Marketing Mix Modeling and Media Optimization

Silver → Gold Dataset

Author: Vandana Bhumireddygari

1. Business Objective

The objective of this project is to build a **Marketing Mix Model (MMM)** to:

1. Quantify the impact of **Network TV**, **Cable TV**, and **Digital** marketing spend on **True Viewership** and Revenue.
2. Incorporate realistic media dynamics such as **carryover (adstock)**, **saturation (diminishing returns)**, and **viewership inertia (lag effects)**.
3. Use the fitted model(s) to:
   a. **Simulate** the impact of a +20% increase in digital spend over a selected period.
   b. **Optimize media allocation** across channels for a given weekly budget to maximize predicted viewership.

The analysis starts with the **Silver-level episode dataset**, transforms it into a **Gold / weekly dataset**, and then fits increasingly sophisticated models.

2. Data and EDA (Exploratory Data Analysis)

2.1 Data Description

The **Silver-level dataset** contains episode-level records with:

- **Media variables**
  - Network_TV_Spend
  - Cable_TV_Spend
  - Digital_Spend
  - Impressions
- **Outcome variables**
  - True_Viewership

    - Revenue
- **Context / metadata**
    - Show
    - Season
    - Episode
    - Air_Date
    - Holiday (binary)
    - LeadIn_Bonus (indicator or numeric intensity)

This dataset is later aggregated to a weekly level to create the **Gold / weekly dataset** used for the final MMM.

2.2 Data Quality Checks

Key data quality checks performed:

1. **Missing values**
    a. Verified no problematic missingness in core modeling variables (*_Spend, True_Viewership, Revenue, Holiday, LeadIn_Bonus).
    b. Any minor missingness in non-critical fields was either ignored or excluded from modeling.
2. **Duplicates**

```
df.duplicated().sum()
```

    a. Checked for duplicated rows at the episode level.
    b. No meaningful duplicates were retained; any accidental duplicates would be removed before modeling.

3. **Data types**
    a. Air_Date converted to datetime.
    b. Spend and numeric fields coerced to numeric using pd.to_numeric(..., errors="coerce").
    c. Categorical variables like Show and Season kept as object types for later encoding.
4. **Outliers and distributions**
    a. Visual checks using histograms / boxplots for:
        i. Network_TV_Spend, Cable_TV_Spend, Digital_Spend

        ii. True_Viewership, Revenue, Impressions

   b. As expected, media spends are **right-skewed** (a few high-spend weeks).
   c. Viewership and revenue also show natural variability, with some high-performing episodes.

2.3 Exploratory Insights

From the EDA:

- **Spend vs Viewership**

Episodes with higher **Network TV** and **Cable TV** spend tend to have higher viewership. Digital also correlates positively with viewership, but its direct marginal effect is smaller in the simple baseline model.

- **Show and Season Effects**

Different shows and seasons operate at different base levels of viewership. This motivates the use of **show dummies** and/or **lagged viewership** to capture baseline popularity.

- **Holiday & Lead-in Effects**
    - Holiday episodes and episodes with LeadIn_Bonus tend to have higher viewership.
    - These variables are included as **control variables** in later models.

3. Gold / Weekly Dataset and Feature Engineering

To better match MMM practice and reduce noise, episodes are aggregated to a **weekly granularity**.

3.1 Weekly Aggregation

Using Week_Number, the data is aggregated as:

```
weekly_df = df.groupby("Week_Number").agg({
  "Network_TV_Spend": "sum",
  "Cable_TV_Spend": "sum",
  "Digital_Spend": "sum",
  "Holiday": "max",        # if any episode is a holiday, week is marked as holiday
  "LeadIn_Bonus": "mean",    # average lead-in across episodes in the week
  "True_Viewership": "mean", # average weekly viewership
  "Air_Date": "min",        # representative date
  "Show": "first",         # representative show (if needed)
  # plus any additional fields as needed
}).reset_index()
```

Rationale:

- Summing **spend** and **impressions** captures total weekly investment and exposure.
- Using mean True_Viewership provides a smoothed weekly outcome.
- Holiday and LeadIn_Bonus capture weekly conditions.

3.2 Adstock Transformation (Carryover)

To model **carryover effects** (past spend still influencing current week), adstock is applied

- decay controls how quickly media effect decays over time.

A decay **grid search** compares values {0.1, 0.3, 0.5, 0.7, 0.9}.

Best decay based on AIC/BIC from your notebook:

Best decay: 0.7

$R^2 \approx 0.979$, Adj. $R^2 \approx 0.960$ for the adstock + lag model with decay 0.7.

3.3 Saturation via Hill Function (Diminishing Returns)

To capture **diminishing returns** at high spend levels, a Hill-type transformation is used:

```
def hill_transform(x, alpha=1.2, gamma=20000):
    """
    Hill / saturation transform:
    - alpha controls curvature (how quickly it saturates)
    - gamma controls the half-saturation level
```

```
  """
  return (x**alpha) / (x**alpha + gamma**alpha)
```

This maps large spends into a bounded [0,1)-like scale with diminishing marginal impact.

3.4 Lagged Viewership

To capture **viewership inertia** / auto-correlation:

```
weekly_df["Lag_Viewership"] = weekly_df["True_Viewership"].shift(1)
```

- Rows with NaN lag (first week) are dropped for lagged models:

```
weekly_df_lag = weekly_df.dropna(subset=["Lag_Viewership"]).copy()
```

4. Modeling Approach and Results

You build several models, increasing in complexity.

4.1 Episode-level Baseline Model (Silver)

Key results

- **R-squared:** 0.496
- **Adj. R-squared:** 0.490
- **Coefficients (approximate):**
  - Network TV Spend ≈ 4.56 ($p < 0.001$)
  - Cable TV Spend ≈ 4.21 ($p < 0.001$)
  - Digital Spend ≈ 1.66 ($p < 0.001$)
  - LeadIn_Bonus: positive and significant at ~5%
  - Holiday: positive but not statistically significant

Interpretation:

- All three channels show **positive, significant associations** with viewership.
- At the episode level and in this linear specification, **Network and Cable TV appears more impactful per unit spend** than Digital.

- However, this model:
    - Ignores carryover and saturation.
    - Does not control for lagged viewership.
    - Is still at a noisy episode level.

This motivates moving to weekly MMM-style models with adstock, saturation, and lags.

4.2 Weekly Adstock Model (No Lag, decay tuned separately)

First, you fit a weekly model with adstocked spends (no lag):

```
X = weekly_df_ad[[
  "Network_TV_Spend_adstock",
  "Cable_TV_Spend_adstock",
  "Digital_Spend_adstock",
  "Holiday",
  "LeadIn_Bonus"
]]
y = weekly_df_ad["True_Viewership"]

X_const = sm.add_constant(X)
adstock_model = sm.OLS(y, X_const).fit()
```

From the notebook:

- **R-squared:** 0.863
- **Adj. R-squared:** 0.787

Interpretation:

- Moving to a weekly, adstocked specification already improves fit versus the episode-level model.
- However, there is still unmodeled persistence in viewership, which is handled in the next model.

4.3 Weekly Adstock + Lag Model (Decay = 0.7)

You extend the model by adding Lag_Viewership:

```
X_cols_lag = [
  "Network_TV_Spend_adstock",
  "Cable_TV_Spend_adstock",
  "Digital_Spend_adstock",
  "Holiday",
  "LeadIn_Bonus",
  "Lag_Viewership"
  # + optional show dummies if used
]

X_lag = weekly_df_ad[X_cols_lag].apply(pd.to_numeric, errors="coerce").dropna()
y_lag = weekly_df_ad.loc[X_lag.index, "True_Viewership"]

X_lag_const = sm.add_constant(X_lag)
lag_model = sm.OLS(y_lag, X_lag_const).fit()
print(lag_model.summary())
```

Using the tuned decay **0.7**, the model summary shows:

- **R-squared:** 0.979
- **Adj. R-squared:** 0.960

Interpretation:

- Adding **Lag_Viewership** dramatically improves fit, showing strong persistence in viewership week to week.
- Adstocked spends capture the **carryover** of marketing efforts.
- Due to small sample size (only 14–15 weeks), individual p-values can be unstable, but the model captures overall dynamics well.

4.4 Saturation (Hill) + Lag Model

You then try an alternative where spends enter via **Hill saturation**:

```
X_sat = weekly_df_sat[[
  "Network_TV_Spend_sat",
  "Cable_TV_Spend_sat",
  "Digital_Spend_sat",
  "Holiday",
  "LeadIn_Bonus",
  "Lag_Viewership"
]].apply(pd.to_numeric, errors="coerce").dropna()

X_sat_const = sm.add_constant(X_sat)
hill_model = sm.OLS(y_sat, X_sat_const).fit()
```

From the notebook:

- **R-squared:** 0.967
- **Adj. R-squared:** 0.938

Interpretation:

- The lag term (Lag_Viewership) is strongly significant (coef ≈ 0.75, $p < 0.001$), confirming viewership persistence.
- The Hill-transformed media spends capture nonlinearity and diminishing returns, though their individual coefficients are less stable due to few observations.
- Overall fit is strong but slightly lower than the final combined model.

4.5 Final Model: Adstock + Saturation (“Adsat”) + Lag

Your **final chosen model** combines adstock and saturation (often referred to as “adsat”) along with lag:

Features (conceptually):

- Network_TV_Spend_adsat

- Cable_TV_Spend_adsat
- Digital_Spend_adsat
- Holiday
- LeadIn_Bonus
- Lag_Viewership

The OLS summary from your notebook:

- **R-squared:** 0.984
- **Adj. R-squared:** 0.970
- **Observations:** 14 weeks

Key coefficients (approximate):

- Intercept: large negative (reflecting centering/scaling of adsat variables)
- Adsat media coefficients: positive for Network and Digital, negative for Cable in this specific fit, but none strongly significant individually because of small N and multicollinearity.
- Holiday: moderately negative, borderline significant.
- LeadIn_Bonus: positive, borderline significant.
- Lag_Viewership: positive but not highly significant in this specific adsat specification.

Interpretation:

- The model captures **98%+ of the variance** in weekly viewership.
- Interpretation of individual coefficient signs should be **cautious** due to:
    - Small sample size (14 observations),
    - High correlation among transformed media variables.
- Practically, the model is suitable as a **scenario simulator and optimizer**, not as a perfect causal inference engine.

5. Simulation: +20% Digital Spend

The assignment requires simulating the impact of increasing **Digital spend by 20%** for a chosen period.

5.1 Setup

From your notebook:

```
sim_df = weekly_df.copy()

# Example: Weeks 5 to 10
period_weeks = range(5, 11)

# Apply +20% digital spend for selected weeks
sim_df.loc[sim_df["Week_Number"].isin(period_weeks), "Digital_Spend"] *= 1.20
```

Then you **rebuild the adstock + saturation transforms** for sim_df using decay=0.7, apply the **final adsat model (as_model)**, and generate:

- Predicted_Actual: predictions under original spends
- Predicted_Sim: predictions under +20% digital spend in weeks 5–10

5.2 Plot

The plot in your notebook shows:

- **Black points**: actual True_Viewership by week
- **Solid line**: original predicted viewership
- **Dashed line**: simulated predictions with +20% digital spend
- **Yellow vertical band** (weeks 5–10): period where digital spend was increased

5.3 Interpretation

- During weeks **5–10**, the **simulated prediction line sits above the original prediction line**, indicating that increasing digital spend leads to **higher predicted viewership** in those weeks.
- Outside this range, the two prediction lines coincide, confirming that the intervention is localized.
- Due to saturation and adstock, the **uplift is not linear**:
  - Weeks with already high digital spend show **smaller incremental gains**.
  - Weeks where digital is below saturation show relatively **larger relative gains**.

You can optionally quantify uplift by computing:

- Total predicted viewers **before** vs **after** the +20% Digital scenario over weeks 5–10.
- Percentage lift in viewership for that period.

6. Media Mix Optimization

The final task is to recommend an **optimal media allocation** for a fixed weekly budget.

6.1 Optimization Problem

For a given week (you use **Week 8** in the notebook):

1. Compute the **total media budget**:

week = 8

2. Enumerate all allocations of that total spend across the three channels in **10% steps**, where each fraction is in {0, 0.1, 0.2, ..., 1.0}.
3. For each candidate mix (ntv, ctv, dig):
    a. Convert fractions back to dollar spends (ntv * total_spend, etc.).
    b. Rebuild the **adsat features** for that week.
    c. Use the **final adsat model (as_model)** to predict viewership.
    d. Keep track of the best-performing mix.

6.2 Result

From the notebook output:

Best media allocation: (0.6, 0, 0.4)
Predicted viewership: 97113162.00984901

So the best allocation for Week 8 under this grid search is:

- **60% Network TV**

- **0% Cable TV**
- **40% Digital**

with a predicted viewership of approximately **97.1 million**.

6.3 Interpretation and Caveats

- Within the tested grid (10% steps), the model suggests **shifting spend away from Cable** towards **Network TV and Digital** yields higher predicted viewership for that particular week.
- This reflects how the **final adsat model** balances:
  - Carryover effects,
  - Saturation,
  - And the underlying coefficients.
- However:
  - The optimization is **local** to Week 8 and the learned coefficients.
  - The model is trained on **only 14 weeks**, so the recommended mix should be viewed as a **directional insight**, not a guaranteed optimum in production.

7. Final Conclusions & Recommendations

7.1 Modeling Conclusions

1. **Data & EDA**
   a. The Silver dataset is generally clean and rich, with well-defined media, viewership, and contextual variables.
   b. Viewership and revenue respond to media spend and scheduling (holidays, lead-ins).
2. **Baseline Episode Model**
   a. Simple OLS on episode-level data explained ~50% of the variation in viewership ($R^2 \approx 0.50$).
   b. All three media channels show **positive, significant** associations with viewership.
   c. Network and Cable have higher marginal effects per unit spend than Digital in this specification.
3. **Weekly MMM with Adstock and Lag**

    a. Aggregating to weekly level and adding **adstock** and **lagged viewership** dramatically improved fit (R^2 up to ~0.98).
    b. This confirms the importance of **carryover** and **inertia** in media response.

4. **Saturation Effects**
    a. Hill transformation captures diminishing returns.
    b. It shows that very high spends yield reduced incremental viewership.
5. **Final Adsat + Lag Model**
    a. Combines adstock, saturation, and lag.
    b. Achieves $R^2 \approx 0.984$ and Adj. $R^2 \approx 0.970$ on 14 weeks.
    c. Best used as a **scenario and optimization tool** rather than for micro-level causal interpretation of individual coefficients.

7.2 Business Insights

1. **Media Effectiveness**
    a. All channels matter, but the model suggests that, for the tested week, **Network TV + Digital** is more efficient than spending on Cable.
2. **Digital Strategy**
    a. A **+20% Digital spend** in weeks 5–10 increases predicted viewership during those weeks.
    b. Because of saturation, digital should be increased **strategically**, not blindly.
3. **Scheduling**
    a. LeadIn_Bonus and Holiday effects suggest that **scheduling** (strong lead-in shows, holiday specials) can materially improve viewership, and should be considered alongside media spend.
4. **Optimal Mix Example**
    a. For Week 8 and a fixed budget, the model's optimal mix (under 10% grid steps) is:
        i. 60% Network, 0% Cable, 40% Digital
        ii. Predicted viewership ≈ 97.1M.

7.3 Limitations and Future Work

- **Sample size:** Only ~14 weeks at the weekly MMM level; more data would stabilize coefficient estimates.
- **Granularity:** The model predicts weekly viewership; finer-grained optimization (by show, daypart, or creative) would need more granular data and/or hierarchical models.

- **Other channels:** Non-paid drivers (e.g., PR, social, word of mouth) are not explicitly modeled.
- **Model form:** Future work could explore:
    - Bayesian MMM,
    - Regularization (Ridge/Lasso) for stability,
    - Hierarchical models by show.