MMM Mediation Model

This document provides a structured understanding of the steps, methods, and results from the executed notebook (M3.ipynb). It covers dataset details, preprocessing, modeling pipeline, cross-validation results, feature importance, and scenario analysis.

1. Dataset Overview

The dataset contains 104 weekly records over 2 years with 12 variables:

- Paid media spends: Facebook, Google, TikTok, Instagram, Snapchat

- Direct levers: Emails, SMS

- Controls: Average price, Promotions, Social followers

- Target: Revenue

2. Exploratory Data Analysis

Key insights from descriptive statistics and visualizations:

• Revenue is highly skewed with extreme spikes (min=1, max≈398k). Log transform applied.

• Facebook spend is stable baseline; Google, TikTok, Instagram, Snapchat are sparse with bursts.

• Average price has small range (≈87–113) and negative correlation with revenue.

• Promotions are rare but align with spikes in revenue.

• Emails and SMS have consistent scale and moderate correlation with revenue.

• Seasonality: annual peaks detected (period=52 weeks).

• Autocorrelation: revenue shows up to 5-week persistence.

3. Feature Engineering & Preprocessing

Transformations and additions applied:

• Revenue capped at 99th percentile and log1p transformed.

• Adstocked features (lag-1 with decay 0.5) for spends.

• Zero-spend indicators for sparse channels.

• Log-transformed spends to capture diminishing returns.

• Trend and Fourier seasonal features (sin, cos).

• Revenue lag features up to 5 weeks.

• Price squared term for non-linearity.

• Promotion interactions with adstocked social channels.

4. Causal Framework

Google spend is modeled as mediator between social/display channels and revenue.

Two-stage approach:

1. Stage A — Predict Google spend from lagged social features + trend (Ridge regression).

2. Stage B — Predict revenue using predicted Google (google_hat), social total, controls, and lags.

5. Model Training & Validation

Time-series expanding window cross-validation (initial 52 weeks train, 8-week horizon).

Three model settings tested:

- LightGBM (default params)

- LightGBM (reverted params for stability)

- Ridge regression (Stage B)

Cross-validation Results (LightGBM default)

Fold	Val Start	Val End	RMSE Log	R2 Log	R2 Orig
1	2024-09-15	2024-11-03	3.39	-0.13	0.54
2	2024-11-10	2024-12-29	3.58	0.39	-0.53
3	2025-01-05	2025-02-23	3.68	-1.4	-2.28
4	2025-03-02	2025-04-20	1.43	0.7	0.23
5	2025-04-27	2025-06-15	3.92	-0.26	-0.82
6	2025-06-22	2025-08-10	1.59	-50.59	-969.48

Cross-validation Results (LightGBM reverted)

Similar instability observed but slightly different gain distribution.

Cross-validation Results (Ridge Stage-B)

Fold	Val Start	Val End	RMSE Log	R2 Log	R2 Orig
1	2024-09-15	2024-11-03	2.87	0.19	0.36
2	2024-11-10	2024-12-29	4.81	-0.09	-0.55
3	2025-01-05	2025-02-23	3.74	-1.48	-2.19
4	2025-03-02	2025-04-20	1.7	0.58	-0.12
5	2025-04-27	2025-06-15	5.07	-1.1	-0.86
6	2025-06-22	2025-08-10	2.0	-80.39	-3128.84

6. Feature Importance & Coefficients

LightGBM top drivers (gain): average_price, social_total_was, social_total_ad1, seasonal terms, sms_send, revenue lags.

Ridge regression coefficients show:

- Positive: social_total_was, sms_send, sin_1, promotions × Instagram

- Negative: average_price, average_price_sq, rev_lag2, promo × TikTok, promo × Snapchat.

7. Scenario Analysis

Simulated interventions:

- +10% average price → large revenue drop (−1.13M to −4.7M depending on model).

- Removing promotions → revenue drop (−0.25M to −3.4M depending on model).

8. Recommendations

• Strong negative effect of price: maintain competitive pricing.

• Promotions critical for spikes: use strategically but avoid over-reliance.

• Social spend effectiveness mediated through Google is weak in data.

• SMS consistently contributes positively.

• Models show instability due to small dataset; validate with experiments or more data.

• Use Ridge or simpler models for interpretability, LightGBM only for exploratory insights.

9. Final Results Summary

The following summarizes the end results across models and analyses:

Cross-Validation Performance:

- LightGBM (default): Mean RMSE (log) ≈ 2.93, Mean R2 (orig) ≈ -162.1 → unstable, overfit.

- LightGBM (reverted params): Mean RMSE (log) ≈ 2.97, Mean R2 (orig) ≈ -189.6 → similar instability.

- Ridge regression: Mean RMSE (log) ≈ 3.37, Mean R2 (orig) ≈ -522.0 → more interpretable but poor fit due to small dataset.

Top Drivers Identified:

- Negative: Average Price, Average Price^2, revenue lags (especially lag2).

- Positive: Social total (presence), SMS sends, seasonal terms, promotions (interacting with social).

Scenario Analysis:

- +10% Average Price → Revenue drops by −1.13M (LightGBM reverted) to −4.7M (Ridge).

- Removing Promotions → Revenue drops by −0.25M (LightGBM default) to −3.4M (Ridge).

Interpretation:

The dataset size (104 weeks) and sparsity of social channels limit robust causal inference. Nonetheless, consistent patterns show strong negative elasticity to price, positive contribution of SMS, and promotional spikes as key revenue drivers. Google mediation effect was weak in observed data. Ridge model preferred for interpretability despite lower predictive accuracy.