

2025112

INDEX

PROBLEM DESCRIPTION

DATASET DESCRIPTION

DATA PRE-PROCESSING

DATA MERGING/ FEATURE ENGINEERING

DATA VISUALISATION

IMPACT ANALYSIS

RESULT

RISK ANALYSIS

KEY PERFORMANCE INDICATORS

MARKETING MIXED MODELLING WITH ROBYN

OPTIMISATION APPROACH

OPTIMISATION PROBLEM FORMULATION

OPTIMAL BUDGET ALLOCATION

FUTURE IMPROVEMENTS

BUSINESS STRATEGY

ANNEXURE

- Literature Referred
- Objective Functions- ROBYN
- Model Specification for Ridge Regression
- Key Performance Indicators

PROBLEM DESCRIPTION



BACKGROUND

ElectroMart (Ontario, Canada) aims to optimize its **marketing budget** after a significant spend last year. The CFO is concerned about its impact on revenue. The task **involves analyzing marketing variables, assessing ROI, and recommending budget allocation**. Available data includes customer orders, media investments, sales calendars, NPS scores, product hierarchies, and weather data. The goal is to identify **KPIs driving revenue, maximize ROI**, and determine **optimal product & channel strategies**. **Deliverables** include **presentations, code files, and dashboards** for insights and recommendations.

OBJECTIVES



- **Performance Driver Analysis:** Identify key KPIs driving revenue growth.
- **Marketing ROI Impact:** Quantify the effect of each commercial lever on revenue.
- **Budget Optimization:** Allocate marketing spend for maximum returns next year.
- **Product Targeting:** Select focus products/categories for campaigns with rationale.
- **Channel Strategy:** Determine optimal marketing channels per category with justification.

TECH STACK USED

Programming Languages: Python, R; Libraries and Frameworks: NumPy, Pandas, Matplotlib, Seaborn, Scipy, Scikit-learn, Robyn, GurobiPy; Tools: Google Colab, Jupyter Notebook, Git, GitHub.

DATASET DESCRIPTION



 MAIN CONSUMER DATA: <p>Analysis of July 2023–June 2024 order data provides insights into consumer behavior, product performance, and operations. The dataset includes FSN ID, Order ID, Customer ID, delivery performance, GMV, units sold, payment types, and procurement SLA for time-based evaluation.</p>	 MARKETING AND SALES DATA: <p>The dataset tracks marketing investments across advertising mediums, including campaign calendars, NPS metrics, and stock values. It enables correlation analysis between customer satisfaction, marketing strategies, and business performance, providing insights into promotion effectiveness over time.</p>
 PRODUCT INFORMATION: <p>The SKU Mapping File details product hierarchy with FSN ID, Super Category, Category, Sub-Category, and Vertical. This taxonomy enables in-depth product performance analysis and category-level insights for strategic decisions.</p>	 DATA GENERATION: <p>The dataset includes detailed meteorological reports for Ontario, Canada, and a comprehensive calendar of Canadian holidays. These contextual parameters support seasonal pattern analysis and enhance location-specific predictive modelling.</p>

DATA PRE-PROCESSING



Dataset Overview & GMV Imputation

The dataset comprises 4,900 consumer records, including GMV (Gross Merchandise Value) data, with some missing values. These were imputed using the median GMV of the respective product vertical category for the same month, ensuring alignment with market trends. Given the dataset spans 21,000 unique product verticals, this aggregation-based imputation strategy maintains data consistency and analytical reliability.



Weather Data Retrieval & Integration

The Meteostat API follows established data processing standards, including rigorous data cleaning, transformation, and integration procedures. These measures ensure that the weather data retrieved aligns with industry best practices, reducing inconsistencies and enhancing data quality. By using external data sources alongside GMV imputation, this preprocessing strategy ensures that the final dataset is robust, reliable, and ready for further analysis.



Data Cleaning & Validation

Weather data gaps were addressed by sourcing from the Meteostat API, which aggregates reliable meteorological data from institutions like NOAA and DWD. This ensures the integrated weather data is accurate, well-maintained, and suitable for analysis.

STATISTICAL TESTING OR HYPOTHESIS TESTING

Null Hypothesis: GMV follows Weibul Distribution			Aggregation & Market Segmentation	
Metric	Kolmogorov-Smirnov (K-S) test	Log-Transformed Data	Metric	Value
K-S Statistic	0.999	0.999	F-Static	28.07
p-value	0.0	0.0	p-Value (uncorrected)	8.87e-14
Skewness: 139.96 Kurtosis: 59502.904			p-Value (GG-corrected)	2.86e-7
			Generalized η^2	0.552
			Sphericity Assumption	Violated

Conclusion

- GMV data demonstrates extreme non-normality, justifying Weibull distribution as there is strong evidence against Null Hypothesis.
- Product-based market segmentation shows statistically significant explanatory power ($F(4,68)=28.07$, $p<0.001$) in aggregated temporal analysis



Three time-frequency-based datasets were created to align with analytical requirements:

Hourly Dataset	
Base	Customer order data (hourly frequency)
Key mergers	Mapped product hierarchy (category/sub-category/vertical) using FSN_ID from SKU metadata
Added feature	Added discount percentage feature: $\text{Discount\%} = 100 \times (\text{Product_MRP} - \text{List_Price}) / \text{Product_MRP}$

Daily Dataset	
Aggregation	Hourly data → Daily resolution
Methodology	<ul style="list-style-type: none">Supply chain features from hourly data are averaged daily (sla, delivery_cdays, delivery_bdays, procurment_sla)Continuous operational features are better represented by central tendency over 24hr cyclesAs validated by hypothesis testing, revenue was distributed into columns by product category while retaining total daily GMV.
Key mergers	<ul style="list-style-type: none">Sale_Day: Binary flag for promotional daysHoliday: Binary flag for public/national holidaysIntegrated weather data from pre-processed weather data which was already daily in nature.

Monthly Dataset	
Aggregation	Daily category-segmented revenue → Monthly totals
Methodology	As validated by hypothesis testing, revenue was distributed into columns by product category while retaining total daily GMV.
Key mergers	<ul style="list-style-type: none">Channel-wise marketing investments (TV, Digital, OOH)Net Promoter Score (NPS) trendsStock index performance data

Design Rationale

- Hierarchical Structure:** Hourly → Daily → Monthly alignment enables granular-to-macro analysis.
- Hypothesis-Driven:** Product category segmentation ensures revenue attribution aligns with market dynamics.
- Operational Context:** Averaged supply-chain metrics preserve daily operational realities without distortion from hourly spikes.

DATA VISUALISATION



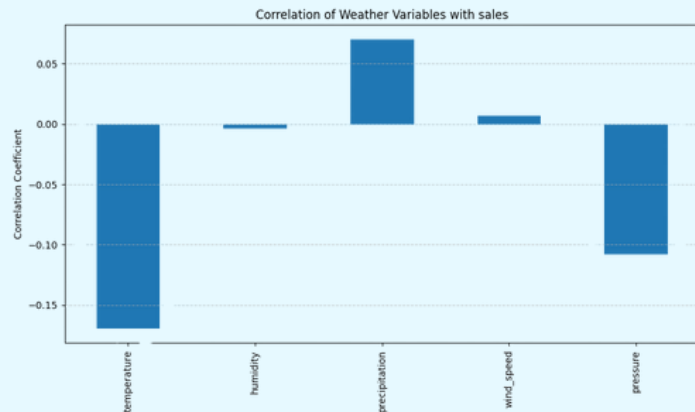
Weather Influence

Analysis indicates that weather variables minimally impact sales:

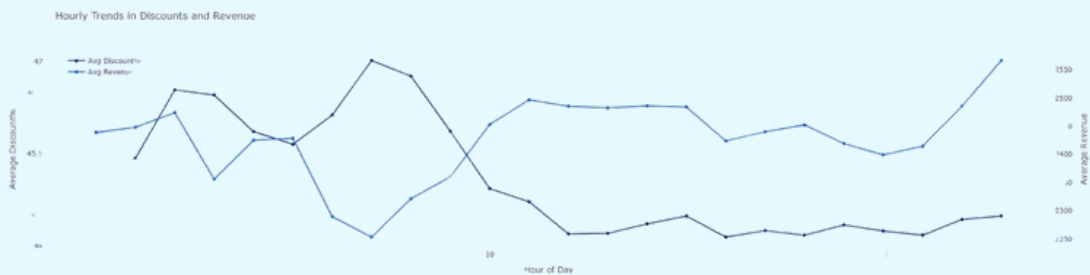
- Temperature and pressure: Weak negative correlations with sales (approximately -0.15), suggesting slight decreases in sales as these variables rise.

- Humidity, precipitation, and wind speed: Correlations near zero, indicating negligible relationships with sales.

These findings imply that weather conditions do not significantly influence sales variations; thus, incorporating weather variables into predictive sales models may offer limited value.



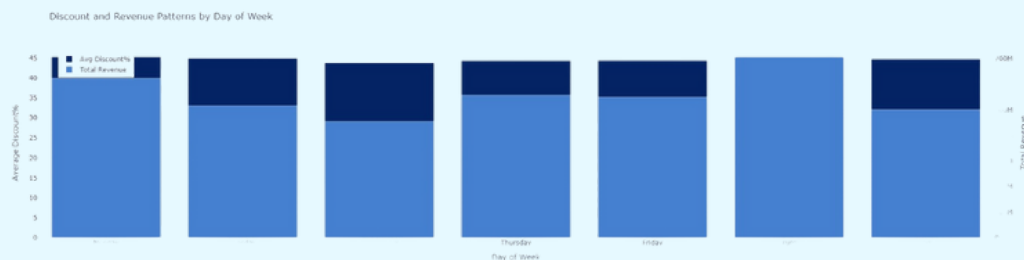
Discount-GMV Hourly Trend



The second graph shows daily fluctuations in discounts and revenue. Discounts peak in the early morning and late evening, while revenue spikes at 4–5 AM and late night, dipping midday. This suggests customer behavior impacts sales more than discounts alone.



Discount-Revenue Patterns

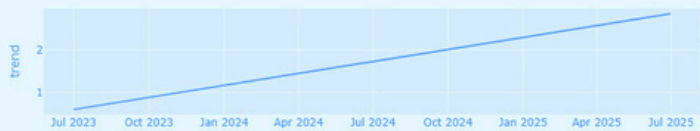


The third chart shows that average discounts remain steady at 40–45% throughout the week, while total revenue peaks on Saturdays and Fridays and dips slightly on Sundays and weekdays like Monday. This indicates that higher weekend sales are likely due to increased consumer activity, rather than varying discount levels.

DATA VISUALISATION
FORECASTED ANALYSIS



Trend Analysis



GMV shows a steady upward trend from 0.5 in mid-2023 to a forecasted value of 2.5 by mid-2025. This growth is likely driven by market expansion, improved customer engagement, and stronger marketing efforts.

Holiday Effects



Holidays have a cyclical impact on GMV, with noticeable spikes during festive periods like Christmas and New Year, driving increased consumer activity and sales. Conversely,

some holidays, such as observed at late 2023 and late 2024, show dips, likely due to reduced operational activity or shifts in consumer behavior.

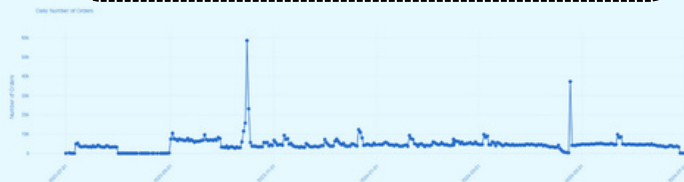
Seasonality



GMV follows a cyclical pattern, with peaks in November and early January, driven by major shopping sales.

Some holidays, such can be seen at late 2023 and late 2024, show dips, likely due to reduced operational activity or shifts in consumer behavior.

Number of order fluctuations

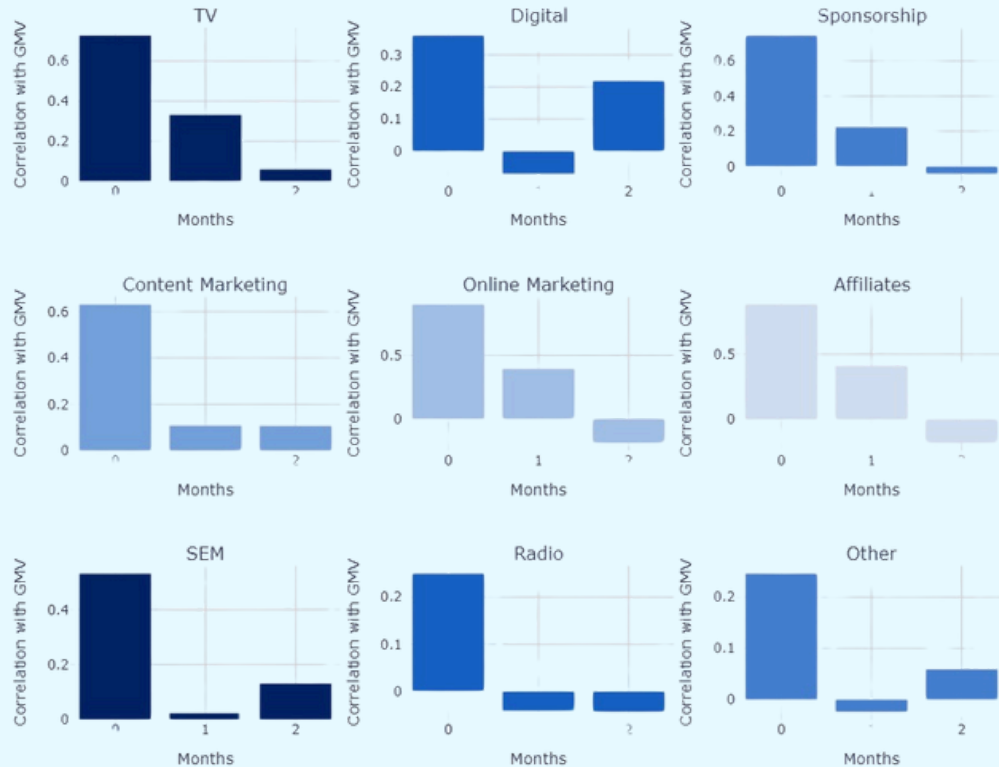


Daily orders saw major spikes in Nov 2023 and May 2024 , which is due to seasonal sales or promotions. Outside these peaks, orders remained stable between 5,000-10,000. The post-spike declines suggest these sales leads temporary surges not sustained growth.



CROSS-CORRELATION ANALYSIS

Impact Lag for Marketing Channels



Computing cross-correlation between monthly marketing expenditures and GMV helps capture the temporal impact of spending, distributed across multiple months:

- **Lag 0 (Same Month Effect):** The immediate effect of spending on GMV within the same month.
- **Lag 1 (One-Month Delay):** The effect of spending in a given month on the following month's GMV.
- **Lag 2 (Two-Month Delay):** The effect of spending in a given month on GMV two months later.

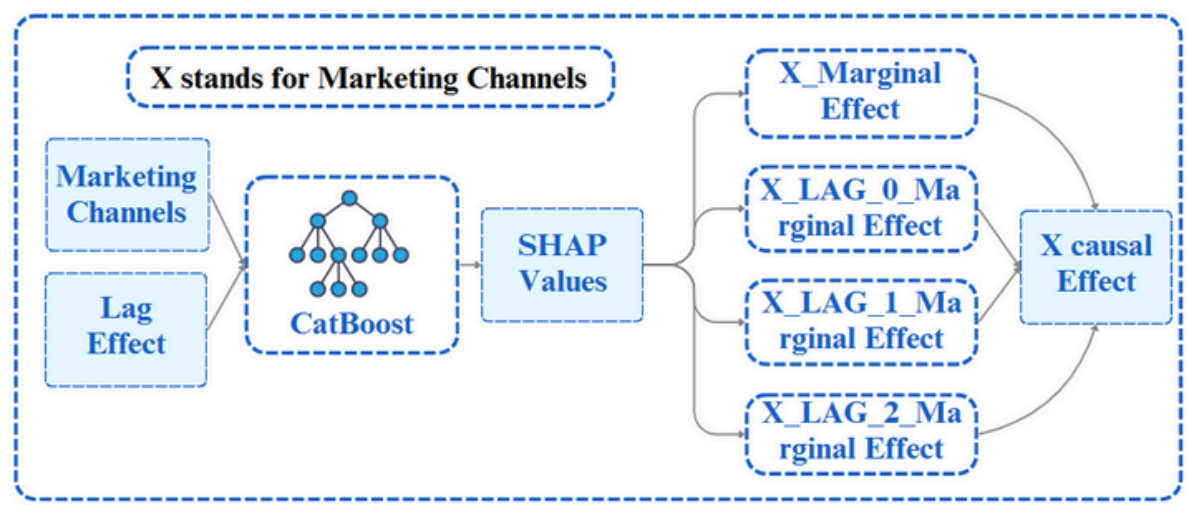
The resulting correlation values generalize the lagged impact across the year. Hence, these correlation values are used as 'lag weights'.

QUANTITATIVE IMPACT ON REVENUE

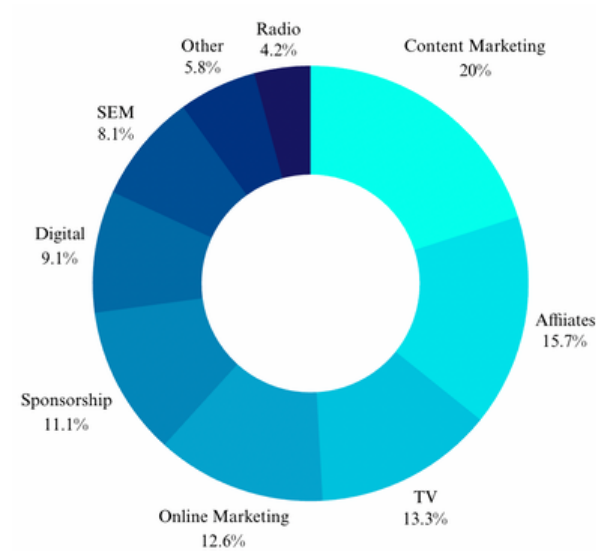
Lag Feature Engineering: Using the derived 'lag weights', the actual investment cost for each marketing channel is distributed into lagged components. This ensures that the observed impact of each investment reflects its multi-month influence, rather than attributing all effects solely to the month of spending.

Modeling and Feature Importance using SHAP: The lagged features, along with the original data, are incorporated into a CatBoost model. The SHAP (Shapley Additive Explanations) framework is then employed to quantify the contribution of each feature to the model's predictions in percentage terms.

IMPACT ANALYSIS

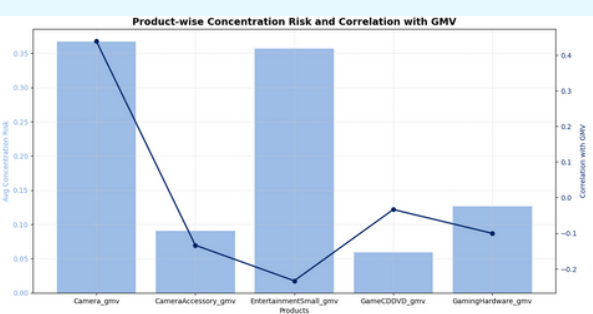


RESULTS

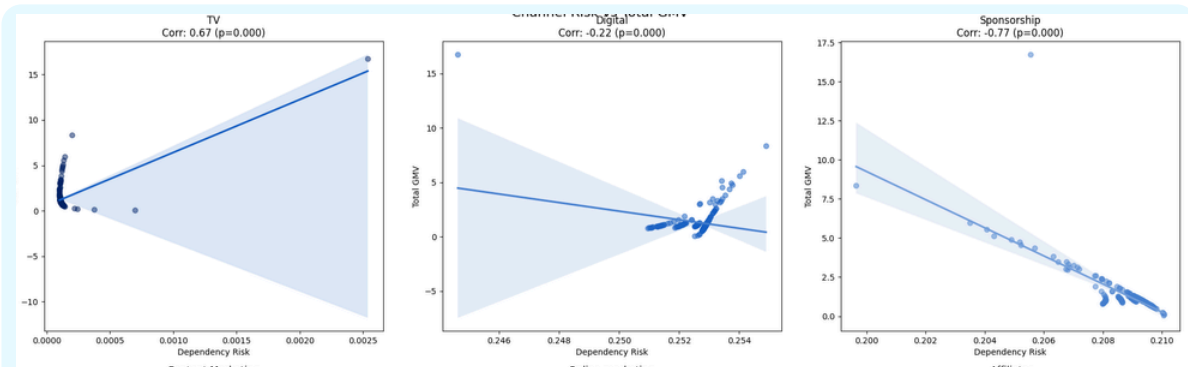


Deriving Causal Impact: Since the SHAP-based contributions represent marginal effects, we sum the SHAP values across all lag features for each channel to approximate the true causal effect. The final quantized percentage contribution, as presented in the results table, represents the estimated causal effect of each marketing investment on revenue. This approach ensures a robust and interpretable measure of marketing effectiveness by integrating statistical correlation,

RISK ANALYSIS



Dependency Measurement: Identifies over-reliance on specific product categories.
Revenue Impact: Negative correlation with GMV suggests that over-concentration may limit revenue potential.
Strategic Implication: A diversified product portfolio can mitigate risks and balance revenue streams.



Risk Share Calculation:

Each channel's dependency risk is measured using the risk share: $r_i = \frac{\text{Investment}_t}{\text{Total Investment}_t}$

The overall risk score is then computed as: Overall Risk Score = $\sum r_i^2$ which follows the Herfindahl-Hirschman Index (HHI) methodology to quantify concentration risk.

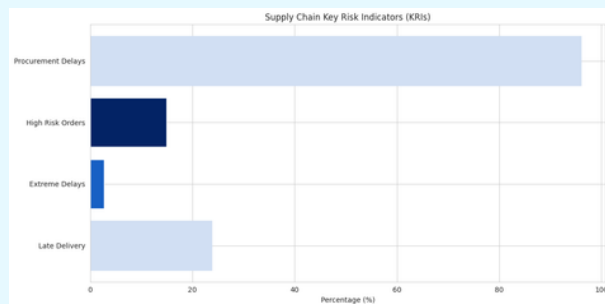
Significant Results:

- TV (Moderate Risk) → Balanced spending but fluctuating impact on revenue. Requires periodic optimization.
- Digital (Stable) → Low risk, consistent returns, strong correlation with revenue. Maintain strategy.
- Sponsorship (Stable) → Predictable performance, minimal risk. Continue investment with periodic reviews.

KEY PERFORMANCE INDICATORS

Supply Chain key Risk Indicators

$$\text{Supply Chain Efficiency} = \frac{\text{SLA Compliance} \times \text{Product Procurement SLA}}{\text{Delivery B Days} + \text{Delivery C Days}}$$



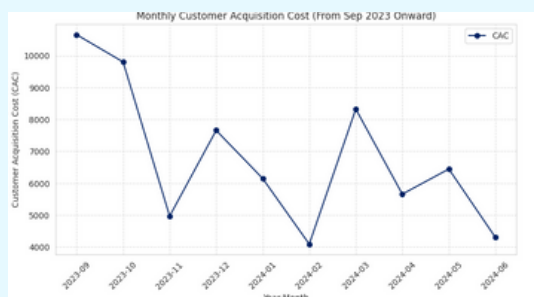
Return on Ad-Spend

$$ROAS = \frac{\text{Revenue from Ads}}{\text{Cost of Ads}}$$



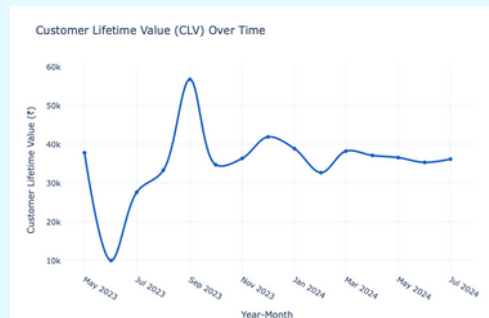
Customer Acquisition Cost

$$CAC = \frac{\text{Total Sales and Marketing Expenses}}{\text{Number of New Customers Acquired}}$$



Customer Lifetime Value

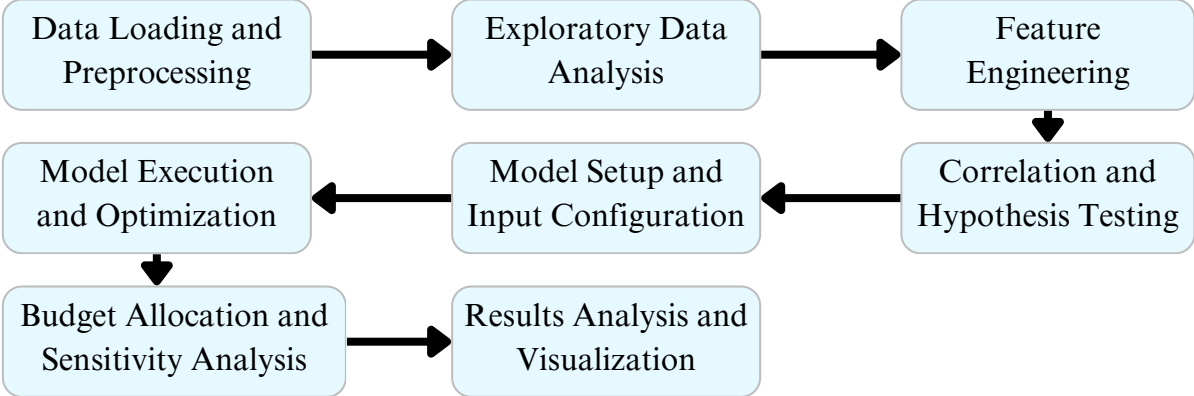
$$CLV = AOV \times \text{Purchase Frequency} \times \text{Customer Lifespan}$$



MARKETING MIXED MODELLING
WITH ROBYN



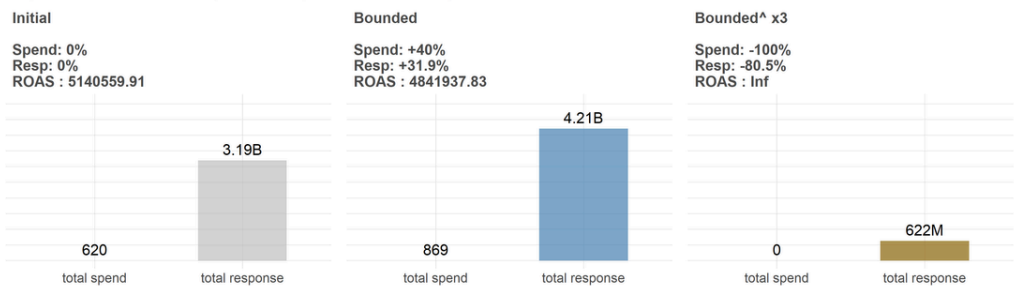
Robyn MMM Process Flowchart



One Pager Report

Budget Allocation Onepager for Model ID 1_190_4
Adj.R2: train = 0.7055, val = 1, test = 0.8231 | NRMSE: train = 0.0974, val = 5e-04, test = 0.1288 | DECOMP.RSSD = 0.0063 | MAPE = NA
Simulation date range: 2023-08-01 to 2024-07-01 (12 months) | Scenario: max_response

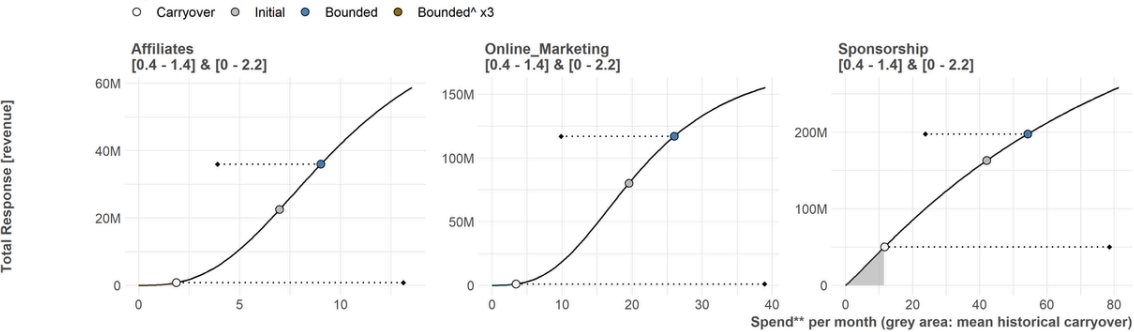
Total Budget Optimization Result (scaled up to 12 months)



Budget Allocation per Paid Media Variable per Month*

		Initial					Bounded					Bounded^ x3				
Paid Media	Affiliates	5.1	9.9%	8.5%	4.4M	6.64M	7.2	9.9%	10.3%	5.02M	6.11M	0	0%	1.5%	0	1.74M
	Online_Marketing	16	31.2%	30.2%	4.97M	6.52M	23	31.2%	33.4%	5.18M	4.49M	0	0%	1.8%	0	1.01M
	Sponsorship	30	58.9%	61.3%	5.35M	3.03M	43	58.9%	56.4%	4.63M	2.61M	0	0%	96.7%	0	4.34M
		abs.mean spend	mean spend%	mean response%	mean ROAS	mROAS	abs.mean spend	mean spend%	mean response%	mean ROAS	mROAS	abs.mean spend	mean spend%	mean response%	mean ROAS	mROAS

Simulated Response Curve per Month

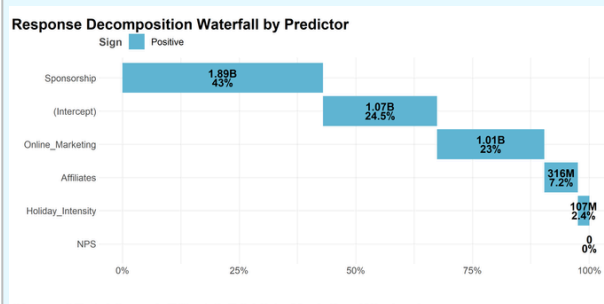


MARKETING MIXED MODELING WITH ROBYN



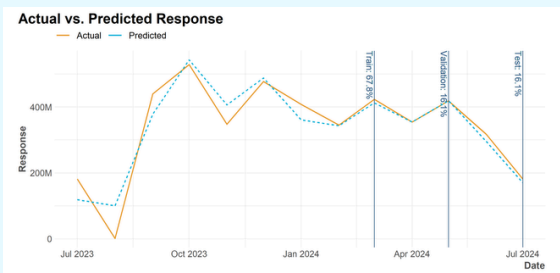
Response Decomposition Waterfall by Predictor

This chart breaks down the contribution of each predictor to revenue performance. This helps identify key drivers for top-line performance.



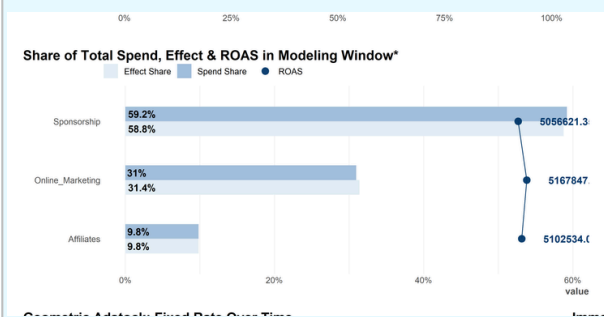
Actual vs. Predicted Response

This line chart compares actual revenue with predicted revenue over time.



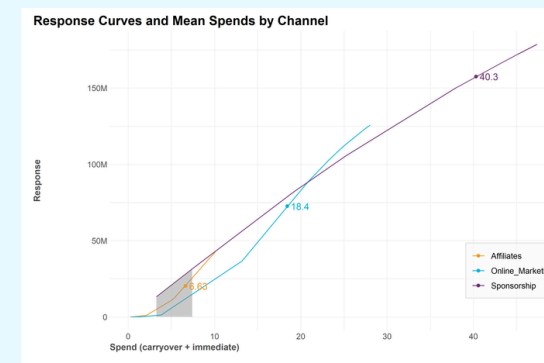
Share of Total Spend, Effect & ROAS in Modeling Window

This bar chart shows the share of spend, effect size, and ROAS for each channel during the modeling period. This guides decisions on optimizing marketing budgets for maximum ROI.

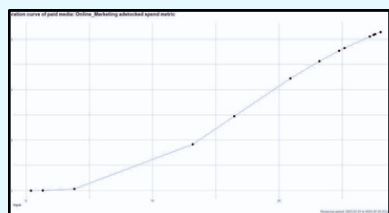


Response Curves and Mean Spends by Channel

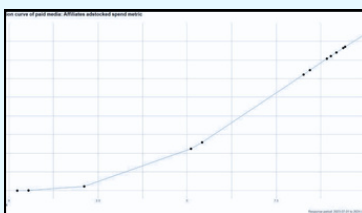
This curve plots the relationship between spend and response for each channel.



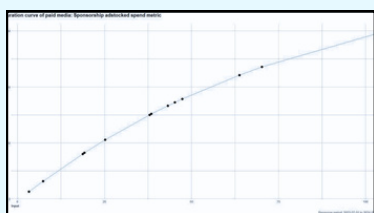
Saturation Curve Analysis- Diminishing Returns Of Channel Revenue



Online Marketing



Affiliates



Sponsorship

OPTIMISATION APPROACH



This marketing mix optimization aims to maximize overall revenue by optimally allocating the marketing budget across different marketing channels and product categories, subject to constraints on spending changes, total budget, and channel-specific limits. The optimization is modelled as a Mixed-Integer Second-Order Cone Programming (MISOCP) problem and solved using Gurobi.

SPLINE INTERPOLATION FOR DAILY DATA

Cubic spline interpolation is applied to convert monthly marketing spend data into daily granularity while ensuring smooth transitions and preserving total monthly spending. Each segment between two months is modelled as a cubic polynomial:

$$f_m(t) = a_m(t - t_m)^3 + b_m(t - t_m)^2 + c_m(t - t_m) + d_m$$

where coefficients are determined by enforcing interpolation and total spend preservation:

$$f_m(t_m) = S_m \quad \int_{t_m}^{t_{m+1}} f(t)dt = S_m$$

RESPONSE PARAMETER ESTIMATION

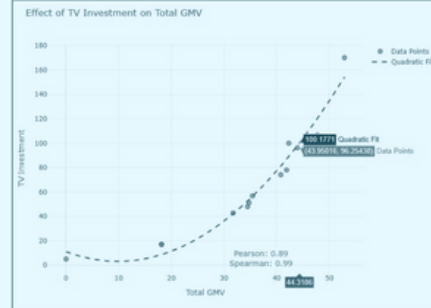
Ridge regression is employed to estimate the response parameters in our marketing mix model, this helps to mitigate multicollinearity issues while preventing over-fitting. For each marketing channel i , we estimate quadratic response parameters (θ_i , ϕ_i , ψ_i) using ridge regression:

$$Revenue_i = \theta_i s_i^2 + \phi_i s_i + \psi_i + \sum_{k \in \text{features}} \beta_k F_k$$

Since GMV's dependence on marketing channel spends follows a quadratic fit, so we proposed a quadratic objective function for optimisation problem. To account for differential impact across product categories, we estimate separate response parameters for each product-channel combination. For each product category p and channel i :

$$Revenue_{p,i} = \theta_{p,i} s_i^2 + \phi_{p,i} s_i + \psi_{p,i} + \sum_{k \in \text{features}} \beta_{p,k} F_k$$

The model integrates SLAs, sales, NPS, stock, holidays, weather, and product discounts to optimize revenue estimation as external features.



OBJECTIVE FUNCTION

The objective is to maximize the total revenue across all product categories and channels, incorporating extra features:

$$\max \sum_{i \in \text{channels}} \left(\theta_i (s_i^{\text{total}})^2 + \phi_i (s_i^{\text{total}}) + \psi_i + \sum_{k \in \text{features}} \beta_k F_k \right)$$

OPTIMISATION PROBLEM FORMULATION



DECISION VARIABLES

For each marketing channel i :

- x_i : Change in marketing spend from baseline s_i
- $s_i^{\text{total}} = s_i + x_i$: Total spend after optimization.
- $z_{L,i}, z_{R,i}$: Binary variables indicating decrease or increase
- $z_{LR,i}$: Binary variable indicating any spend change

Here, σ_i is the standard deviation and μ_i is the mean marketing spend for channel i

RISK FACTOR CALCULATION

The risk factor for each marketing channel i is calculated based on the measurement of volatility of spend relative to its mean:

$$\text{Risk Factor}_i = \frac{\sigma_i}{\mu_i}$$

Risk Level	Normal Period	Strong Sales Period
Very High Risk	(0.90, 1.10)	(0.85, 1.15)
High Risk	(0.80, 1.20)	(0.75, 1.25)
Medium Risk	(0.70, 1.30)	(0.65, 1.35)
Low Risk	(0.60, 1.40)	(0.55, 1.45)
Very Low Risk	(0.45, 1.55)	(0.40, 1.60)

Risk-based scaling factors for channel bounds

A higher risk score indicates greater variability in marketing spend and is associated with tighter budget constraints.

CONSTRAINTS

Budget Limit Constraint:

$$\sum_{i \in \text{channels}} x_i \leq (\text{budget_limit_pct} - 1) \times \sum_{i \in \text{channels}} s_i$$

Spend Adjustment Bounds:

$$l_i \leq s_i^{\text{total}} \leq u_i, \forall i \in \text{channels}$$

Binary Spend Adjustment Constraints:

$$l_i^L \cdot z_{L,i} + l_i^R \cdot z_{R,i} \leq x_i \leq u_i^L \cdot z_{L,i} + u_i^R \cdot z_{R,i}, \forall i \in \text{channels}$$

$$z_{L,i} + z_{R,i} \leq 1, \forall i \in \text{channels}$$

Maximum Changes Constraint:

$$\sum_{i \in \text{channels}} z_{LR,i} \leq \text{max_changes}$$

ITERATIVE IMPROVEMENT

1. Use historical data to estimate response parameters across marketing channels and product categories.
2. Solve the optimization problem using past allocations to determine the optimal spend.
3. Update the dataset, using each optimized allocation as the next period's baseline for continuous refinement.

REVENUE IMPROVEMENT

The percentage revenue improvement per product category is then given by:

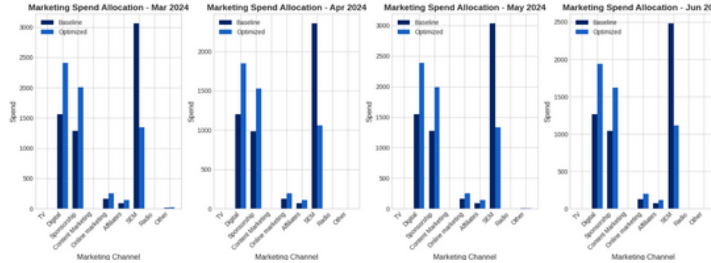
$$\frac{\text{Optimized Revenue} - \text{Baseline Revenue}}{\text{Baseline Revenue}} \times 100\%$$

For baseline GMV, historical records of past allocations are utilised.

$$\text{Revenue}_p = \sum_{i \in \text{channels}} \left(\theta_{p,i} (s_i^{\text{total}})^2 + \phi_{p,i} (s_i^{\text{total}}) + \psi_{p,i} + \sum_{k \in \text{features}} \beta_{p,k} F_k \right)$$

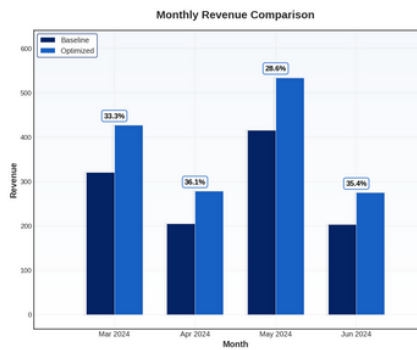
Similarly percentage revenue improvement of each product is also included

OPTIMAL BUDGET ALLOCATION



The optimization results show significant reallocation of budget across channels, with highest percentage increases in the most efficient channels and decreases in underperforming ones.

MONTHLY REVENUE COMPARISON



REVENUE ACROSS EACH PRODUCT CHANNEL



IMPACT OF CHANNEL SPENDS ON PRODUCT REVENUE

We fit XGBoost for each product segment under its budget, using historical sales and marketing channels as inputs. This identifies the most impactful channels for each segment.

XGBoost's Prediction Model:

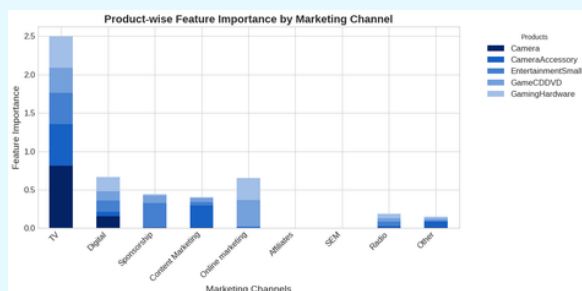
$$\hat{y}_p = f_p(X; \theta_p)$$

\hat{y}_p : predicted GMV

θ_p : model parameters related to product segments

$$\text{Gain}(F) = \frac{1}{T} \sum_{t=1}^T \sum_{n \in \text{nodes in tree } t} \text{Gain}(F, n)$$

Gain(F) measures a feature's contribution to improving predictions. These help pinpoint the most effective marketing channels for each segment

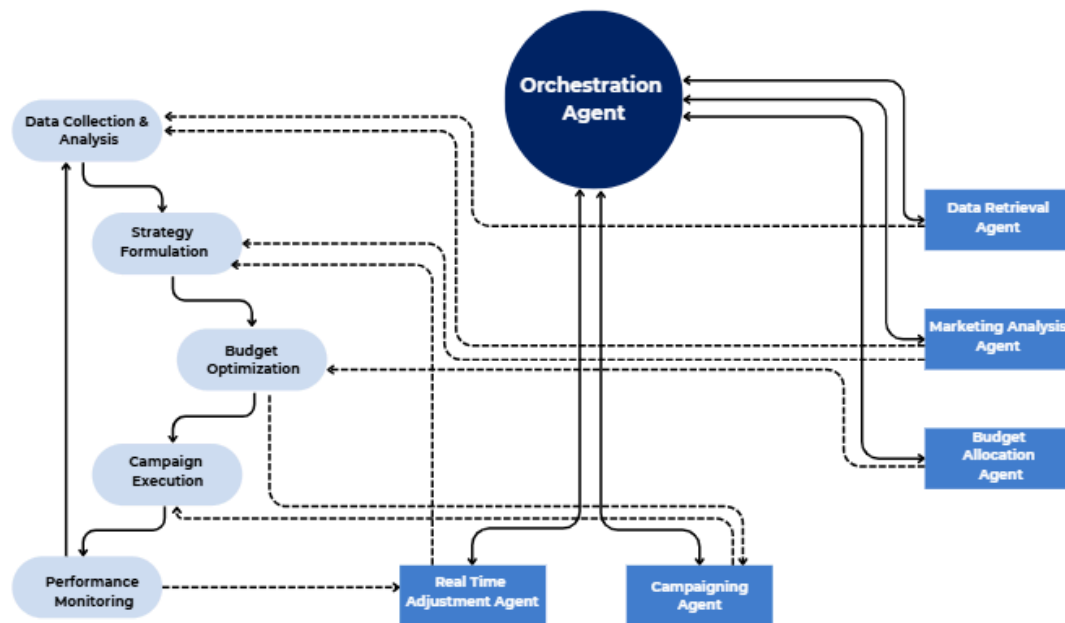


TV boosts Cameras & GamingHardware, Digital/Online aid GamingHardware & GameCDDVD, Affiliates impact GameCDDVD, while Sponsorship/Content help Camera Accessories & EntertainmentSmall. Radio & Other have little effect.

FUTURE IMPROVEMENT



AGENTIC FRAMEWORK



WHAT AGENTIC ?

Agentic refers to an intelligent, autonomous system powered by a network of specialized AI agents, orchestrated to collaboratively optimize budget allocation. Each agent focuses on a specific task—such as data analysis, strategy formulation, budget management, campaign execution, or real-time adjustments—working together to dynamically adapt to evolving conditions and deliver efficient, data-driven outcomes.

WHY AGENTIC ?

The framework excels in scenarios requiring multi-dimensional trade-offs (e.g., balancing cost vs sustainability vs compliance), offering:

Scalability: Agents handle large datasets and real-time updates efficiently.

Specialization: Each agent focuses on specific KPIs, ensuring high accuracy in domain-specific analysis.

Automation: Reduces manual intervention by autonomously integrating insights into a unified scoring system.

Dynamic Adaptation: Responds to evolving regulations or market conditions without reprogramming.

The Orchestrator agent centrally manages interactions with users and other agents. It retrieves historical data via the Data Retrieval Agent and gains insights from the Marketing-Analysis Agent. The Budget Allocation Agent uses these insights along with web and academic resources to plan strategies. The Campaigning Agent handles media and outreach, while the Real-Time Adjustment Agent monitors performance and updates strategies. The Orchestrator oversees all communication between agents.



SWOT ANALYSIS

Strength

- High-Performing Marketing Channels: TV and content marketing drive GMV; optimize further.
- Improved Customer Acquisition Efficiency: Lower CAC enhances profitability.
- Positive ROAS Recovery: Stabilized post-Aug 2023, with spikes in March & June 2024.

Weaknesses

- Channel Dependency Risk: Overreliance on SEM and Sponsorship leads to GMV fluctuations.
- Product Concentration Risk: Dependence on specific categories like cameras and gaming limits revenue growth.
- Supply Chain Inefficiencies: Procurement delays and high-risk orders negatively impact customer experience

Opportunities

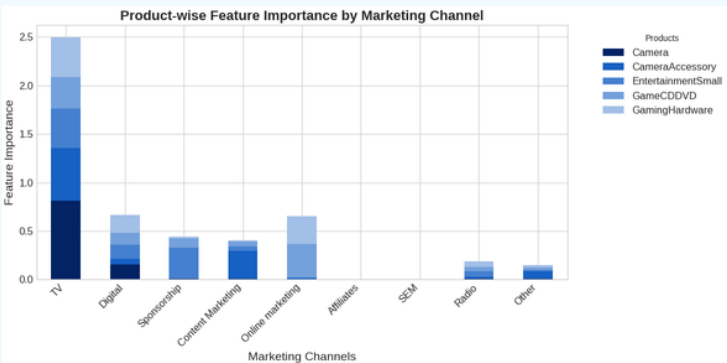
- Seasonal Sales Boost: Targeted strategies can increase sales by 35% in peak seasons.
- Diversified Portfolio: Expanding products improves margins by 20%.
- Weather-Based Marketing: Bad weather boosts engagement, enabling adaptive strategies.

Threats

- Declining CLV Trend: Post-May 2024 decline suggests retention challenges.
- Raw Material Price Volatility: 18% cost increase due to supply chain disruptions impacts margins.
- Competitive Landscape Evolution: Global competition, inflation, and labor shortages pose strategic risks.

PROPOSED STRATEGY

- **Maximize TV & Digital** – Allocate the highest budget to TV (Feature Importance: 2.5) and high-ROAS digital ads to drive GMV growth.
- **Optimize Sponsorships & Affiliates** – Strengthen influencer partnerships and affiliate marketing (Feature Importance: ~0.6) to diversify acquisition and mitigate channel dependency risk.



- **Cut Low-ROI Channels & Boost Retention** – Reduce SEM & Radio spend (Low Feature Importance <0.2) while enhancing CLV through loyalty programs and weather-triggered campaigns (42% engagement lift).

ANNEXURE

LITERATURE REFERRED



The article **A Unified Framework for Marketing Budget Allocation (Kui Zhao, Junhao Hua, Ling Yan, Qi Zhang, Huan Xu, Cheng Yang)** introduces a two-stage, data-driven approach to optimize marketing budgets in dynamic online environments. It combines a semi-black-box model (enhancing logit demand curves with neural networks) to predict market reactions and optimization techniques to allocate budgets while meeting constraints like ROI targets and expenditure limits. Validated through Alibaba's real-world use and A/B testing, the method is scalable and effective for balancing performance goals with business rules in both continuous and discrete budget allocation scenarios.

The paper **A Nonlinear Optimization Model of Advertising Budget Allocation (Sung-Hyuk Park, Kitae Kim, Minhyung Lee, Dongwook Shin)** across Multiple Digital Media Channels presents a nonlinear model to optimize advertising budgets across multiple digital media channels, aiming to improve performance (e.g., conversions or revenue) under budget constraints. Using data from a 12-channel e-commerce business and a logarithmic function to account for diminishing returns, the model—solved with the MOSEK solver—achieved up to 33.97% performance gains over traditional methods. While effective, the study is limited by its focus on one company's data and lack of consideration for seasonality or inter-channel relationships, which future research aims to address.

The blog **Open Source Battle for MMM: Robyn vs LightweightMMM (Andy Kozak)** compares Meta's Robyn and Google's LightweightMMM, two open-source Marketing Mix Modeling tools. Robyn uses R with a frequentist approach using Prophet's Generalized Additive Models, while LightweightMMM uses Python's JAX for Bayesian modeling. Robyn demonstrates strength in documentation and community support, while LightweightMMM offers better code quality and GPU acceleration. Both tools contribute to marketing science, with Robyn slightly ahead in democratization.

Links

- [1] <https://arxiv.org/pdf/1902.01128>
- [2] <https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/ea6310e7-8ef8-4335-9625-b82b11042797/content>
- [3] <https://www.forvio.com/resources/blog/open-source-battle-for-mmm-robyn-vs-lightweightmmm>

OBJECTIVE FUNCTIONS- ROBYN



Robyn implements three objective functions as the goals for hyperparameter optimization:

- **NRMSE (Normalized Root Mean Square Error):** Also referred to as the prediction error. Robyn allows time-series validation by splitting the dataset into train, validation, and test sets. When fitting without time-series validation, the training error (`nrmsetrain`) is the objective function for evolving iterations. With time-series validation, the validation error (`nrmseval`) becomes the objective function, while `nrmsetest` is used to assess out-of-sample predictive performance.

- **DECOMP.RSSD (Decomposition Root Sum of Squared Distance):**

Also referred to as the business error and a key innovation of Robyn. It represents the difference between share of spend and share of effect for paid media variables. This metric is controversial due to media ROAS convergence, but it helps rule out models with extreme decomposition and aids in narrowing down model selection.

- **MAPE.LIFT (Mean Absolute Percentage Error for experiments):**

Activated during calibration, this is known as the calibration error. It minimizes the difference between predicted effect and causal effect, making it another key innovation of Robyn.

Using CatBoost and SHAP for Categorical Features

CatBoost simplifies SHAP interpretation by handling categorical features natively, avoiding the complexity of one-hot encoding. Unlike other models, where SHAP values must be manually aggregated, CatBoost provides a single SHAP value per categorical feature, making insights clearer.

Key Takeaways:

- **CatBoost** handles categorical features without transformation.
- **SHAP** values are easier to interpret compared to one-hot encoding.
- Limitations exist in SHAP aggregations for categorical features.
- **Beeswarm plots** help visualize individual feature impact.

Model Specification for Ridge Regression



MMM uses regression modeling, which aims to derive an equation that describes the dependent variable. The model aims to assign a coefficient to each independent variable, where only the variables that are statistically significant stay in the model.

In very simple terms, the following model shows how the KPI is affected by changes in all the factors you have data for:

$$KPI_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \dots + \beta_n x_n$$

- KPI_t : The KPI at the time (t) you want to model.
- β_0 : The base performance, or what performance would be if all other factors were at their minimum.
- β : The coefficients, or what a change in the variable (x) means for the KPI.

$$\sum_{t=1}^n (y_t - f_{\beta}(x_t))^2 + \lambda \sum_{j=1}^p \beta_j^2, \text{ where } \beta_j \text{ is the weight of variable } x_j$$

$$y_t = \text{Intercept} + \beta_j \times \frac{x_{decay,t,j}^{\alpha}}{x_{decay,t,j}^{\alpha} + \gamma^{\alpha}} + \beta_{hol} \cdot hol_t + \beta_{sea} \cdot sea_t + \beta_{trend} \cdot trend_t + \dots + \beta_{ETC} \cdot ETC_t + \varepsilon$$

pendent variable y_t is on the left. The equation is enclosed in a dashed blue box. Below the box, it says "Main components of the function:". The term $\frac{x_{decay,t,j}^{\alpha}}{x_{decay,t,j}^{\alpha} + \gamma^{\alpha}}$ is labeled "S-Curve component for each media (j)". The terms $\beta_{hol} \cdot hol_t + \beta_{sea} \cdot sea_t + \beta_{trend} \cdot trend_t + \dots + \beta_{ETC} \cdot ETC_t$ are labeled "Holiday, Seasonality and Trend effect*". An arrow points from the text "Independent Variables" to the right side of the equation.

1. Adstock transformation: $X_{decay,t,j} = X_{t,j} + \theta_j \cdot X_{decay,t,j-1}$

2. S Curve transformation: $S \text{ Curve } (x, j) = \beta_j \times \frac{x_{decay,t,j}^{\alpha}}{x_{decay,t,j}^{\alpha} + \gamma^{\alpha}}$

where: y_t = revenue at time t

t = time index of dependent and independent variable (week)

j = media index (e.g. FB, TV, OOH) and $\beta, \alpha, \gamma, \theta$ = regressor specific to each media j

γ implemented on the S - Curve is a transformed γ where $\gamma_{tran} = \text{quantile}(X_{decay,j}, \gamma)$

β_{ETC}, ETC_t = further independent variables to be added to the model (e.g. competitor, promotions)

ε = Error term (accounting for all the other factors not addressed in the model)

Key Performance Indicators



1) Supply Chain key Risk Indicators

- **Procurement Delays (Major Issue):** Nearly 100% contribution, indicating a critical risk in sourcing materials.
- **Late Delivery (Moderate Concern):** Around 20-25%, possibly caused by procurement inefficiencies or logistics issues.
- **High-Risk Orders & Extreme Delays (Lower Impact):** High-risk orders are around 10-15%, while extreme delays are minimal but still notable.
- **Actionable Takeaway:** Focus on streamlining procurement processes to mitigate the biggest bottleneck in the supply chain.

2) Return on Ad-Spend

- **Fluctuating Ad Investments:** Significant variation in investment, with peaks in Sep and Oct 2023 and moderate spending afterward.
- **ROAS Recovery Post-August 2023:** Initially low in August, ROAS improved and stabilized, with noticeable spikes in March and June 2024.
- **No Direct Correlation Between Investment and ROAS:** Higher spending doesn't always translate to higher returns, suggesting optimization opportunities in ad strategy.

3) Customer Acquisition Cost

- **High CAC in Sep–Oct 2023** suggests aggressive customer acquisition efforts, possibly through expensive marketing campaigns.
- **Declining trend after Nov 2023** indicates improved efficiency in acquisition strategies or reduced marketing spend.
- **Spikes in Dec 2023 and Mar 2024** may imply seasonal effects or campaign-driven increases in acquisition costs.

4) Customer Lifetime Value

- **Peak CLTV in December 2023** suggests strong seasonal demand or successful marketing campaigns.
- **Sustained high CLTV** from Jan–May 2024 indicates customer retention but requires monitoring for long-term stability.
- **Sharp decline after May 2024** signals possible churn or reduced spending, necessitating corrective strategies.