Marketing Mix Model

Springboard Data Science Bootcamp Capstone 3 Project

Presented by: Rebeca Mahr (Spring 2021)

Business Problem

Can a marketing mix model be developed to quantify the impact of marketing media investments on unit sales?

Marketing Mix Variables

Baseline	Incremental	Long-Term
$ Price	Advertising	Competition Effects
Seasonality	Promotions	
Macro-factors	Discounts	

Data Wrangling

Data Source	kaggle
Sales Dataset (sales level)	■ Date ■ Sales_name ■ gmv_new ■ units ■ product_mrp ■ discount ■ product_category ■ product_subcategory ■ product vertical
Media Dataset (monthly level)	■ Year ■ Month ■ Total Investment ■ TV ■ Digital ■ Sponsorship ■ Content Marketing ■ Online Marketing ■ Affiliates ■ SEM ■ Radio ■ Other

Data source: https://www.kaggle.com/datatattle/dt-mart-market-mix-modeling

Data Cleaning

Created pandas datetime variables

₹ → $ Converted currency to USD

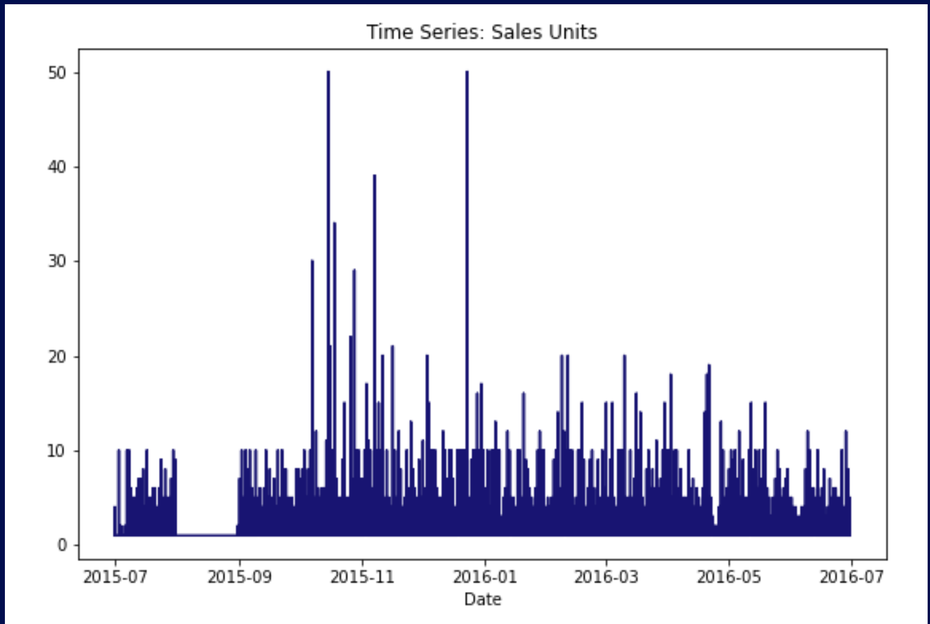
Filled missing categorical data with ‘other’

= Renamed variables for consistency

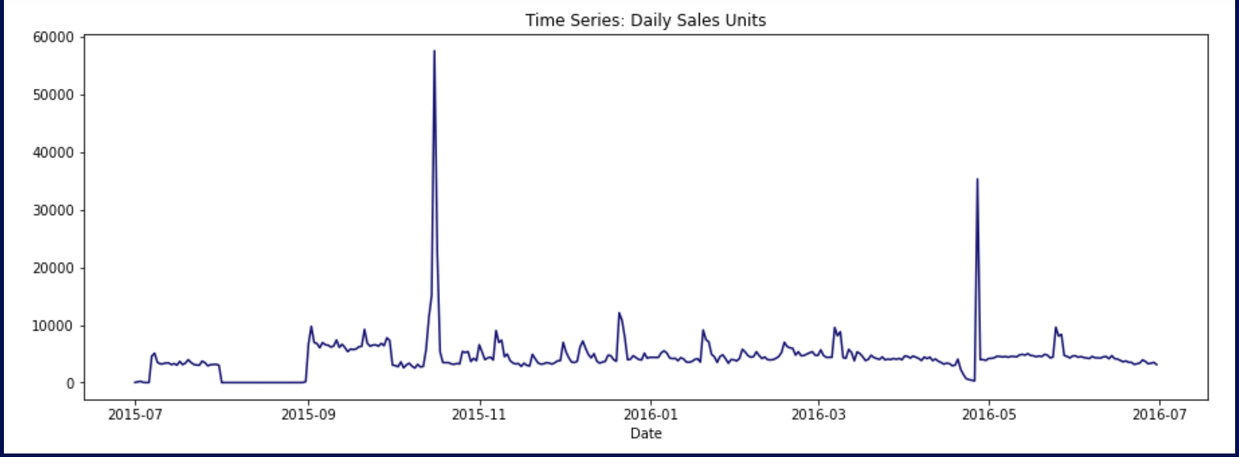
EDA

How should the data be aggregated for modeling?

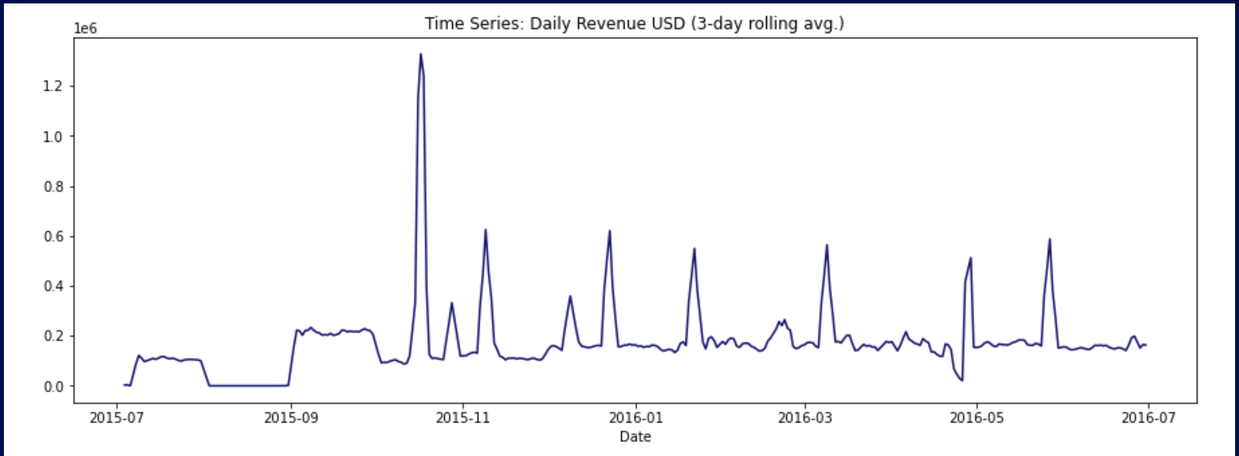
Sales Level



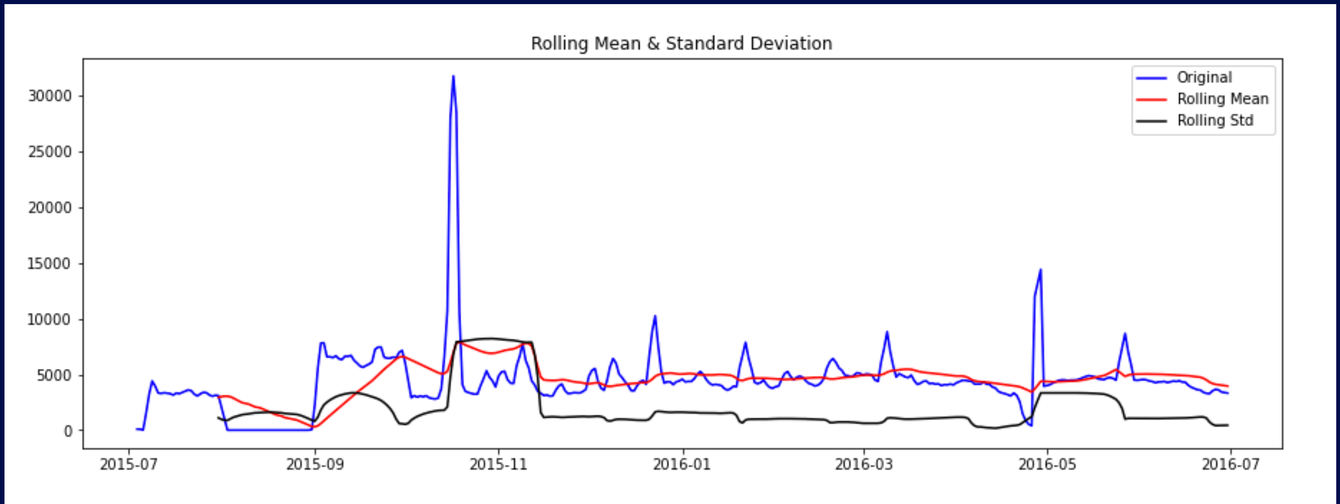
Daily Level



✓ 3-Day Rolling Avg. Level



Is there seasonality that needs to be adjusted for prior to modeling?



Results of Dickey-Fuller Test:	
Test Statistic	-3.777941
p-value	0.003141
#Lags Used	12.000000
Number of Observations Used	345.000000
Critical Value (1%)	-3.449447
Critical Value (5%)	-2.869954
Critical Value (10%)	-2.571253

p-value 0.003 < 0.5

T-Statistic -3.78 < CV (1%) -3.45

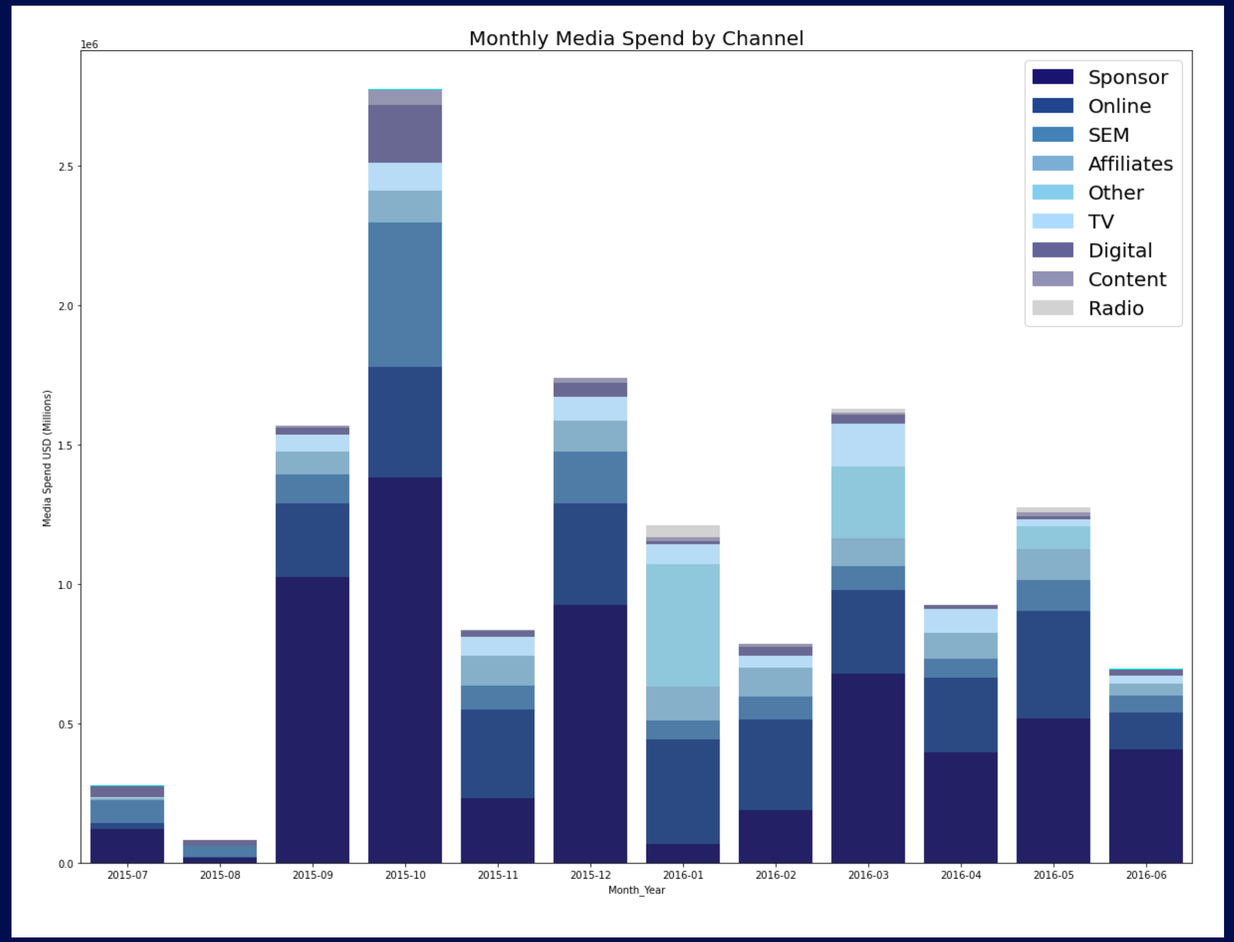
T-Statistic -3.78 < CV (1%) -2.87

T-Statistic -3.78 < CV (1%) -2.57

⬇

Stationary: No Seasonality

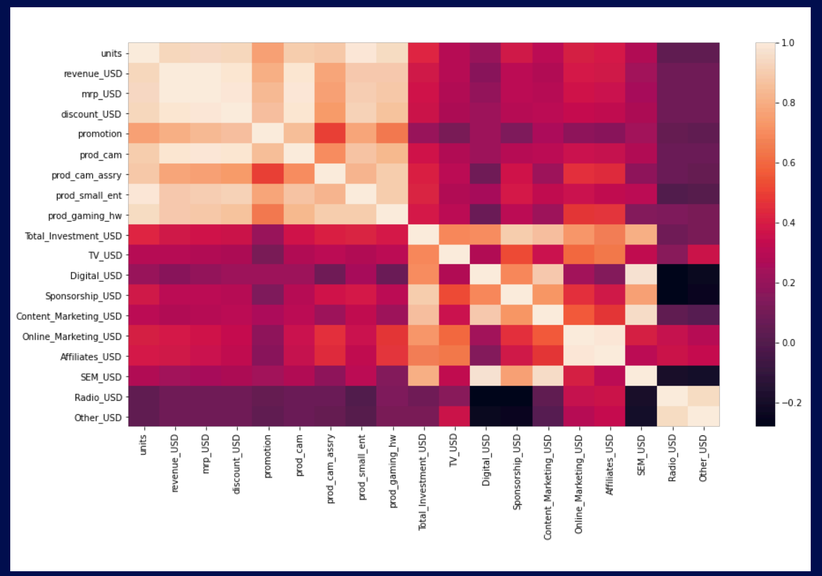
How was the media budget allocated across marketing media?



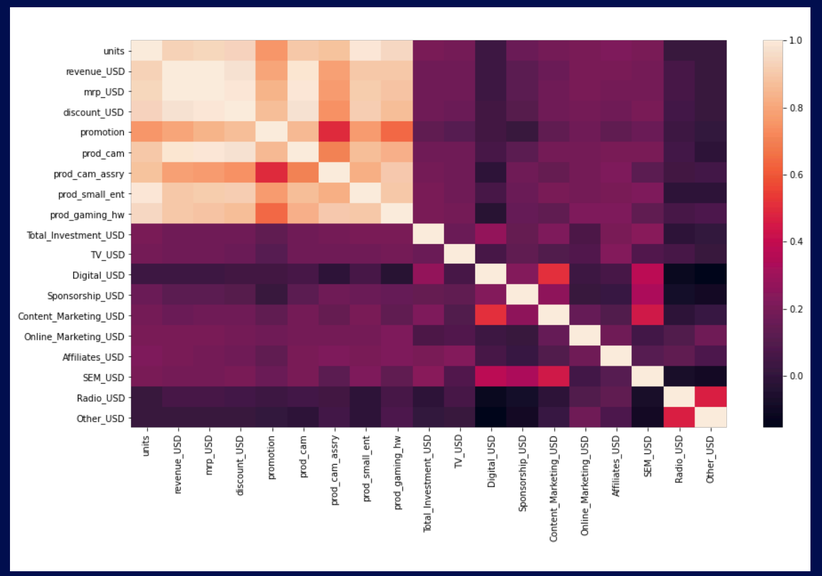
Sponsorship	43.2%
Online Marketing	22.9%
SEM	10.8%
Affiliate	7.3%
Other	5.7%
TV	5.2%
Digital	3.5%
Content Marketing	0.9%
Radio	0.6%

How should the monthly media spend be divided for merging?

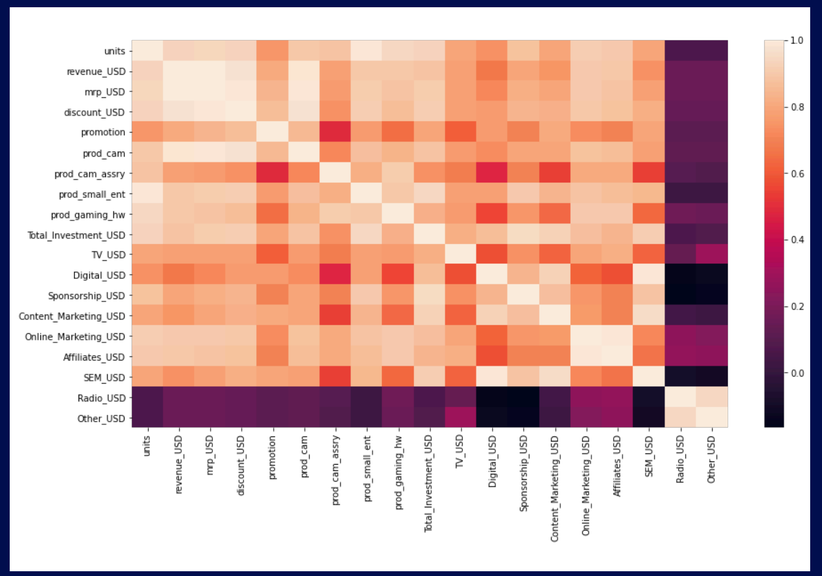
Split evenly by days / month



Split randomly across days / month



Split proportionately to sales / month



✓ ✓ ✗

Which variables should be used for modeling?

Dependent Variables

- ✓ Daily Unit Sales
- ✗ Daily Revenue

Independent Variables

- ✓ Price (mrp)
- ✓ Discounted Amount
- ✓ Promotion
- ✗ Sales by Product Category
- ✗ Total Marketing Spend
- ✓ TV Spend
- ✓ Digital Spend
- ✓ Sponsorship Spend
- ✓ Content Marketing Spend
- ✓ Online Marketing Spend
- ✓ Affiliate Marketing Spend
- ✓ SEM
- ✓ Radio
- ✓ Other

Preprocessing

Addressing Multicollinearity

Option 1 VIF Scores

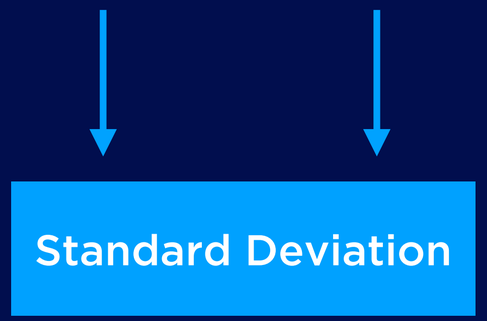
	w. Total Investment	w/o Total Investment
Total Investment	6,030,152	n/a
Sponsorship	1,399,732	29.29
Online Marketing	263,870.7	2,384.27
SEM	89,315.25	1,190.04
Other	69,849.15	487.65
Affiliates	40,672.8	2,104.27
Digital	20,621.59	593.57
TV	17,747.88	126.45
Radio	1,815.94	458.59
Content Marketing	1,018.13	61.32
mrp	378.65	374.15
discount	353.19	350.57
promotion	5.53	5.5

Train / Test Split

70% | 30%

Scaling

USD Amounts Sales Amounts



Modeling

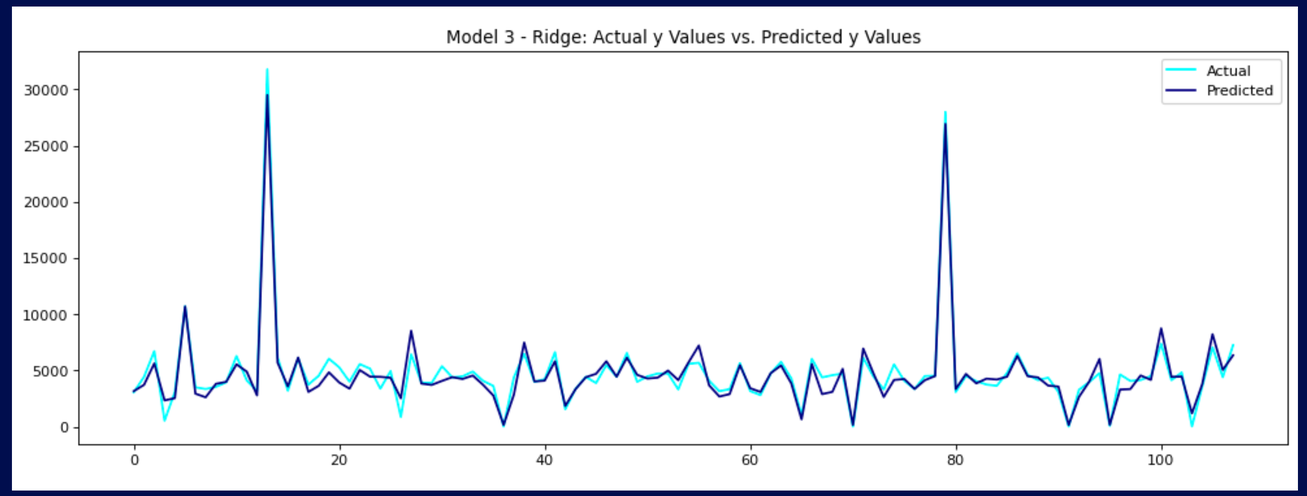
	Model	Dataset	r2	r2_rank	MAE	MAE_rank	rMSE	rMSE_Rank	Avg_Rank
OLS Linear Regression	1	1	0.96097	2	586.07757	4	758.6619	2	2.7
OLS Linear Regression	2	2	0.9121	7	810.8975	10	1138.6845	7	8.0
Ridge Regression	3	1	0.9611	1	575.8927	3	757.1106	1	1.7
Ridge Regression	4	2	0.9231	5	781.9295	9	1064.8538	5	6.3
Lasso Regression	5	1	0.9317	3	684.1494	7	1003.5886	3	4.3
Lasso Regression	6	2	0.9264	4	686.0254	8	1041.7917	4	5.3
Random Forest Regression	7	1	0.8951	8	549.1776	1	1243.9442	8	5.7
Random Forest Regression (Optimized)	8	1	0.9143	6	558.0814	2	1124.3996	6	4.7
Random Forest Regression	9	2	0.8085	9	646.0023	5	1680.3987	9	7.7
Random Forest Regression (Optimized)	10	2	0.7944	10	665.423	6	1741.2403	10	8.7

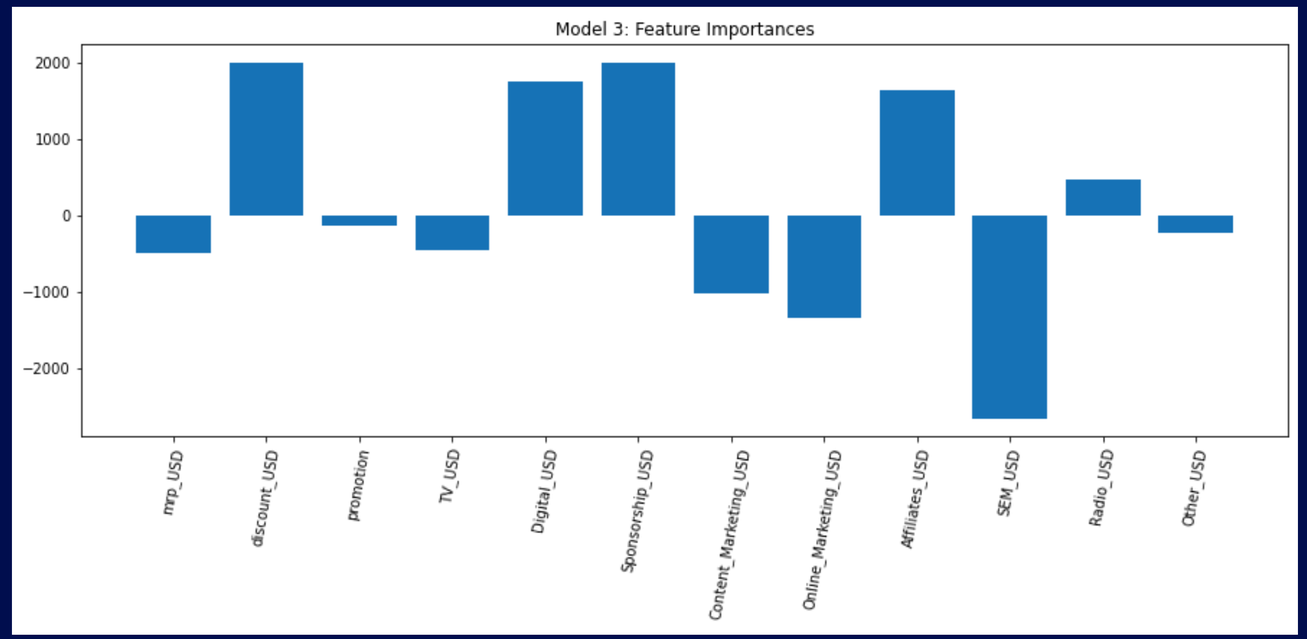
Selected Model

Ridge Regression Model
Option 1 Dataset

Hyperparameters
- alpha (penalty term): 0.1

Evaluation Metrics
- r-squared: 0.9611
- MAE: 575.8927
- rMSE: 757.1106





Conclusion

Which marketing efforts had the greatest impact on unit sales?

SEM spend

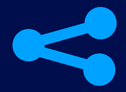
2,654

Sponsorship spend

2,011

Digital spend

1,766



Affiliate Spend

1,652

Online marketing spend

1,333

Content marketing spend

1,025

Limitations

Monthly-level media investment data

Only one year of data

Missing long-term impact variables

Unable to measure adstock / lag effects

Thank you!