Rebeca R. Mahr
June, 2021

Marketing Mix Model

Springboard Data Science Bootcamp Capstone 3 Project

Table of Contents

Background **1**

Data Wrangling **2**

Exploratory Data Analysis (EDA) **3**

Preprocessing & Baseline Model Development **12**

Modeling & Hyperparameter Tuning **14**

Model Selection & Conclusion **22**

Limitations **24**

References **25**

Background

Marketing is key to building brand awareness and driving business objectives. With a multitude of marketing media options available such as TV, radio, and various online options including social media, email, search, and digital video, businesses can and should advertise using a combination of media to reach their unique customer segments. However, it can be difficult to identify which marketing media are most effective at driving business goals. A marketing mix model utilizes regression techniques to measure the impact of marketing and advertising efforts on business goals and inform the efficient use of marketing funds (Kumar, 2017).

Marketing mix models can vary in complexity depending on the variety of data a business has available and the scope and resources budgeted for modeling. They may include a few different types of independent variables including baseline variables, incremental variables and long-term impact variables. Baseline variables are factors that impact the business goal without any marketing or advertising efforts such as price, seasonality, and macro factors like GDP and inflation. Baseline variables serve as controls in the regression model. Incremental variables are the marketing and advertising media metrics or spend including TV, radio, digital, search, etc. as well as sales and promotions. Long-term impact variables are related to competition such as competing businesses and products (A Complete Guide to Marketing Mix Modeling, n.d.). The dependent variable is typically the business goal which can be unit sales, revenue, services sold, or conversions for digital objectives.

For this project, a marketing mix machine learning model was developed to quantify how marketing media investments contributed to unit sales of electronics products within one fiscal year. Seasonality and price were explored as potential baseline variables, promotions, discounts and marketing spend on various different marketing media were explored as incremental variables, and unit sales and revenue were explored as potential dependent variables. Due to a lack of information about the business and its products, long-term impact variables were not explored.

Data Overview

This project utilized two Kaggle datasets sourced from DT Mart (Miglani, 2020). They included sales and marketing investment data from an electronics retailer between July of 2015 and June of 2016. The datasets used were the 'MediaInvestment.csv' and the 'Sales.csv' files. The media investment dataset included aggregate monthly media spends for each month. The sales dataset included sales-level data for each day of the fiscal year.

<u>Data Dictionary</u>

- Dataset 1 – Sales.csv (n= 1,578,079):
  - Date: YYYY-MM-DD
  - Sales_name: Indicates whether a promotion was applied to the sale and the name of the promotion if one was applied
  - gmv_new: Revenue / amount paid
  - units: number of units sold
  - product_mrp: maximum retail price of the sale
  - discount: amount discounted from product_mrp
  - product_category: business category of product
  - product_subcategory: business subcategory of product
  - product vertical: business vertical of product
- Dataset 2 – MediaInvestment.csv (n=12):
  - Year: Year of campaign
  - Month: Month of campaign
  - Total Investment: total amount spent on media (millions)
  - TV: total amount spent on TV advertising
  - Digital: total amount spent on Digital advertising
  - Sponsorship: total amount spent on sponsorship advertising
  - Content Marketing: total amount spent on targeted content marketing
  - Online marketing: total amount spent on online marketing
  - Affiliates: total amount spent on third-party affiliate marketing
  - SEM: total amount spent on search engine marketing
  - Radio: total amount spent on radio advertising
  - Other: total amount spent on other forms of advertising

Data Wrangling

The data wrangling step of this project included loading and inspecting the datasets, assigning appropriate variable types, renaming variables, imputing missing values, and converting currency to US dollars.

Sales Dataset:

- In order to analyze the data for seasonality, the date variable was converted to a pandas datetime variable type.
- Inspecting the product prices and promotions, it appeared that the business was likely in India and the currency was in Indian Rupees. To make the findings easier to interpret, the currency was converted to US dollars based on the highest currency exchange rate from Indian Rupee to USD in 2015 which was $0.0163 (Indian Rupee to US Dollar Spot Exchange Rates for 2015, n.d.).
- Missing values in the product_vertical variable of the sales dataset were identified having the value "\N". These missing values were imputed as "OtherSpeaker" based on the value in the "product_subcategory" being "Speaker".

Media Investment Dataset:

- To ensure consistency in variable naming and avoid compatibility issues, spaces in variable names were converted to underscores.
- For consistency with the sales data, all currency values were converted to US dollars.
- Missing values likely indicated that the media type was not used in the corresponding month. Therefore, missing values were imputed with 0.
- To allow for merging with the sales data, a pandas datetime variable was created using the 'Year' and 'Month' variables and assigning the day as 1 (the first day of the month).

Data Wrangling Code Path:
rrmahr / Marketing_Mix_Model / a_Notebooks / a_Data_Wrangling.ipynb

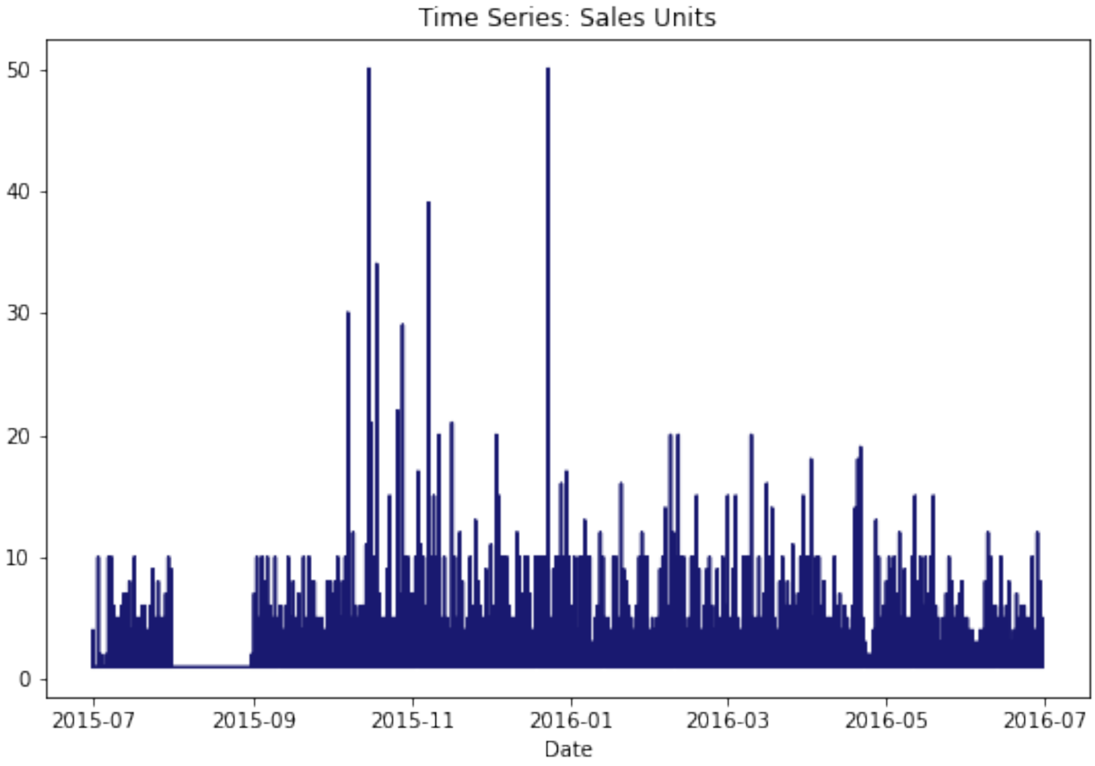
Exploratory Data Analysis (EDA)

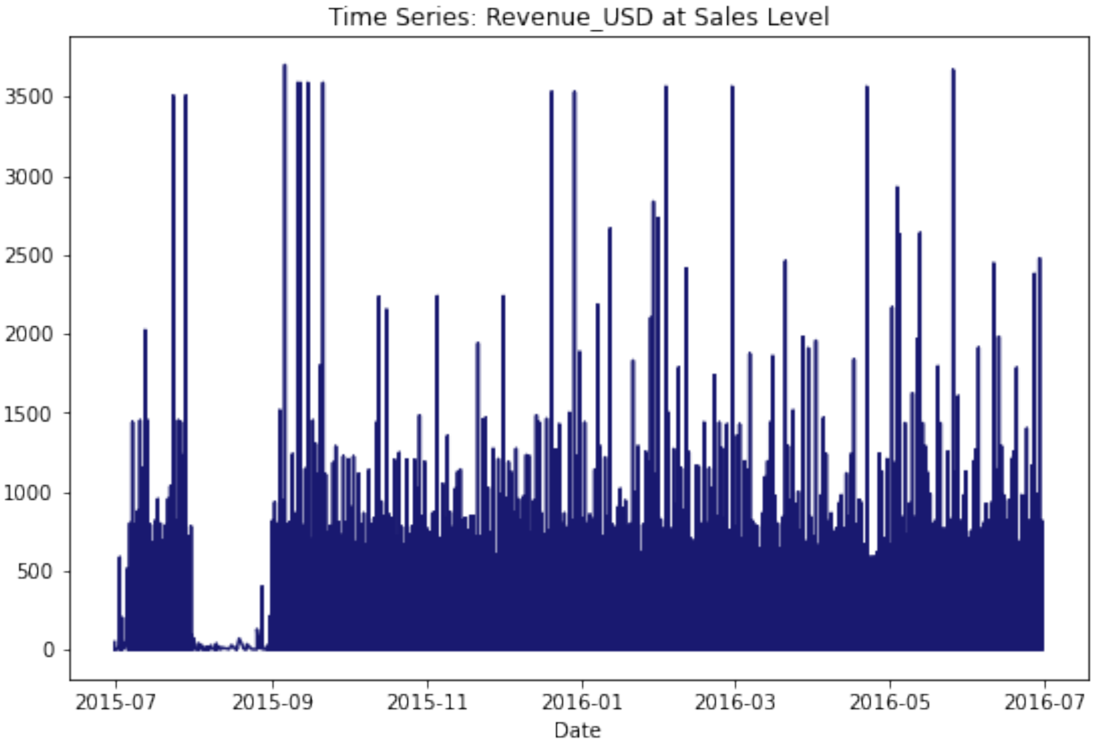
Exploratory data analysis was conducted to answer several key questions that would inform the development of the marketing mix model.

Question 1: How should the data be aggregated for modeling?

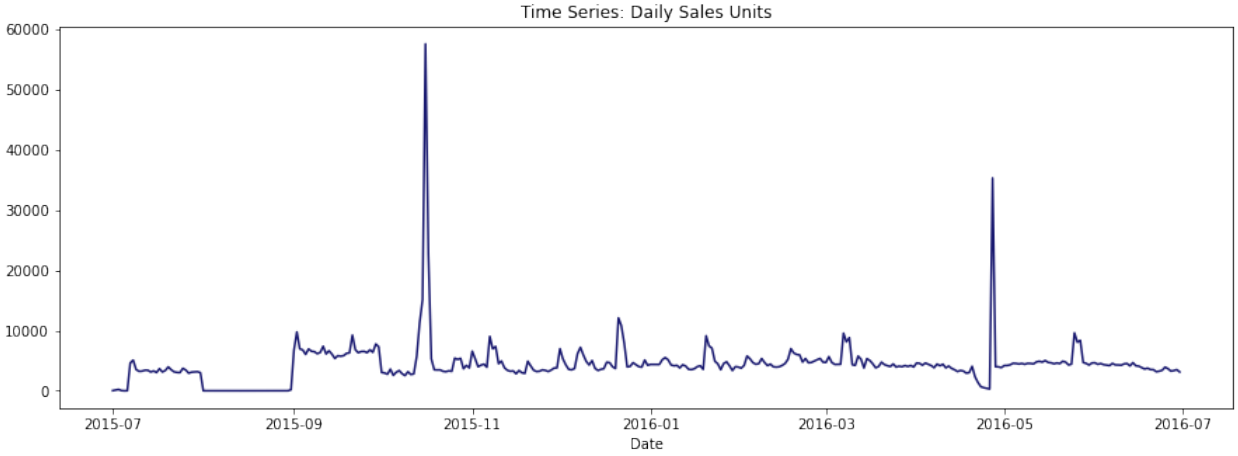
Typically, data used in marketing mix models are aggregated at the weekly level to minimize large variations in sales that are common at the daily level. However, that best practice is based on a minimum of two years of data. Since this project's datasets only contained one year's worth of data, another aggregate level was needed.

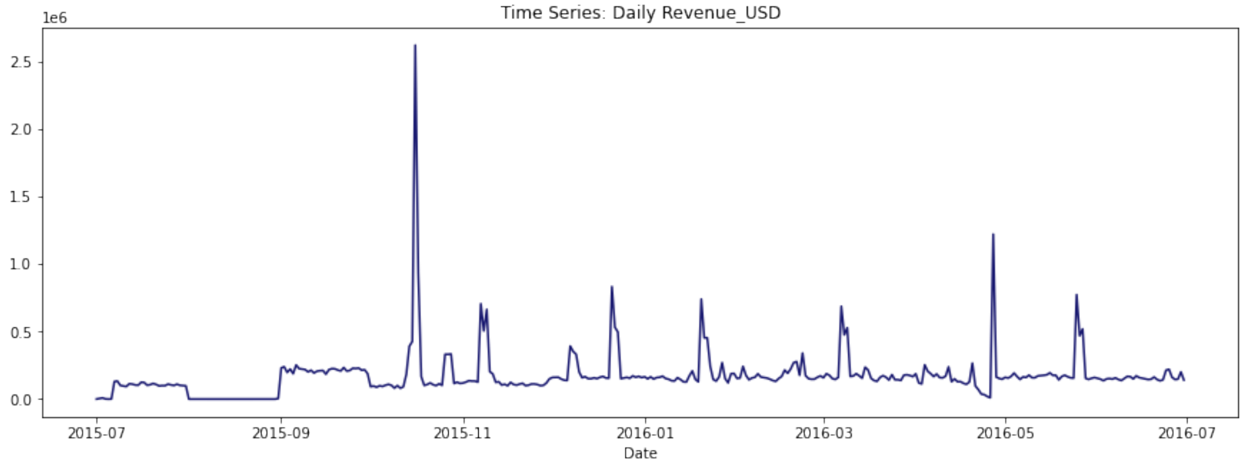
The potential dependent variables, units and revenue, were first explored at the sales level. On average, each sale included 1.02 units with a standard deviation of 0.25 while the average revenue for each sale was $40.12 with a standard deviation of $91.93. When plotted over time, as shown in the figures below, there was a lot of variation at the sales level and some large outliers that may have been due to holidays or may have just been random occurrences.



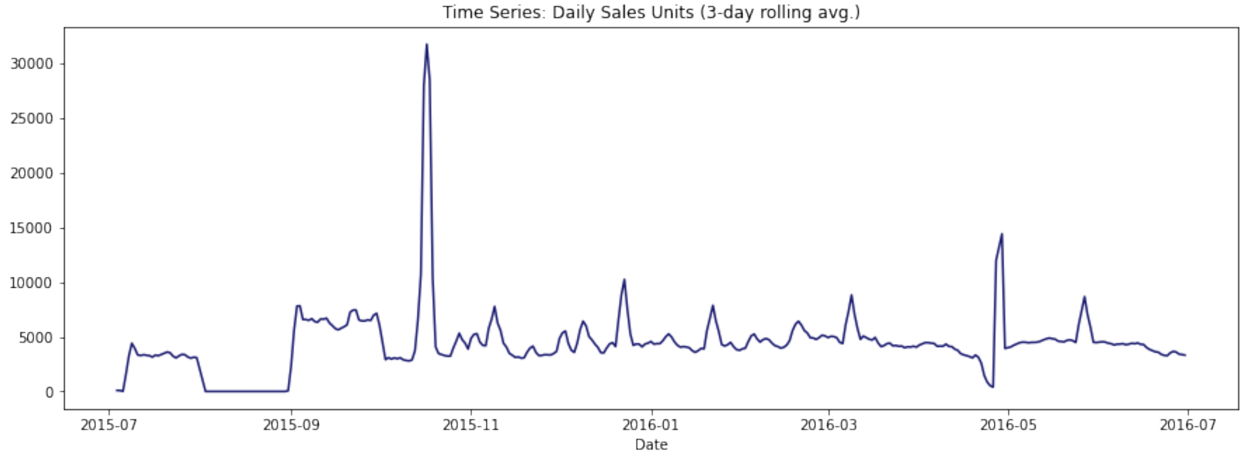


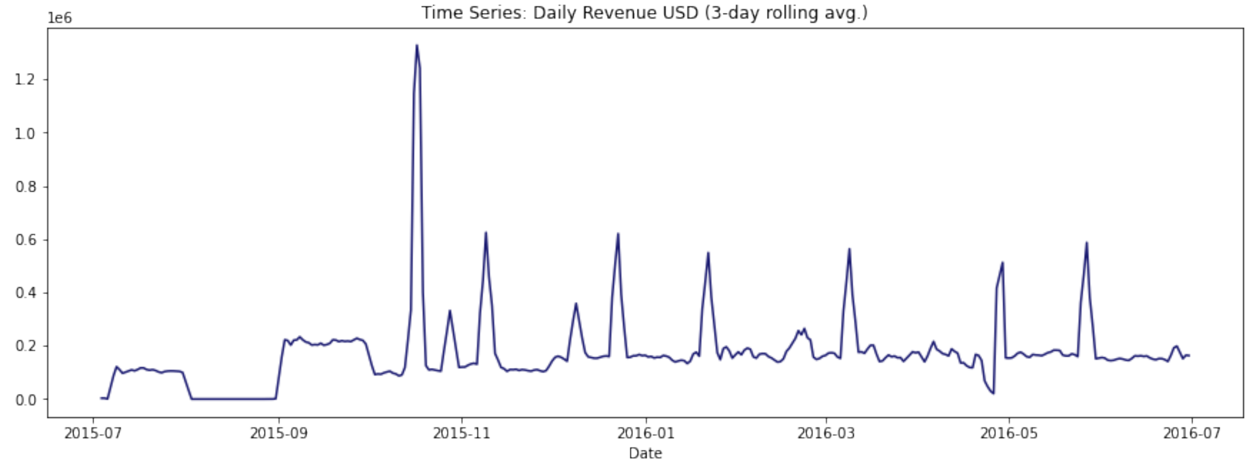
At the daily level, the average number of units sold was 4,479.15 with a standard deviation of 3,957.66 units while the average revenue was $175,887.80 with a standard deviation of $185,805.80. Time-series plots at the daily level showed less variation, but there were still several outliers.





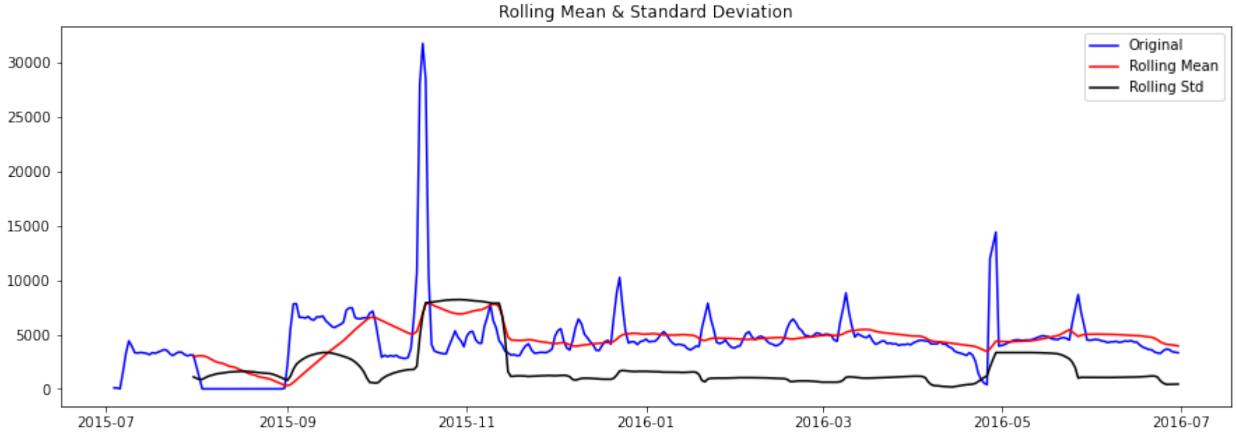
In order to further minimize daily variations, the data was transformed using a 3-day rolling average. Using the 3-day rolling average, the spread of the data decreased with the average number of sales units being 4,494.82 units with a standard deviation of 3,037.37 units and the average revenue being $176,412.20 with a standard deviation of $140,109.50. As shown in the time-series plots below, the 3-day rolling average resulted in less variation over time which better fits the weekly aggregate standard for marketing mix models without losing too much data.

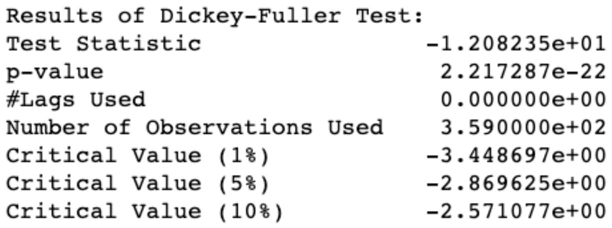




<u>Question 2: Is there seasonality that needs to be adjusted for prior to modeling?</u>

Seasonality in time-series analysis refers to “repeating patterns or cycles of behavior over time” (Anish, 2020). Seasonality can appear in the form of annual events such as holidays where sales in certain categories peak around the same time each year or in other time increments such as time of day or day of the week. Because a time-series with seasonality is not stationary, an augmented Dicky-Fuller unit root test was conducted using statsmodels’ tsa.stattools.adfuller to test whether the data was stationary over time (Perktold, Seabold, & Taylor, 2021).
A plot showing a monthly rolling mean, standard deviation and the results of the test are shown in the figures below. While there were some variations in the plot of the rolling mean and standard deviation, the plots did appear to be stationary over time. The results of the test also indicated the data was stationary over time with a p-value well under .05 and the test statistic of -12.08 being less than the critical values at 10%, 5%, and 1%. According to these findings, seasonality was not deemed to be a baseline variable to be adjusted for in this dataset and model.

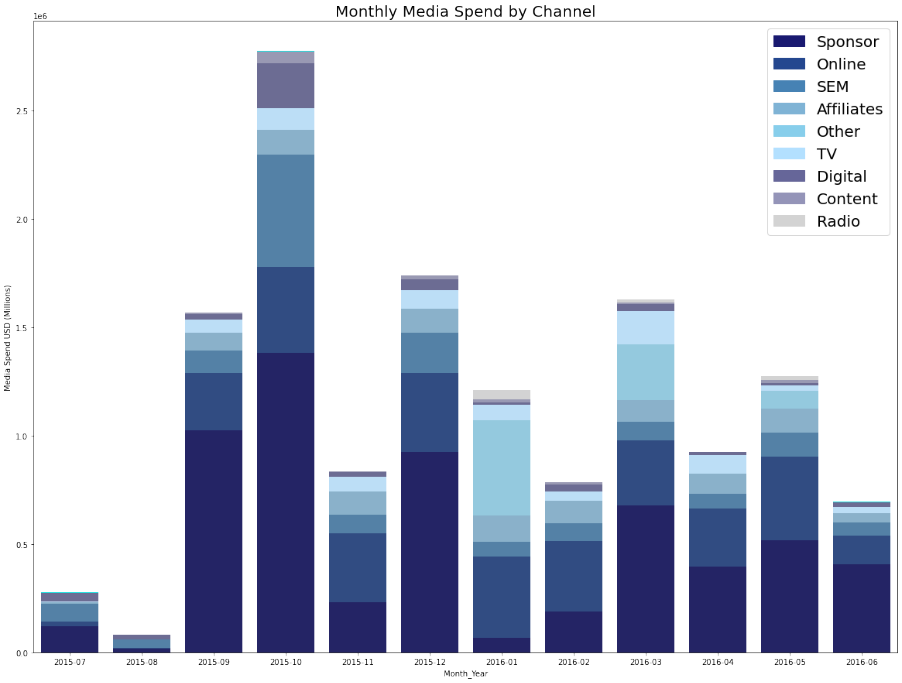




```
Results of Dickey-Fuller Test:
Test Statistic                 -1.208235e+01
p-value                         2.217287e-22
#Lags Used                      0.000000e+00
Number of Observations Used     3.590000e+02
Critical Value (1%)            -3.448697e+00
Critical Value (5%)            -2.869625e+00
Critical Value (10%)           -2.571077e+00
```

<u>Question 3: How was the media budget allocated across marketing media?</u>

Understanding how the media budget was allocated across marketing media helps to shed light on modeling results. The expectation or hypothesis would be that media with a higher budget allocation should have a greater impact on the business goal or, at least, produce a positive return on investment for media which are more expensive. Shown below is the monthly media budget allocation.

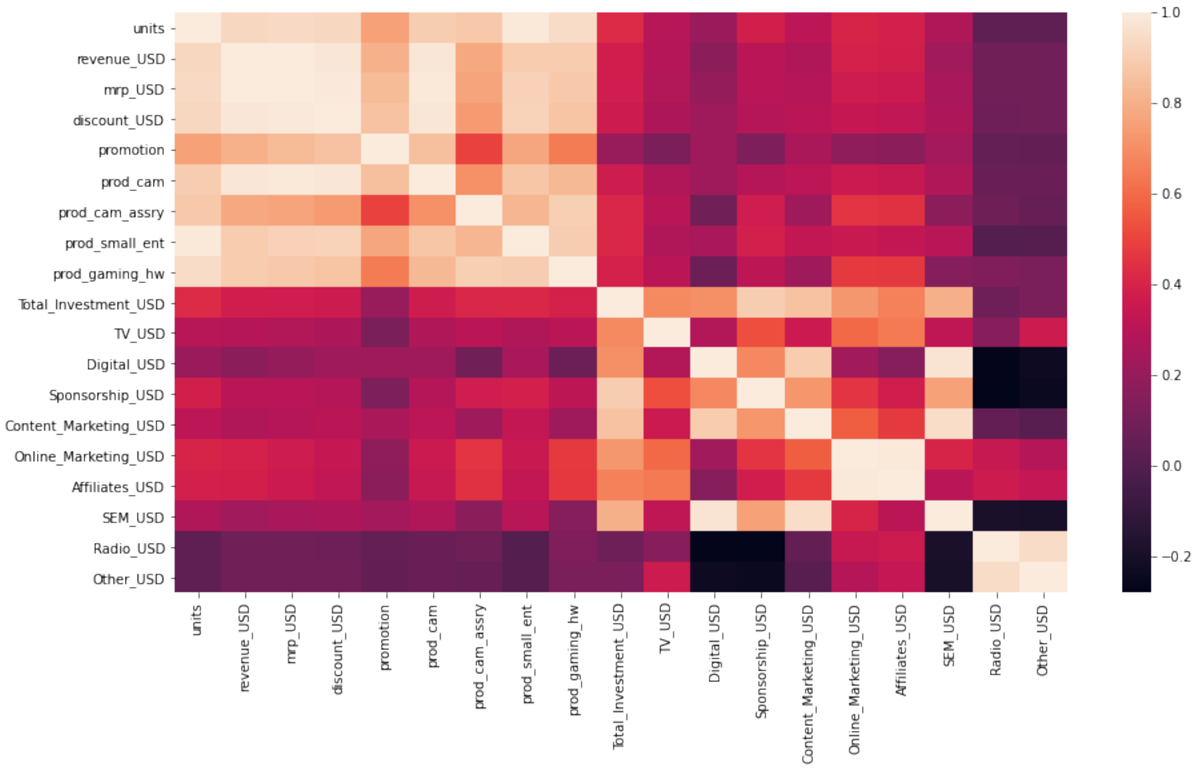


<u>Question 4: How should the monthly media spend be divided for merging?</u>

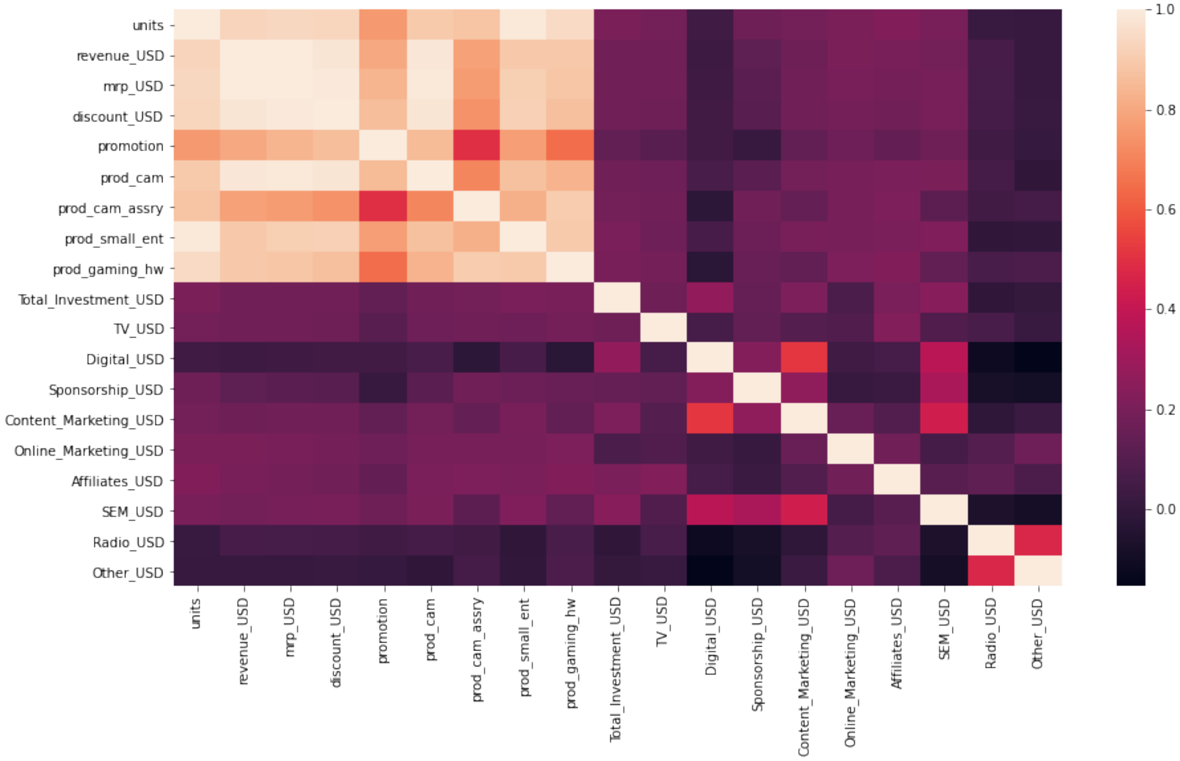
Because the marketing media spend data was collected at the monthly level, it could not be merged directly with the daily sales data. Three approaches for dividing the monthly media spend were employed and evaluated based on their relationship to unit sales and revenue. Options 1 and 2 described below were used for modeling.

Option 1:

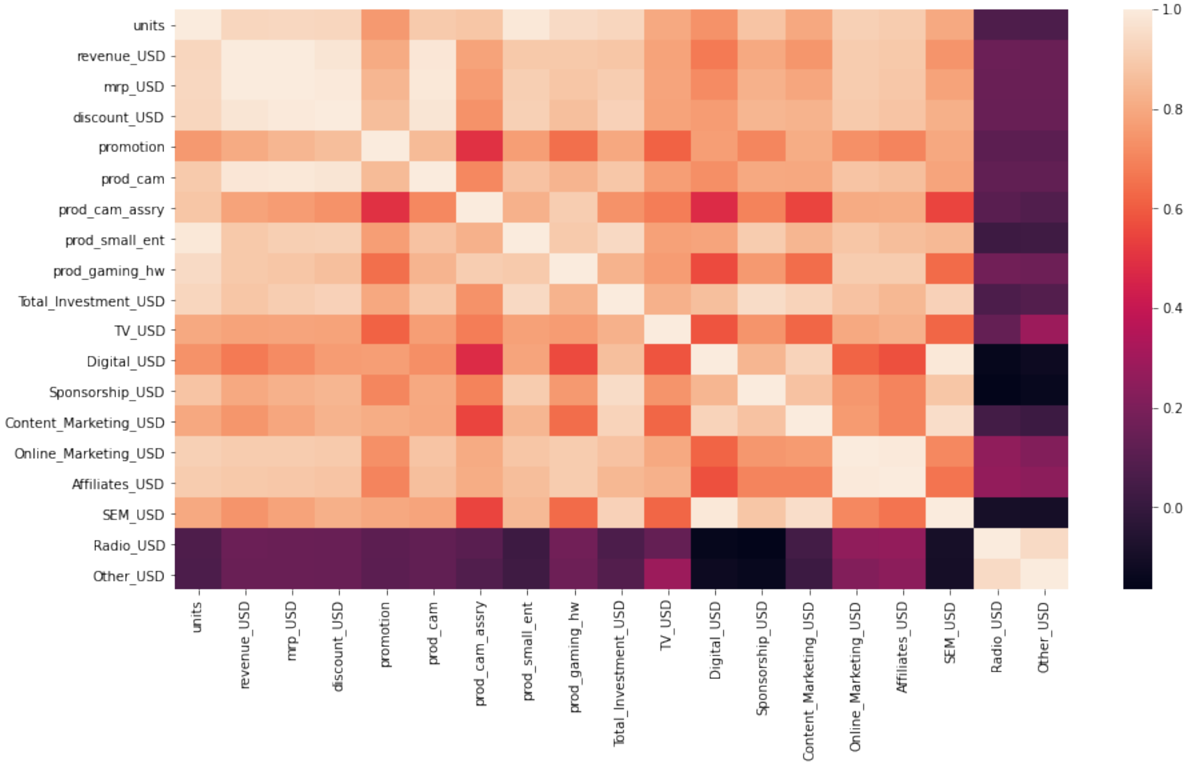
The media investment data were divided evenly by the number of days in each month in the sales data and then merged with the 3-day rolling average sales data. An analysis of correlations was conducted on the merged dataset. Of the incremental variables, daily unit sales and daily revenue were most highly correlated with the number of promotions, the online marketing spend, the affiliate marketing spend, and the sponsorship marketing spend.



Option 2:
The media investment data were divided randomly across the days of each month in the sales data and then merged with the 3-day rolling average sales data. Splitting randomly across the month resulted in much lower correlations with media spend. Of the media spend variables, daily unit sales and daily revenue were most highly correlated with the affiliate spend, the online marketing spend, and the search engine marketing spend.



Option 3:
The media investment data were divided proportionately to the number of sales per day per month and then merged with the 3-day rolling average sales data. Logically, splitting monthly the media spend proportionately to the number of unit sales resulted in very high correlations with daily unit sales. These high correlations were likely not representative of the true relationship between media spend and units sales, so option 3 was not used for modeling.



<u>Question 5: Which variables should be used for modeling?</u>

- Dependent Variable:
  - Either daily number of unit sales or daily revenue could be utilized as the dependent variable for modeling.
  - Based on the time-series plots, there appeared to be less variation from day to day in unit sales as compared to revenue.
  - Additionally, because this project did not have macro-economic data to control for the impact of macro factors, the daily unit sales metric was a more stable dependent variable to use for modeling.
- Independent Variables:
  - Because the sales by product category variables are subdivisions of unit sales, the product category variables were not selected for modeling.
  - Because of multicollinearity between the total media investment and each individual media spend variable, the total media investment variable was also not selected for modeling. See the preprocessing step for details on the multicollinearity analysis.
  - All remaining baseline and incremental variables were selected as independent variables for modeling.

EDA Code Path:
rrmahr / Marketing_Mix_Model / a_Notebooks / b_EDA.ipynb

Preprocessing & Baseline Model Development

Final preprocessing steps conducted included testing data for multicollinearity, splitting the data into training and testing datasets and scaling the independent variables.

Multicollinearity
Multicollinearity was tested by the variation inflation factor (VIF)[1] using statsmodels.stats.outliers_influence.variance_inflation_factor (Variance Inflation Factor, n.d.). In the option 1 dataset, Total_Investment_USD, the sum of all media spend, had a very high VIF score of 6,030,152. While the VIF score for total media investment in the option 2 dataset was much lower at 1.91, Total_Investment_USD was not selected for modeling for both models because of its direct relationship with the other media spend variables. Reassessing the VIF scores excluding Total_Investment_USD resulted in lower VIF scores for all variables although it was still high for the online marketing spend, affiliate marketing spend, and search engine marketing spend using the option 1 dataset. Despite having high VIF scores for some variables, all remaining independent variables were selected for modeling and models that control for high multicollinearity were developed and evaluated.

Train-Test Split
Both the option 1 and option 2 datasets were split into training and test sets in order to evaluate model performance using "unseen" data. The splits were conducted using scikit-learn's model_select.train_test_split (Pedregosa, 2011) with 30% of each of the option 1 and option 2 datasets reserved in the test set.

Scaling
Because the independent variables varied in terms of magnitude and measurement units, the independent variables were scaled using scikit-learn's preprocessing.StandardScaler (Pedregosa, 2011).

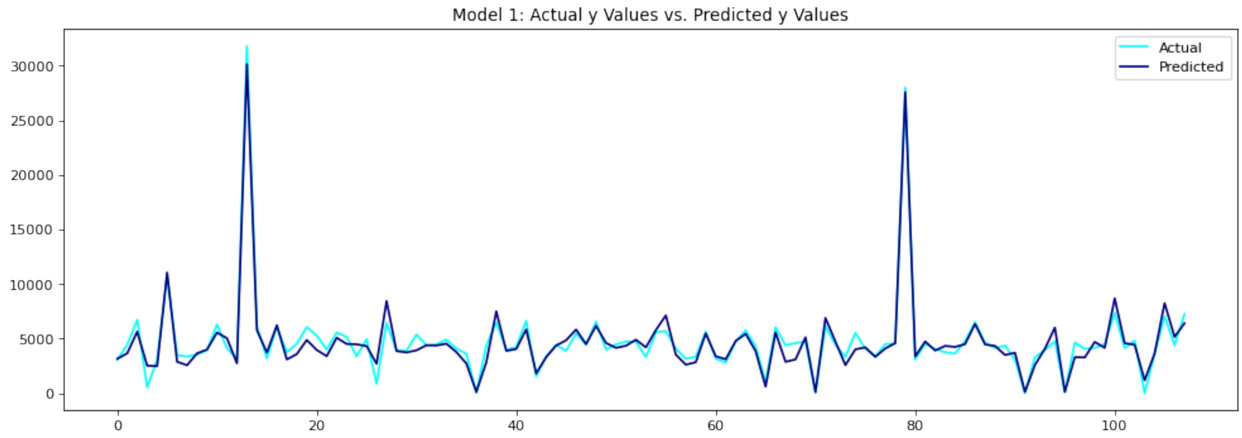
Baseline Model Development
Baseline models were developed for both the option 1 and option 2 datasets conducting ordinary least squares (OLS) linear regression using scikit-learn's linear_model.LinearRegression (Pedregosa, 2011). Each of the models developed were evaluated based on their r-squared score, mean absolute error (MAE), and root mean squared error (rMSE). To quantify the impact of each of the independent variables on unit sales, feature importances were calculated by multiplying the standard deviation of each

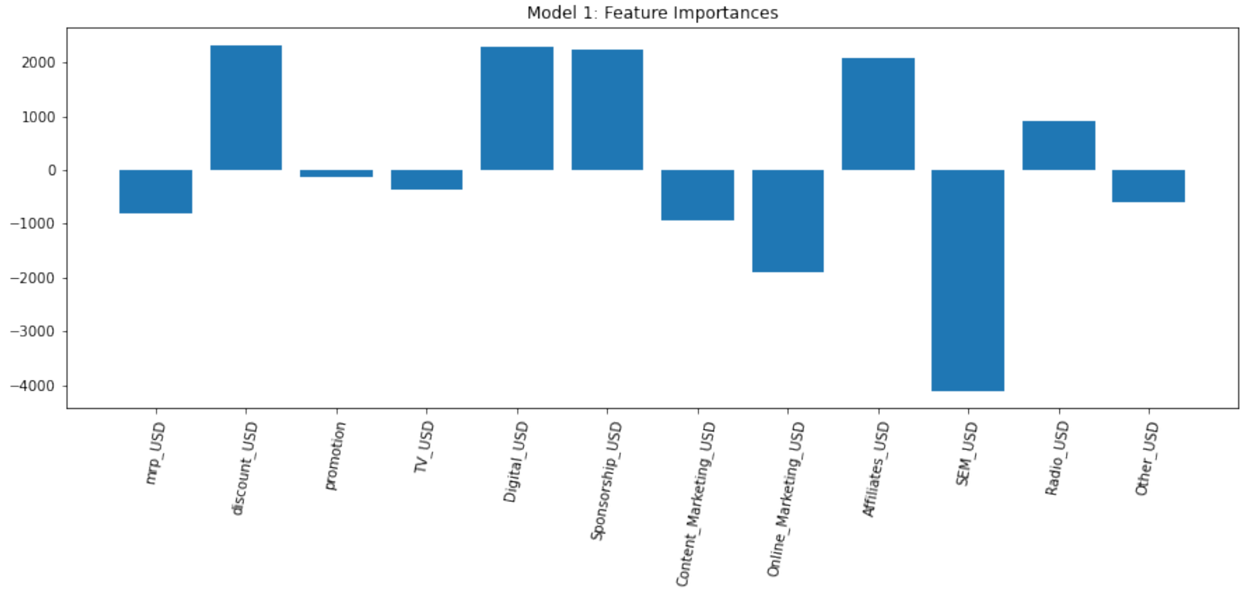
[1] "VIF is used to identify the correlation of one independent variable with a group of other independent variables" (Pulagam, 2020).

independent variable by their corresponding regression coefficient. See model results below.

Model 1: Option 1 OLS Linear Regression

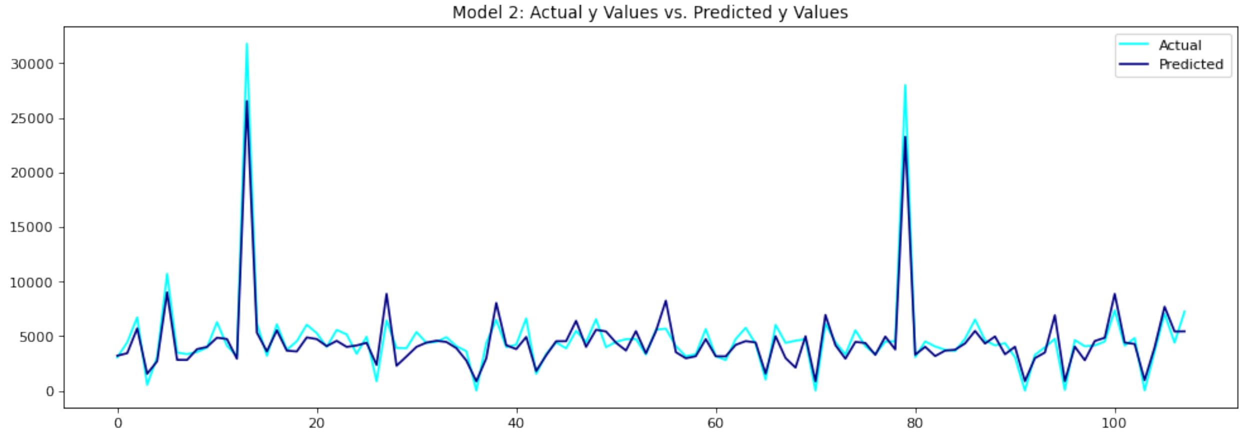
r-squared: .96097
MAE: 586.07757
rMSE: 758.6619

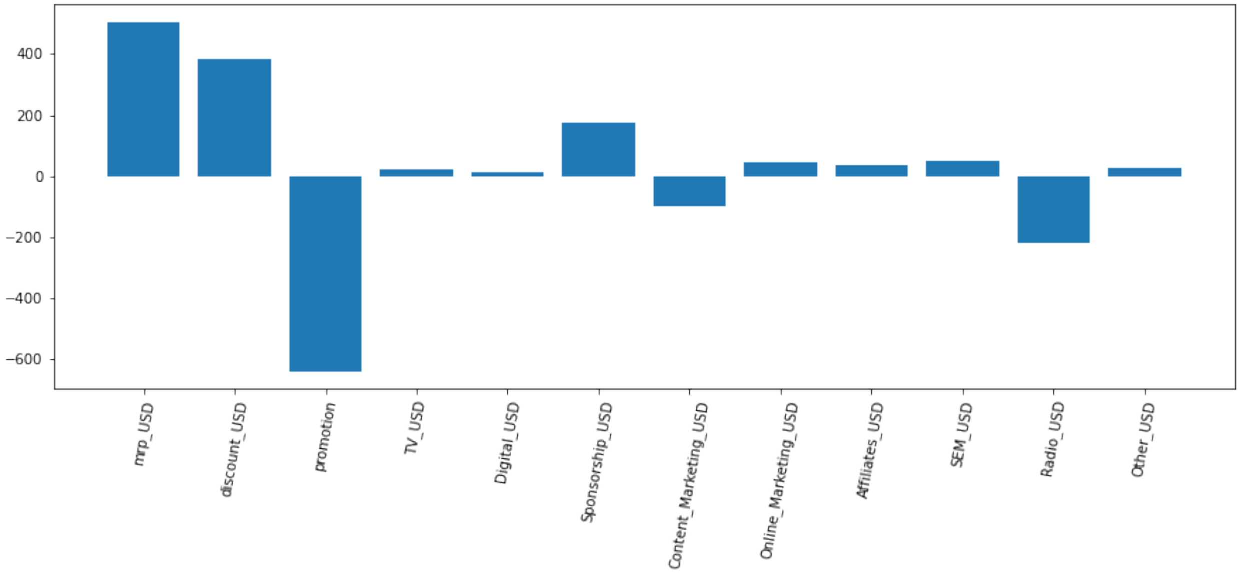

Model 1: Actual y Values vs. Predicted y Values


Model 1: Feature Importances

Model 2: Option 2 OLS Linear Regression

r-squared: .9121
MAE: 810.8975
rMSE: 1138.6845





Preprocessing & Baseline Model Development Code Path:
rrmahr / Marketing_Mix_Model / a_Notebooks / c_Preprocessing_Baseline_Models.ipynb

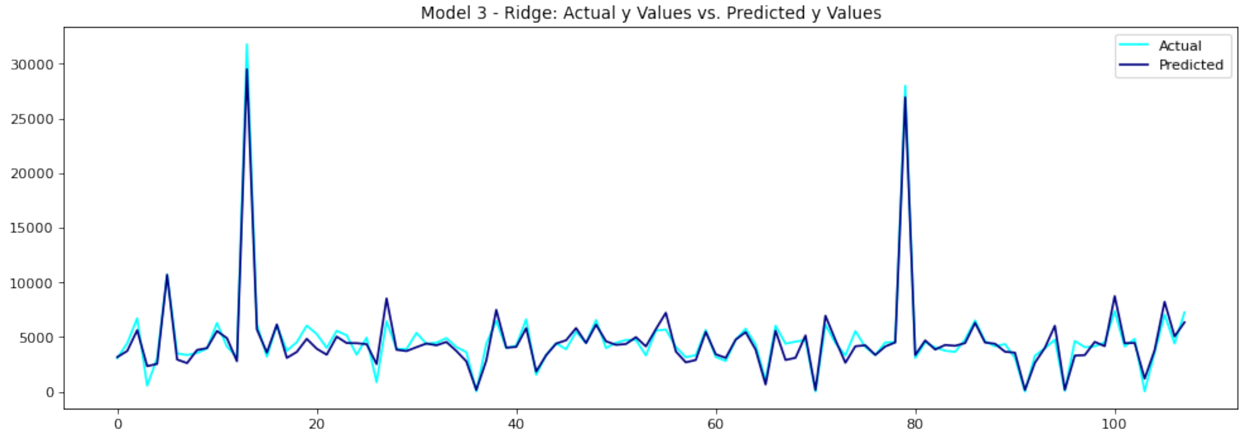
Modeling & Hyperparameter Tuning

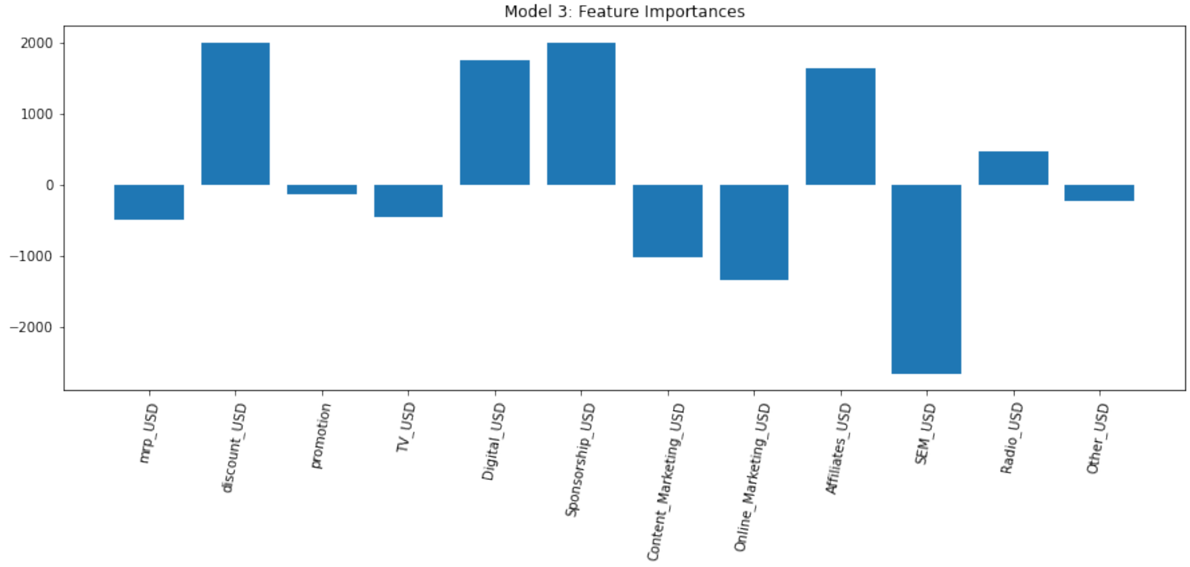
Eight additional models were developed and evaluated including two ridge regression models, two lasso regression models, and four random forest regression models. The ridge regression and lasso regression models were selected for their ability to address multicollinearity by adding a penalty term to the regression model. Ridge regression models were developed for both the option 1 and option 2 datasets conducting 5-fold cross-validation to select the best value for alpha, the penalty term, using scikit-learn's linear_model.RidgeCV and linear_model.Ridge (Pedregosa, 2011). Similarly, lasso regression models were developed for both the option 1 and option 2 datasets conducting 5-fold cross-validation with a tolerance for optimization of 0.0005 to select the best value for alpha using scikit-learn’s linear_model.LassoCV and linear_model.Lasso (Pedregosa, 2011). Random forest regression models were also developed for both the option 1 and option 2 datasets for their handling of outliers since the 3-day rolling average data still had some large outliers. The random forest models were developed using scikit-learn’s ensemble.RandomForestRegressor and were optimized using model_select.RandomizedSearchCV with 3-fold cross-validation (Pedregosa, 2011). Each of the model’s results are summarized below. For the random forest regression models, the feature importances were calculated based on the Gini importance.

Model 3: Option 1 Ridge Regression

Hyperparameters:
alpha: 0.1

r-squared: .9611
MAE: 575.8927
rMSE: 757.1106

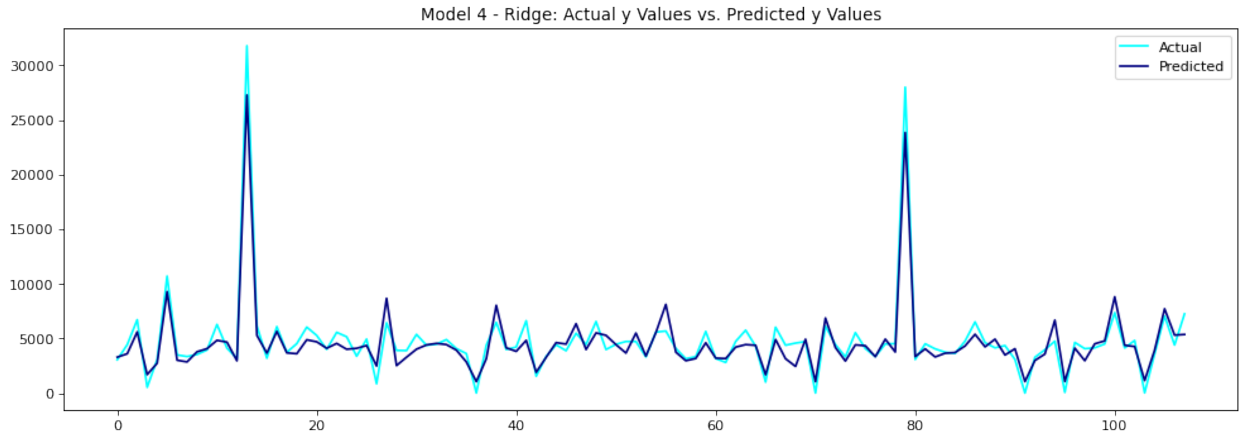


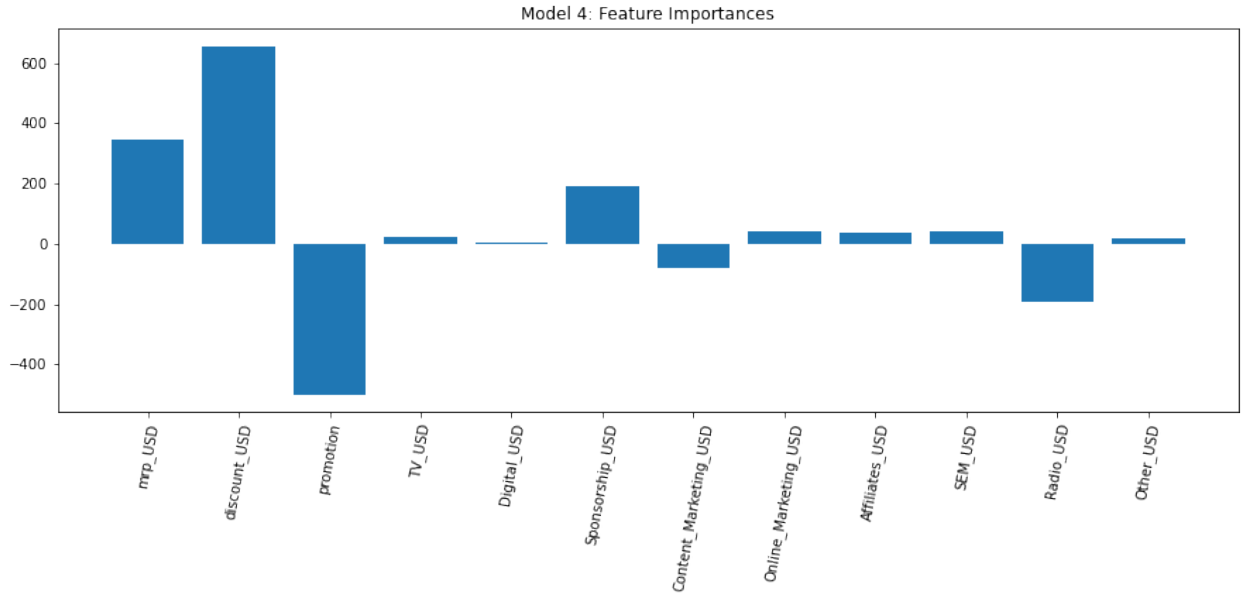


Model 4: Option 2 Ridge Regression

Hyperparameters:
alpha: 10

Performance Metrics:
r-squared: .9231
MAE: 781.9295
rMSE: 1064.8538


Model 4 - Ridge: Actual y Values vs. Predicted y Values

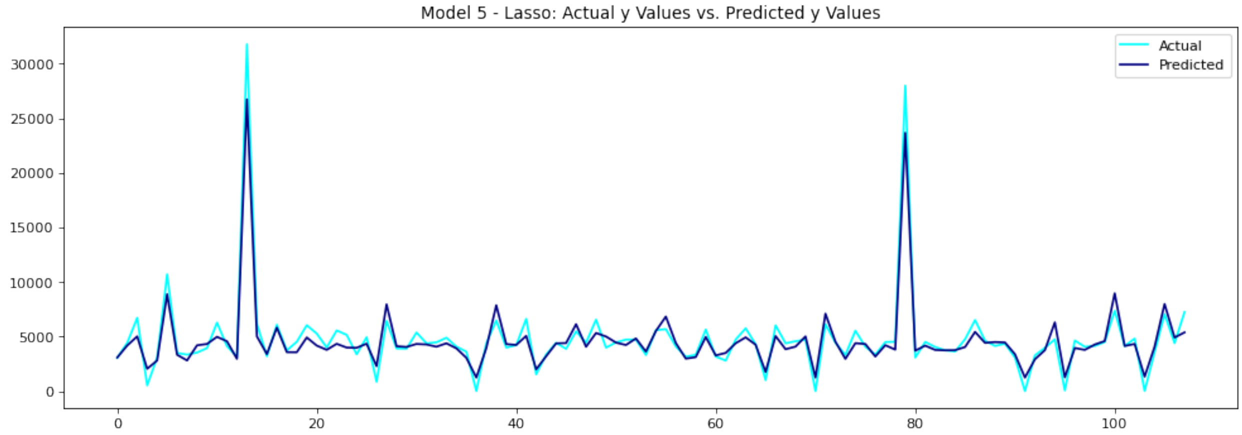

Model 4: Feature Importances

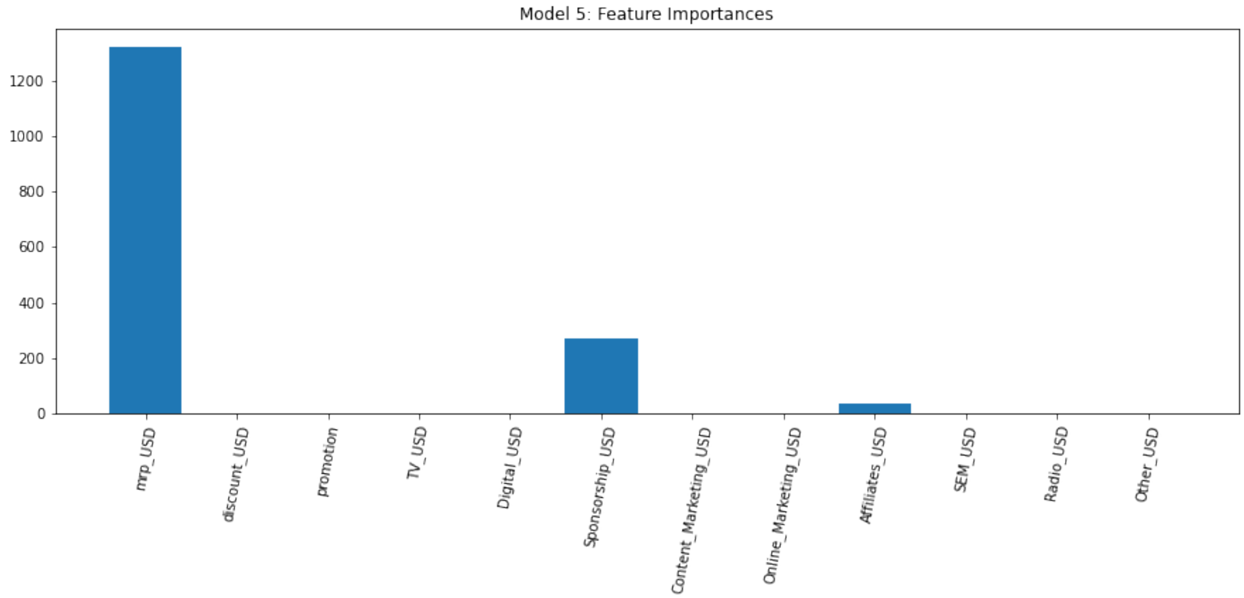
Model 5: Option 1 Lasso Regression

Hyperparameters:
alpha: 134.5518

Performance Metrics:
r-squared: .9317

MAE: 684.1494
rMSE: 1003.5886

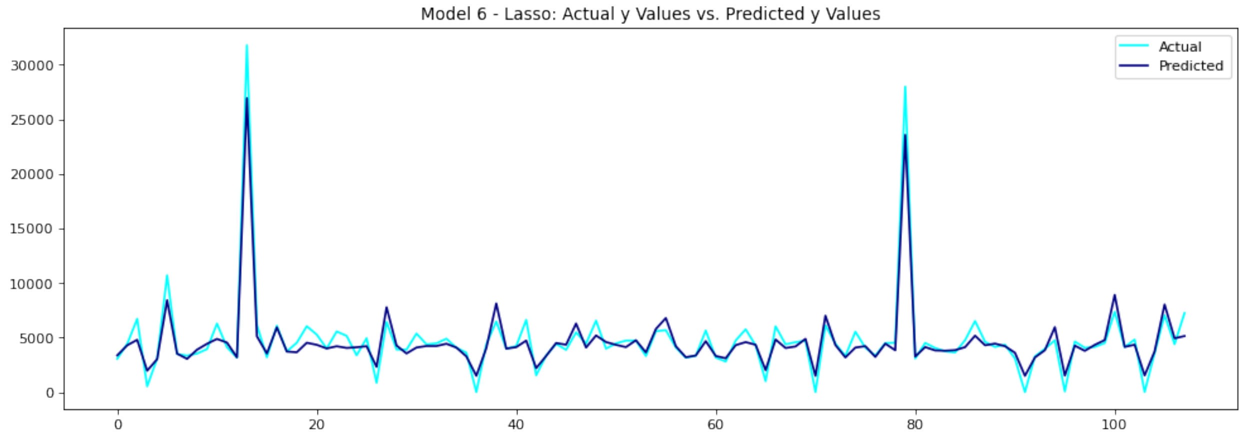

Model 5 - Lasso: Actual y Values vs. Predicted y Values

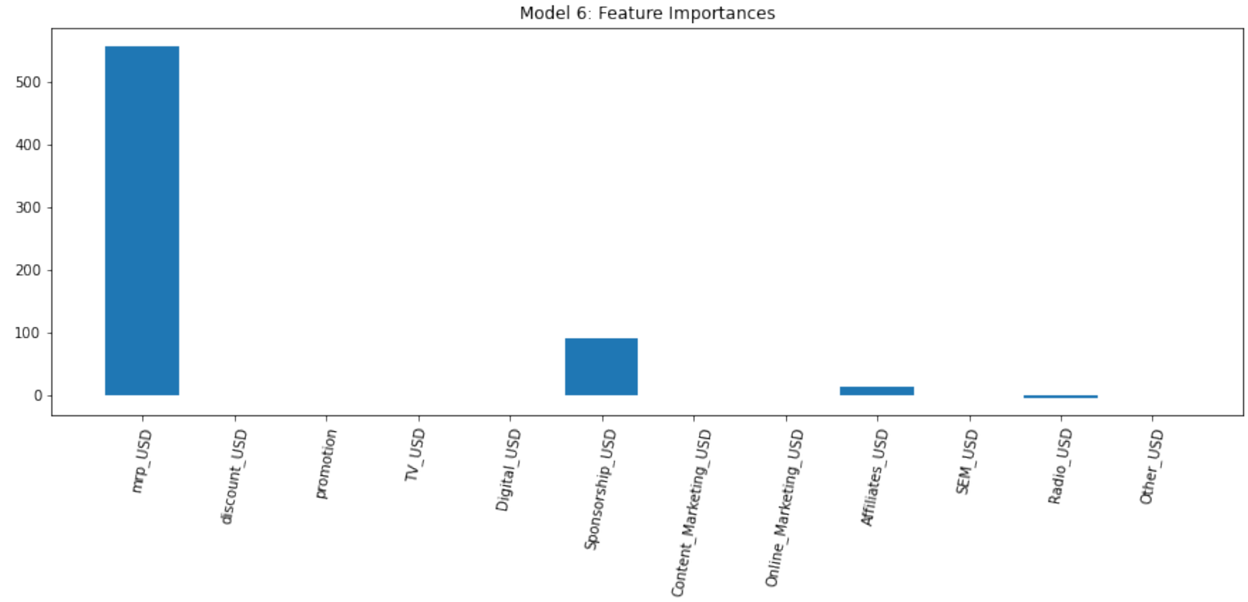

Model 5: Feature Importances

Model 6: Option 2 Lasso Regression

Hyperparameters:
alpha: 134.5518

Performance Metrics:
r-squared: .9264
MAE: 686.0254
rMSE: 1041.7917

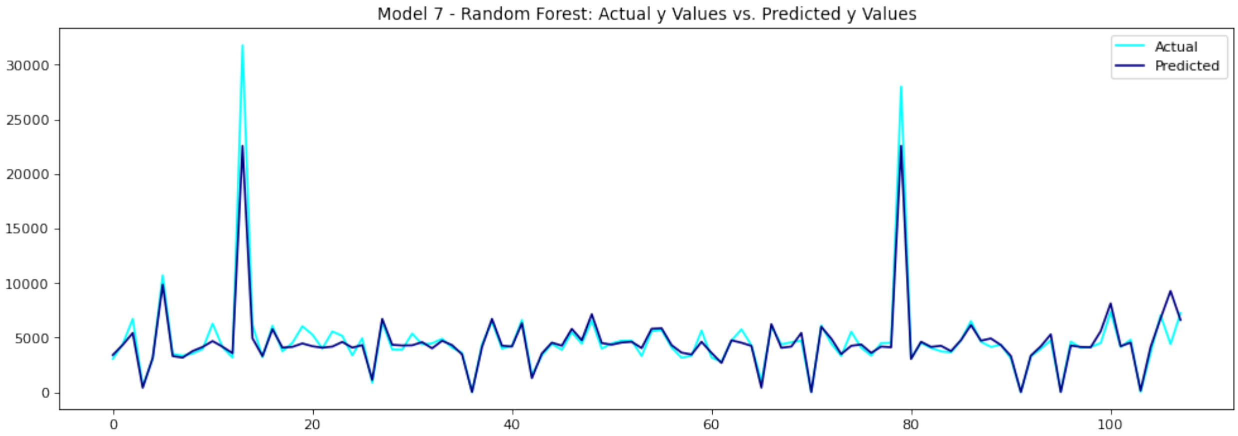


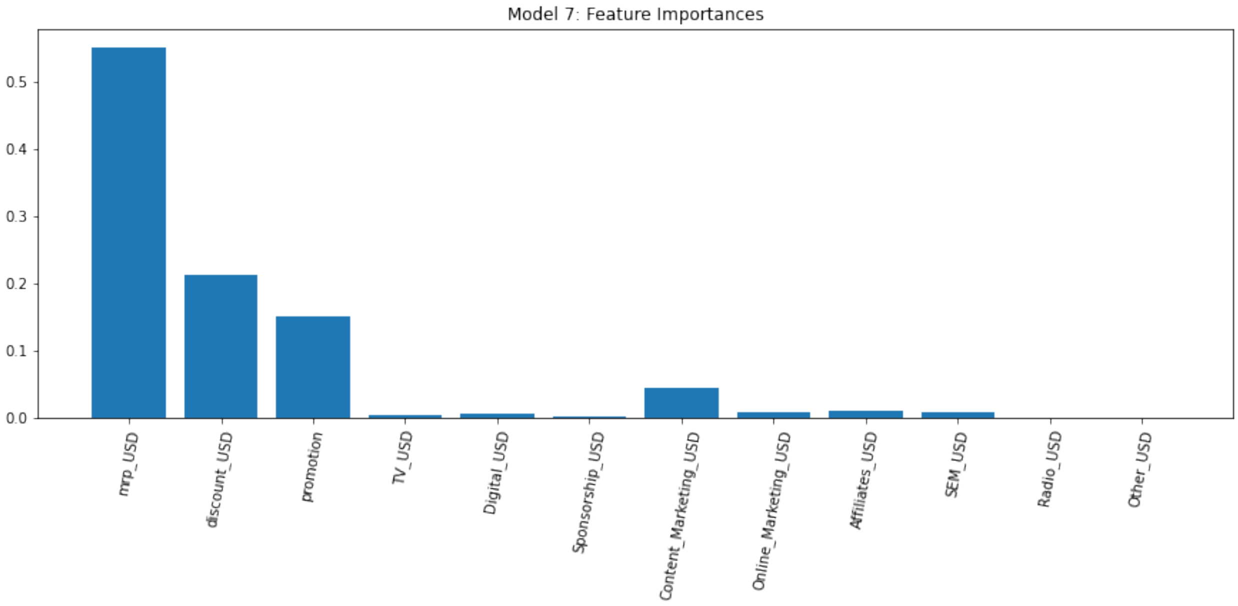


Model 7: Option 1 Random Forest Regression w. Default Hyperparameters

Hyperparameters:
bootstrap: True
max_depth: None
max_features: auto
min_samples_leaf: 1
min_samples_split: 2
n_estimators: 100

Performance Metrics:
r-squared: .8951
MAE: 549.1776
rMSE: 1243.9442

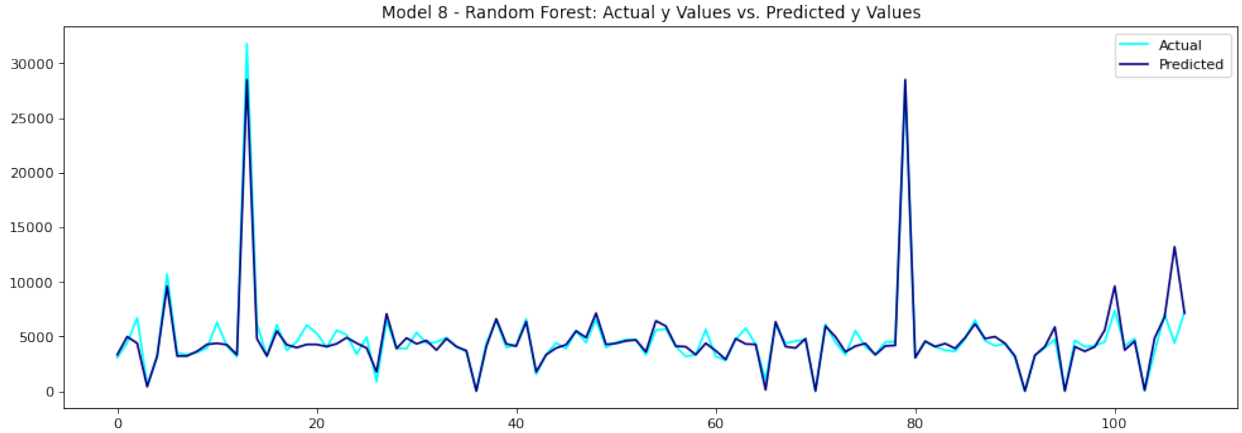


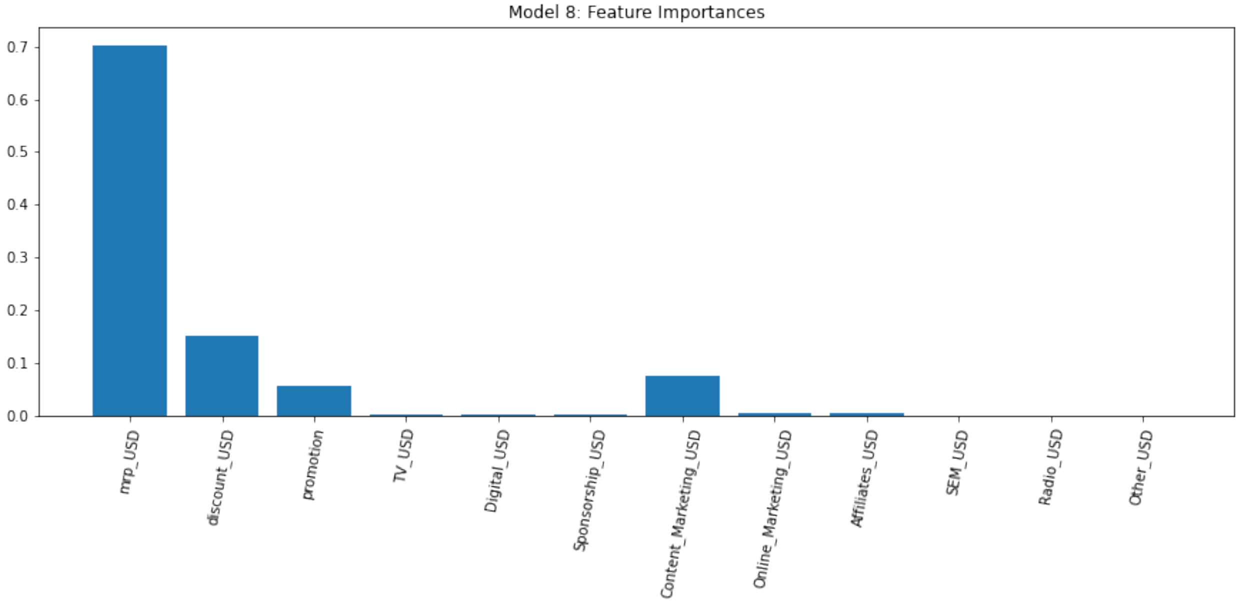


Model 8: Option 1 Random Forest Regression w. Hyperparameter Tuning

Hyperparameters:
n_estimators: 1783,
min_samples_split: 5,
min_samples_leaf: 1,
max_features: auto,
max_depth: 40,
bootstrap: False

Performance Metrics:
r-squared: .9143
MAE: 558.0814
rMSE: 1124.3996

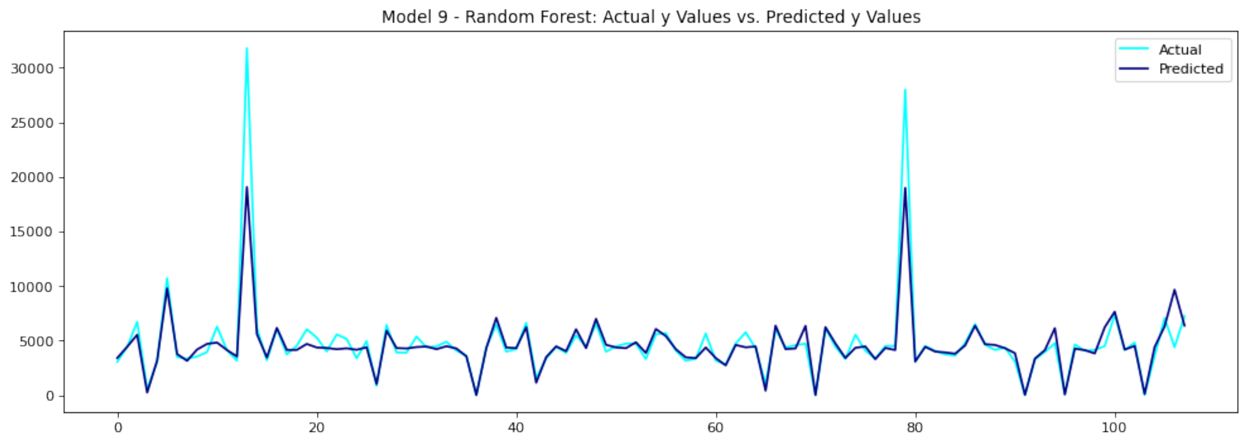


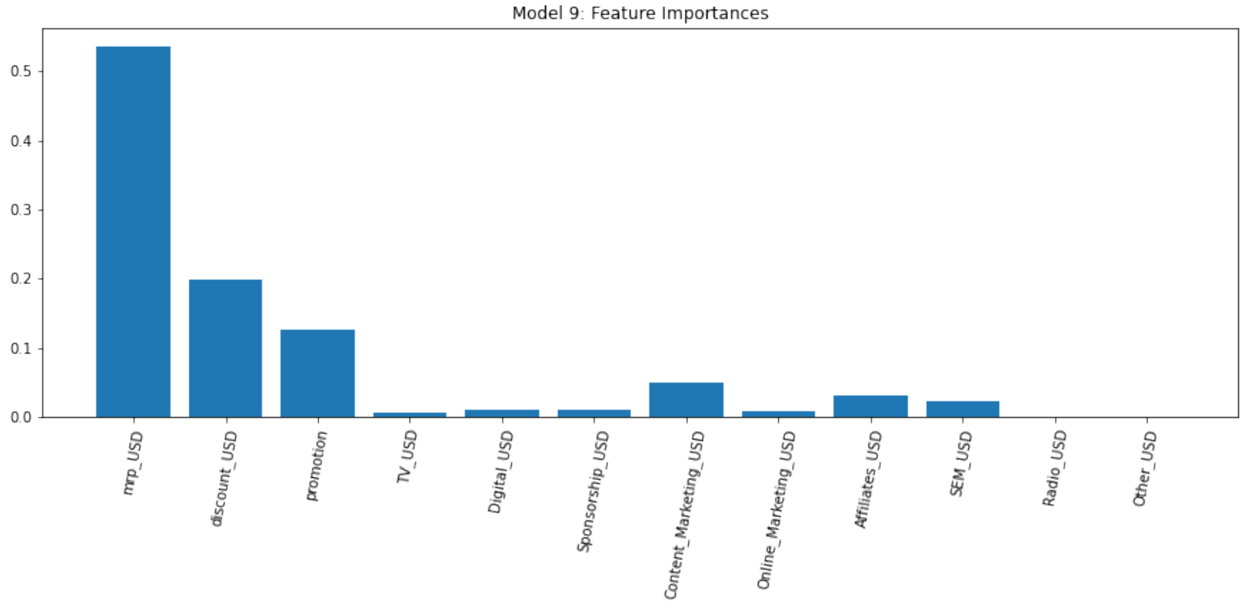


Model 9: Option 2 Random Forest Regression w. Default Hyperparameters

Hyperparameters:
bootstrap: True
max_depth: None
max_features: auto
min_samples_leaf: 1
min_samples_split: 2
n_estimators: 100

Performance Metrics:
r-squared: .8085
MAE: 646.0023
rMSE: 1680.3987

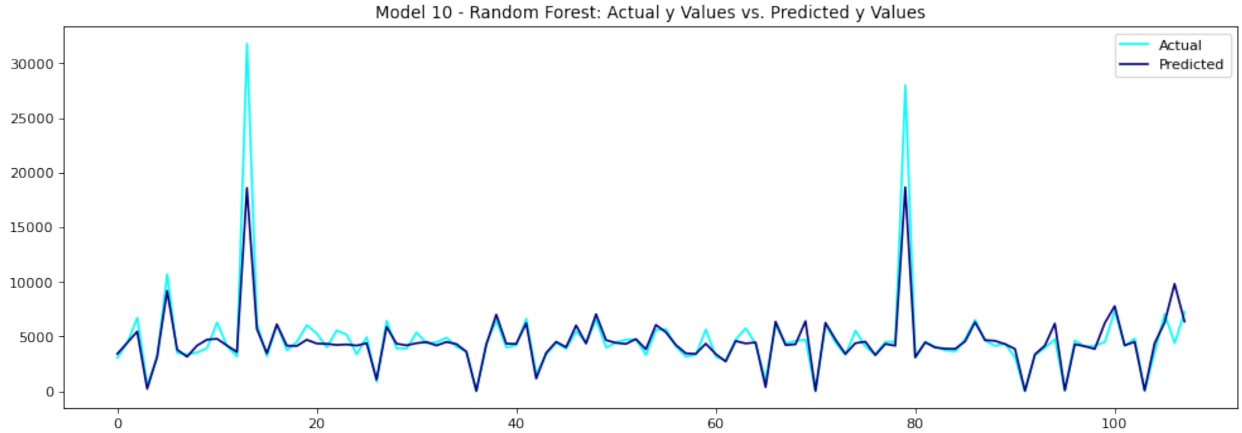


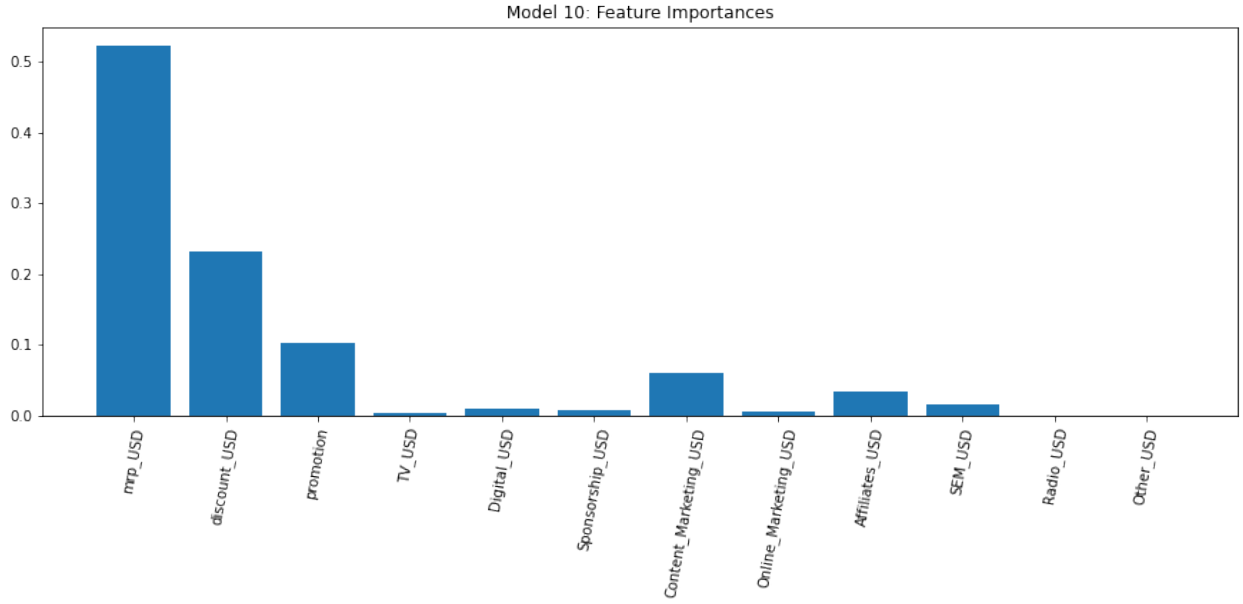


Model 10: Option 2 Random Forest Regression w. Hyperparameter Tuning

Hyperparameters:
n_estimators: 266,
min_samples_split: 2,
min_samples_leaf: 1,
max_features: auto,
max_depth: 110,
bootstrap: True

Performance Metrics:
r-squared: .7944
MAE: 665.4230
rMSE: 1741.2403





Modeling & Hyperparameter Tuning Code Path:
rrmahr / Marketing_Mix_Model / a_Notebooks / d_Modleling_Hyperparameter_Tuning.ipynb

Model Selection & Conclusion

Each of the models were compared based on their r-squared, MAE, and rMSE regression evaluation metrics as shown in the table below.

Model	Type_Dataset	r2	r2_rank	MAE	MAE_rank	rMSE	rMSE_Rank	Avg_Rank
1	OLSReg_1	0.96097	2	586.07757	4	758.6619	2	2.7
2	OLSReg_2	0.9121	7	810.8975	10	1138.6845	7	8.0
3	RidgeReg_1	0.9611	1	575.8927	3	757.1106	1	1.7
4	RidgeReg_2	0.9231	5	781.9295	9	1064.8538	5	6.3
5	LassoReg_1	0.9317	3	684.1494	7	1003.5886	3	4.3

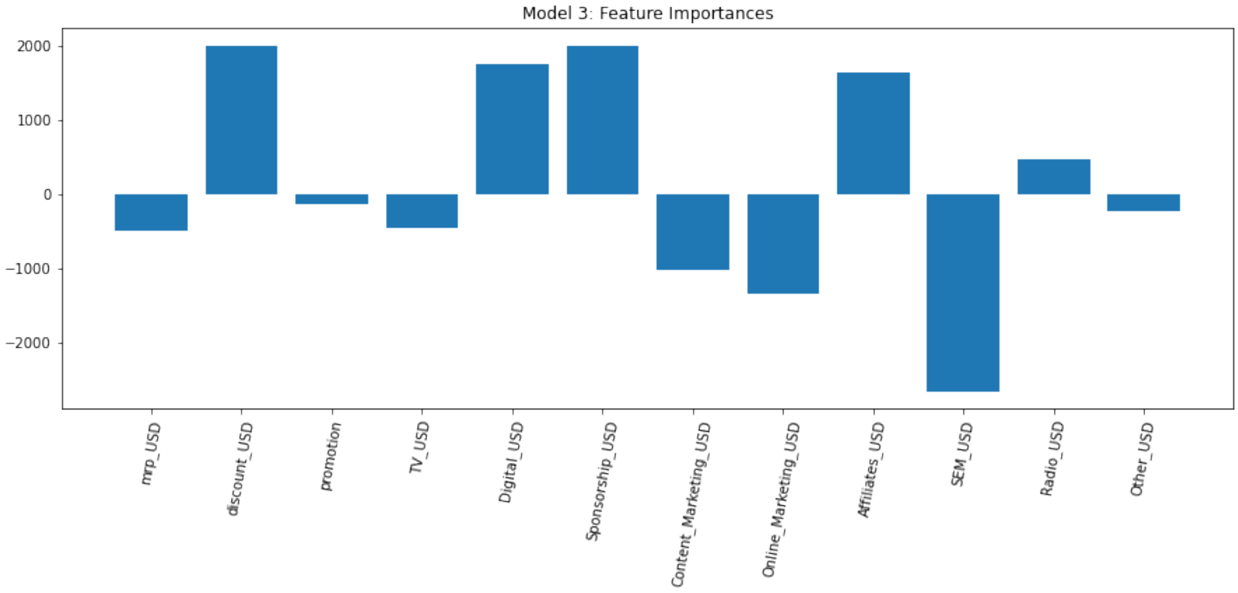
6	LassoReg_2	0.9264	4	686.0254	8	1041.7917	4	5.3
7	RandomForestReg_1	0.8951	8	549.1776	1	1243.9442	8	5.7
8	RandomForestReg_1	0.9143	6	558.0814	2	1124.3996	6	4.7
9	RandomForestReg_2	0.8085	9	646.0023	5	1680.3987	9	7.7
10	RandomForestReg_2	0.7944	10	665.423	6	1741.2403	10	8.7

Overall, the highest performing model was model 3 which utilized a ridge regression method with a penalty term of 0.1 on the option 1 dataset where monthly media investments were split evenly across the number of days in each month in the sales dataset. Model 3, predicted 96.1% of the variance in daily unit sales with an average absolute error rate of about 576 daily unit sales (~12% of daily unit sales on average).

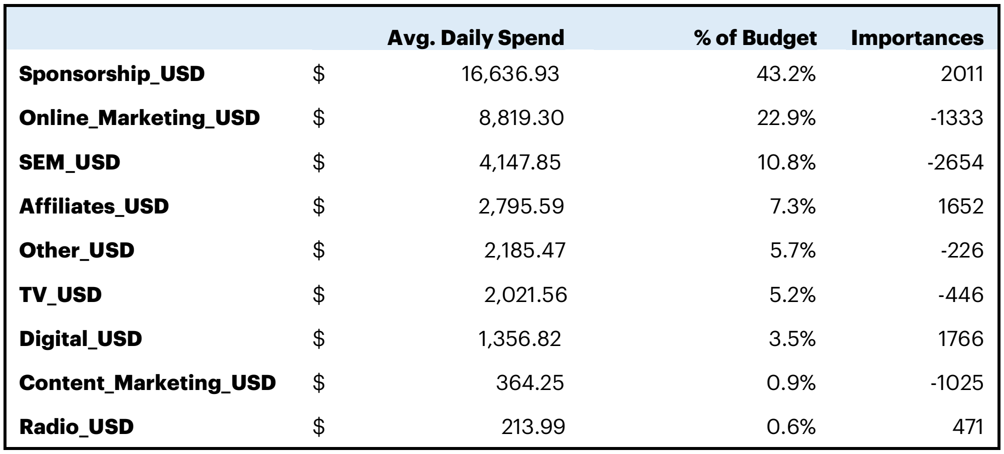
Applying the Model to the Business Problem

According to the feature importances calculation which indicates the average change in daily unit sales expected from a one standard deviation increase in the media investment, the media investments with the strongest impact on daily unit sales were:

- Search engine marketing spend: 2,654 *decrease* in daily unit sales
- Sponsorship marketing spend: 2,011 increase in daily unit sales
- Digital marketing spend: 1,766 increase in daily unit sales
- Third-party affiliate marketing spend: 1,652 increase in daily unit sales
- Online marketing spend: 1,333 *decrease* in daily unit sales
- Content marketing spend: 1,025 *decrease* in daily unit sales



In the 2015-2016 fiscal year, the media spend budget was allocated as shown in the table below:



		Avg. Daily Spend	% of Budget	Importances
Sponsorship_USD	$	16,636.93	43.2%	2011
Online_Marketing_USD	$	8,819.30	22.9%	-1333
SEM_USD	$	4,147.85	10.8%	-2654
Affiliates_USD	$	2,795.59	7.3%	1652
Other_USD	$	2,185.47	5.7%	-226
TV_USD	$	2,021.56	5.2%	-446
Digital_USD	$	1,356.82	3.5%	1766
Content_Marketing_USD	$	364.25	0.9%	-1025
Radio_USD	$	213.99	0.6%	471

This shows that the sponsorship, third-party affiliate, digital, and radio (although to a lesser extent) marketing efforts had a positive return on investment in the 2015-2016 fiscal year. Therefore, the business recommendation would be to decrease the search engine marketing, online marketing, and content marketing budgets in the next fiscal year and reallocate those funds across the sponsorship, third-party affiliate, digital, and radio marketing budgets. Based on their low impact and negative relationship with unit sales, eliminating TV and other marketing channels might be considered as well. A next step would be to develop an optimization model to quantify exactly how much to decrease the SEM, online, and content marketing budgets and how much to increase the sponsorship, third-party affiliate, digital, and radio marketing budgets.

Limitations

1. Monthly-level media investment data:
   One of the biggest challenges in developing the models for this project was that the media investment data was at the monthly level while the sales data was at the daily level. Having media investment data at the daily level would help to identify a one-to-one relationship between media investment and daily sales. Since that data was not available, the model results may not reflect the daily relationship between media investment and unit sales as accurately.

2. Time-series length:
   Although the sales dataset had over 1.5 million rows of data, this data only spanned across one fiscal year. In order to properly measure and control-for seasonality, at least two years of data would be necessary and five years would be ideal.

3. Missing long-term impact variables:
   Because the Kaggle dataset did not include information about the business such as company name and brand names for the products, identifying and controlling-for

long-term impact variables such as competition and halo and cannibalization effects was not possible.

4. Missing media campaign flighting dates:
   Because the media investment dataset was missing flighting dates, it was not possible to measure adstock effect which is "the memory effect carried over from the time of first starting advertisements" (A Complete Guide to Marketing Mix Modeling, n.d.). Transforming the data to account for the adstock effect would more accurately reflect the relationship between unit sales and media investments.

References

A Complete Guide to Marketing Mix Modeling. (n.d.). Retrieved from LatentView: https://www.latentview.com/marketing-mix-modeling/

Anish, A. (2020, 11 24). *Time Series Analysis*. Retrieved from Medium: https://medium.com/swlh/time-series-analysis-7006ea1c3326#:~:text=Trend%3A%20The%20linear%20increasing%20or,be%20explained%20by%20the%20model.

Indian Rupee to US Dollar Spot Exchange Rates for 2015. (n.d.). Retrieved from Exchange Rates UK: https://www.exchangerates.org.uk/INR-USD-spot-exchange-rates-history-2015.html#:~:text=Average%20exchange%20rate%20in%202015%3A%200.0156%20USD.

Kumar, R. (2017, 12 20). *Market Mix Modeling (MMM) -- 101*. Retrieved from Towards Data Science: https://towardsdatascience.com/market-mix-modeling-mmm-101-3d094df976f9

Miglani, A. (2020, 09 06). *DT MART: Market Mix Modeling*. Retrieved from Kaggle: https://www.kaggle.com/datatattle/dt-mart-market-mix-modeling

Pedregosa, e. (2011). Scikit-learn: Machine Learning in Python. *JMLR 12*, 2825-2830. Retrieved from scikit-learn: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html?highlight=tfidfvectorizer#sklearn.feature_extraction.text.TfidfVectorizer

Perktold, J., Seabold, S., & Taylor, J. (2021, 02 02). *statsmodels.tsa.stattools.adfuller*. Retrieved from statsmodels: https://www.statsmodels.org/stable/generated/statsmodels.tsa.stattools.adfuller.html

Pulagam, S. (2020, 06 06). *Towards Data Science*. Retrieved from How to detect and deal with Multicollinearity: https://towardsdatascience.com/how-to-detect-and-deal-with-multicollinearity-9e02b18695f1

Variance Inflation Factor. (n.d.). Retrieved from statsmodels: https://www.statsmodels.org/stable/generated/statsmodels.stats.outliers_influence.variance_inflation_factor.html