Using Machine Learning in Media Mix Modelling

Accuracy and usability

By Alexander Oude Elferink and Sean de Hoon

This research was funded by Facebook

Introduction

Finding the optimal media mix is the holy grail for advertisers. In order to determine this optimal mix, advertisers make use of different methods. Digital attribution analyses for example provide insights into the contributions of different digital advertising investments to online conversions and investments can be optimized. Attribution models range from last click, arguably a flawed model, to algorithmic, at the very least less flawed. Across digital and non-digital media, various types of experiments can be used to determine effectiveness for tactics, campaigns or entire media channels. Such experiments include brand lift studies with a control and experimental group or geo-tests where regions are subjected to different types or levels of advertising. Although experiments represent one of the purest ways to measure advertising effectiveness, they are not always a feasible solution. A method which provides insights into both digital and non-digital media effectiveness, without relying on experimentation, is media mix modelling.

Media mix modelling (MMM) has been around for decades under a number of different monikers. Sales data, or some other KPI, are modelled against media and marketing data, as well as other factors such as the weather or economic trends, at a daily, weekly or monthly granularity. The analyses mostly employ some sort of regression technique, ranging from simple linear regression to more advanced time series analyses. Analysts or scientists, whether at universities, working in-house advertiser side or working for consulting firms or media agencies, have used either packaged modelling tools such as EViews or IBM's SPSS or (packages in) open source programming languages like R and Python. Although there is undoubtedly a huge variation in how well these models have been specified and interpreted, the basic methodology they all use is the same. The ultimate goal is often also the same: to explain how media and external factors impact sales in the past and to help optimize media investments in the future through scenario planning.

Over the last decade, use of machine learning models in academia and industry has skyrocketed. Machine learning is a field of data science where algorithms automatically learn without explicit programming or human intervention. Regression models are a part of this field, however the term tends to be reserved to refer to more advanced techniques and we will use the term machine learning in this paper as a reference to techniques that combine different models or iterative processes. One reason why the use of machine learning is skyrocketed is

the increased availability of relatively cheap processing power. Another reason why the use of machine learning models has skyrocketed, is that predictions from these models routinely outperform more traditional methods (such as regression). In Kaggle competitions, where teams compete to produce the best prediction engine for a specific problem, almost all winning teams over recent years are using methods such as gradient boosting.

It is surprising then that machine learning models are not widely used in media effectiveness research. However, it is reassuring to see some advertisers and agencies in the industry picking up on the latest technology and machine learning is widely applied in other areas in marketing such as building look-a-like audiences. In this study we examine how machine learning models may be employed for MMM. We compare machine learning models to more traditional regression models for MMM from two perspectives. First, we look at the accuracy of the models and their ability to explain the impact of past media investments and to predict the impact of future investments. Second, we look at the usability of machine learning models. Since not all marketing science teams will consist of a group of people with a PhD in mathematics, usability and the extent to which models are understandable and actionable for clients are big factors in the potential use the methodology has in the media industry.

Background

The exact reason why marketing science teams are not using machine learning is not clear and probably differ from team to team. We see several possible driving factors.

First, teams may lack expertise to use machine learning models and feel that moving away from more traditional models requires too big of an investment in resources. These teams may be using methods that have proven themselves over the years and may also be sceptical about ‘changing a winning team’.

Second, applying a new methodology may also be problematic because it changes the status quo with long lasting client relationships. With a new methodology, the contributions for different media channels may change and even a small change may cause clients to ask questions about the extent to which they can trust the findings or the extent to which they should have trusted the findings of the earlier methodology. Complicated discussions between clients and MMM providers may arise when different results are found using different methodologies.

Third, the substantive differences in the standard output of machine learning models versus traditional methods may play a role. ML approaches do not provide straightforward performance measures and increase complexity in decision making and scenario planning for marketers. Traditional methods focus on uncovering the impact of different variables (including media investments and other external factors) on a KPI such as sales. Model fit statistics like the coefficient of determination (R-squared) are used to indicate how well the resulting model explains historical patterns in the data. The impact of different variables is often given in a format like: ‘an increase of 100 TV GRPs is related to 300 (incremental) units sold’. Machine learning models on the other hand, are generally developed to make predictions about the future, based on a collection of signals or features (explanatory variables). The performance of these models is usually tested by looking at their predictive power on a new dataset (the ‘test’ set). The individual contributions of features do not tend to be a main concern, but most models do provide the opportunity to look at feature importance. Feature importance, although indicative of whether features have influence on the KPI being modelled, does not provide

insight into the precise impact of increasing for example TV investment with 100 gross rating points. This makes optimizing media investment in the future more difficult, as the output from machine learning models has to be reworked somehow to come up with the same kind of outputs.

Although all of these reasons may contribute to the lack of machine learning models being used in MMM work, the former may provide more accurate predictions than the latter. In order to test this hypothesis, we will use machine learning models in an MMM study for PepsiCo, The international Food & Beverage company. As mentioned above, we will not only focus on accuracy of the machine learning models, but also on the usability. We will compare three (sets) of approaches. Our baseline is a regression method, which uses the Prophet package in Python[1]. We compare our baseline to two (sets of) machine learning approaches. The first ML approach is one we will refer to as *full ML*. The full ML approach will utilize two machine learning models and rework the standard output to produce the necessary insights for future media planning. The second ML approach is one where we use a Genetic Algorithm.

The second machine learning approach is one in which we optimize variable and transformation selection. Part of traditional MMM work is to determine what kind of adstock[2] and diminishing returns[3] apply to a given media type. The selection of the best fitting transformation is generally done through simple linear regression or by sequentially adding or removing independent variables, often referred to as stepwise regression[4]. This means that for media type 1 (e.g. TV) the best transformations are selected based on a model in which they are used to predict sales (or some other KPI). Once the best transformation is selected, media type 2 (e.g. display advertising) is introduced or tested separately and the best transformation for that one is selected. This process is repeated until each of the different media types is included in the model. The inherent problem with this approach is that both simple linear regression and sequential selection may not lead to the best overall model, instead all of the different transformations should be optimized simultaneously. Unfortunately, even with a very limited number of transformations (e.g. 10) and variables (e.g. 6), the total number of different combinations is still 466 billion (60^6). At a speed of 100 models per second, running all of the different combinations would take roughly 14 years. This could be achieved faster with parallelisation and cloud computing resources, but things would still take a lot of resources (i.e. time and money). Therefore, we will use a machine learning model to optimize variable selection, specifically a Genetic Algorithm, and examine whether this produces models with better accuracy than a model with sequential transformation selection.

Methodology

Data collection and processing

For this study we used sales data for two brands from the multinational PepsiCo: Lay's®, a snack brand and Quaker®, a cereal brand. For both brands we received sales data from two major supermarkets in The Netherlands: supermarket chain A and supermarket chain B for a period of two years. All sales data were delivered at a weekly level by stock keeping unit

[1] "Prophet - Forecasting at scale" https://facebook.github.io/prophet/ Accessed 25 March 2020.
[2] "Advertising adstock" https://en.wikipedia.org/wiki/Advertising_adstock Accessed 25 March 2020.
[3] "Diminishing returns" https://en.wikipedia.org/wiki/Diminishing_returns Accessed 25 March 2020.
[4] "Stepwise regression" https://en.wikipedia.org/wiki/Stepwise_regression Accessed 25 March 2020.

(SKU). For Lay's® specifically, we considered the following SKU's: Oven Baked (excluding Oven Baked Crispy Thins, Crunchy Biscuits and Veggie), snacks (Hamka's, Grills, Mamamia's, Mixes, Wokkels, Poppables, Sunbites, Pomtips, Sticks and Stax), Sensations (excluding Streetmix and Coated Peanuts), Core (Lay's Chips, Superchips, Deep Ridged and Light Chips) and Bugles. Quaker® was split on SKU level into Cruesli and Quaker. With two brands per supermarket we have a total of four different dependent variables. We decided to model the two supermarkets separately, in order to gain insights into how the sales for each are impacted differently by media investments. Sales data only concerned sales made in physical supermarkets and not online.

Four categories of data are used in the models to explain sales patterns over time, namely media, weather, competitor and promotion variables. Media data that was included relates to Out Of Home (OOH) advertising spend, Google search and Google display impressions, as well as clicks and spend. Gross Rating Points (GRPs) and spend data for TV and Radio are used. Facebook advertising impressions and spend (Facebook + Instagram) were collected by means of the Facebook MMM feed (Annalect is a global Facebook MMM partner) and Snapchat impressions and spends were collected through Snapchat Business Manager.

To get local weather data we used the 'knmy' Python package, that utilises the Royal Dutch Meteorological Institute (KNMI) API for fetching and parsing weather data observations from KNMI's 48 automated weather stations. Gathered variables included temperature in Celsius, sun and rain measured in hours and also rain measured in millimetres on a daily level. Competitor data was included as total gross spend for Lay's® and Quaker® separately and collected through Nielsen.

Lastly promotion data was collected through a custom-made web scraper using Python packages 'selenium' and 'BeautifulSoup'. For each week, we fetched: product names, product prices before promotion and product prices during promotion from online supermarket folder data for supermarket chain A and supermarket chain B. After scraping, additional feature engineering was done to also account for promotions among competitors of Lay's® and Quaker®. For Lay's®, these were differentiated between 'internal' and 'external' competitor promotions, to separate a possible cannibalisation effect by other PepsiCo brands from competitor promotion effects.

Feature engineering adstock

After data collection, we split the data by brand and supermarket, leading to four data sets. After this first split, all four sets were split further into a 70% train versus a 30% test data set. As the last step, before starting adstock feature engineering, we filtered promotion and weather features. This left us with media and competitor data that could undergo adstock transformations, which refer to the delayed - additional - impact of advertising. Following an earlier study[5] Weibull transformations were used, since they improved MMM performance in terms of R-squared and Mean Absolute Percentage Error (MAPE).

Properties that were used to create all the different transformations include:

- Decay type and strength: Weibull or traditional exponential decay

[5] "Modelling adstock using Weibull transformations" https://github.com/annalectnl/weibull-adstock/blob/master/adstock_weibull_annalect.pdf . Accessed 17 Jan. 2020.

- Window: length of the adstock effect in weeks
- First week correction: decreased ad effect of the first week or not

By utilising the Weibull cumulative distribution function[6], adstock transformations were calculated with the two following equations:

$$\lambda = \frac{window}{(-ln(0.001))^{1/k}}$$

$$adstock = e^{-(lag/\lambda)^k}$$

Here *k and window* parameters refer to resulting adstocks, whereas *lag* represents the number of weeks after the ad was published. Examples of transformations can be seen below.

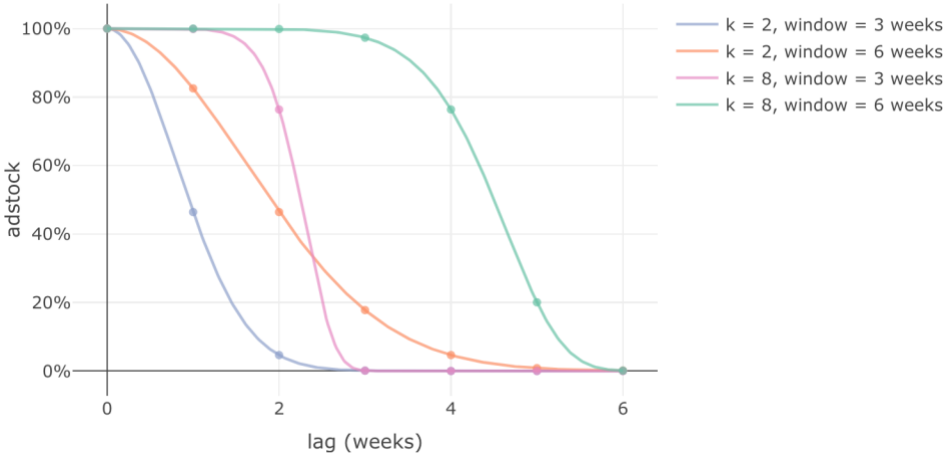

Figure 1. Weibull adstock transformation examples. Parameter 'k' defines the curve shape and parameter 'window' defines the week on which the adstock will wear out to nearly zero.

Also, when applying Weibull adstock transformations, we accounted for non-linear advertising effects on sales; when making initial spends on media, advertising effects might be limited until a certain threshold. We also accounted for the tendency that advertising effects diminish, as ad spends will reach a point of saturation. To do so and get the desired s-curve shape, we incorporated a Sigmoid function[7], as well as normalisation, when making the Weibull adstock transformations. The corresponding equations and examples of s-curves (Figure 2) can be seen below.

$$scale = \underline{x} = \frac{\Sigma\ adstock\ vector}{n\ adstock\ vector\ elements}$$

$$s - curve\ adstock = \frac{1}{1 + b * (\frac{adstock}{scale})^c}$$

[6] "Weibull distribution - Wikipedia" https://en.wikipedia.org/wiki/Weibull_distribution . Accessed 17 Jan. 2020.
[7] "Sigmoid function" https://en.wikipedia.org/wiki/Sigmoid_function. Accessed 23 Jan. 2020.

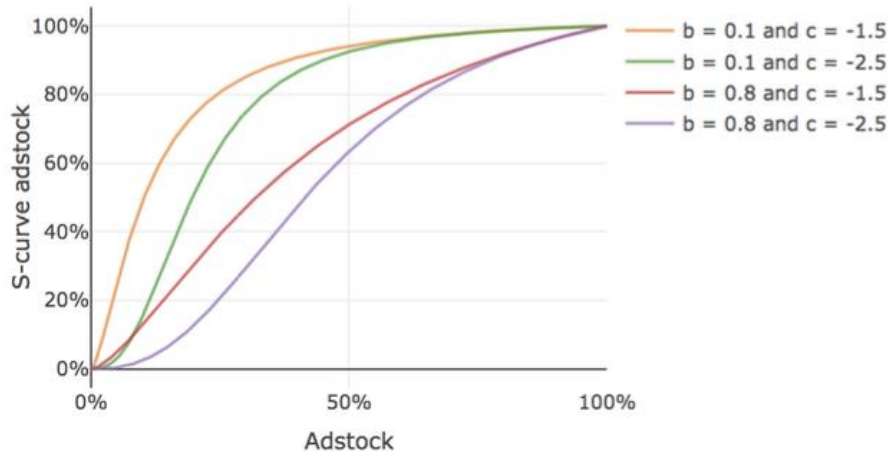

Figure 2. S-curve shape examples. Parameter 'b' determines the saturation effect and parameter 'c', declares the diminishing adstock effects.

To generate different Weibull transformations, we created a grid with all unique combinations of Weibull transformations. Parameter values that we considered are: Weibull k parameter: 2, 3, 5, 8; Window: 2, 4, 6, 8; first week correction: 50%, 100%; s-curve b parameter: 0.1, 0.8; s-curve c parameter: -1.5. How the best Weibull transformation was found per feature, will be extricated in the next paragraph.

Transformation selection

In order to determine the best Weibull adstock transformation for each feature, univariate statistical testing was done. Since this study examines sales, a continuous dependent variable, univariate linear regression tests were done. For each original feature, all associated Weibull transformations were assembled in batches, after which individual feature effects would be determined. These effects were quantified in two steps. First, correlations were calculated between each feature and sales. Second, these were converted into F-values and then to p-values. Based on these values, the best performing transformation was selected. This means that after this step we have one transformed feature instead of several different transformations. For paid social for example, we have one transformed spend variable and one transformed impression variable. In the feature selection step described below, we determine whether to include the spend or the impression variable in the final model. It is worth mentioning here that the methodology described here was used for the baseline Prophet model and the *full ML* models, but not for the GA model. The GA method selects the best transformation in the modelling procedure.

Feature selection

Feature selection was done in four steps. We first calculated the absolute Pearson correlations per feature. Second, the same univariate statistical test method as described above was used.

Third, recursive feature elimination[8] was used. In fact, it builds a model on the entire set of features and then computes the importance per feature. It removes the least important feature, re-builds the model, computes feature importances and so on. Fourth, feature importance was approximated with bagged decision trees[9]. These trees created subsets of the training data randomly with replacement. After sub setting, all decision trees were trained and in the end its feature estimations were averaged, so this ensemble of decision trees would be more robust than just a sole decision tree.

After applying all methods, each method generated a series of feature rankings (scores). These were condensed in one final ranking by summarising all scores per feature. The lower the ranking (score) was, the more likely it was the feature would be important for the modelling phase. To get an idea of how the feature rank reduction matrix with scores looks like, an example is illustrated in Table 1.

Feature names	Correlation rank	Univariate rank	Recursive rank	Importance rank	General rank
display impressions cruesli	62	55	29	54.5	200.5
display impressions quaker	33	22	67	28	150
display impressions total quaker	36	38	15	11	100
display spend cruesli	59	43	11	17	130
display spend quaker	28	20	47	54.5	149.5
display spend total quaker	27	30	17	24	98

Table 1. Feature reduction matrix example (supermarket chain A - Quaker Model).

Table 1 shows that for display variables the total spend variant has a lower general ranking (98) than the total impression ranking (100); the lower the ranking the better. Hence for these display variables, display total spend was selected for modelling. After the best variant per feature was selected, Variance Inflation Factor (VIF) checks were done for the final feature combinations; a threshold of five was taken. For all data sets the feature category 'search' was left out of scope since it caused multicollinearity (VIF > 5). Furthermore, Facebook and Instagram features were combined to a 'paid social' feature.

Modelling

To model both brands (Lay's® and Quaker®) among two supermarkets (supermarket chain A and supermarket chain B) we used the eXtreme Gradient Boosting (XGB)[10] and Random

[8] "Feature Selection in Python - Recursive Feature Elimination" https://towardsdatascience.com/feature-selection-in-python-recursive-feature-elimination-19f1c39b8d15 . Accessed 27 Jan. 2020.
[9] "Decision Tree Ensembles - Bagging and Boosting" https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9 . Accessed 27 Jan. 2020.
[10] "XGBoost Documentation" https://xgboost.readthedocs.io/en/latest/index.html Accessed 20 March 2020.

Forest (RF) methodology from Python's scikit-learn package[11] and a Genetic Algorithm that was built from scratch. For our baseline model we used Prophet[12].

eXtreme Gradient Boosting (XGB) and Random Forest (RF) both belong to ensemble learning methods and can perform regression or classifications tasks, by combining outputs from individual decision trees. This is actually the goal of models that belong to the ensemble family: combining predictions from multiple estimators in one learning algorithm, to improve generalisability over one model. The difference between XGB and RF is that XGB builds decision trees one by one, where each following tree helps to correct errors that were made by the previously made tree. RF on the contrary, trains each tree independently, using random samples of the data. The advantage though, is that this randomness can make the RF method more robust to more 'noisy' data and thus is less likely to overfit.

In addition to these two methods, a Genetic Algorithm (GA) was developed. GA's are part of the larger class of Evolutionary Algorithms and are inspired by the process of natural selection[13]. Just like many other GA's, our algorithm consists of six steps. The algorithm starts off by initialising a population of 'solutions'. This population is a collection of different sets of features. Since in MMM we want to explain the impact of each media channel on sales, our GA selects one feature for each feature category, e.g. one in OOH, one in TV, one in Facebook etc. The goal is to evolve this population to a population of better solutions. In order to do so the GA evaluates fitness of each solution within the population. Depending on what kind of metric the GA optimizes for, fitness should be maximised or minimised. The latter is our case, as Mean Absolute Percentage Error (MAPE) was our key metric. To retrieve this metric per solution the GA utilises Ordinary Least Squares (OLS) to fit training data and predict on test data. The best solution (set of features) is chosen to represent the so-called 'parents'. Then the crossover step: the process where the characteristics or features for the 'offspring' are determined. As our GA selects only the best solution, all characteristics from parents will pass over to the offspring. A subsequent step is mutation, where some characteristics or features are randomly changed to another feature, i.e. another transformation or base variable. The final evolution step in the GA will create the new population that consists of 50% parents and 50% mutated offspring. All these steps are referred to as a generation. Conditional on the number of generations the GA will iterate over these steps.

Shapley Values

To interpret the impact of features in the XGB and RF methodologies we looked at Shapley Values in the package SHAP. These provide two major advantages. First, global interpretability - Shapley values indicate how much a feature contributes to the dependent variable i.e. sales. Therefore, it provides something more than feature importance, as it also provides effect directions. Second, Shapley values also support local interpretability - this refers to each observation having its own Shapley values. In the context of this study it implies that for example media and promotional impact can be calculated for each week.

[11] "1.11. Ensemble methods" https://scikit-learn.org/stable/modules/ensemble.html Accessed 20 March 2020.
[12] "Quick Start | Prophet" https://facebook.github.io/prophet/docs/quick_start.html Accessed 23 March 2020.
[13] "Genetic Algorithm" https://en.wikipedia.org/wiki/Genetic_algorithm Accessed 20 March 2020.

SHAP is built on game theory that has an additive attribution property. From a game theory perspective[14] it is about fairly distributing gains and costs of several players when working together. So, it is about assigning pay-outs to players depending on their contribution to the total pay-out. Shapley values themselves are calculated by simulating all possible combinations between features for that data point. Important to note is that the Shapley value is *not* the difference between the prediction including and excluding a feature, it concerns the average of the marginal contributions across all permutations. The sum of all feature Shapley values for a data point are equal to the difference between the actual prediction and baseline (i.e. mean) prediction.

We will explain Shapley values with a marketing related example. Consider you have trained a machine learning model to predict sales for a specific brand over time. At a certain point in time weekly sales are predicted to a total of €85.000. To reach that amount they used the respective media channels: TV (€10.000), Facebook (€2.000) and Search (€500). In this particular week, no In-store promotions were used (Figure 4). We know that the baseline prediction for all weekly data points is €95.000. The question is: 'How much did each channel contribute to sales?'

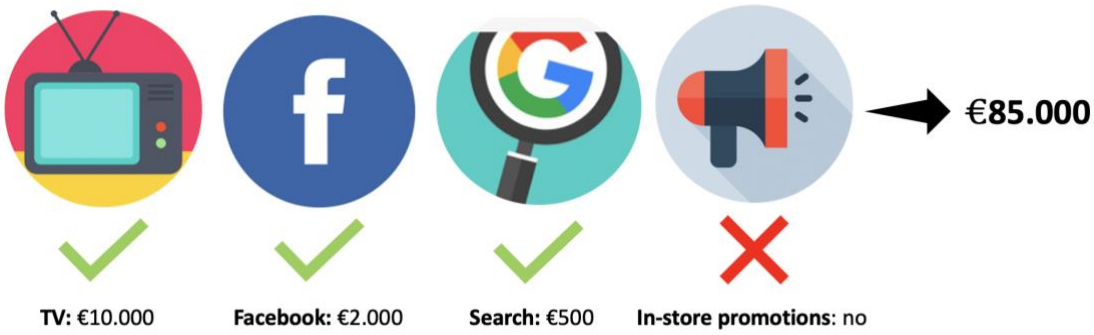

Figure 4. Example of predicted total weekly sales at one point in time, when TV (€10.000), Facebook (€2.000), Search (€500) and no In-store promotions were used to reach a total weekly sale of €85.000.

In our example TV, Facebook, Search and In-store promotions worked as players together to achieve the prediction of €85.000. In order to ascertain what the channel contributions are, the goal is to explain the difference between the actual prediction (€85.000) and the baseline prediction (€95.000), i.e. a difference of -€10.000. An answer could be that: TV contributed €20.000, Facebook contributed €4.000, Search contributed €1.000 and no In-store promotions contributed -€35.000. These contributions add up to -€10.000, which is the actual prediction minus the baseline of predicted weekly sales[15].

[14] "Shapley Value" https://www.investopedia.com/terms/s/shapley-value.asp . Accessed 12 March 2020.
[15] For further explanation of Shapley values, please see the appendix.

Results

To evaluate model accuracy, we used the Mean Absolute Percentage Error (MAPE). Overall it can be deduced from Table 2 that our Genetic Algorithm outperforms all the other methods, with a 20% lower MAPE. With MAPE as the key metric, we wanted to elaborate on model interpretability.

Model (test) data	**Prophet**	**XGB**	**RF**	**Genetic Algorithm**
supermarket chain A - Quaker	*19.5%*	*15.49%*	*15.26%*	*14.42%*
supermarket chain A - Lay's	*13.7%*	*8.81%*	*9.54%*	*9.08%*
supermarket chain B - Quaker	*9.67%*	*10.84%*	*11.45%*	*8.96%*
supermarket chain B - Lay's	*8.64%*	*16.25%*	*12.93%*	*11.14%*
Average	*12.87%*	*12.84%*	*12.29%*	*10.90%*

Table 2. Modelling results with mean absolute percentage error (MAPE) as a key metric.

Model Usability

We will compare two media variables and one promotion variable across Prophet and the machine learning models that are built, i.e. the eXtreme Gradient Boosting (XGB), Random Forest (RF) and Genetic Algorithm (GA). For the sake of brevity, we only discuss the supermarket chain B Quaker model. When considering Prophet outcomes (Figure 7), we see that the relative contributions for these variables are in respective order paid social (5.91%), promotions (2.86%) and TV (2.8%).

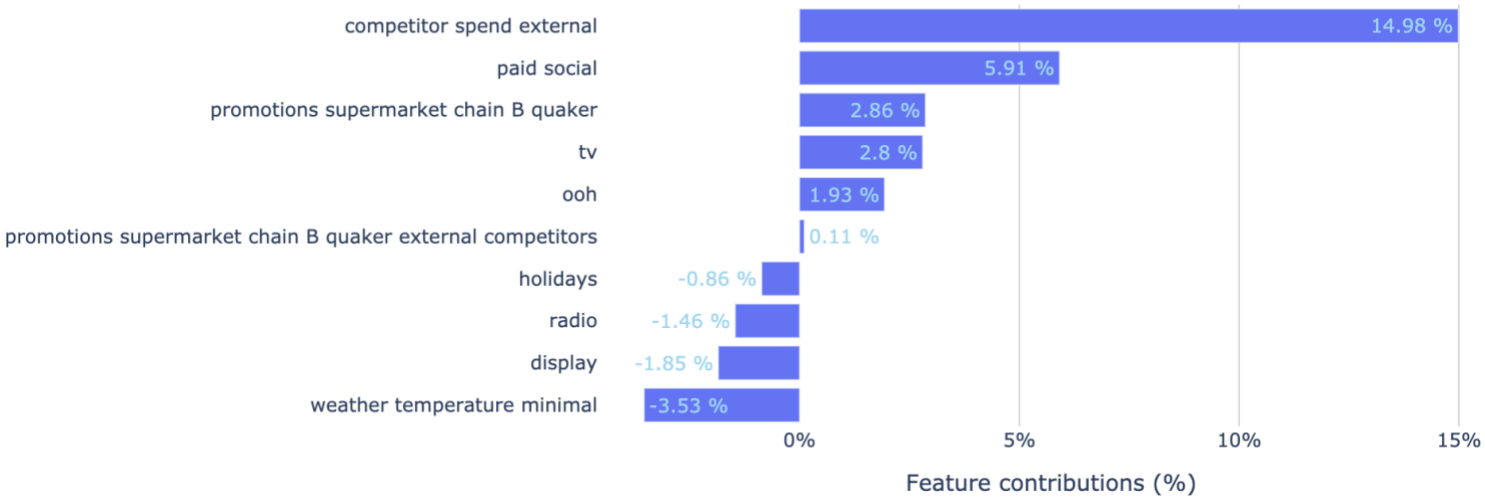

Figure 7. Contribution plot of the Prophet method for the supermarket chain B Quaker model.

Bar plots below list the most important variables in descending order for the XGB and RF methods. The top variables contribute more in the model than the bottom ones. To get these contributions, Shapley values were converted to absolute values, means were taken per feature and their relative contributions were calculated to respective baseline models.

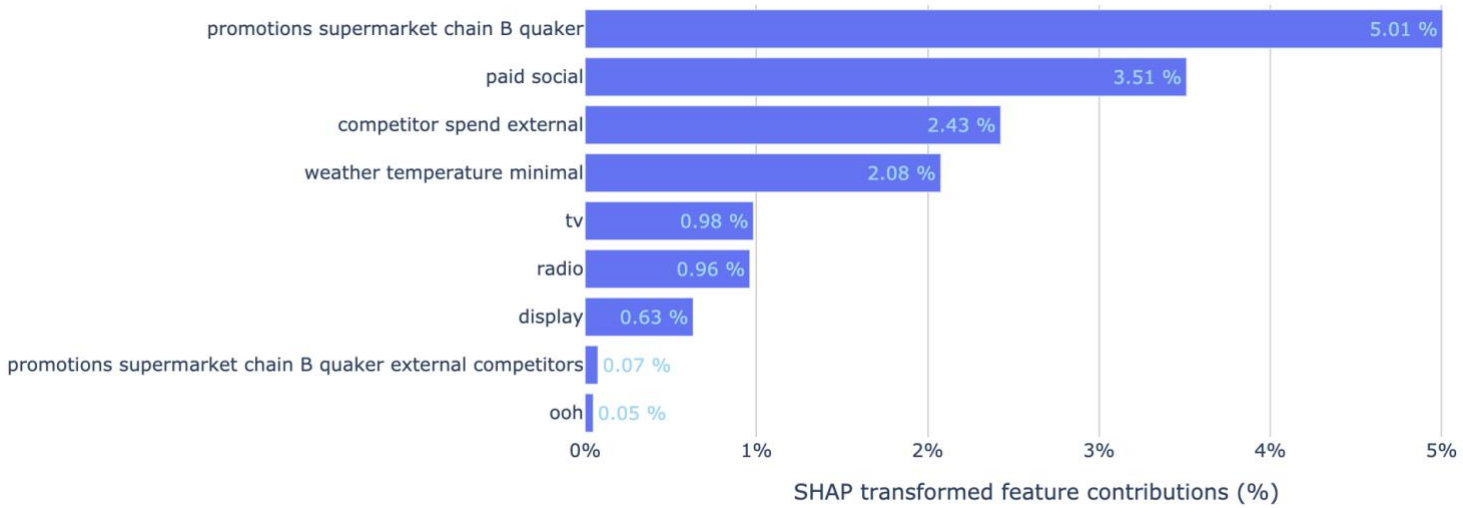

Figure 8. SHAP transformed contribution plot of the XGB method for the supermarket chain B Quaker model.

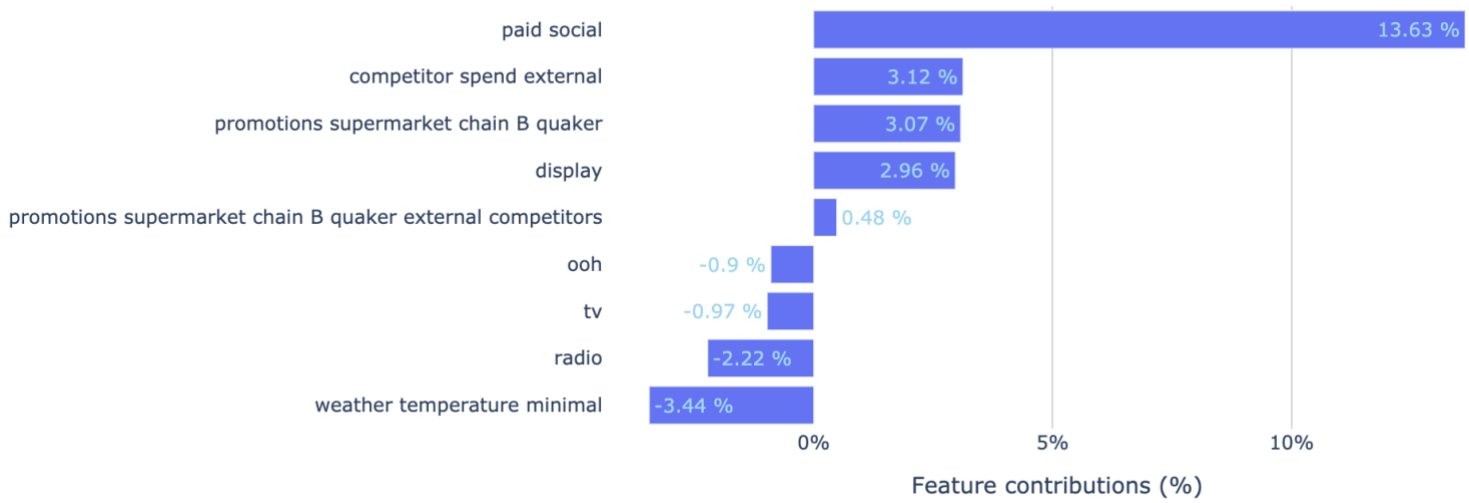

Figure 9. SHAP transformed contribution plot of the RF method for the supermarket chain B Quaker model.

Figures 8 and 9 show that promotions have the highest predictive power in these models, whereas paid social contributed more in the Prophet model. For all models, we do see that paid social seems to contribute more to sales than TV. Contribution sizes differ significantly across the models: promotions contributed 2.86% in the Prophet model and around 5% in the full ML models, while paid social contributes respectively 5.91%, 3.51% and 2.23%.

Considering the GA, paid social contributes most to weekly sales with an estimate that is much higher than all of the other models at 13.63%. The impact of promotions is smaller than the full ML models, but comparable to the Prophet results. The slight negative contribution of TV is likely to have a statistical origin. It is beyond the scope of this research to ascertain why we find this relation, but it can be stated that it is highly unlikely TV advertising has this effect on sales.

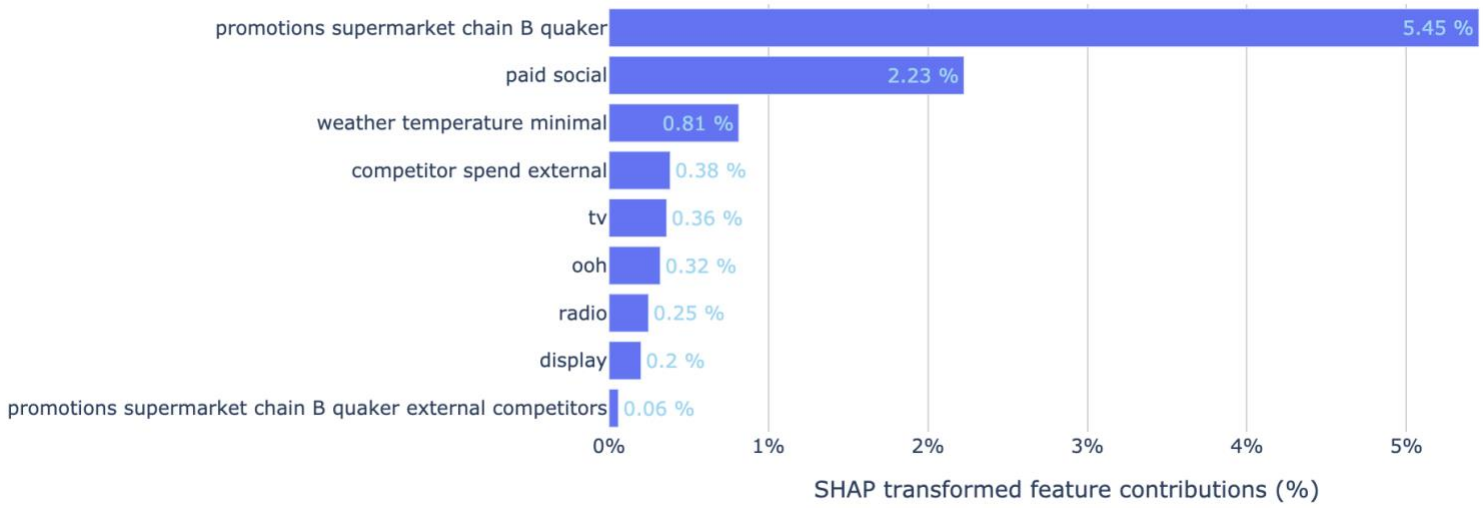

Figure 10. Contribution plot of the Genetic Algorithm method for the supermarket chain B Quaker model.

In terms of the applied transformations, paid social and TV are compared, since the promotion variable did not undergo any transformations. Given that we varied values for the Weibull transformations between the Weibull k parameter, Window, first week correction and the s-curve b parameter, we point out its differences in Table 3.

	Weibull k parameter	*Window (weeks)*	*First week correction*	*S-curve b parameter*
Prophet, XGB and RF				
Paid social (impressions)	8	2	50%	0.8
TV (grps)	5	8	100%	0.8
Genetic Algorithm				
Paid social (impressions)	8	8	100%	0.8
TV (grps)	8	4	50%	0.8

Table 3. Weibull adstock transformations per modelling methodology for supermarket chain B Quaker.

Three main distinct differences can be distinguished when we compare both variables. For paid social we see that adstock decay strength is equally strong (8), but for paid social within Prophet, XGB and RF the adstock effect (window = 2) will wear out after 3 weeks, instead of 9 weeks within the Genetic Algorithm. Secondly, we see that within the Genetic Algorithm TV has a stronger decay effect: 8 compared to 5 for Prophet, XGB and RF. Although the decay effect is less strong for Prophet, XGB and RF, the wear-out effect of TV takes longer (window = 8), whereas this wear out effect is 5 weeks for the Genetic Algorithm. For the Genetic Algorithm, TV does not directly have a decreased ad effect on the first week, whereas this does hold for paid social. The contrary holds for Prophet, XGB and RF when it comes to first week corrections.

Discussion and Conclusions

In this research, we set out to examine how, if at all, machine learning could be utilized in media mix modelling. We consider machine learning models, in this case, to refer to types of models that have gained traction over the past decade, such as random forests and gradient boosting, and not to more traditional approaches like regression. We assess the utility of machine learning models both objectively and subjectively. Objectively, we will compare whether machine learning models may improve the predictive capabilities of the media mix models. In the end, MMM models are focused on predicting the impact that media investments have on sales, or some other KPI, so any method that could improve these predictions may

be interesting to use. However, when it comes to MMM there is also a more subjective question to answer, which is whether machine learning is easily used and will produce the necessary output. It could be that even though machine learning models are more accurate, their usability is so much less, that a more traditional approach is still preferred. We will start off by discussing the objective differences between the approaches.

In the results section, we saw that on average both the *full ML* approach and the *Genetic Algorithm* approach performed better in terms of the mean absolute percentage error (MAPE). That being said, there were quite substantial differences between the two. The Genetic Algorithm approach performed best out of the two methodologies employing machine learning. As noted above, many of the winning entries in machine learning competitions on Kaggle use gradient boosting or random forests, which is why we expected these methods to outperform the Genetic Algorithm approach. When we look at the contributions of the different variables across the methodologies, we see that there are substantial differences in the contributions. The observed differences are inherent to the fact that the approaches differ in both feature selection and modelling. When it comes to feature selection for Prophet, features were selected sequentially. The Genetic Algorithm in contrast, was able to select features in parallel. As a result, selected features underwent different Weibull transformations (Table 3) and have different corresponding effect sizes. Moreover, Prophet accounted for seasonality and a holiday feature including vacations or other special events. The Genetic Algorithm approach did not account for these. In the Prophet model, this means that some of the effect of media variables is probably captured by the seasonality, while in the GA some of the seasonal effect is probably captured by the media variables. This is the consequence of advertisers seasonally planning their media investments and can only be solved by experimenting with different media planning (more about experiments below). As the accuracy of the models is only one part of the question we are trying to answer, we will now turn to the other part of that question, namely the usability of each of the models.

When it comes to the usability of the different models, we need to consider the reason MMM analyses are generally performed, namely scenario testing and planning. Advertisers and media agencies use MMM to establish the impact of media investments and more importantly, develop and compare possible media plans for the future in terms of their impact on sales. This means that the ideal MMM method is well suited for scenario planning and testing. Looking at this prerequisite, the baseline model is probably most well suited. In our case the baseline model was a model estimated in Prophet, which is essentially a time series regression model. Prophet provides effect sizes and contributions for each of the features included in the model, much like a standard linear regression model would also have. Taking into account the outputs provided, the full ML models (XGB and RF) do not provide effect coefficients and contribution to the KPI. Instead, it provides a view on variable importance and direction (positive or negative), not an effect size directly. In order to create the same or at least similar outputs as you would get in a traditional regression model, one method to resort to is Shapley values. However, this restricts the usability of ML models due to the additional work required and the interpretation differs from coefficients from a traditional regression model. With regards to the Genetic Algorithm approach, the output is the same as a regression model so in terms of usability, it is not compromised. This approach will require a different set up compared to the Prophet or full ML models, since it is automating an evolutionary process to improve the models through iterations. Still, we feel that the Genetic Algorithm represents the most promising way forward for MMM. One aspect none of these methodologies can resolve is whether the model makes sense, for instance is it unlikely that TV has a negative

impact. This is when humans have to intervene to make sure the model makes sense which will be used for decision making, to validate such models we look to using external tests outside of modelling such as experiments.

Finally, a word on the use of lift studies or geo-experiments. In this research, we have examined different statistical analyses based on their accuracy and their usability. Ultimately though, it would be good to be able to validate some of the findings against a ground truth. In earlier work for McDonald's a geo test was conducted between different regions testing the impact of Facebook and Instagram campaigns against in store sales, whilst also building an MMM model (using the Prophet methodology) with the relevant variables including Facebook and Instagram[16]. The MMM model showed Facebook and Instagram advertising contributed 2.8% of in-store sales, whereas the geo test showed 3.2% lift of in-store sales was due to Facebook and Instagram. Despite the small difference, we concluded that the MMM model was accurate in establishing the impact of Facebook and Instagram advertising on in-store sales. This is an example of using an experiment (a ground truth) as a form of validation to MMM. In the current study, we do not have an experiment available as a ground truth, but if we did, this would have presented another and perhaps the most effective way to determine which of the methodologies is most accurate.

[16] https://www.facebook.com/business/success/mcdonalds-netherlands Accessed 26 March 2020.

Appendix 1. Shapley explanation continued

So, let's dive further into this example and see what the marginal contribution of no in-store promotions is when it is added to the *coalition* of TV, Facebook and Search. We simulate a situation where only TV (€10.000), Search (€500) and no In-store promotions are in the coalition and that we just randomly draw another spend value for Facebook. In this case the Facebook value (€2.000) was replaced by (€3.000). Subsequently we predict weekly sales for this combination; €90.000. Afterwards, the value 'no' is removed for In-store promotions and is randomly replaced by 'yes'. Now when we predict weekly sales for the coalition of TV (€10.000), Facebook (€3.000), Search (€500) and 'yes' for In-store promotions, a total of €120.000 on weekly sales is predicted. Hence, implied is that the contribution of 'no' In-store promotions is equal to: €90.000 - €120.000 = -€30.000. This computational process is repeated for all possible *coalitions*. After all computations, the Shapley value will be the average of all the marginal contributions to all possible coalitions.

Figure 5. Example sample repetition to estimate the contribution of In-store promotions with value 'yes' to the prediction when added to the coalition of TV (€10.000), Facebook (€2.000), Search (€500).

To illustrate what all possible coalitions would be to determine the Shapley value for 'no' in-store promotions, all coalitions are shown in Figure 6. The first row shows a coalition without any feature values. The second, third and fourth row depict different coalitions that are increasing in size and a separated by a pipe "**|**".

To summarise all possible coalitions that Figure 5 shows:

- No feature values
- TV
- Facebook
- Search
- TV + Facebook
- TV + Search
- Facebook + Search
- TV + Facebook + Search

For each coalition, weekly sales are predicted with and without the feature value of 'no' In-store promotions, after which the difference is taken in order to get the marginal contribution. The average of all these marginal contributions is the corresponding Shapley value.

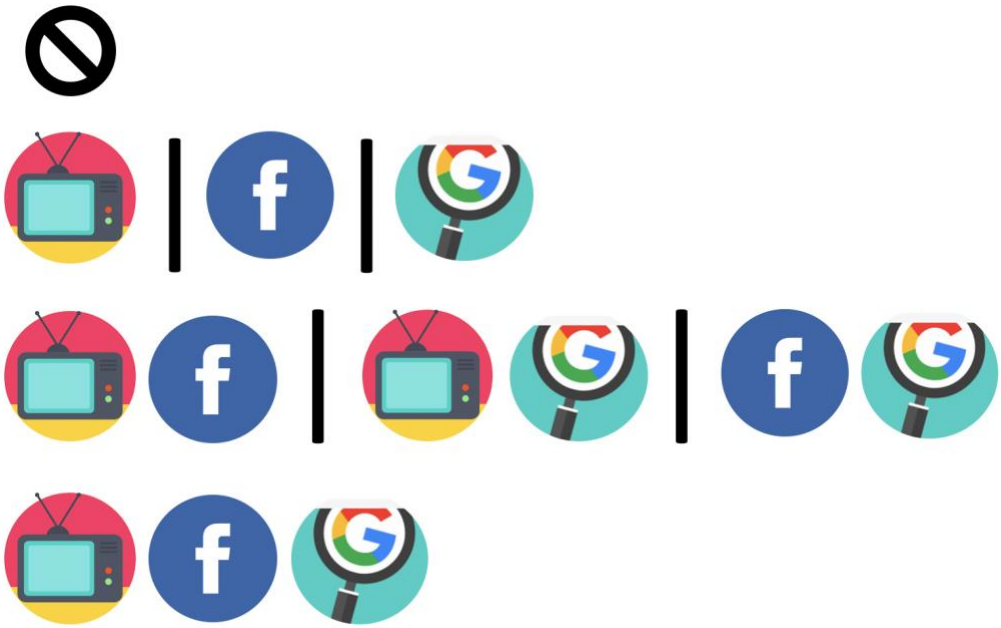

Figure 6. All eight coalitions that are required to compute the exact Shapley value of no In-store promotions feature value.