International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 07 | July - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Identifying the Base Sales Contribution for MTA Model

Sandnya Dalvie, Sameer Siddiqui, Neelima Ratra

Abstract:

Multi-touch attribution (MTA) is a technique for measuring marketing effort that considers all the consumer touchpoints and assigns a certain percentage of credit to each channel so that marketers can evaluate the importance of each touchpoint in generating a potential sale. MTA is a bottom-up strategy that uses cookies to give users'-level perceptions of every channel, format, and ad creative before the sale event.

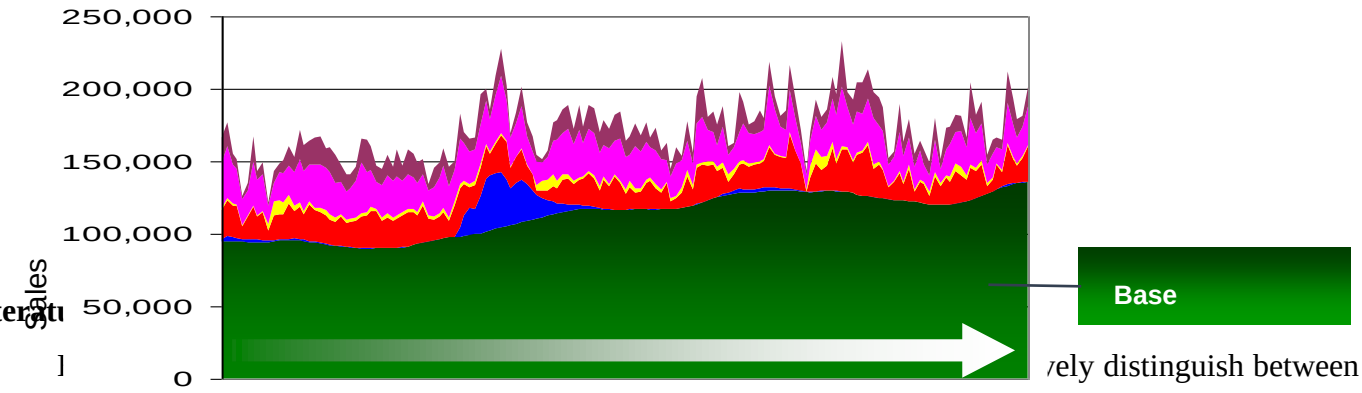
The strategy underlines the fact that not all purchases are a direct result of marketing efforts but can also arise from other non-obvious factors when capturing customer event logs to attribute sales. This highly sophisticated automated solution considers factors like seasonality, specific macroeconomic factors (like change in economic index, inflation, or incentive to promote sales like lower loan interest rates), government subsidies to avoid falling in the pitfall of assigning all the credit entirely to Digital Marketing drivers. MTA then assigns sales credit to each of the touch points using a highly effective Deep Learning algorithm.

The goal of the article is to build a statistical process for determining which sales are driven by actual marketing efforts and which sales are generated without spending much on product advertisements, campaigns, or other non-promotional activities. It provides a detailed explanation of a variety of scenarios for predicting the Base Sales percentage for online leads and sale produced by modelling approaches.

Problem Statement:

Customer Event Logs (Clicks and Impressions) captured at most granular level help in attributing sales and credits to the Digital channels. While running MTA models, it is assumed that all sales are driven from Marketing Effort on the Digital Channels. However, it is not a fair assumption. Some of the sales might have come from Brand Awareness, Economic factors complimenting Marketing efforts. The sales which are also driven due to other external factors like certain macro-economic factors, change in economic index, inflation, or incentive to promote sales, lower loan interest rates, government subsidies & competitor launches contributes to **Base Sales**.

The framed problem is to identify a robust analytical process which can decompose and identify the sales occurring due to marketing effort and the other one due to the above-described external factors.



Literatul ely distinguish between sales driven by marketing efforts and external aiding factors, as they primarily focus on fragmented aspects like trends, seasonality, and sales forecasting.

The primary focus of the paper we referred to "**Sales Prediction using Linear Regression**" is to predict future sales based on historical data using the technique of linear regression. The paper introduces a user-friendly Graphical User Interface (GUI) that enhances the visualization of the sales prediction process. Specifically, the prediction models developed in this study aim to establish the relationship between sales and advertising across various media channels, including TV, newspaper, and social media, by employing simple linear regression techniques.

Similarly, in the article referred "**Regression Forecasts with Seasonality**" the approach involves representing seasonality through dummy variables and applying a simple linear regression for forecasting purposes.

The paper titled "**Sales Prediction Based on ARIMA Time Series and Multifactorial Linear Model**" by Ziru Zhang that we studied explores the use of statistical methods and the R programming language to develop prediction models for sales forecasting. The study considers multiple factorial linear regression and time series models. The prediction results of both models are compared using the root mean square error (RMSE), and their respective advantages and disadvantages are analyzed.

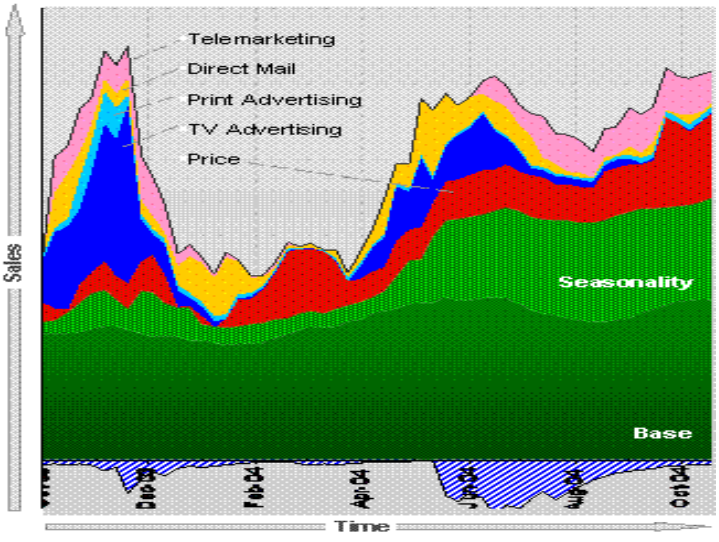
In their paper "**Dynamic seasonality in time series**" Mike.K.P.So and Ray S.W.Chung explore a new class of time series models that capture dynamic seasonality. They compare that with the traditional Holt-Winters method which assumes that all seasons have the same smoothing parameters. However, it has been noted that this approach may not be suitable for time series with periodically stationary behavior. The authors propose an alternative approach that goes beyond traditional seasonal models, which mainly focus on the dynamic mean and variance processes.

The research we explored "**In praise of Prais-Winsten: Evaluating autocorrelation methods in interrupted time series analysis**" studies the use of interrupted time series analysis in evaluating public health interventions while considering temporal correlation. The study assesses the performance of the Prais-Winsten method, Newey-West method, and ARMA modeling, commonly used for correlated time series data. Using pneumococcal vaccine data and simulated autocorrelation scenarios, the study finds that both the Prais-Winsten and ARMA methods demonstrate lower mean square error. Notably, the Prais-Winsten method generally provides better coverage, making it a favorable choice for managing autocorrelation in interrupted time series analysis.

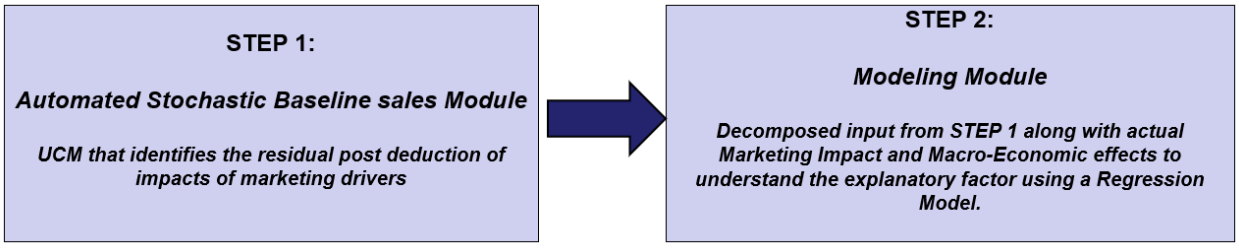
It is important to note that the above papers do not consider the impact of specific economic conditions that may influence sales for a brand. Additionally, they do not provide a detailed step-by-step segregation of sales influenced by marketing efforts versus sales influenced by external factors like seasonality or autocorrelation using machine learning techniques in a unified process. These aspects led us to areas for further research and exploration.

Proposed Solution:

The proposed solution tries to establish an Automated statistical process which can identify what sales come from Marketing Efforts and what come from base (external factors) for each product. It is a two-step process. **The first step** involves an Automated Baseline Creation Module that uses UCM (Unobserved Components Model) to decompose components like trend, seasonality, and cyclic behaviors in the data. Below is an illustrative representation of decomposed sales.



The second step uses the identified UCM baseline stochastic sales to further understand the implicit effect of macro-economic factors (like loan interest rates, consumer index, etc.) on actual sales. Removing the effect due to external factors will indicate the actual marketing efforts using a regression model for each product. Collectively the process will be called as "**Optimal Base Model Identification Process**".



Parameters to Validate the Models:

The output that is expected from the linear regression model should be such that :

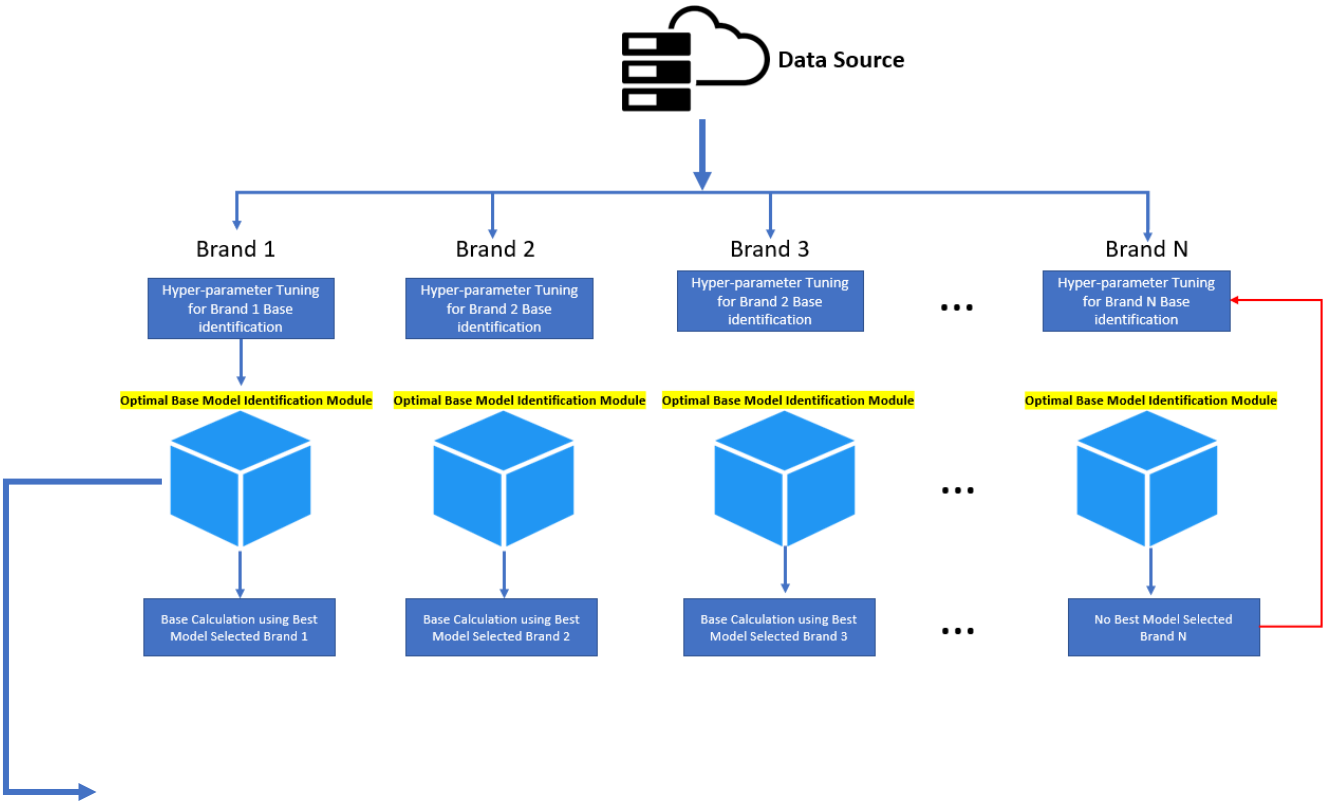
1. The R^2/ Adjusted R^2 of the model should be greater than the threshold value that is acceptable to the business and can be tuned as one of the hyperparameters , as it indicates better model fit.
2. The Durbin Watson statistic is used to identify autocorrelation. A value of this statistic is expected to be in a range acceptable to the industry standards which indicates absence of autocorrelation. In case where, a value is below or above the acceptable threshold range a simple linear regression model should be replaced with Prais-Winsten Regression model which will be explained in one of the cases below.
3. P-value of the variables in the model also play a very important role in identifying how significantly does the variable relate to sales. A lower P-value indicates your model fits the data well.
4. The regression estimates obtained for the marketing efforts should never be negative as we're identifying sales using them and sales can never be negative.

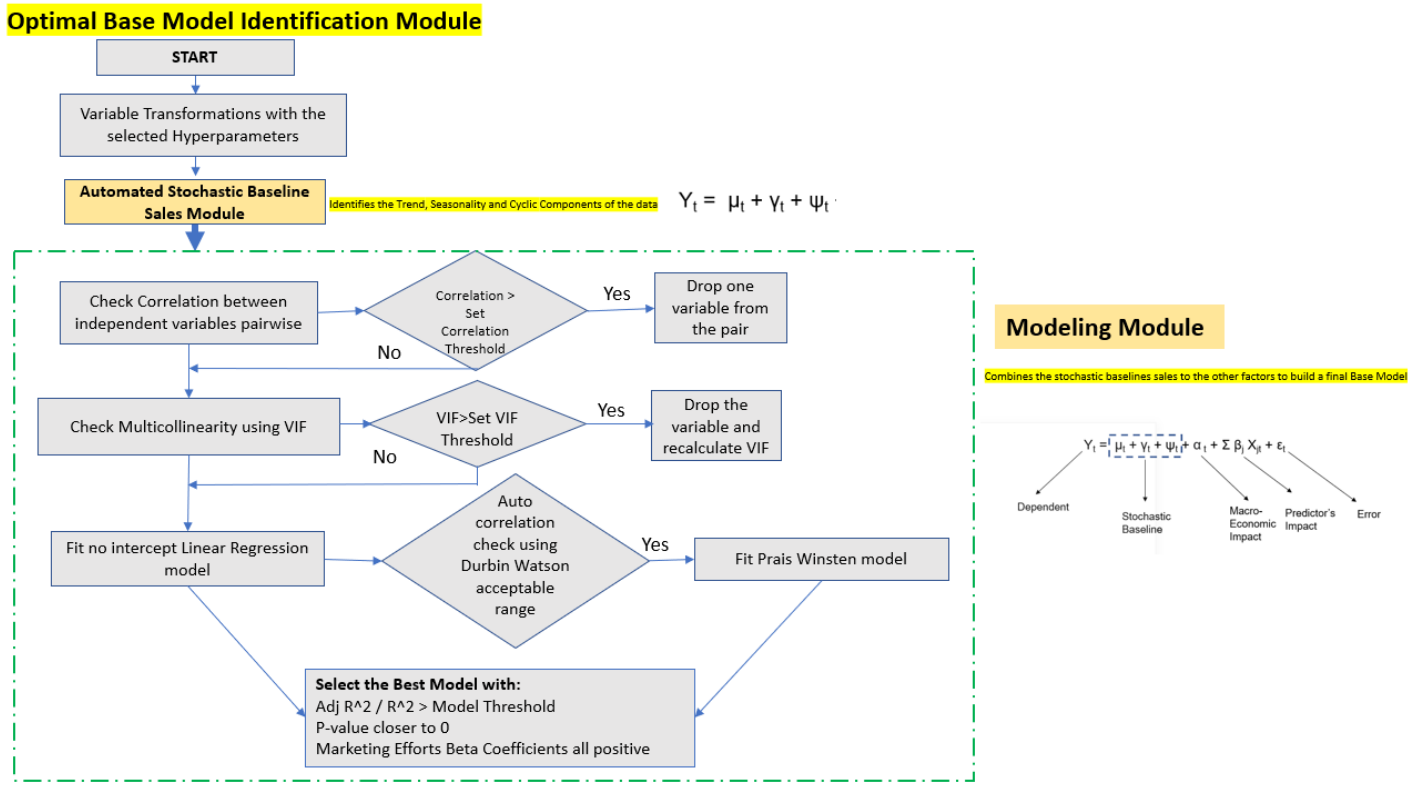
Uniqueness of the Solution:

The uniqueness of the solution lies in the fact that we have tried establishing a composite statistical process that decomposes every factor in an analytical and logical way to better understand the actual marketing contributions towards the sales and those that come from non-promotional activities. This **Optimal Base Model Identification** process is a container designed in a way that it can run in an automated way for multiple brands using the concept of parallelization. The baseline sales will be calculated for each product/brand by

segregating signal at product level and arriving at marketing effort at each individual product. This will enable the client to identify and plan further marketing strategies/campaigns for their products.

Described below is the unique work flow we have designed for this process followed by the detailed workflow of the **Optimal Base Model Identification** process:





Validation and Experiments:

To validate the Proposed solution ~~the~~ following Steps were followed in the **Optimal Base Model Identification Module**. These steps can run in an automated framework for multiple products/brands in parallel. It takes the input as data from say "**N**" brands and runs each of the below Steps for each the "**N"** brands in parallel.

STEP 1

Data Understanding and Hyper-parameter tuning:

- **Tuning AdStock and Lag for Data Transformation:**

  The aggregated marketing efforts are transformed to include Lag and AdStock factors.

  Transformed Marketing Efforts = Original Value + AdStock*Lag

  Lag = Original Value from the previous row for that variable in the data

  The values of Lag and AdStock depend on the brand, the domain knowledge and expertise after a keen data understanding.

- **Permissible R^2 , Correlation,VIF and Durbin Watson statistic ranges for Best Model Selection**
  Based on the data and business understanding the ranges of the important model parameters can be decided or tuned.

STEP 2

Data Transformations:

1. Applying Lag and AdStock transformations to the data decided in STEP 1.
2. Data Normalization can be done using the following formula:
   Normalized marketing efforts for a week =
   round((marketing efforts for a week /Max(marketing efforts throughout all weeks)*100)

STEP 3

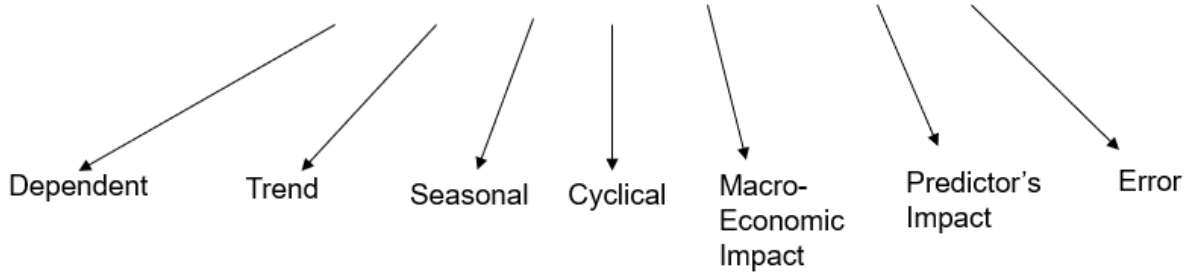
Automated Stochastic Baseline sales Module:

This step identifies the "Stochastic base" sales that happen without any marketing or promotional activities using Unobserved Component Model (UCM).

UCM is a time-based model that performs a time series decomposition into components such as trend, seasonal, cycle, and gives real time baseline sales that happens without any marketing efforts. The algorithm uses Kalman filter for estimation of unobserved components. The intelligent feature of this approach is that it accounts for the parameters that have influenced sales that we are not aware of, or we do not have an accurate way to track them.

The model is automated in a way that for every product it identifies and decomposes these undefined factors using the following equation:

$$Y_t = \mu_t + \gamma_t + \psi_t + \alpha_t + \Sigma \beta_j X_{jt} + \varepsilon_t$$



- The model components μ_t, γ_t , ψ_t , α_t and $\Sigma \beta_j X_{jt}$ are assumed to be independent of each other

- $\mu_t + \gamma_t + \psi_t$ is the "stochastic base" of the dependent series that is predicted in this Step which goes further as an input in the Modelling Module.

STEP 4

Model pre-requisites check:

The "stochastic base" from the previous step is now combined to the actual marketing efforts and the macro-economic factors.

But before we proceed to the Modelling Step we ensure some assumptions of Regression Model be checked.

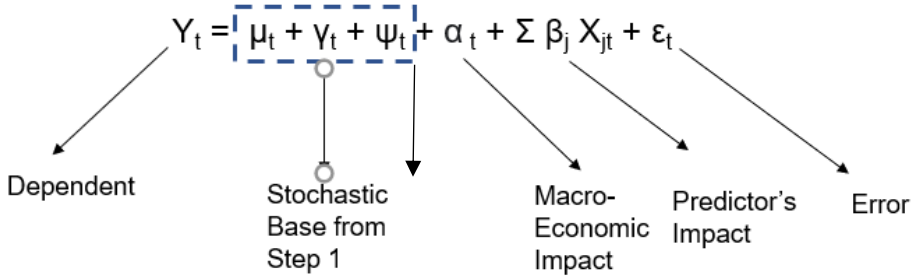
1. Checking Correlation between the independent variables (Pair Wise)
   a. If any pair of variables have a correlation greater than the set correlation threshold we retain only one variable from that pair.
2. Checking for multi-collinearity between the independent variables. (Using VIF)
   a. If any variable has a VIF value greater than the set permissible value that variable is excluded in model building.
   b. Removal of variables using VIF is done one by one.
   c. VIF is calculated for all the variables first, the variable with the highest VIF is removed from the list of independent variables.
   d. VIF is then re-calculated for the remaining independent variables and Step c is then repeated till all the variables have a VIF of less than the set permissible value.
3. Checking the Durbin Watson statistic for autocorrelation. When the Durbin-Watson statistic is not in the pre-defined range set in the hyper-parameters it indicates presence of autocorrelation. In this case we use the Prais-Winsten model in R.

STEP 5

Model Building Module:

Once the assumptions of linear regression are checked we move ahead with model building.

Here we now have the stochastic base and add the marketing efforts along with the effective macro-economic factors. And the final equation is as follows:



We use a simple linear Regression Model without considering the intercept in Python in cases where data follows all the assumptions using the statsmodel.api library.

In the case where there is presence of autocorrelation in the data a Prais-Winsten Regression model needs to be used available in R with the in-built library prais.

The model with an acceptable R^2, significant P-values, and positive Beta co-efficients for the marketing efforts is the best selected model.

If either or all the above model checks are not met for a product/brand the process goes back to STEP 1 to better tune the hyper-parameters and re-runs the entire process iteratively till the optimal results are achieved.

STEP 6

Base Sales Calculation:

Once the model has been built and the co-efficient value for the marketing efforts has been obtained, following steps are used to calculate Base Sales:

1. Multiply the week-wise or month-wise Actual Marketing Efforts with the model co-efficient to get the Estimated sales that come from Marketing Efforts for each week or month.
2. Subtract these Estimated sales that come from Marketing Efforts from the Total sales recorded during that week. This Value is the Base sales for that week / month.
3. Calculate total sales, estimated sales that come from Marketing Efforts and Base sales by adding them throughout all weeks / months.
4. Base Sales Percentage = (Total Base sales / Total sales) * 100

STEP 7

Incremental sale calculation for MTA:

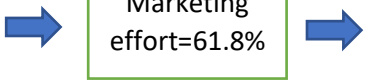
After the base sales percentage has been calculated in the previous step, marketing effort is calculated which is the actual sales resulted by investment on the marketing side.

Marketing Effort=Total Sales-Base sales

Below is the graphical representation of the actual flow of how marketing efforts are calculated.

As an example, we will use illustrative data. Beginning from the step 6, we assume that the calculated base sales 38.2%. With the above formula, marketing effort will come out as (100-38.2) 61.8%.

Channel	**MTA Sales**
Display Ads	$1,200.00
Paid Search	$900.00
Video Ads	$2,600.00
Email Campaigns	$3,478.00
Facebook	$267.00
Twitter	$134.00
Google Ads	$567.00



Channel	**MTA Incremental Sales**
Display Ads	$741.60
Paid Search	$556.20
Video Ads	$1,606.80
Email Campaigns	$2,149.40
Facebook	$165.01
Twitter	$82.81
Google Ads	$350.41

Future Directions

Use of real time macro-economic variables that might better explain the Base sales.

Use of current scenarios inflation/demonetization/wars/natural calamities that might affect sales in any way.

Conclusion

We have been able to establish a step wise analytical process that will help any client better understand the effectiveness of their strategies along with the external factors that could aid the sales of their product.

Once they become aware of the sales contributions that are coming from the base versus those because of their marketing efforts using this statistical process, it will enable them to plan the product marketing and investments with added intelligence.

References

https://www.researchgate.net/publication/365874983_Sales_Prediction_using_Linear_Regression

https://real-statistics.com/multiple-regression/multiple-regression-analysis/seasonal-regression-forecasts/#:~:text=Three%20dummy%20variables%20are%20required,than%20the%20number%20of%20periods).

https://www.researchgate.net/publication/369462420_Sales_Prediction_Based_on_ARIMA_Time_Series_and_Multifactorial_Linear_Model

https://www.sciencedirect.com/science/article/abs/pii/S0167947313003253

https://www.researchgate.net/publication/367767257_In_praise_of_Prais-Winsten_An_evaluation_of_methods_used_to_account_for_autocorrelation_in_interrupted_time_series

https://goois.net/chapter-9-measuring-success-with-mmm-mta-and-promotional-lift-highly-effective-marketing-analytics.html

https://www.home.neustar/blog/multi-touch-attribution-key-terms-to-know

https://etav.github.io/python/vif_factor_python.html#:~:text=The%20Variance%20Inflation%20Factor%20(VIF,if%20it%20were%20fit%20alone.

https://www.statisticshowto.com/durbin-watson-test-coefficient/

https://cran.r-project.org/web/packages/prais/prais.pdf