Analysing the Challenges and Opportunities in Media Mix Modelling using Sales and Media Time Series Data

Saumya Gupta
Dalarna University
Borlänge, Sweden
h20saugu@du.se

***Abstract*—Media Mix Modelling is a statistical analysis technique widely used by many businesses for the past two decades for evaluating the effectiveness of different media spends on sales volume or sales revenue. They use the inferences obtained in driving optimal budget allocations. Having been commonly used for so many years, advertisers and modellers ought to know the various challenges and opportunities involved in the technique, facilitated by a clear understanding of the fundamental concepts and methodologies. This analysis uses encrypted time-series data of media spending and sales information and shows how linear regression analysis on such data can extract valuable insights for efficient decision-making. Besides pointing out the favourable aspects during the process, our analysis will show the difference between data-driven model results and those based on modellers' choices in the model building with the uncertainties involved in such results.**

Keywords-Media Mix Modelling; Linear Regression Analysis; Multicollenearity; Time-Based Cross-Validation; Time-Series Media Spend and Sales Data; Media Contribution; Return on Investment

I. INTRODUCTION

Marketing Mix Modelling or Media Mix Modelling (MMM) are two very similar terms, both referring to the technique where businesses or precisely the marketing teams of these businesses attempt to model the relationship between various marketing variables and the outcome of the corresponding sales through regression analysis. Typically, on the predictors' side, we have media variables – online (display ads, websites, blogging, social media, e-mail marketing and pay-per-click advertising (PPC)) or/and offline (television, radio, newspaper print collateral) – and control variables, including seasonality, weather (precipitation and temperature), gross domestic product (GDP), inflation and market competition information that accounts for the external factors affecting sales. In addition, businesses usually measure marketing variables in spending or specific activities, such as gross rating points (GRPs) for television. On the response side, we have sales volume or revenue. Sales have two components – base (fixed sales due to brand equity) and incremental (sales generated from marketing activities). Businesses aggregate this information over years of recording and logging to prepare time-series data of different frequencies (weekly, daily, or monthly), suitable for analysis. Many metrics such as media channel contributions denoting media impact on sales, return on investments (ROIs) or return on advertising spending (ROAS) are calculated. They then use these metrics for budget allocation decisions - such as keeping the total budget constant, shifting the money from a low ROI media channel to a high ROI channel. It is the fundamental workflow of how MMMs work for businesses. These decisions bring in enhanced performances increasing profits, prominence and eventually, customer base.

However, all of it assumes that these models provide valid results. During the modelling process, the modeller faces numerous challenges, due to which the model results, instead of the data, depend significantly on the choices made by the modeller in handling those challenges. Therefore, before performing such regression analysis for making important marketing decisions, modellers must understand the common problems that are part of modelling such relationships.

This analysis aims to help the modeller understand these problems and what causes them. Knowing this, they can make informed decisions according to their business settings. This report will walk the modeller through a multilinear regression analysis on a time series data of media spending and sales volume information. The report will call attention to the setbacks faced and handled for the specific data set throughout the modelling process. In the end, it will list the challenges in detail. In this way, the modeller understands the modelling process, step-by-step; understand which steps are problematic and do not miss out on the details.

The author has organised the rest of the report as follows. Section II peeks at the scholarly works on this topic. The regression analysis experiment and the handling of the encountered difficulties in the modelling process are present in section III. Section IV highlights the opportunities in the technique by listing the insights from the model inferences. Section V lists the challenges in more detail. Finally, Section VI gives the concluding statements for this analysis.

II. Literature Review

The marketing mix concept was introduced to the world by Neil H. Borden in [5]. Consequently, the process became a part of demand modelling, where we predict demand for a product selling in a particular division/region over a specific amount of time. Nevertheless, over decades, the technique has not been used for prediction but distinguishing one media channel from another in terms of impact on sales. Moreover, the demand modelling literature does not provide any concrete guidance on the functional form of the model or the control variables to use [2]. Due to these ambiguities and partly the uncertainties from the complex sales response processes, modellers often face many inherent challenges in modelling the required relationship.

Another major problem is multicollinearity. Mainly, we use linear regression analysis to do MMM, and one of its significant weaknesses is that its results are not reliable when predictors are highly correlated, something prevalent amongst marketing variables. [8] concludes that Shapley value regression to be one of the best methods to resolve this adversity and proposes a direct implementation of a simplified approach towards calculating Shapley values.

III. Experiment

This section explains the media mix modelling experiment for this analysis and highlights the setbacks and solutions in the process. Fig. 1 explains the workflow of the experiment.

A. Data Set

To fulfil both our objectives, we use encrypted sample data of media spending, with corresponding recorded sales volume, available for public use on Kaggle. It is a time-series medium frequency weekly data, containing marketing spends for five media channels – Google search impressions, Facebook impressions, e-mail impressions, YouTube (paid and organic) views and affiliate impressions. Data is present from Jan 2018 – Feb 2020 (113 weeks) for 27 firm divisions. However, since the data is encrypted, we do not know whether these represent the firm sub-divisions or different products marketed and sold by the firm. Nevertheless, this analysis holds for both kinds. In total, we have 3051 observations or patterns with ten variables or fields for our analysis.

Initial data checks and enquiries with the data owner reveal that YouTube (YT) views are also present in an aggregated form, i.e., an almost summation ("almost" due to encryption, discussed in a later section) of paid and organic views, in a different column, constituting one of the ten columns. We discuss handling YouTube views in Section III-D. The data set has no missing values.

B. Data Pre-processing

1) Rectifying Incorrect Encryption: For whatever reason, the data has 27 division names encrypted to 26 alphabets of the English language, wherefore, there exist wrong encryptions of two divisions with the same letter 'Z'. Therefore, we rectify the naming of one of them to 'AA'.

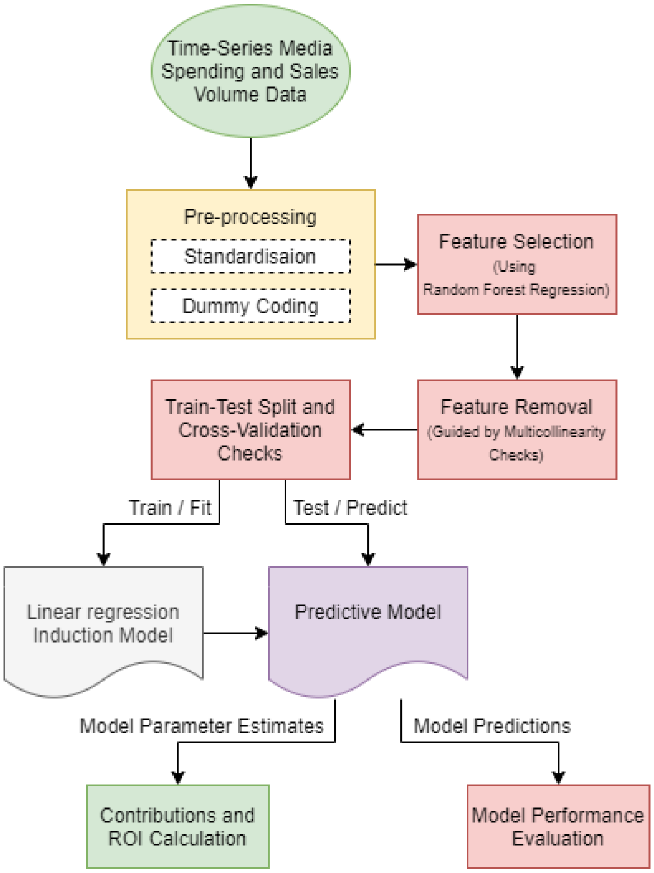

Figure 1. The media mix modelling process.

2) Standardisation: Since spending in different media channels is of widely varying ranges, standardisation is necessary. For time-based data, meaningful information is present within the feature values from one observation to another. At the same time, we want to be able to compare values within different features. Hence, we use the standard scaler from scikit-learn [3] that removes the mean and scales the data to unit variance feature-wise independently. We standardise all media spends and sales volume.

3) Dummy Coding: To use the division information in the regression analysis, we dummy-code the division names column. It results in 27 more columns representing a division each (1 – spends done at that division; 0 – spends not done at that division).

4) Sorting: Any splitting in the data for cross-validation (CV) during model selection or final model building should give importance to the time-based order. Hence the whole data is sorted chronologically, i.e., based on weekend dates.

C. Exploratory Analysis

Correlation calculations in Fig. 2 show high correlations between spends of Google, Facebook and e-mail impressions

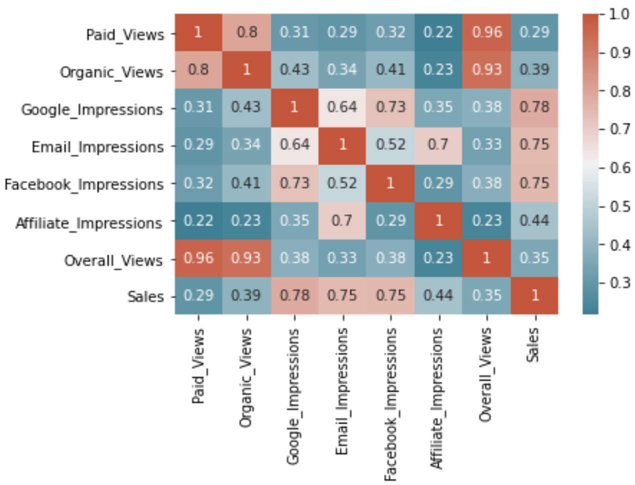

Figure 2. Heatmap shows pairwise correlations between media channels spending and sales volume.

and sales volume. Among the predictors, sufficiently high positive correlations are present between paid and organic YT views, Facebook and Google impressions, and e-mail impressions and affiliate impressions. At the time of model building, we check if these correlations induce the multicollinearity phenomenon in the model. It is important to note that we have high correlations between paid/organic views and overall views. We expect it because the latter almost sums the former two. Here, overall views, instead of being the exact sum of paid and organic views, represent the sum with some error. In this analysis, we will take this as an opportunity to use overall views, paid views and organic views as three highly correlated marketing channels to help the reader walk through the multicollinearity problem with an example in hand. Hence, instead of treating overall views as the sum of paid and organic views and hence a redundant variable, we will use it as a different marketing channel with a high correlation with paid/organic views.

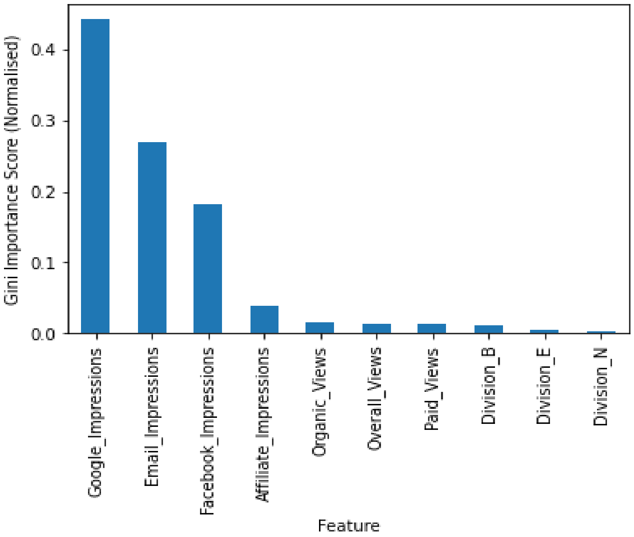

Figure 3. Top 10 features by Gini importance scores calculated by the random forest regressor.

Preliminary analysis shows that Google search, e-mail and Facebook impressions are key media channels contributing to sales. However, YT views and affiliate impressions do not correlate a lot to sales.

D. Modelling

At this point, we have 34 predictors (7 media channel spends + 27 dummy coded features denoting division involvement in an observation) for our response – sales volume; all features standardised. In this subsection, we fit a linear model and generate the coefficients for predictors.

1) Feature Importance and Feature Selection: We use the random forest regression to extract essential features. Also known as Gini importance, the importance score for a feature is the total reduction in the criterion (here, Gini score) brought by the feature. The more the score, the more the importance. Our analysis picks up the top 10 features with the highest importance shown in Fig. 3. Notice the normalised importance scores produced by the random forest regressor, which means scores for all 34 features sum up to 1.

Fig. 3 shows that only features denoting division B, E and N have high enough importance to be considered significant. It means that whether any spending at divisions other than B, E and N does not matter much in predicting sales volume. In addition, as expected, Google, e-mail and Facebook impressions have high importance and contribute to ~ 89.4% of the total importance scores. On the other hand, YT views and affiliate impressions contribute to 2.8% and 3.9% of the total score, respectively, making them insignificant compared to the previously mentioned channels.

2) Multicollinearity: Notice that many features suggested by random forest regression have high correlations between them. We check if these variables induce the multicollinearity phenomenon by calculating the variance inflation factor (VIF), which measures the increase in the variance of parameter estimates for each variable in consideration. VIF value greater than 5 means the collinearity is high enough to lead to parameter estimates with significant standard errors. As expected, we find very high VIF scores (> 2000) for paid, organic, and overall views (all belonging to YT). It means that we must choose one feature from the three. We move forward by choosing organic views given its highest Gini importance of the three.

3) Train-Test Split and Cross-Validation: Intuitively, models should not be trained on future data and be used to predict entities of the past. Hence, with the remaining essential features, we perform a time-based split for training and further testing. Studies usually use elbow checks and other similar methods to investigate the optimal number of training points. This analysis does not perform any such checks considering the main objective at hand. Therefore, we choose 90% of the initial dates for training and the remaining 10 per cent for testing to give sufficient data for training.

We perform cross-validation within this train set to check how well the linear model with the selected feature set gets trained by some data and then predicts data it has not seen. It is crucial since we can now get the best estimates possible of the model's ability to learn and predict. To perform validation, we use the time-based CV solution produced for public use by O. Herman-Saffer in [6]. It offers extensive customisations for train-validation windows, split date, and train and test period, keeping in mind the time-based train-test splitting approach where each split should have higher test indices than before. Results show that the model performs well. The average mean squared error on test predictions is 0.21, and that on train predictions is 0.11. The R-squared score is 0.75 on test predictions, where on train predictions, it is 0.87. It is correct since we expect the training score to be a little better than test scores. The linear model with the selected feature set performs well; hence, we now train the model on all training data. We use scikit-learn for the same.

4) Negative Parameter Estimates: Fitting the optimal least-squares (OLS) model on the 90% training split results in a good fit but poses specific problems with interpretability. The linear regression model produces negative intercept and negative coefficients for organic views and affiliate impressions. Modellers often visualise the intercept as the base sales, which is sales without media. Negative intercepts say that without media, the sales volume is negative, which is absurd. Similarly, negative parameter estimates for YT organic views show that the increase in spending for organic views leads to a decrease in sales volume, which is again illogical on uncomplicated grounds.

As far as interpreting intercept is concerned, Jim in [4] explains that the y-intercept has no real meaning, and one should not try to attribute one to it. Hence, we move forward with the calculated negative y-intercept. We remove the predictors with negative coefficients and retain only those with positive parameter estimates in the model to handle the negative parameters. Nevertheless, this problem needs further investigation for the best solution. For example, can we transform the predictors, such as with a log transformation, to eliminate the negative signs, or could there be some correlations that we missed and is impacting the results, probably the relationship between the channels and the divisions? However, finding the best solution is beyond the scope of this analysis.

5) Final Model Results: Table I shows the predictor coefficients estimated by the final model with significance values through t-test. We use the statsmodel [9] to get the results to look like model summaries presented in the R language. It shows the VIF scores for estimates, showing no multicollinearity. Coefficients for all the predictors are positive. A p-value of 0 (~ 0) tells us that the predictor holds a significant relationship with the response. Minor standard errors tell us that all inferences made on these estimates are reliable.

TABLE I. MULTIPLE LINEAR REGRESSION THROUGH OPTIMAL LEAST-SQUARES - MODEL SUMMARY

	coef	***std err***	***P > |t|***	***VIF***
Intercept	-0.0629	0.009	0.000	1.2
Google_Impressions	0.3756	0.014	0.000	2.8
Email_Impressions	0.3359	0.015	0.000	2.6
Facebook_Impressions	0.2112	0.013	0.000	2.2
Division_B	0.4819	0.064	0.000	1.9
Division_E	0.3040	0.048	0.000	1.1
Division_N	0.4669	0.047	0.000	1.1

6) Contribution and Return on Investment (ROI): To acquire inferences from the final model, we need to calculate media channel contributions. Equation (1) shows that contribution/impact from a media channel M_i is the multiplicative product of the media channel spend vector S_i and the corresponding predictor coefficient estimated by the model.

$$C_i = \beta_\iota \times S_\iota \qquad (1)$$

We calculate contributions to facilitate understanding which campaigns or media channels work better than the other. In other words, what are the media channels that contribute more to sales (revenue or volume based on the response) and, in a way, have more impact in general?

ROIs for predictors are the same as the beta estimates for predictors given by the model. The study in [7] shows that a constant ROI implies a linear model since increasing the spending by x in one channel would increase the sales by the ROI times that amount. For division variables, spending on a channel in that division would increase the sales by its ROI amount.

IV. RESULTS

This section explains the model performance results and opportunities in MMMs in the form of the various inferences that will answer the key questions we intended to find the solutions to in the first place.

We use the final model to perform prediction on the 10 % test data. Model achieves a mean absolute error (MAE) of 0.36 (0.22 for train), a mean squared error (MSE) of 0.44 (0.19 for train) and r-squared of 0.6 (0.80 for train). The model performs well compared to the test MSE of 1.11 of the simplistic model: prediction by averaging. Prediction by averaging is when the prediction outcome for every observation is simply the average value of the observations. The final model has a 155% better MSE score as compared to the simple average model. Mean absolute percentage error (MAPE) is another error measurement statistic. Given the negative comments in the various literature about it being a poor metric, we do not use it. Check [1] for one such example.

We also check the model performance glancing at the sales predictions for different division and check for any

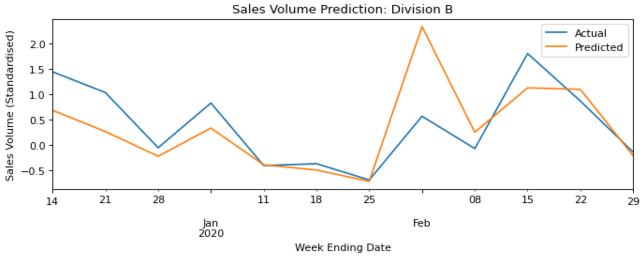

Figure 4. Actual versus predicted sales volume for spends in division B test split, estimated by the optimal least-squares linear regression model. Note that the y-axis has standardised sales volume instead of the actual scale values for simplicity.

overfitting signs. For example, Fig. 4 shows the actual versus predicted sales volume for division B, the division with the highest total sales volume.

After the model performance, Fig. 5 shows the quarter-on-quarter media impact on sales. In other words, the percentage of total contribution to sales. While contributions from Facebook remain low throughout the two years' time, Google contributions kick off in the first quarter and remain in the 40 – 70 % range throughout. Initially, high contributions from e-mails also reduce and remain in the 30 – 50 % range. In short, it shows the vast contributions to sales from Google search impressions spending.

Fig. 6 shows the ROIs of media channels along with the actual total spends on the media channel. Investments on Google search impressions have the highest returns, and that on Facebook has the lowest. We know that the firm/company already spends according to the ROIs evaluated by the model from the actual spending. If they were spending on a lesser ROI media channel, the analysis would guide them to change the maximum budget allocation to high ROI channels.

Moreover, we calculate the ROIs for the division variables in the model. Division B has the highest ROI. Often in media mix modelling, it is beneficial to understand which divisions generate better sales. Hence, these conclusions are very well a part of media mix modelling results.

Once businesses get hold of the beta coefficients for their media channels, they can answer many such questions, for example, optimising spends to maximise ROI and revenue maintaining a constant total budget, which is the next step and out-of-scope for this analysis.

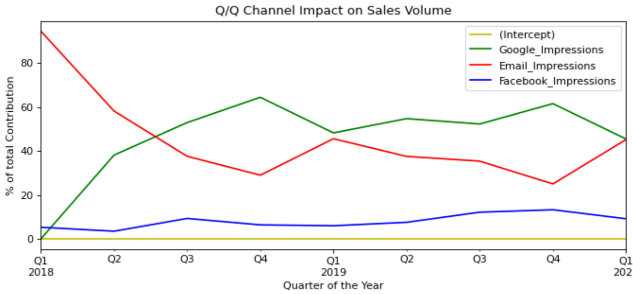

Figure 5. Quarter-on-quarter contribution impact on sales volume from Google search, e-mail, and Facebook impressions. Intercept here has no associated real meaning.

V. Challenges

This section discusses the several issues encountered during the experiment discussed in the previous two sections.

A. *Adequately Correlated Predictors*

Our experiment encountered highly correlated predictors - paid and organic YT views ($\rho = 0.8$). Often marketing decisions involve intentional correlated budget allocations across various media channels. Eventually, this spending could correlate with other marketing variables too. Hence, media channels are bound to have a sufficient correlation between them.

Using these highly correlated variables together in a multivariate regression analysis could lead to high variance in the estimated parameters and phenomenon where other predictors can linearly predict one predictor with considerable accuracy (multicollinearity). Thus, such models steer wrong budget allocations leading to low profits or even customer losses, and would not predict well for the spending patterns that deviate from the learned relationships. We dealt with this problem by removing only those correlated variables with high VIF scores, leading us to lose YT paid views spend information. The challenge arises when the modeller must use the ignored channels in the analysis, as per the requirements.

B. *Negative Parameter Estimates*

Our experiment removed the predictors with negative coefficient estimates – YT organic views and affiliate impressions. It is an odd observation for media channels where their actual correlation with sales is positive. In our experiment, the ignored variables showed positive coefficients in the univariate models.

There is no hindrance if the problem does not demand positive coefficients, but it does stand as a significant challenge for modellers when they only want positive coefficients, not non-negative coefficients. There are mechanisms to force non-negative coefficients, mostly yielding zero values, which is unwanted.

C. *Model Uncertainty*

It is important to note that, for this analysis, all predictors have linear relationships with the response. The data used for the analysis does not contain predictors with non-linear impacts – such as TV GRPs, for which analyses usually calculates ad-stock transformations with diminishing returns and carry-over effects. Also, the data here consists only of incremental sales and not base sales. Our experiment gives a model with good CV results. Nevertheless, whether we can rely on the model estimates is the question. The study in [2] uses a typical real-world weekly MMM data set for the US and creates five acceptable models from different discussions. They found that all models fit the data well but lead to different conclusions about budget allocations. The models differ in the variables used for predictors along with

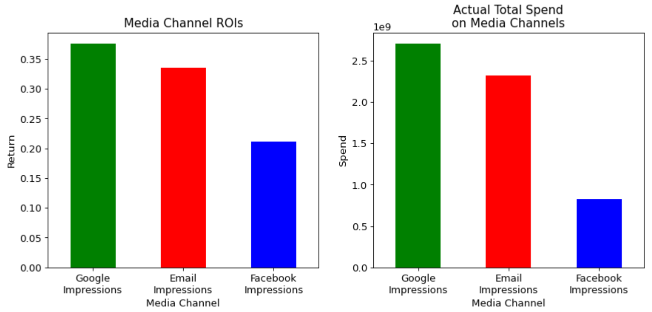

Figure 6. Media channel ROIs calculated using the OLS linear regression model (left); total actual spending for the media channels (right).

the time subset of data used. It shows that different hypothesis could lead to different conclusions, and it is essential to investigate every aspect of it.

In this experiment, we look at linear regression CV results to validate our model. However, intermediate steps could be performed differently based on the hypotheses specific to organisations to check how the model concludes – such as observing different train test splits, using different time-based subsets for the complete analysis, and using log-based transformations for predictors or interactions between them.

VI. Conclusion

Before proposing a media mix modelling solution for finding the impact and effectiveness of different marketing inputs or campaigns, modellers ought to clearly understand the underlying ideas and methodologies behind the modelling, what they want to achieve, and the explanations an MMM model can provide. Besides, the modeller must also know the challenges and opportunities involved. In our analysis, using time series data on different online marketing channels spending and sales volume, we explain how multivariate regression analysis on panel data can help get answers to questions related to impacts and ROIs. During the process, we show the challenges that the modeller encounters. While we can handle some of these, as done in our analysis, we can regard others as pitfalls to the approach. Said that there are many opportunities too in the form of insights offered by the model's results. While choosing amongst all MMMs, organisations should keep in mind the proportion of the model results that are data-driven to the model results driven by modellers' decision to handle the challenge.

An extension to this analysis can be possible using an extensive dataset with both incremental and sales component of data with varieties in predictors (online and offline media spends/activity and external factors/control variables). Future works could also include a comparative study of the solutions already present in other literature for the challenges to MMMs mentioned in this report.

References

[1] C. Tofallis, "A better measure of relative prediction accuracy for model selection and model estimation," *Journal of the Operational Research Society*, vol. 66, no. 8, pp. 1352-1362, 2015.

[2] D. Chan and M. Perry, "Challenges and Opportunities in Media Mix Modeling," Google, New York, 2017.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss and V. Dubourg, "Scikit-learn: Machine learning in Python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825-2830, 2011.

[4] J. Frost, "How to Interpret the Constant (Y Intercept) in Regression Analysis," Statistics By Jim, 1 May 2018. [Online]. Available: https://statisticsbyjim.com/regression/interpret-constant-y-intercept-regression/. [Accessed 21 May 2021].

[5] N. H. Borden, "The Concept of the Marketing Mix," *Journal of Advertising Research*, pp. 2-7, June 1964.

[6] O. Herman-Saffar, "Time Based Cross Validation," Towards Data Science, 20 January 2020. [Online]. Available: https://towardsdatascience.com/time-based-cross-validation-d259b13d42b8. [Accessed 21 May 2021].

[7] R. Wigren and F. Cornell, "Marketing Mix Modelling: A comparative study of statistical models," 2019.

[8] S. K. Mishra, "Shapley value regression and the resolution of multicollinearity," Munich Personal RePEc Archive, 2016.

[9] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.