Bayesian time series for Marketing Mix Modeling

Accuracy, usability and utilisation of prior knowledge

By Alexander Oude Elferink

This research was funded by Facebook

Introduction

Marketing Mix Modeling (MMM) has been one of the most used methods for decades to estimate the impact of media, of developments in the market among competitors, of economic uncertainty and to account for impact of weather, social and cultural trends on revenue. Traditionally, brands employ it to estimate how each factor contributes to (most often) sales, over time. Marketing Mix Modeling (MMM) is a more commonly used methodology in either a marketing or media context compared to e.g. attribution analysis, simply because you can consider a wider variety of factors. Not just when it comes to external factors, but also with regard to the type of media channels which can be distinguished, i.e. both offline and online channels. Analyses come in a variety of forms. From simple linear regression to Long short-term memory (LSTM), a type of artificial recurrent neural network used in the field of deep learning. Despite these differences the final goal is generally the same: unfolding and explaining the impact of media and other factors on sales in the past, to help optimise media investments in the future by means of scenario planning.

In this study we researched the application of Bayesian modeling for MMM and what benefits may arise from taking a Bayesian approach. To assess this, we compare our Bayesian approach to a so-called 'frequentist' approach (a more traditional regression model) by viewing them from two perspectives. The first perspective is about model accuracy and usability, while the second focus is on the implementation of prior knowledge in Bayesian modeling and its implications for the results of analyses. The fact that prior knowledge can be accounted for in Bayesian analysis is one of the main reasons why it has gained popularity in recent years. Previously prior knowledge tended to come in the form of experience from marketers or sales managers, whereas nowadays, prior knowledge comes increasingly more in data driven results derived from lift studies and experiments. We expect that by being able to include prior knowledge, Bayesian MMM will be more accurate and more useful than a frequentist approach.

Background

As described in our earlier research on Machine Learning in MMM[1], use of machine learning has skyrocketed in academia and industry alike. Partially because of increased availability of computing power and partially due to their high predictive power. However, you can argue this predictive power comes at a cost, when it comes to for example explainability. Machine learning methods are often called 'black-box models', because what happens inside them is hard to discern. Furthermore, machine learning models tend to be prone to overfitting, since maximum likelihood estimation[2] is often at the heart of these approaches. Suppose for example that you toss a coin five times and you have to estimate the probability of tossing heads. According to maximum likelihood estimation the probability would be: the amount of heads divided by the total coin toss events. In the end when you have tossed four times heads and one time tails, it would mean that the maximum likelihood is 0.8. We know this is not accurate since a fair coin has only two possibilities, either a heads or tails and hence the probability is 0.5. We also know that if we keep flipping the coin, we could end up with a more realistic tossing probability. This repetitive flipping represents the main difference between frequentist versus Bayesian approaches[3]; frequentist approaches estimate one point - in this context of measuring the impact of a (media) feature over time on sales - and Bayesian approaches estimate a distribution for the possible impact of that feature. A frequentist approach therefore gives you the impact of a certain feature as a number, whereas the Bayesian approach gives you a distribution. This is an added benefit, as it introduces the concept of (un)certainty, something that is not always easy to quantify in frequentist approaches, as it provides a range of numbers of what the impact is likely to be. As mentioned above, Bayesian approaches also have the property of being able to include prior knowledge. Prior knowledge[4] can come from earlier research or from related topics and they serve to inform the current analysis by providing an idea about how the probability distribution will look like. When thinking of our coin flipping example this can come in the form of a fair coin having a probability of 0.5 being head or tails. In MMM research this comes in knowledge of industry domain experts, some academic grounded principles or analyses and experiments that were conducted in the past.

Depending on how the knowledge is incorporated in modeling, weakly-informative, informative or non-informative priors can be distinguished. For a weakly-informative prior one might include the mean outcome of the experiments and some uncertainty around the outcomes in the form of a standard error. However, for the informative prior, the prior is strictly numerical and contains most certainty. Practically this could imply an informative experimental prior refers to experiments generally having shown five percent in sales uplift for a certain marketing factor, and one also highly believes it will be around five percent for the same marketing factor in the future. Non-informative priors principally concern either a flat distribution or a normal

[1] "Using Machine Learning in Media Mix Modeling" https://github.com/annalectnl/ml-in-mmm/blob/master/ML%20in%20MMM%20-%20final%20-%20edited.pdf . Accessed 23 September 2020.

[2] "A Gentle Introduction to Maximum Likelihood Estimation for Machine Learning" https://machinelearningmastery.com/what-is-maximum-likelihood-estimation-in-machine-learning/ . Accessed 23 September 2020.

[3] "The Frequentist vs Bayesian Debate" https://medium.com/datadriveninvestor/chapter-5-machine-learning-basics-part2-69721bf70c7f . Accessed 23 September 2020.

[4] "Prior probability" https://en.wikipedia.org/wiki/Prior_probability Accessed 3 November 2020.

distribution centered around zero. Hence, this form of prior accounts for most uncertainty compared to weakly-informative and informative priors.

Bayesian approaches are becoming increasingly popular as frequentist approaches do not always offer the practical output that is needed to optimise marketing or media effectiveness. As computational power no longer an issue and Bayesian's properties to view feature impact as a distribution, to include prior knowledge in the form of e.g. experiments, make the approach very appealing from a data science as well as a business perspective.

This study examines the benefits of using a Bayesian time series model compared to a frequentist approach for PepsiCo, The international Food & Beverage company. Focus is not solely on comparing both approaches in terms of model accuracy, but also on usability and assessing the added value of implementing prior knowledge in the Bayesian approach. Broadly speaking, we will compare three approaches. As a baseline model we will use the Prophet package in Python[5], to represent our frequentist approach.

For the Bayesian approach two variants are compared: one excluding prior knowledge and one including prior knowledge. As PepsiCo recently performed geo-experiments on the effectiveness of both Facebook and YouTube advertising, we will use their results as priors in our Bayesian approach. A second way of testing the added value of prior inclusion, will be done by imputing the coefficients and standard errors that are found by our Genetic Algorithm when fitting data. We adopted the same Genetic Algorithm as employed in our latest machine learning study as it produced models with higher accuracy (it outperformed Random Forest and eXtreme Gradient Boosting relatively by 20%). Hence we believe that including prior knowledge from the Genetic Algorithm, will amplify performance of our Bayesian approach. For both Bayesian prior models, weakly-informative priors will be used instead of (more) informative priors. Weakly-informative priors indicate that we have an idea of what the impact will look like, without restraining the model to fit the prior exactly, which would require a very informative prior. More informative priors are therefore sound to use when you have a multitude of experiments that have comparable results.

Methodology

Data collection and processing

This study builds on our previous machine learning study and sales data was extended for one snack brand from the multinational PepsiCo: Lay's® among one major supermarket chain in The Netherlands. The related time series now concerns nearly three years instead of two years. All sales data were aggregated from the right stock keeping units (SKU's) to a total in euros at weekly level and were scaled from millions to thousands. To be consistent in sales aggregation for the time series, for Lay's® we distinguished the following SKU's: Lay's Oven Baked (excluding Lay's Oven Baked Crispy Thins, Lay's Oven Baked Crunchy Biscuits and Lay's Oven Baked Veggie), Snacks (Hamka's, Grills, Mama Mia's, Lay's Mixups, Wokkels, Lay's Poppables, Sunbreaks, Pomtips, Lay's Sticks and Lay's Stax), Lay's Sensations (excluding Streetmix and Coated Peanuts), Core (Lay's chips, Lay's Superchips, Lay's Superchips Deep Ridged and Lay's Light chips) and Bugles. Sales data regard both sales in euros made in physical supermarkets and online.

[5] "Prophet - Forecasting at scale" https://facebook.github.io/prophet/ . Accessed 25 September 2020.

To explain sales over time, four categories of data were used for modeling: media, weather, competitor and promotion data. Incorporated media data relates to Out Of Home (OOH) advertising spend, Google Search and Display impressions, clicks and spend. For TV Gross Rating Points (GRPs) were used. For social media, 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. Besides Facebook, Instagram and Snapchat data, this study also incorporated YouTube data from Google DV360 in order to include corresponding geo-experimentation results as prior knowledge during modeling.

For weather data the 'knmy' Python package was used, utilising the Royal Dutch Meteorological Institute (KNMI) API for fetching and parsing weather data observations from 48 automated weather stations. Variables that were fetched include temperature in celsius, sun and rain measured in hours and also rain measured in millimeters on a daily level. Before modeling all these features were averaged on week level, to nationally capture the Dutch weather. Competitor data was collected through Nielsen and included as total gross spend for Lay's®.

To fetch promotion data, we adjusted the custom made web scraper using Python packages 'selenium' and 'BeautifulSoup' from our previous machine learning study. Now we were not only able to gather product names and prices (before promotion and during promotion) from online supermarket folder data, but also product promotion descriptions. By getting the latter, additional feature engineering resulted in being able to differentiate between basket composition promotions (e.g. getting two bags of chips for two euros instead of three euros) and regular discount promotions. Also, this study differentiated promotions among competitors of Lay's® on 'internal' and 'external' level. Hence we were able to seperate possible cannibalisation effects from competitor promotion effects. Furthermore, another custom made web scraper was created based on 'selenium' and 'BeautifulSoup' to collect dates of soccer matches for the Dutch national team, Ajax and the UEFA Champions League (UCL), to include those dates either in Prophets holidays parameter or as extra dummy features in our Bayesian approach to estimate the impact of soccer for Lay's®. Lastly, it should be noted that we accounted for the impact of COVID-19 as epidemics and pandemics are historically been known for human responses that are integral to increased buying behavior[6]. Through creating a mean search index proxy built on using multiple related COVID-19 keywords using a Google Trends API, we captured topicality around COVID-19 and its impact.

Feature engineering adstock

Once data collection and cleaning was completed, data was split into a train (70%) and test (30%) data set. Media and competitor data underwent adstock feature engineering. Following an earlier study[7] on adstock transformations, exclusively Weibull transformations were used, as they improved accuracy in terms of R-squared and Mean Absolute Percentage Error (MAPE).

[6] "Psychological underpinning of panic buying during pandemic (COVID-19). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7202808/ Accessed 17 November 2020.
[7] "Modeling adstock using Weibull transformations" https://github.com/annalectnl/weibull-adstock/blob/master/adstock_weibull_annalect.pdf . Accessed 29 September 2020.

In line with our machine learning study, a grid with all unique combinations of Weibull transformations were created. 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.

Genetic feature selection

As mentioned in the background section, a Genetic Algorithm (GA) was used to determine the best adstock transformation for a certain feature. This is justified as it not only improves model accuracy but also feature selection speed. It's higher speed derives from applying parallel fitting instead of sequential testing. GA's are commonly used within the field of computer science and operations research[8] for optimisation and search problems. The term Genetic Algorithm itself, is a metaheuristic that refers to the process of natural selection and belongs officially to the larger class of Evolutionary Algorithms. Our GA consists of six steps (see Figure 1), which are further described in our latest machine learning paper[9].

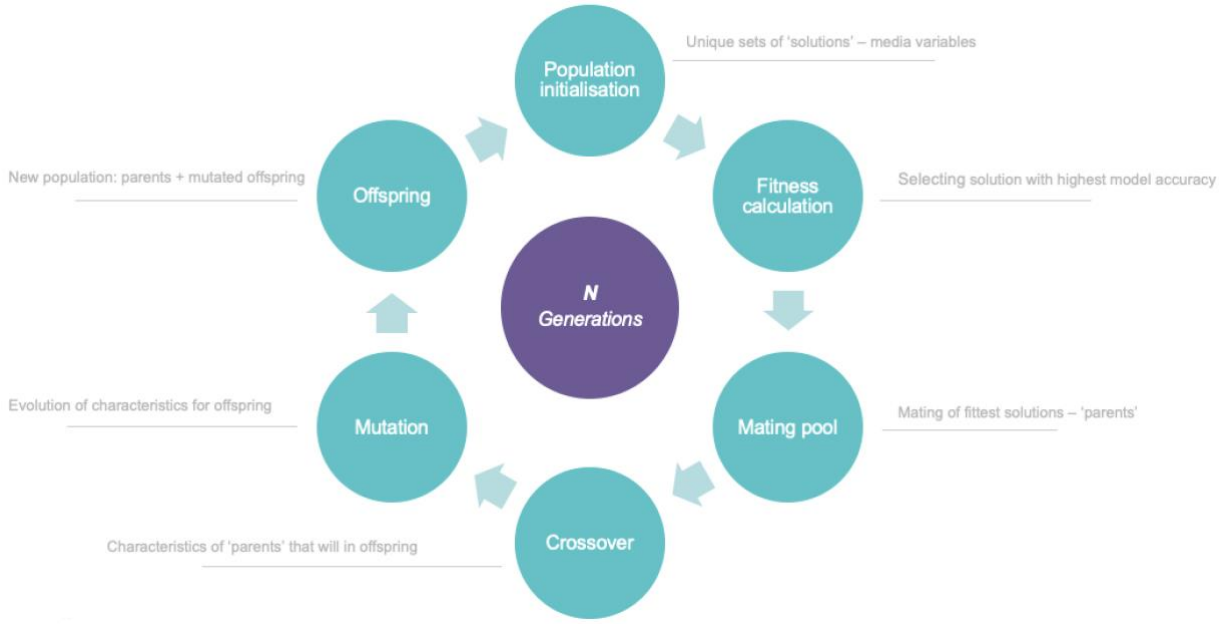


Figure 1. Genetic Algorithm (GA) process for feature selection.

Once the GA completed all 2000 generations and thus selected the best variant per feature, our GA optimised our total feature set from a MAPE of 14,08% to 11,29% (Figure 2). Before starting modeling, we checked for multicollinearity by doing Variance Inflation Factor (VIF) checks on the final set of features. Taking a threshold of five, no indication of multicollinearity had been detected. Note that Facebook and Instagram features were merged to a 'paid social' feature.

[8] "Genetic Algorithm" https://en.wikipedia.org/wiki/Genetic_algorithm Accessed 29 September 2020.
[9] "Using Machine Learning in Media Mix Modeling" https://github.com/annalectnl/ml-in-mmm/blob/master/ML%20in%20MMM%20-%20final%20-%20edited.pdf . Accessed 16 November 2020.

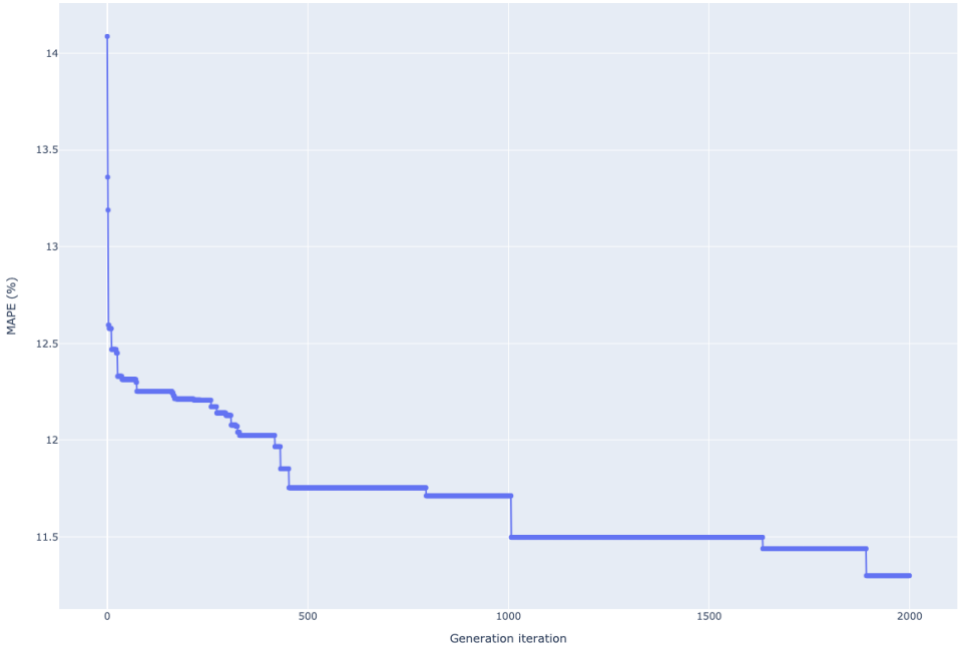


Figure 2. Genetic Algorithm (GA) fitness performance (MAPE) over 2000 generations.

Modeling

For modeling Lay's®, a frequentist approach and a Bayesian approach were applied. For the frequentist approach, also our baseline model, Prophet[10] was used. For the Bayesian approach PyMC3[11] was utilised. PyMC3 is a state-of-the-art probabilistic Python programming framework that allows for Bayesian inference based on user-defined probabilistic models. The approach is inherently Bayesian given its fundament on Bayes Theorem[12] and such one can specify priors to inform and constrain models and get uncertainty estimations in the form of posterior distributions. The distributions are generally very flexibly estimated by using Markov Chain Monte Carlo (MCMC) sampling algorithms. However, one major drawback of pure MCMC sampling is that it is often slow as it is computationally expensive. In this study we have therefore made use of the more recently developed variational inference algorithms, that are nearly as flexible as MCMC, but much faster as they rely on Theano[13]. To be more precise: Automatic Differentiation Variational Inference (ADVI)[14] was used, that instead of drawing samples from posterior distributions, fits a distribution (e.g. normal) to posteriors turning a sampling problem into an optimization problem by calculating gradients.

[10] "Quick Start | Prophet" https://facebook.github.io/prophet/docs/quick_start.html Accessed 5 October 2020.
[11] "Probabilistic Programming in Python using PyMC3" https://www.researchgate.net/publication/308715485_Probabilistic_programming_in_Python_using_PyMC3 Accessed 6 October 2020.
[12] "Bayes theorem" https://en.wikipedia.org/wiki/Bayes%27_theorem Accessed 9 October 2020.
[13] "Hands-On Theano: One of the Most Powerful Scientific Tools for Python" https://medium.com/towards-artificial-intelligence/hands-on-theano-one-of-the-most-powerful-scientific-tools-for-python-f023a1929f57 Accessed 7 October 2020.
[14] "Automatic Differentiation Variational Inference" https://www.jmlr.org/papers/volume18/16-107/16-107.pdf Accessed 8 October 2020.

As described before, our Bayesian approach encompasses in total three model variants. First, a no prior model. Second, a prior model with coefficients and standard errors found by our GA for all media plus promotion features. Third, another prior model with the geo-experiment results of Facebook and YouTube executed for the SKU Bugles. With sales lift results of respectively: 1,5% and 3,9% in a period of six weeks. Note that we assumed normal and bounded normal distributions for modeling. For which the latter were used in the prior models for media and promotions features.

Results

Before assessing model accuracy for our Bayesian models, we checked for model convergence using the Gelman Rubin diagnostic[15]. The Gelman Rubin diagnostic only showed values of one's or close to one for all models, hence convergence was statistically confirmed. The Bayesian Fraction of Missing Information (BFMI)[16] was calculated to quantify the quality of all our sampling procedures. All BFMI values were well above 0.2, so there was no indication of poor sampling. It can be derived from Table 1, that our Bayesian PyMC3 models performed better than Prophet when it comes to model accuracy, measured with the Mean Absolute Percentage Error (MAPE). Specifically the no prior model did best with a MAPE of 8.65% on test data.

Model (test) data	**Prophet**	**Bayesian GEO priors**	**Bayesian GA priors**	**Bayesian No priors**
major supermarket chain - Lay's	*10,92%*	*10,61%*	*10,28%*	*8,65%*

Table 1. Modeling results of Prophet versus PyMC3 with MAPE as model accuracy metric.

Model Usability: social advertising and promotions

In this section, we will specifically compare the impact of paid social and YouTube given the geo-experiment. We investigated how impact changes when we include or exclude prior knowledge in modeling. Furthermore, as there was interest in comparing the impact of basket composition promotions versus regular discounts, these will also be compared. To make comparison fair, models were trained on all data available and decomposition was done on the geo-experimentation period, i.e. week 35 to 40 in 2019. Figure 3 shows all contributions of compared features across used modeling methodologies. A decomposition overview of all other features per methodology can be found in Appendix 1.

For our baseline model Prophet (Figure 3) we found relative contributions of 6,72% for paid social, YouTube (4,59%), regular discounts (3,78%) and basket composition promotions (2,36%).

[15] "General Methods for Monitoring Convergence of Iterative Simulations" http://www2.stat.duke.edu/~scs/Courses/Stat376/Papers/ConvergeDiagnostics/BrooksGelman.pdf Accessed 20 October 2020.
[16] "Diagnosing Suboptimal Cotangent Disintegrations in Hamiltonian Monte Carlo" https://arxiv.org/pdf/1604.00695.pdf Accessed 20 October 2020.

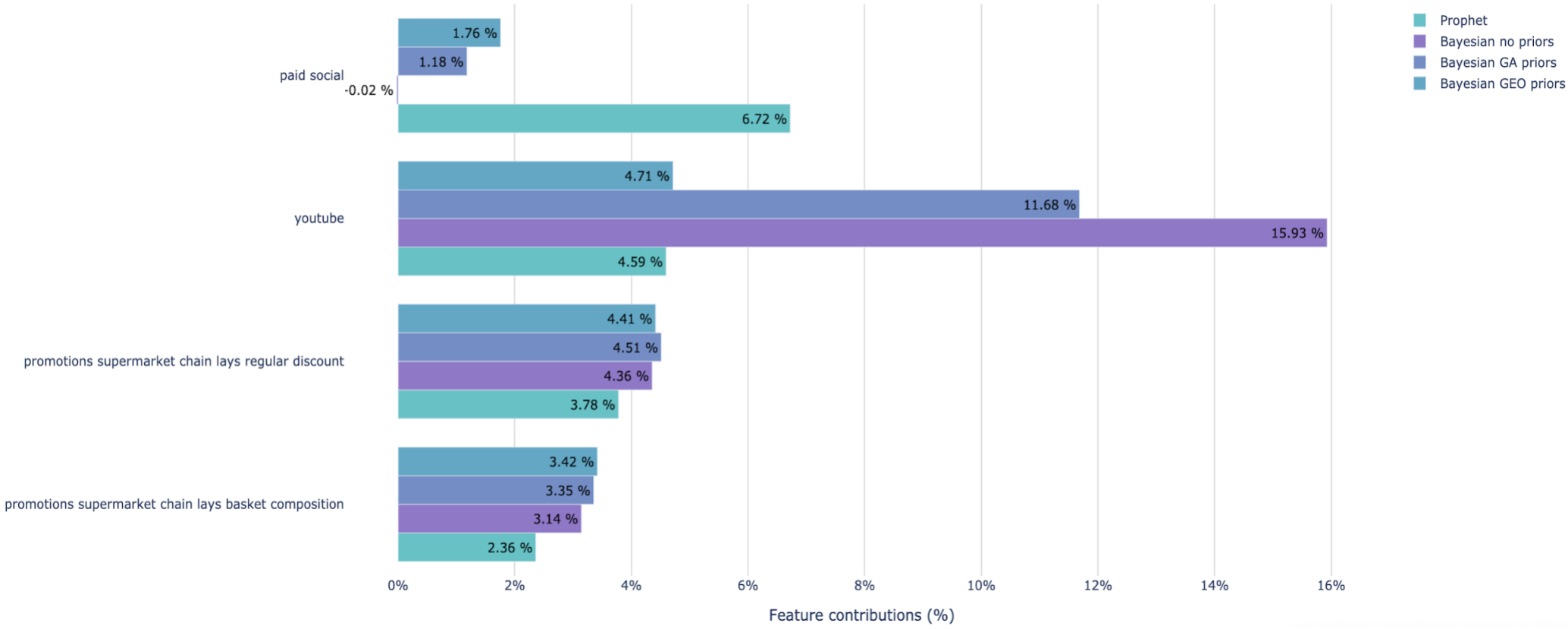


Figure 3. Contribution overview of all compared features (paid social, YouTube, regular discount promotions and basket composition promotions) across utilised methodologies for the major supermarket chain Lay's models.

In the Bayesian no prior model (Figure 3), YouTube has the largest contribution share in the four features we compare (15,93%). Regular discounts follow with 4,36%, basket composition promotions 3,14% and paid social -0,02%. In the Bayesian GA prior model (Figure 3) regular discounts, and basket composition promotions show a contribution with about the same magnitude of respectively 4,51% and 3,35%, whereas YouTube contributes 11,68% and paid social 1,18%. In the Bayesian GEO prior model we also find that YouTube has the largest contribution (4,71%) although it shrinked by more than half. Of other contributions, regular discounts represent 4,41%, basket composition promotions 3,42% and lastly paid social roughly doubled to 1,76%.

In general, it appears that the Bayesian models devote more impact to YouTube than the baseline Prophet model. In contrast, Prophet devotes more impact to paid social. Regular discounts and basket composition promotions are approximately in the same range across all models. This can be elucidated due to high seasonality of the promotions that match sales peaks. Moving from models and their contributions without prior knowledge to models with priors, we see that the impact of YouTube shrinks and that of paid social increases.

Prior knowledge

Prior knowledge was used in two forms to test for its added value. The first form relates to coefficients and standard errors found by the GA that were imputed as priors. In addition, earlier geo-experiment results on paid social and YouTube advertising for the SKU Bugles, were used. The latter is what we consider a stronger and more informative prior than the former. During the experiment of six weeks (week 35 - 40 in 2019), different combinations of switching on Facebook and YouTube advertising were executed over all 12 Dutch provinces. Each week had a different advertising combination: four provinces Facebook only, four provinces YouTube only and four provinces with Facebook and YouTube switched on (see Figure 4 for the experiment design).

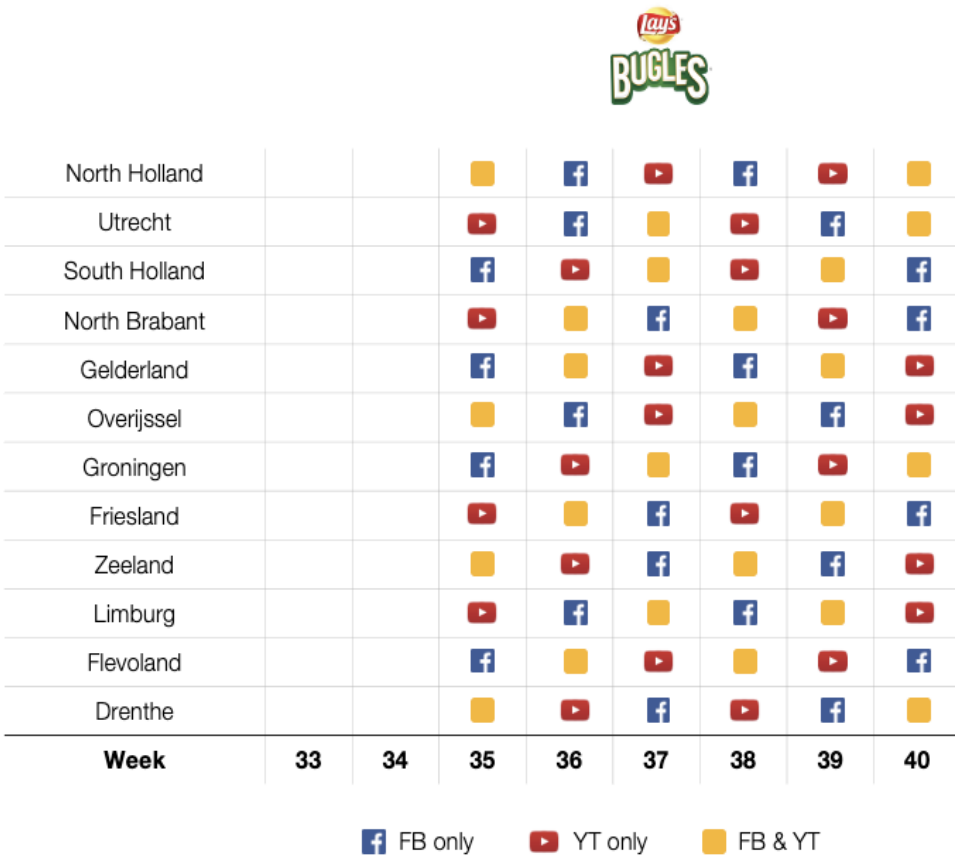


Figure 4. Geo-experiment design on Facebook and YouTube advertising for SKU Bugles.

Effects of having Facebook or YouTube switched on were determined by comparing them to when being switched off. This comparison was made by province. To have one national uplift effect for Facebook and YouTube, effects were weighted by province according to their respective sales share and subsequently averaged. By doing so, the results showed that Facebook had an uplift of 1,5% in sales, whereas YouTube showed an 3,9% uplift in sales, both uplifts were statistically significant. To include the geo-experiment results as priors in the Bayesian GEO prior model, first respective sales shares of Facebook (1,5%) and YouTube (3,9%) were calculated per week over the geo-experimentation period. Afterwards, their weekly shares of total sales were divided by corresponding weekly adstock transformations in order to get weekly media contributions. Over the six week total of Facebook (paid social) and YouTube contributions, both the mean and standard deviation were calculated to subsequently use these as priors into prior parameters mu and sigma.

By looking at the sampled posterior distributions of all four compared features in the model usability section (i.e. paid social / Facebook, YouTube, regular discount and basket composition promotions) across our Bayesian no prior, GA prior and GEO prior model, we can reflect the powerful concept of uncertainty that these models bring.

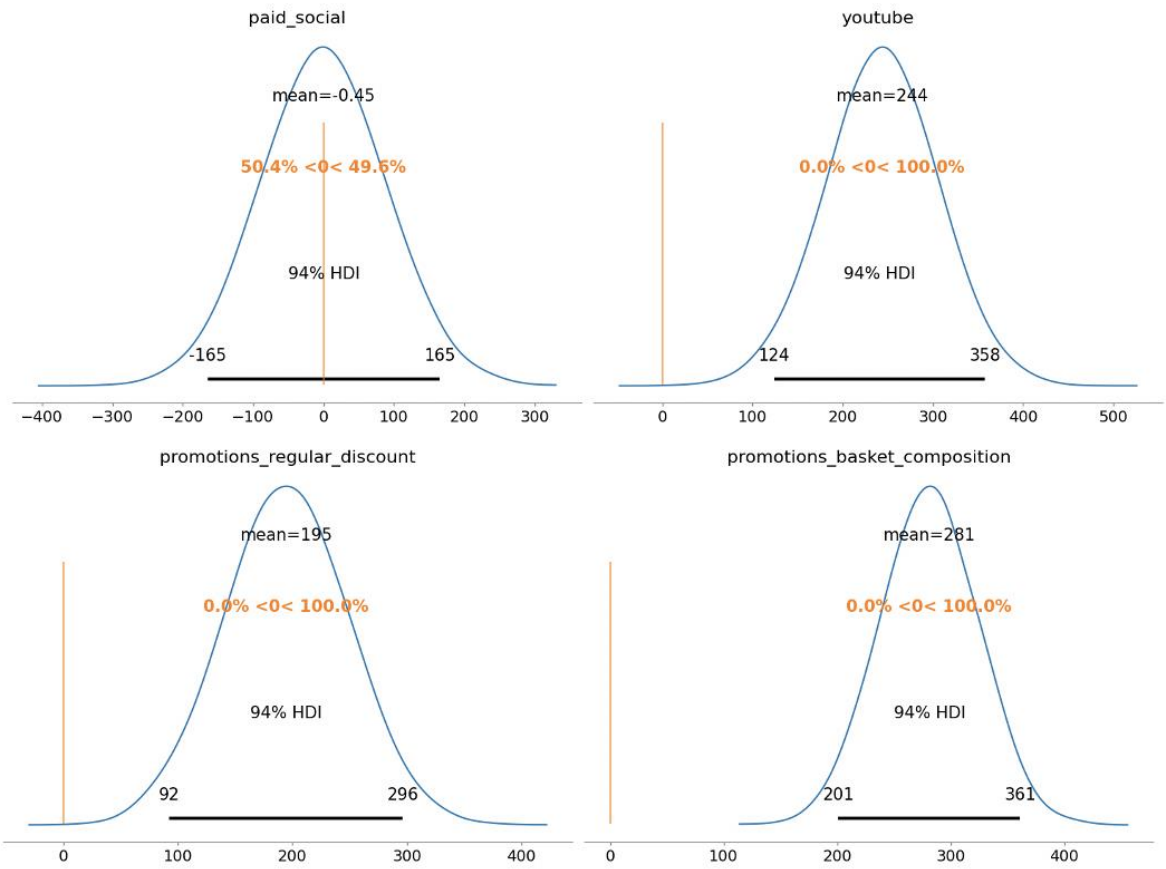


Figure 5. Posterior distributions of Facebook (paid social), YouTube, regular discount and basket composition promotions for the Bayesian no prior method for the major supermarket chain Lay's model

Posterior distributions of the Bayesian no prior model (Figure 5) show that three out of four features have a 94% Highest Density Interval (HDI) certainty that the coefficients are positive. For paid social the impact is 50,4% more likely to be negative. Also interesting to see is that basket composition promotions have the largest coefficient with a mean of 281. Figure 6 shows roughly the same for the Bayesian GA prior model. However, comparing the ranges of the 94% HDI certainty, we see that they are considerably smaller. Especially for Facebook there is now a non negative posterior distribution. After basket composition promotions, regular discount promotions now have the second largest coefficient instead of YouTube.

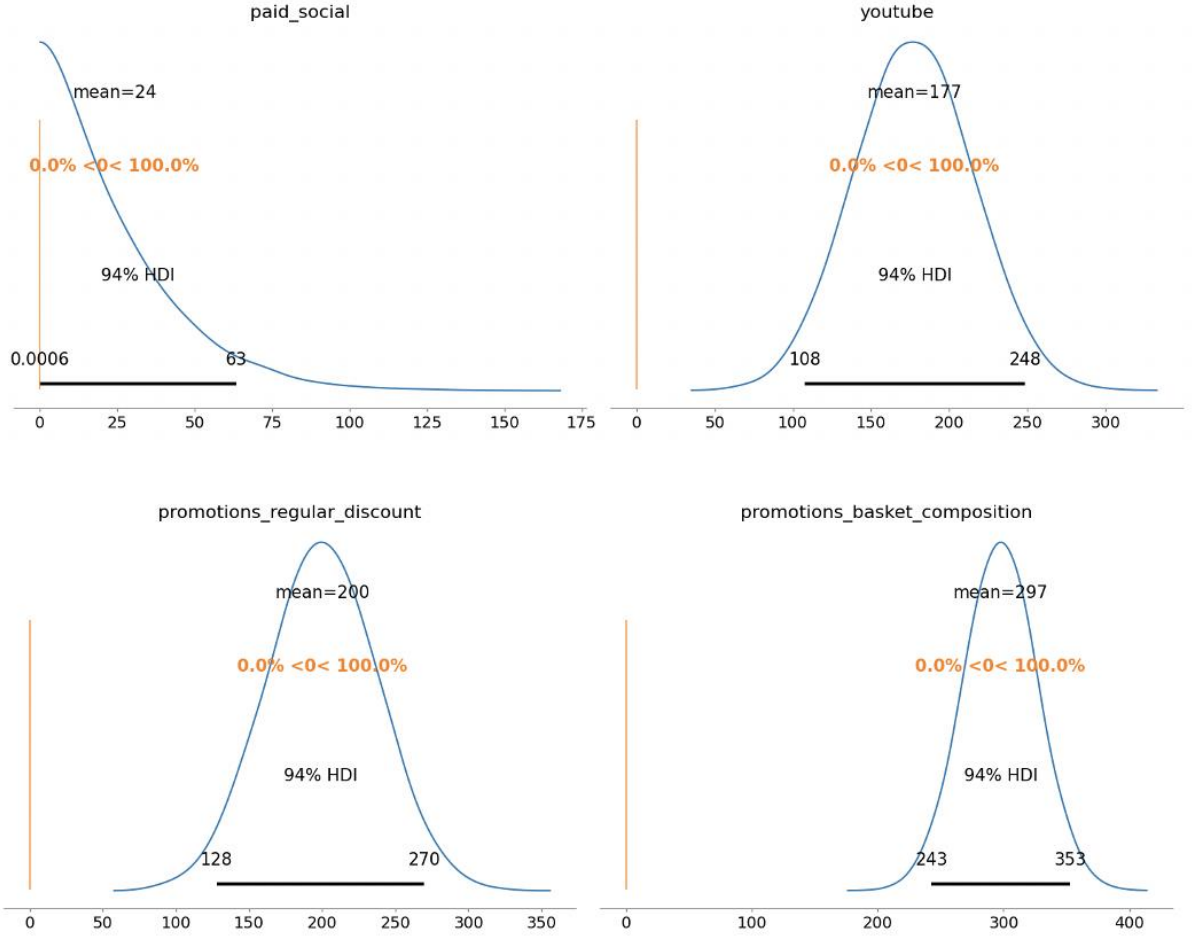


Figure 6. Posterior distributions of Facebook (paid social), YouTube, regular discount and basket composition promotions for the Bayesian GA prior method for the major supermarket chain Lay's model.

The posterior distributions of Bayesian GEO prior model (Figure 7) show that by using a stronger and more informative prior (i.e. geo-experiment results) can drastically reduce uncertainty around feature impact of both Facebook and YouTube. The 94% HDI certainty intervals for the regular discount and basket composition promotions however increased somewhat. Still, the order of feature impact remains the same as in the Bayesian GA prior model.

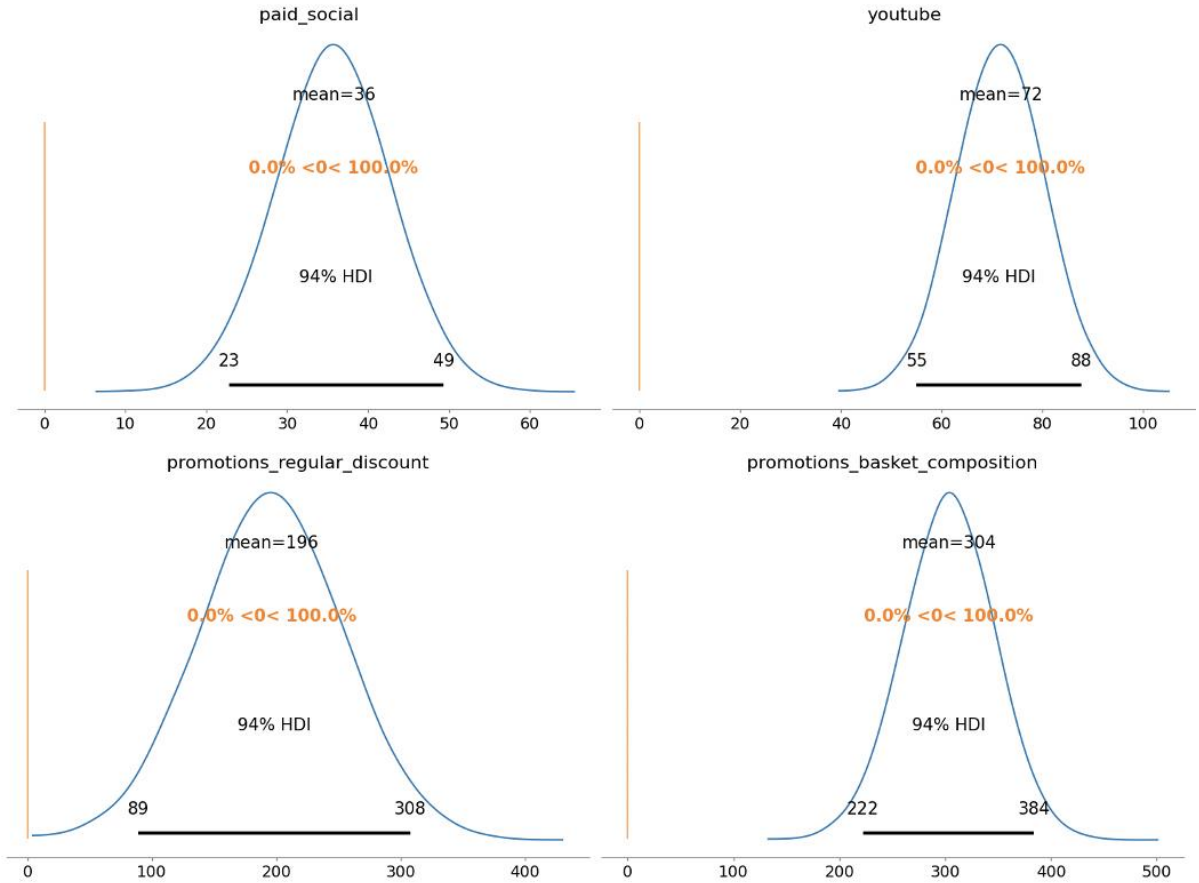


Figure 7. Posterior distributions of Facebook (paid social), YouTube, regular discount and basket composition promotions for the Bayesian GEO prior method for the major supermarket chain Lay's model.

When considering modeling results of Prophet and all Bayesian models, it can be deduced that going from a no prior model to a stronger weakly-informative prior model, uncertainty around features can be drastically reduced. Our Bayesian GEO prior model showed the most realistic and comparable results on scales of contributions for Facebook (1,76%) and YouTube (4,71%) versus the 1,5% and 3,9% geo-experiment results respectively. It should be noted that although regular discount promotions showed higher sales contributions than basket composition promotions, basket composition promotions have a stronger impact when comparing all associated Bayesian posterior distributions. Considering that two out of six weeks had regular discount promotions and only one week basket composition promotions, while having nearly the same contributions (overall ~1% difference), the latter form of promotion has twice as much impact.

Discussion and Conclusions

The research goal was to assess potential benefits of taking a Bayesian approach versus a frequentist approach in MMM. We compared both approaches on scales of model accuracy and usability. Also we assessed the added value of having the possibility to include prior knowledge in Bayesian modeling. When it comes to prior knowledge, two forms of weakly-informative priors can be distinguished in this study. Instead of using informative priors, weakly-informative priors were used for Bayesian modeling as these tend to be closer to reality given they account for more uncertainty than informative priors. For example stating that sales uplift will be strictly 5% in the future for a certain media channel would be unrealistic, as the world around us keeps changing in a multitude of ways. Moreover, other Bayesian meta analysis[17] shows that weakly-informative priors are more useful than non-informative priors and generally guard against overfitting. The two forms of weakly-informative priors in this study refer to: 1.) coefficients and standard errors that were found by our Genetic Algorithm (GA) and 2.) the geo-experiment results on Facebook and YouTube advertising for Lay's® SKU Bugles. To properly examine the added value of utilising these forms of prior knowledge, results were not only compared to the Prophet frequentist approach but also to the Bayesian non prior model. We expected that the ability of including prior knowledge would make Bayesian MMM more accurate and useful.

Results showed that taking a Bayesian approach performed better than a frequentist approach (Prophet) on model accuracy measured with the mean absolute percentage error (MAPE). The Bayesian no prior model emerged to be the most accurate model with a MAPE of 8,65%, which was not our hypothesis. Theoretically it can be stated that it is not fully unexpected, since using priors restricts parameter space for features more than using no priors and just assuming normal distributions. When further considering the decomposition of contributions per model over the geo-experimentation period, we see that the Prophet model attributes a higher contribution to paid social, while the Bayesian models show more impact on YouTube. The compared promotion variables: regular discount promotions and basket composition promotions, show overall similar contributions. Having high seasonality in the promotions that match sales peaks clarify this. To get the best grasp of media contributions, we recommend experimentation within media planning. That in general contribution differences exist, is of course inherent on that e.g. Prophet does not allow for prior knowledge plus it accounted for separate holiday parameters, whilst the Bayesian models accounted for holiday effects through one parameter. Having less features for model estimation, could lead to less potential noise and thus may explain why the Bayesian models perform better than Prophet. This is the case when the features try to capture a similar pattern and hence cause multicollinearity.

With regard to usability, both the frequentist approach (Prophet) and the Bayesian approach are well suited for MMM since both provide coefficients and contributions to sales. This is what is fundamentally needed for MMM, since this supports the main reason of executing MMM: scenario planning and testing. By default Prophet returns 80% uncertainty intervals for its forecast, i.e. in this case sales predictions. This is important when the purpose is to do precise planning and testing. However, these uncertainty intervals cannot be used to say that the probability is 80% of sales having certain values. The 80% probability is a property

[17] "Bayesian Meta-Analysis with Weakly Informative Prior Distributions" https://osf.io/7tbrm/download?format=pdf Accessed 3 November 2020.

of the rule that was used to create the interval, not the interval itself. So this 80% probability relates to the reliability of the estimation procedure[18]. The same holds for p-values. In contrast, with Bayesian inference one can derive exactly the desired interpretation: a Bayesian 94% interval for a feature means precisely that there is a 94% probability that the feature values are in that interval[19]. Having said that, this is a great advantage of the utilised Bayesian approach. Not only gives associated models directly an uncertainty interval of sales in the form of a 94% Highest Density Interval (HDI) certainty distribution, but also do so for all other features. Hence one can inspect the certainty of feature impact or put differently, see what the probability is of a feature taking a range of values and also infer whether feature impact is strictly positive, negative or a mix of both. Moreover, these joint distributions allow the researcher to examine trade-offs among parameter estimates for multiple features. Other advantages are that one can update model knowledge in the form of priors and forces to better reflect on either similarities and differences of other research, account for asymmetric distributions and not have to assume normal distributions as often holds for frequentist approaches[20] and the ease of making multiple comparisons in modeling by hierarchical modeling that subsequently allows for sharing information across groups that are compared; i.e. "shrinkage"[21].

To specifically examine the added value of prior knowledge in modeling, we compared our two forms of priors to having no prior knowledge. As it became apparent, moving from a no prior model to more stronger weakly-informative prior model, did bring contributions for specifically Facebook and YouTube closer to sales uplift results that were found in a previous geo-experiment. The Bayesian GEO prior model showed respective contributions of 1,76% for Facebook and 4,71% for YouTube and most closely resembled the geo-experiment results (Facebook: 1,5% and YouTube: 3,9%). The uncertainty around feature impact is also reduced when using stronger priors; i.e. our geo-experiment results instead of GA priors. Furthermore, we found that basket composition promotions have a higher impact on sales than regular discount promotions. We can emphasise its importance by calculating the probabilities for both types of promotions, when we take a certain reference value in the distributions. Across Bayesian models (Figure 5, 6, 7) we see that basket composition promotions have at lowest, a coefficient of around 200, so let this be the reference value. Subsequently calculating probabilities for regular discount promotions to have at least this impact, shows an average probability of 47,7% versus a 98,6% probability of basket composition promotions (see Appendix 2). Accordingly we quantified that the major supermarket chain should put promotional focus on basket composition promotions for Lay's®.

This research indicated that a Bayesian approach for MMM is worth considering, given the multitude of aforementioned advantages. Mainly the hallmark of viewing either the measured KPI (sales), features and even priors as distributions is a great and useful concept. Hence one can directly calculate the probability from the posterior distribution to test if a certain

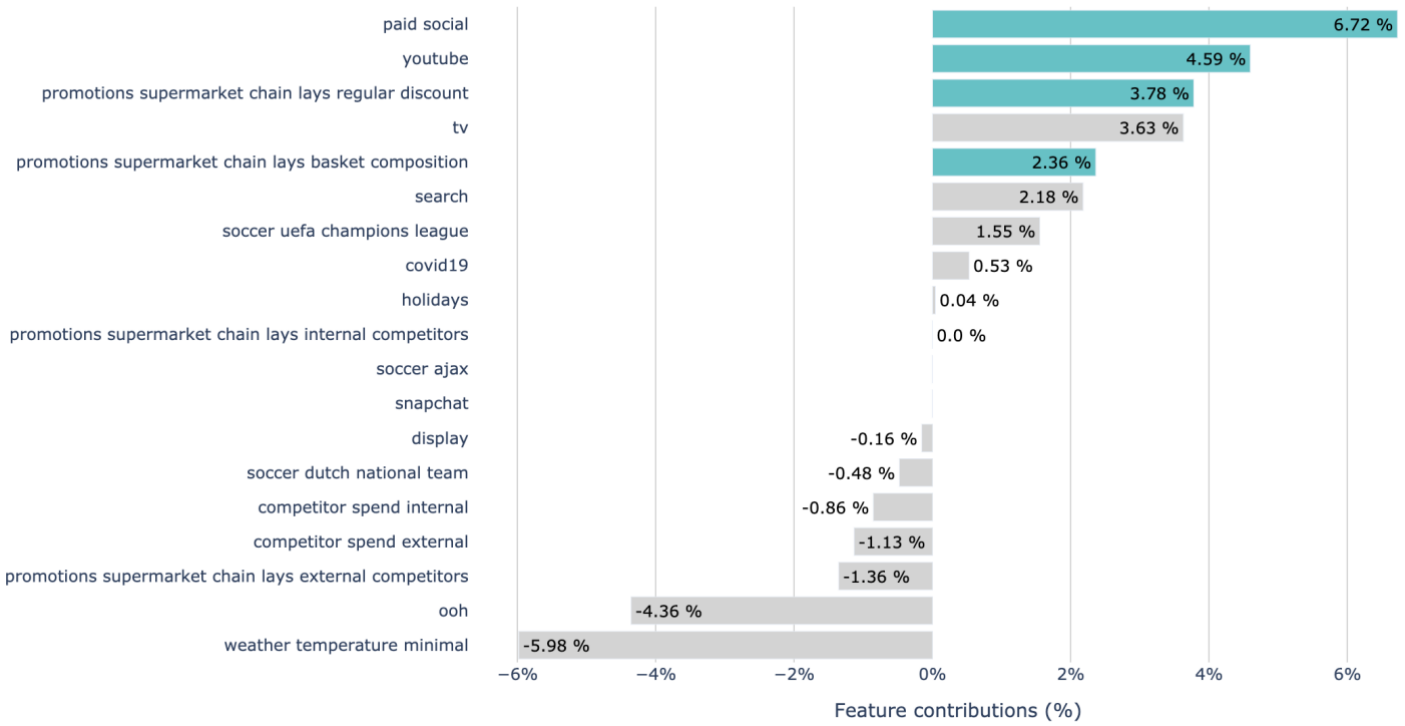
[18] "Confidence interval" https://en.wikipedia.org/wiki/Confidence_interval Accessed 18 November 2020.
[19] "Bayesian statistics: principles and benefits" https://library.wur.nl/frontis/bayes/03_o_hagan.pdf Accessed 18 November 2020.
[20] "A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research" https://www.rensvandeschoot.com/wp-content/uploads/2017/02/2014-JA-RENS-Kaplan.pdf Accessed 18 November 2020.
[21] "The Time Has Come: Bayesian Methods for Data Analysis in the Organizational Sciences" http://www.hermanaguinis.com/ORM2012.pdf Accessed 18 November 2020.

hypothesis on a feature holds. For example in the last paragraph we showed that calculating the hypothesis H_0: $\beta_{regular_discount_promotions} \geqq 200$, allowed us to not only reject the null hypothesis but also to quantify corresponding support for it in the form of a probability. This probability $P(\beta_{regular_discount_promotions} \geqq 200)$, also called the Bayesian p-value, can directly be interpreted as a measure of strength of evidence for the null hypothesis[22]. This makes Bayesian an appealing methodology compared to frequentist approaches, as it can be considered as more 'whitebox' modeling. A sole focus on model accuracy is not a good practice in MMM; considering the Bayesian no prior model had the highest accuracy in terms of statistical rigour, it was not accurate enough when comparing against lift results. In contrast, our Bayesian GEO prior model was only 1,96% less accurate (measured in MAPE) compared to the Bayesian no prior model, though it was closest to our ground truth in the form of the experiments. Maximising model fit to reflect incremental using external evidence like (geo-)experiments are highly favorable, especially if the available quantity of experimental results relates to numerous ones. Hence, we encourage all individuals that can be involved with data driven marketing - from marketing manager, to data scientist, analyst up to c-suite level - to do experiments, since these can serve as ground truth and thus can be costly methods of validation for MMM. Before execution of these experiments one should always first consider a so-called 'dark period', i.e. without advertising, to be able to accurately calculate lift of advertising afterwards. Of course it might be that this is not an option because some type media channels e.g. have an always-on layer or preference is just to not switch off advertising for a concerned channel. If one of these two points are the case, we demonstrated that switching on and off Facebook, YouTube or having a combination of them, can be a good solution too. Finally note that you do not have to arrange everything around experimentation yourself. Experimentation is increasingly more accessible, certainly since big tech companies like Facebook and Google make technical support convenient.

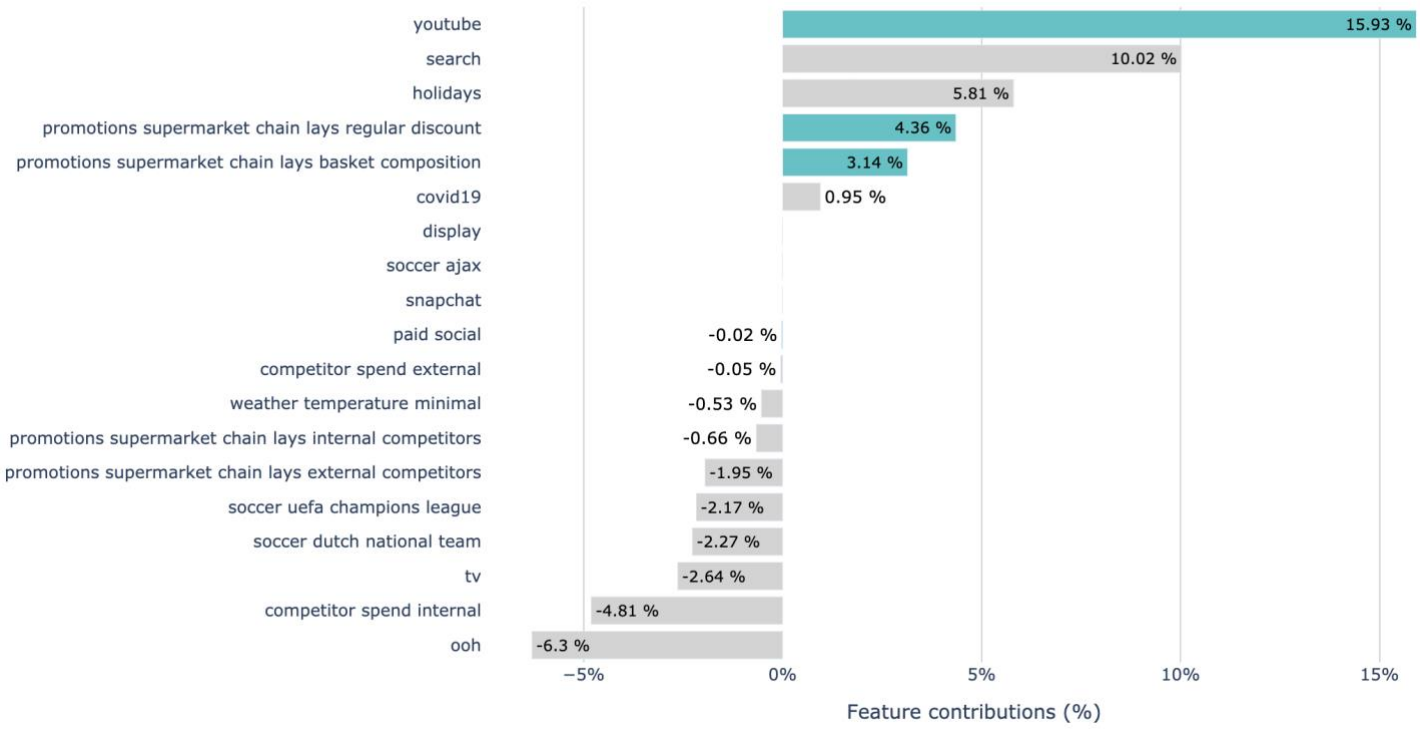
[22] "BANOVA: Bayesian Analysis of Experiments in Consumer Psychology" https://www.rug.nl/feb/organization/departments/marketing/seminar/docs/titleabstractmichelwedel.pdf Accessed 18 November 2020.

Appendix 1. Decomposition overview of contributions per methodology

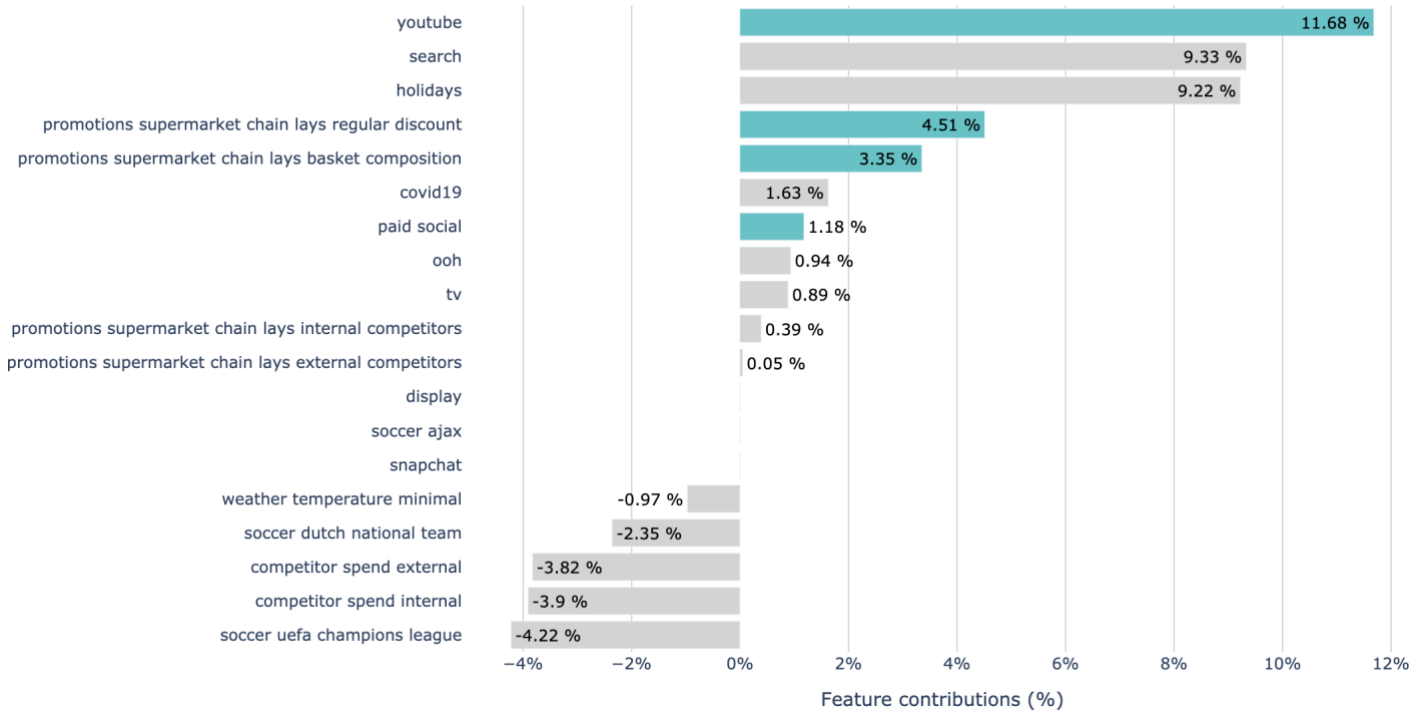
a. Prophet



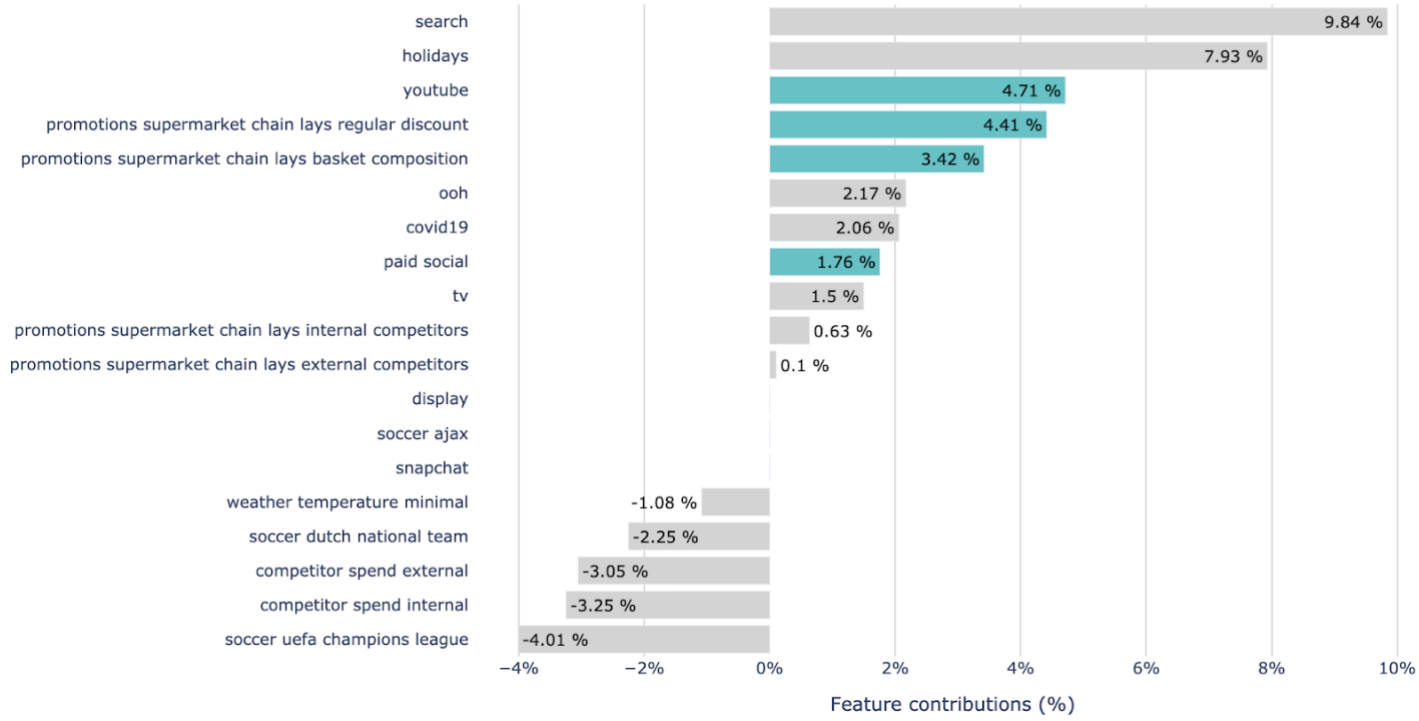
b. Bayesian no priors



c. Bayesian GA priors

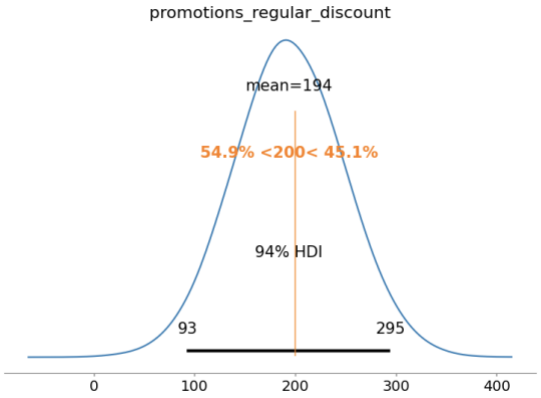


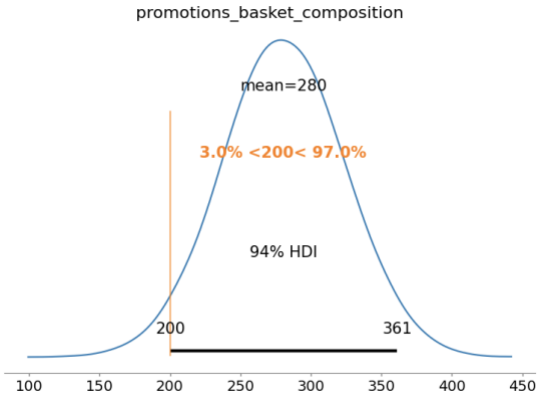
d. Bayesian GEO priors



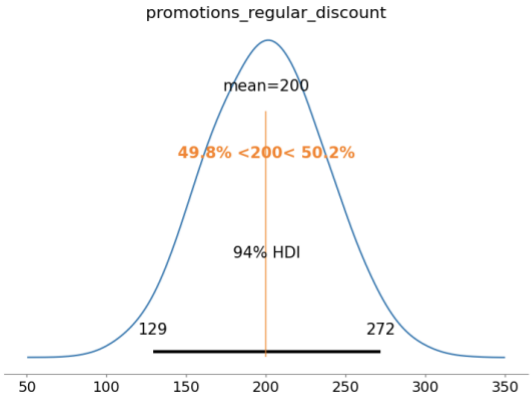
Appendix 2. Bayesian promotion probabilities across models

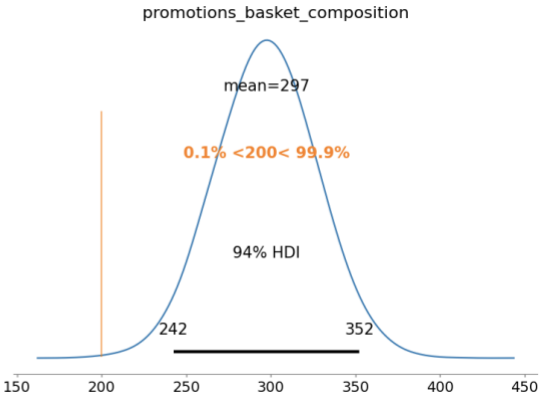
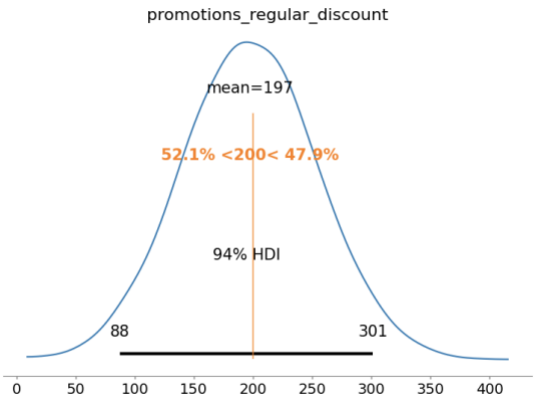
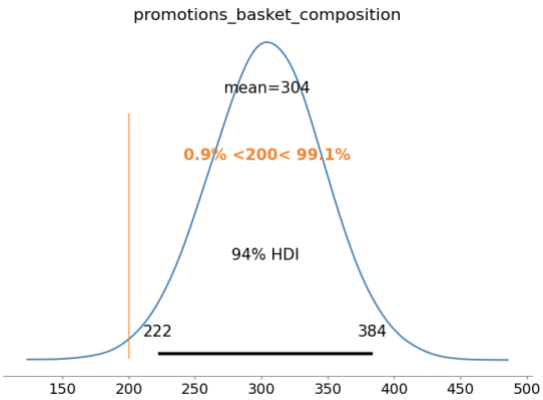
a. Bayesian no priors





b. Bayesian GA priors





c. Bayesian GEO priors