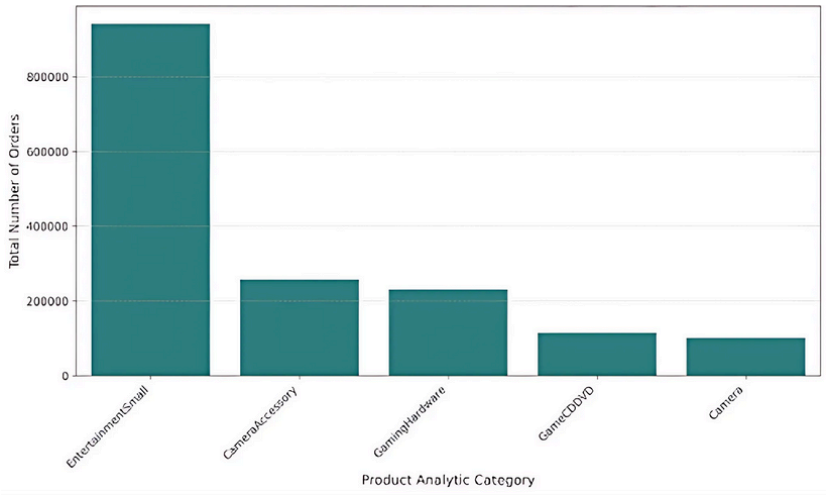


Figure 2: Total Number of Orders per Product Analytic Category

All product categories experience substantial sales increases during promotional periods, especially *EntertainmentSmall*, which exhibits the highest uplift. *EntertainmentSmall* consistently dominates sales in both promotional and regular periods, highlighting strong consumer demand and notable price sensitivity. Additionally, *GamingHardware* and *CameraAccessory* benefit significantly from discount strategies, indicating that targeted pricing adjustments could effectively enhance their overall sales performance.

Marketing Channel Analysis

A data-driven analysis reveals key marketing channel impacts, highlighting high-performing, synergistic, and underperforming areas for strategic optimization.

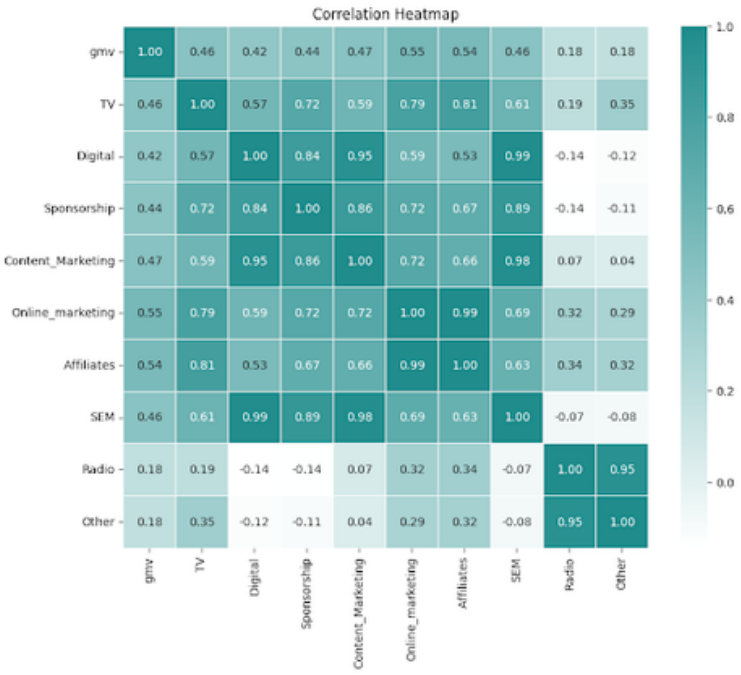


Figure 3: Marketing Channels: Correlation with GMV

Online Marketing (0.91) and Affiliates (0.92) emerge as high-impact channels significantly driving GMV, with TV and Sponsorship offering moderate contributions. Digital, Content Marketing, and SEM demonstrate strong synergy, benefiting from integrated marketing efforts. Conversely, Radio proves ineffective and demands reassessment, while TV and Sponsorship investments should be moderately maintained.

Customer Satisfaction Analysis

Around 80% of customers buy only during sales, showing high price sensitivity and discount-driven behavior. Low repeat purchases highlight retention challenges and weak engagement strategies. Loyalty programs, personalized offers, and post-sale engagement could turn sale-only shoppers into repeat buyers, boosting customer lifetime value (CLV).

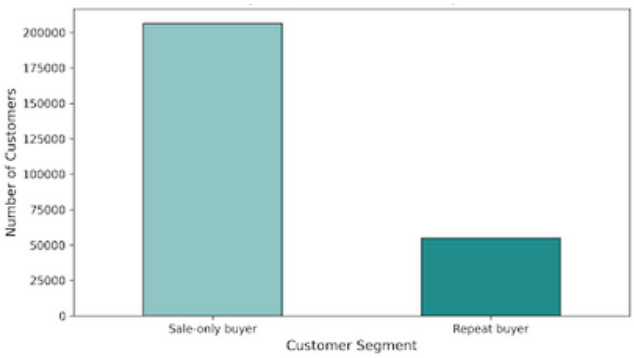


Figure 4: Customer Segmentation

Sales Performance Analysis

Prepaid orders have a higher average price (₹3,704 vs. ₹1,988 for COD), but COD generates more revenue (₹6,553,330 vs ₹482,376). Prepaid orders are delivered (5.22 vs. 5.87 days) and procured (5.01 vs. 5.55 days) faster. The most successful sale ran from October 15-17, while the Christmas sale underperformed, suggesting an earlier shift. Promotions contributed 22.67% of revenue and 17.90% of total sales.

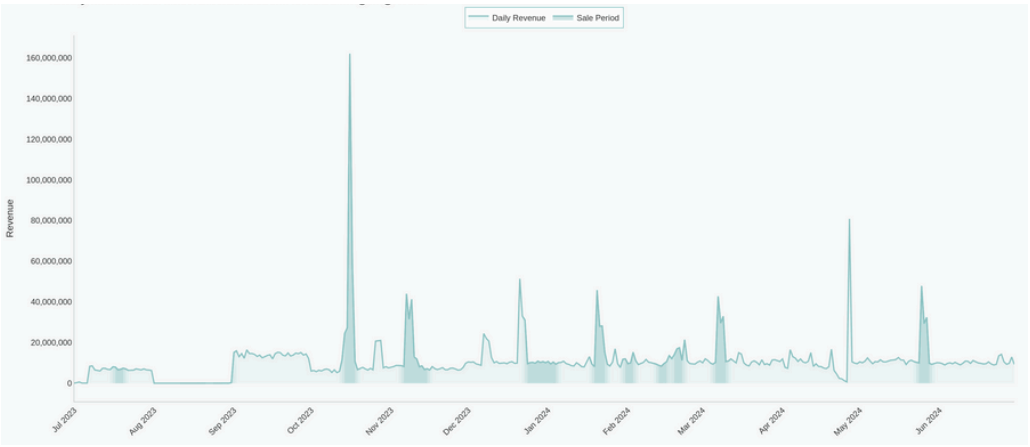


Figure 5: Daily Revenue Trend with Sale Periods Highlighted

The business demonstrated steady revenue growth from mid-2023 to mid-2024, with significant spikes aligning with promotional events, particularly during holidays like October-November 2023 and May-June 2024. Key milestones included achieving revenues of 500M (September 2023), 1B (November 2023), and 2.5B (March 2024). Notably, revenue accelerated after November 2023, peaking sharply between March and July 2024. Despite plateau periods in August-September 2023, baseline revenue consistently remained around ~10M, showing modest yet continuous underlying growth throughout the year.

Product Segmentation

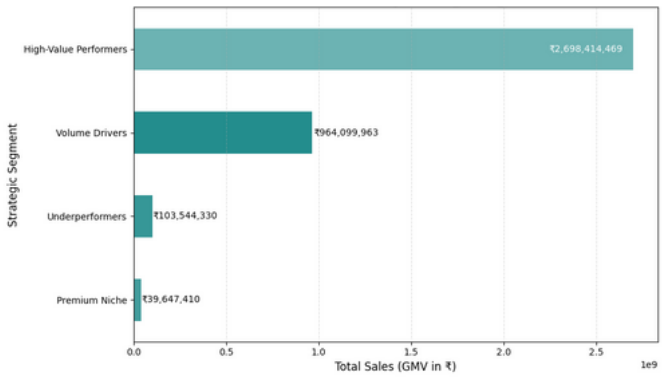
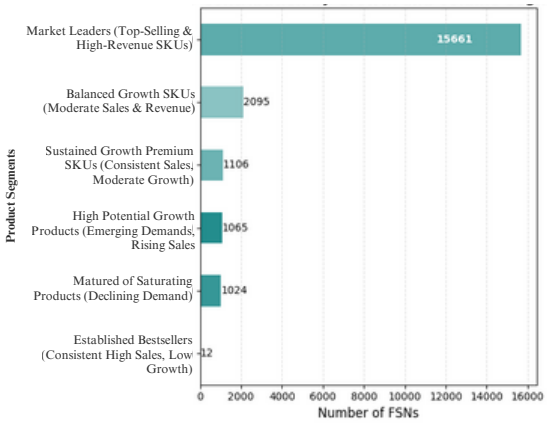


Figure 6: GMV Contribution by Profitability Segment

This report segments products by value vs. volume, profitability, and growth using metrics like GMV, orders, and AOV. Categories include Market Leaders, High-Traffic SKUs, Luxury Products, and Underperformers, highlighting top revenue drivers, high-volume low-value SKUs, and niche premium products.

The analysis, using GMV, AOV, and growth rates, revealed key segments: Balanced Growth SKUs (13,458), Underperformers (14,370), Sustained Growth Premium SKUs (15,661), Market Leaders (2,123), and High-Value Performers (1,793). Recommendations focus on leveraging Market Leaders, nurturing growth products, and addressing underperformers for strategic decisions.

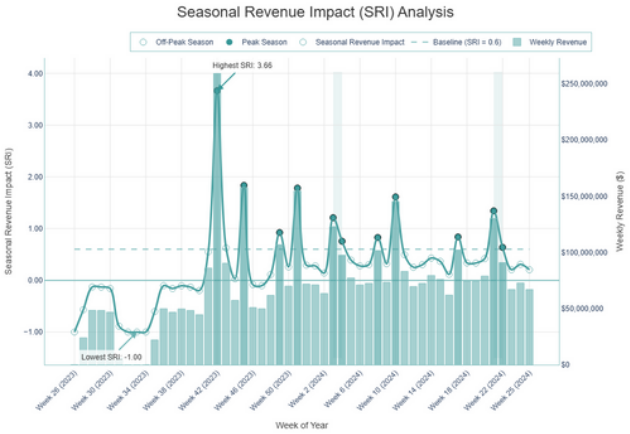


3. KPIs

1. Seasonal Revenue Impact

Inferences:

The Seasonal Revenue Impact (SRI) metric compares weekly revenue to off-peak periods, identifying peak seasons (above the 80th percentile). High SRI signals strong demand, while negative values show inefficiencies, enabling data-driven financial planning and optimized investments.



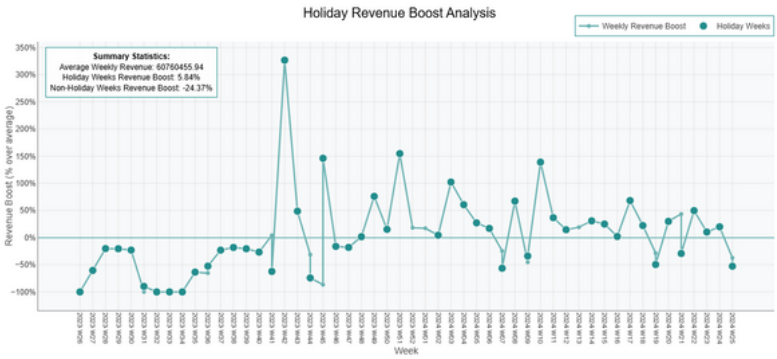
$$\text{SRI} = \frac{\text{Avg. Revenue In Current Week} - \text{Avg. Revenue In Off Peak Season}}{\text{Avg. Revenue In Off Peak Season}}$$

Figure 7:Seasonal Revenue Impact (SRI) Analysis

2. Holiday Revenue Boost Analysis

Inferences:

The *Holiday Sales Boost (HSB)* metric measures revenue changes during holidays, revealing a **5.84%** uplift in holiday weeks and a **24.37%** decline in non-holiday weeks.



$$\text{HSB} = \frac{\text{Avg. Current Week Revenue} - \text{Avg. Weekly Revenue}}{\text{Avg. Weekly Revenue}}$$

Figure 8: Holiday Revenue Boost(HSB) Analysis

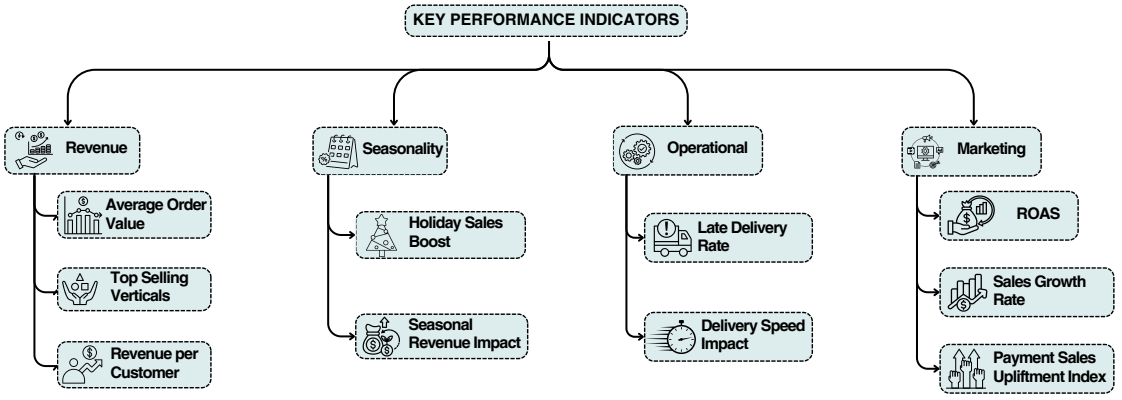
3. Sale Week Revenue Boost Analysis

Inferences:

The Sale Week Revenue Boost (SWRB) metric assesses promotional effectiveness, revealing a **1.23%** drop in sale week revenue and a **10.75%** decline in non-sale weeks.

Figure 9:Sale Week Revenue Boost Analysis

$$\text{SWRB} = \frac{\text{Average Current Week Revenue} - \text{Average Weekly Revenue}}{\text{Average Weekly Revenue}}$$



Key Resultant Areas
Product Portfolio Strategy
Customer Experience Impact
Marketing Channel Strategic Alignment
Customer Retention Strategy
External Market Factor Responsiveness
Marketing Effectiveness

Key Risk Indicators
Average Order Value Fluctuations
Customer Acquisition Cost
Delivery Delay Percentage
Discount Depth Risk
Weather Impact Sensitivity
Channel Dependency Risk

4. KRAs

The *Key Resultant Areas (KRAs)* for *ElectroMart* focus on driving business growth. These include Product Portfolio Strategy to meet customer demand, Customer Experience Impact for enhancing satisfaction and loyalty, and Marketing Channel Strategic Alignment to ensure effective marketing reach. The Customer Retention Strategy helps in building loyalty, while staying agile to External Market Factors like trends and competition. Finally, continuously optimizing Marketing Effectiveness ensures maximum return on investment.

5. KRIs

The *Key Risk Indicators (KRIs)* for *ElectroMart* include Average Order Value Fluctuations, Customer Acquisition Costs, and Delivery Delays, which can affect customer satisfaction and profitability. Discount Depth Risk from excessive discounting and Weather Sensitivity affecting demand are additional risks. Channel Dependency Risks arise from relying too heavily on specific marketing channels. Monitoring these *KRIs* helps ElectroMart mitigate risks, adapt to market changes, and ensure consistent growth and profitability.

6. DATA PREPROCESSING

Data Cleaning

Data Type Correction: Converted numerical strings to proper numeric format.
Handling Missing/Invalid Values: Removed rows with minimal NaNs in GMV and negative GMV values; converted negative values in key identifiers (Customer ID, Pincode, Procurement SLA) to absolute values.
Merging Data Sources: Integrated media investment, stock data, and SKU details via FSN ID.
Deduplication: Removed ~100K duplicate rows; eliminated rows where MRP or GMV was zero.
Final Dataset: Cleaned dataset with **1,643,272** rows, **39** columns; a reduced version with **18,000** rows aggregates daily category orders.

Processing

Chow-Lin method is a regression-based temporal disaggregation technique that converts low-frequency data (e.g., monthly or quarterly) into high-frequency estimates (e.g., daily or monthly) while preserving aggregate consistency. It uses Generalized Least Squares (GLS) to improve estimation accuracy, especially in the presence of autocorrelation. The method leverages high-frequency indicator variables, which are correlated with the low-frequency series, to guide the disaggregation process and ensure the resulting high-frequency data reflects underlying trends in a statistically sound manner.

Assumptions and Limitations:

Assumes a linear link with indicators, homoscedastic or autocorrelated errors, and exact aggregate matching.

Sensitive to indicator quality, cannot model non-linearities, and assumes stable relationships over time.

Mathematical Model:

Model Specification	$y = X\beta + \varepsilon$
Error Structure	$\varepsilon \sim \mathcal{N}(0, \sigma^2\Omega)$
Estimation Method	$Y = Cy$
Aggregation Consistency	$\hat{\beta} = (X^\top \Omega^{-1} X)^{-1} X^\top \Omega^{-1} y$
Disaggregation Output	$\hat{y} = X\hat{\beta}$

Friedrich-Litterman is a regression-based method for temporal disaggregation that uses high-frequency indicators and Bayesian inference using MCMC ensemble sampler to produce aggregate-consistent high-frequency estimates while accounting for residual autocorrelation.

Assumptions and Limitations:

It assumes linearity, stationary errors, and an exact total match, ensuring consistency in relationships. However, it is highly sensitive to errors, struggles with non-linear shifts, and incurs high costs. These limitations make it less effective for complex, dynamic systems requiring adaptability, robustness, and efficient resource utilization.

Model high-frequency data	$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2\Omega)$
Compute Liklehood	$p(Y_L \mid \beta, \rho, \sigma^2)$
Sampling with MCMC	$\{(\beta^{(i)}, \rho^{(i)}, \sigma^{2(i)})\}_{i=1}^{N} \sim p(\beta, \rho, \sigma^2 \mid Y_L)$

Simplistic Feature Introduced :

List Price = GMV / Units.

Discount = (Product MRP - List Price) / Product MRP.

Pay Day introduced as a binary column.

Adstock modeling is a technique used to measure the long-term effect of advertising on sales or brand awareness. It accounts for the decay effect, where the impact of an ad diminishes over time. In Adstock modeling, advertising spend is transformed into a lasting effect, assuming that past ads continue to influence future consumer behavior. This model uses a lagged effect function, where each period's advertising effect is discounted by a constant factor. It helps marketers assess the return on investment (ROI) of advertising campaigns by understanding how long the effects last and optimizing future spending.

$$\text{Adstock}[t] = \sum_{k=0}^{\min(\text{lag},t)} \text{spend}[t-k] \cdot \text{decay}^k$$

7. EXPERIMENTATION

1. XGBoost Regression Optimization

We used XGBoost to optimize marketing budget allocation, reducing spend by 2.13% while increasing revenue. XGBoost captures non-linear relationships by minimizing:

$$\text{Obj}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

with parameters n_estimators=100, max_depth=4, learning_rate=0.1, and subsample=1.0. Our data preparation involved preprocessing revenue, investment, and events data, converting dates, merging datasets, handling missing values, and creating feature matrices with marketing channels and event indicators. XGBoost revealed significant variation in channel effectiveness, which we quantified using:

$$\text{Effectiveness}_{c,m} = \frac{\text{Importance}_c}{\log(1 + \text{Budget}_{c,m})}$$

and for zero-budget channels with positive importance:

$$\text{Effectiveness}_{c,m} = \text{Importance}_c \times 2 \quad \text{if} \quad \text{Budget}_{c,m} = 0 \text{ and } \text{Importance}_c > 0$$

We then applied a dynamic reduction factor:

$$\text{ReductionFactor}_{c,m} = \max\left(0.1, \min\left(0.6, \frac{\text{MedianScore} - \text{Score}_{c,m}}{\text{MedianScore} \times 1.5}\right)\right)$$

to underperforming combinations, reducing their budgets and reinvesting in high-performing channels within defined constraints. We applied key constraints including: historical min/max budget limits per channel, revenue increase cap of 40% per month, and carryover of funds between periods when necessary to optimize the overall marketing ROI.

Optimization

The optimization reduced the annual budget from 846.50 to 838.00, a decrease of 1.00%. This led to a **4.54% increase** in total revenue (173,038,336.00) and an ROI improvement of 5.60%.

Limitations:

1. The model has the potential to overfit to a limited dataset, making it difficult to capture longer-term seasonality and market trends.
2. Potential diminishing returns not fully captured in the model.
3. Capacity constraints in scaling up high-performing channels.

8. MARKETING MIX MODELING

This **Marketing Mix Model** transforms monthly data into daily estimates using the Chow-Lin algorithm for better temporal alignment. Marketing spend and **GMV** relationships are modeled through XGBoost regression with log transformations, while Bayesian Hierarchical Models capture channel uncertainty. **Adstock** transformations quantify time-lagged responses, and differential evolution optimizes budget allocation, projecting an **8% revenue** increase.

Channel effectiveness analysis shows **TV (0.73)** has the highest impact, followed by **Affiliates (0.08)** and **Digital (0.07)**. The Bayesian foundation models both marketing channels and GMV as Normal distributions, incorporating uncertainty and reducing overfitting with limited data points. This approach creates a robust framework that delivers actionable insights while effectively managing marketing response complexity.

Insights:

- TV is a good marketing channel for Entertainment_Small, Gaming_console, CameraStorage, and Camera, while a very bad indicator for amplifier-receiver and home-theatre
- Content Marketing is excellent for CameraStorage but ineffective for other products, while Online Marketing and Affiliates are great for all products, especially for conversions.
- Digital is a bad channel overall, being good only for Home Theatre. It is a bad channel for every other product vertical.

Marketing Levers to Target

Camera Accessory

1. Online Marketing
2. Affiliates
3. Sponsorship

Gaming Hardware

1. TV
2. SEM
3. Online Marketing

Camera

1. SEM
2. Affiliates
3. Digital

Game CDDVD

1. Online Marketing
2. SEM
3. TV

Entertainment Small

1. Content Marketing
2. Online Marketing
3. Affiliates

9. FINAL APPROACH

1. MIDAS-ARMA Model (Mixed Data Sampling-Autoregressive Moving Average)

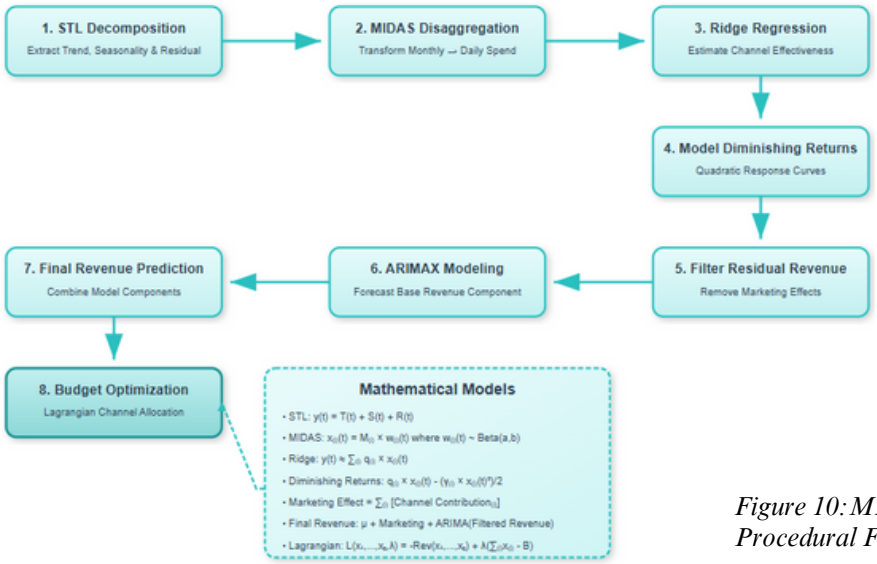


Figure 10: MIDAS-ARMA Procedural Flow

The model decomposes daily revenue y(t) into trend, seasonality, and residuals using STL, focusing on the residual R(t). It uses MIDAS to disaggregate monthly channel spend Mi into daily values $x_i(t) = M_i \times w_i(t)$ via Beta weighting.Ridge regression estimates each channel's effectiveness qi from the relationship $y(t) \approx q_i x_i(t)$, and diminishing returns are modeled with a quadratic penalty:

$$q_i x_i(t) - \frac{\gamma(x_i(t))^2}{2}$$

These contributions are summed and subtracted from R(t) to prepare the input for training the ARIMAX model.Limitations include assumptions of linearity, constant channel effectiveness,stationarity, and simplified diminishing returns, while external factors and potential inaccuracies in spend data are not considered.

Increase in Revenue percentage:**19.8%** (Against actual revenue by training on first 11 months and predicting for the next month by taking next month budget equal to actual)

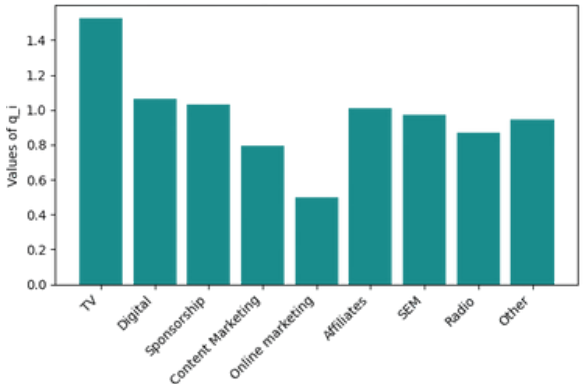


Figure 11 : Marketing Channel Performance

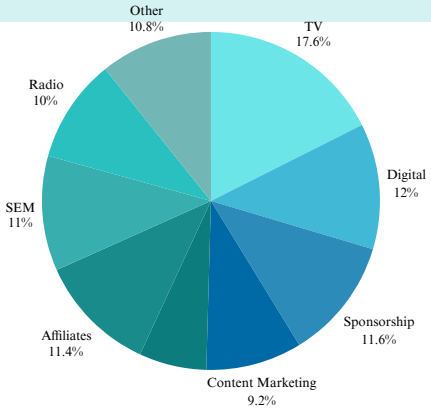


Figure 12:Optimal Budget Allocation (for 42.8 Cr)

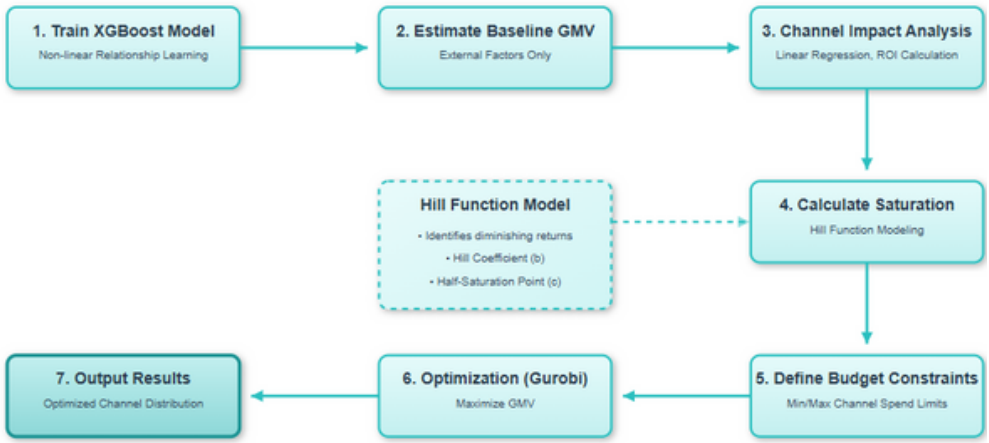


Figure 13: H.O.V.A Model Procedural Flow

The model optimizes marketing spend to maximize GMV by training an XGBoost model to predict GMV based on marketing spends (TV, Digital, Sponsorship, SEM) and external factors (e.g., holidays, sales events, pay days, rolling GMV/units). Baseline GMV is estimated by setting marketing spends to zero. Linear regression measures each channel's impact, and saturation levels are calculated using ROI and modeled with the Hill function. Budget constraints (min/max spends per channel and total budget) are set, and Gurobi optimization is applied to maximize GMV within these limits. The output includes the optimized budget allocation and expected GMV.

Limitations:

The baseline revenue estimation shows the sensitivity of data availability, as even small changes in data can lead to significant changes in revenue dynamics.

Results:

13.60% increase in total revenue and an **11.68% decrease in total investment**.

Year Month	TV	Digital	Sponsorship	Content marketing	Online marketing	Affiliates	SEM	Radio	Other
2024-07	6.32	4.22	10.84	1.14	27.58	8.74	12.99	0.67	1.43
2024-08	0.37	0.25	0.63	0.07	1.61	0.51	0.76	0.04	0.08
2024-09	8.29	5.54	14.22	1.5	36.17	11.46	17.04	0.87	1.87
2024-10	19.35	12.94	33.2	3.5	84.46	26.77	39.78	2.04	4.36
2024-11	12.88	8.61	22.1	2.33	56.23	17.82	26.48	1.36	2.91
2024-12	16.68	11.15	28.61	3.02	72.78	23.07	34.28	1.76	3.76
2025-01	11.58	7.74	19.87	2.1	50.56	16.02	23.81	1.22	2.61
2025-02	9.8	6.55	16.81	1.77	42.75	13.55	20.14	1.03	2.21
2025-03	12.18	8.14	20.89	2.2	53.15	16.85	25.03	1.28	2.75
2025-04	10.45	6.99	17.93	1.89	45.62	14.46	21.49	1.1	2.36
2025-05	11.61	7.76	19.93	2.1	50.69	16.07	23.87	1.22	2.62
2025-06	8.26	5.52	14.18	1.5	36.06	11.43	16.99	0.87	1.86

10. HYPOTHESIS TESTING

The results of the hypothesis tests reveal the following insights:

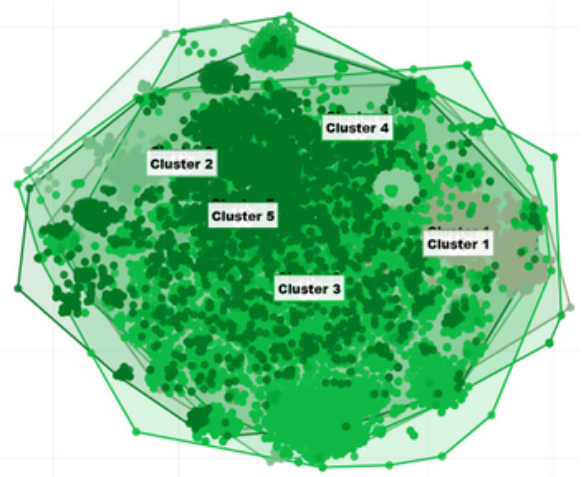
1. Impact of Media Spending on Revenue: The t-test (T-statistic = 3.11, p-value = 0.011) confirms that media spending impacts revenue, positively affecting sales.
2. Effect of Promotions on Sales: The t-test (T-statistic = -7.68, p-value = 1.15e-07) indicates that promotional offers significantly boost sales.
3. Effect of Sale Days on Revenue: On sale days, average revenue (₹20.76 million) more than doubles compared to non-sale days (₹9.00 million), with a significant p-value of 0.0052, highlighting the strong impact of sales on revenue.
Hypothesis testing reveals that among exogenous events, only sale days significantly impact revenue, with an average of ₹20.76 million—over double that of non-sale days (₹9.00 million) and a highly significant p-value of 0.0052. In contrast, holidays and paydays show lower revenues but are not statistically significant, indicating that sales are the primary driver of revenue spikes.

11. CONCLUSION

This report presents a data-driven approach to optimising ElectroMart's marketing budget, aligned with driving revenue growth. Our analysis began with extensive data exploration, including customer orders, media investments, and external factors like holidays and sale events. We transformed monthly investment data into daily insights using temporal disaggregation to capture daily spending fluctuations. To measure the lasting impact of marketing efforts, we applied adstock modelling to quantify the delayed and diminishing effects of advertising over time. Various modelling techniques were then deployed. An **XGBoost** regression model captured non-linear relationships between marketing spend and revenue, while the Hill function helped model saturation and diminishing returns. Complementary approaches such as **MIDAS-ARMA** and Bayesian hierarchical models added value by decomposing daily revenue into trend, seasonal, and residual components and incorporating uncertainty into channel effectiveness estimates.
Beyond the modeling efforts, the report addressed essential business metrics, incorporating **Key Performance Indicators** (KPIs), **Key Resultant Areas** (KRAs), and **Key Risk Indicators** (KRIs). These measures, product segmentation, and marketing channel analysis provided a holistic view of performance. Additionally, agent-based tools ensured that insights could be translated into actionable strategies quickly and effectively. The final optimisation, executed through Gurobi, resulted in a strategic reallocation that reduced the overall marketing spend by 11.68% while boosting revenue by 13.60%. In summary, this project lays out a scalable framework for marketing budget optimisation, which can enhance marketing efficiency and drive sustainable growth in a dynamic marketplace.

ANNEXURE

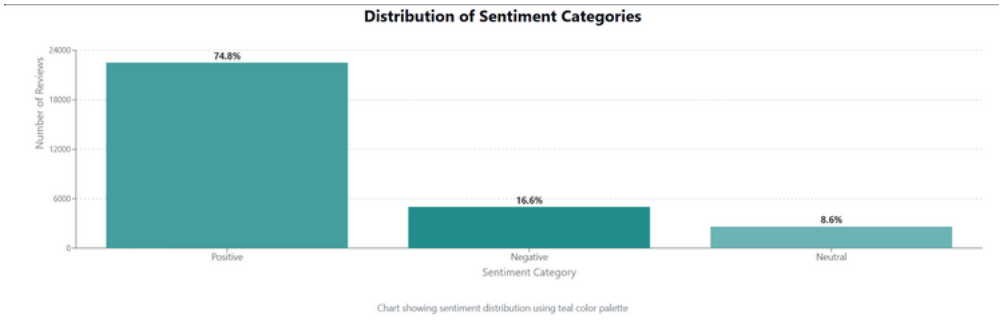
CUSTOMER SENTIMENT ANALYSIS



Customer reviews clustered by product type, not sentiment, reflecting different evaluation criteria per category.

- **Audio Enthusiasts:** Positive reviews on sound quality with occasional critiques
- **Gaming Experience:** Feedback on controllers and interactivity
- **Quick Endorsers:** Brief, mostly positive remarks
- **Quality Analyzers:** Balanced, in-depth quality assessments
- **Image Perfectionists:** Technical camera reviews focused on optical performance

Aspect	Avg. Sentiment	Total Mentions	Positive Mentions	Negative Mentions	Positive %	Negative %
price	0.659	3,492	3,175	245	90.9%	7.0%
packaging	0.637	352	312	36	88.6%	10.2%
quality	0.552	5,705	4,690	864	82.2%	15.1%
delivery	0.654	1,386	1,250	112	90.2%	8.1%
service	0.566	665	544	117	81.8%	17.6%
durability	0.503	50	43	4	86.0%	8.0%



Overall Sentiment Distribution:

74.8% of reviews are positive, 16.6% negative, and 8.6% neutral, indicating high customer satisfaction. This positive skew is typical in e-commerce, where satisfied users are more likely to leave reviews.

TECH STACK OVERVIEW - DATA ANALYTICS

Library	Version	Purpose
Python	3.13.2	Language of choice
Numpy	2.0.2	To perform mathematical operations on arrays
Pandas	2.2.3	Data handling, manipulation and analysis
Matplotlib	3.10.1	Plotting visualizations
Seaborn	0.13.2	For making statistical graphics
Geopandas	1.0.1	For plotting data on geographic maps
Plotly	6.0.1	For Data Visualization
Scikit-learn	1.6.1	For Machine Learning and statistical modelling
Selenium	4.29.0	Used for automating web browsers for scraping purposes
Beautifulsoup4	4.13.3	For pulling data out of HTML code from a website
asyncio	1.6.0	For asynchronous calling of APIs
Pymc	5.21.1	Probabilistic framework for bayesian analysis
arviz	0.21.0	Exploratory analysis of Bayesian models
langchain-core	0.3.45	For chat user experience
langgraph	0.3.11	For building Agentic Workflow
ipython	9.0.2	Interactive computing environment

TECH STACK OVERVIEW - WEB DEVELOPMENT

Library	Purpose
Next.js 15, React 19, React DOM 19	Core frontend framework.
Tailwind CSS 4, Framer Motion, clsx	UI styling & animations.
Chart.js, Recharts, D3.js	Data visualization.
Lucide React, React Icons	Icons & UI components.
Flask 3, Flask-SQLAlchemy, Flask-Migrate	API & database management.
MongoDb	ORM & database support.
TensorFlow 2, Keras, Scikit-learn, PyMC	ML & AI tools.
Axios, Requests, FastAPI (Optional)	API communication.
Vercel	Hosting & deployment.
TypeScript, ESLint, Prettier	Code quality.

AGENTS AND TOOLS

Meridian Agent

A multi-stage system that processes natural language queries by extracting key search phrases, executing web searches via DuckDuckGo, analyzing product databases, and synthesizing findings. It features iterative search refinement based on result quality, conditional routing between information sources, and comprehensive response generation that combines external information with internal product data.

Data Analysis Agent

Transforms natural language questions about data into insights by converting queries into executable pandas code, running analyses, and generating visualizations and reports. Features a robust debugging loop that identifies and fixes execution errors, translates technical results into accessible explanations, and produces professional HTML and PDF reports with embedded visualizations and formatted data.

XGBOOST AND WHY WE USED IT FOR REGRESSION

XGBoost is an advanced gradient boosting implementation that builds sequential decision trees to predict continuous target variables. The algorithm starts with a simple prediction (typically the mean of the target) and iteratively adds trees that correct previous errors. For each iteration, XGBoost fits a new tree by minimizing an objective function:

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t)$$

Where L is the loss function and Ω is a regularization term. XGBoost approximates this using a second-order Taylor expansion involving gradients (first derivatives) and Hessians (second derivatives).
The algorithm builds trees using a greedy approach that evaluates potential splits based on a gain score that balances improvement in the objective function against added complexity. Key technical features include:

- L1/L2 regularization to control complexity
- A learning rate (shrinkage) to scale tree contributions
- Efficient handling of missing values
- Column and row subsampling to reduce overfitting
- Parallel processing for computational efficiency

For each leaf in the final trees, XGBoost calculates optimal weights to minimize the objective function. The final prediction is the sum of all tree outputs, creating a powerful ensemble model that excels at capturing complex relationships in structured data while maintaining good generalization performance.

KEY PERFORMANCE INDICATORS

CATEGORY	KPI	Formula
Revenue	Average Order Value	Total Revenue/Number of Orders
Revenue	Revenue per Customer	GMV/Number of Unique Customers
Marketing	Return on Ad Spend (ROAS)	Revenue Attributed to Advertising/Ad Spend
Marketing	Sales Growth Rate	Revenue increase/ Revenue previous period
Marketing	Payday Sales Uplift Index	(Payday Week GMV - Non-payday week)/Non-payday week GMV
Operational	Late Delivery Rate	Orders Delivered After SLA/Total Orders
Seasonality	Seasonal Revenue Impact (SRI)	(Avg Revenue Current Week - Avg Revenue Off-Peak)/total orders
External Factors	Holiday Sales Boost	(Avg Holiday Revenue - Avg Daily Revenue)/Avg Daily Revenue

KEY RESULTANT AREAS

KRA	EXPLANATION
Product Portfolio Strategy	Ensuring the right mix of products to meet market demand and drive profitability.
Customer Experience Impact	Measuring how product and service delivery influence customer satisfaction.
Marketing Channel Strategic Alignment	Aligning marketing efforts with the most effective customer touchpoints.
Customer Retention Strategy	Initiatives aimed at increasing repeat purchases and loyalty.
External Market Factor Responsiveness	The business's agility in reacting to market trends and disruptions.
Marketing Effectiveness	Evaluating ROI and performance of marketing initiatives.

KEY RISK INDICATORS

KRIs	EXPLANATION
Average Order Value Fluctuations	Monitoring inconsistency in spending per order to detect pricing or demand risks.
Customer Acquisition Cost/CMV	Tracking rising costs in gaining new customers, which can reduce margins.
Delivery Delay Percentage	Identifying fulfillment bottlenecks impacting customer trust.
Discount Depth Risk	Measuring over-reliance on deep discounts that could hurt profitability.
Weather Impact Sensitivity	Understanding how climatic factors influence sales and operations.
Channel Dependency Risk	Assessing over-dependence on single sales or marketing channels.Key Resultant Areas: