Project Work

Data-Driven Marketing: Developing a Strategic Plan for Boosting Product Sales in the Next Year

Fabiano Mangini

Objective

The objective is to develop a data-driven marketing strategy for the upcoming year to maximize product sales. This involves analysing existing data to identify patterns and trends, thereby informing a plan that aligns with and anticipates market dynamics to enhance sales performance.

Table of contents

Introduction

To achieve the goal of enhancing future sales, a careful and methodical approach is needed. For this reason, the analysis here reported has been carried out following the Crisp-dm standard. Thus, the following findings are the result of 5 phases: Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation (Deployment is the only phase which has not been covered). The code relative to these steps is available in the Databricks notebook provided and will not be discussed in this report.

Throughout the course of the data analysis all the required aspects have been explored: relation between variables and target, variables relation with time, external influences, synergy effect, saturation effect and carryover effect.

Findings

Beginning with the relationship between the variables and the target: revenue. The first insight indicates that, for the data provided, the most influent advertising channel has been Tv, followed by Facebook with about half the impact, then Out of home and last Print. However, when considering the efficacy, it has been found that Facebook advertising performs the best, about five times better than the second best: Tv. Print and Out of home follow in this order.

For what regards the variables relation with time, a consistent seasonal trend has been observed. It repeats yearly: starting with a revenue peak in December, reaching a minimum in June and then increasing again until the next December. In part it can be associated with the equivalent trend in advertising spendings, but the same trend remains true also for the competitor sales (which shouldn’t be correlated with our spends).

As regards the external influences, there is a clear indication that the Competitor sales can effectively be used as a source of information for general market trends.

No signs of synergy effects have been found while saturation effects are too small to be relevant.

There is a clear sign of carryover effect which seems to be particularly bound to the Out of home and Facebook advertising channels.

Coefficient analysis

The findings related to the influence and efficacy of the advertising channels stem from the analysis of the learned coefficients of the models trained to predict the revenues. In particular, the model here referred achieves an R-squared score of 0.94 and a Root Mean Squared Error of 173k. These performances are considered high enough to consider the models' explanation of data reliable.

For the sake of clarity, a detailed description of the coefficients meaning must be reported.

Two sets of coefficients have been calculated. The first ones explain the revenue value in terms of features variance and are referred to as normal coefficients, the second ones explain the revenue value in terms of raw feature value and are referred to as informative coefficients.

In other words, the first set better represents the influence of each feature on the target variable given the available data: a normal coefficient gives a measure of how much revenue is created thanks to a standard deviation amount of dollars spent on a specific advertising channel, where the standard deviation is different for each channel, reason why these coefficients are not good for comparisons across different channels.

E.g. a channel coefficient could be large because a lot of money has been spent on that channel and therefore it has influenced revenues a lot, but this doesn't mean that this channel is more efficient than the others in creating revenue (it could simply have a high standard deviation!).

The second set better represents the efficiency of each feature in creating revenue: an informative coefficient gives a direct measure of how much revenue is created thanks to a dollar spent on a specific advertising channel, making it good for comparisons across different channels.

E.g. a channel coefficient could be small because very little amount of money has been spent on that channel and therefore it has not influenced revenues much, but this doesn't mean that this channel is the less efficient than the others in creating revenue.

To make it even more clear: Let's say for a given week the revenue is 101.000 $, if 100.000 $ have been generated thanks to the channel_1 and 1.000 $ have been generated thanks to the channel_2 then, surely, the channel_1 has been more influent overall, but if 10.000 $ have been spent on channel_1 and only 1$ has been spent on channel_2 then channel_2 has been more efficient (10x increase vs 1000x increase).

The informative coefficients can be obtained either by training a model on normalized features and then dividing the learned coefficients by the standard deviation of each feature (standard deviation calculated on the training set) or by training a model on non-normalized features. The first approach is more numerically stable and therefore has been used.

Once this explanation has been given, the obtained coefficients can be analysed.

Normal coefficients:

Revenue = [128k * sTv_C + 8k * sOoh_C + 44k * sPrint_C + 34k * sFacebook_C] + [36k * sTv_P + 56k * sOoh_P – 22k * sPrint_P + 45k * sFacebook_P + 530k * sCompetitor_P] + 201418.35

where sTv_C is the amount spent on Tv advertising in the current week normalized by the standard deviation of the Tv spends in the training data and sTv_P is the amount spent on Tv advertising in the previous week normalized in the same way.

Informative coefficients:

Revenue = [1.48 * Tv_C + 0.03 * Ooh_C + 2.26 * Print_C + 4.81 * Facebook_C] + [0.39 * Tv_P + 0.23 * Ooh_P – 1.18 * Print_P + 4.78 * Facebook_P + 0.26 * Competitor_P] + 201418.35

where Tv_C is the amount spent on Tv advertising in the current week and Tv_P is the amount spent on Tv advertising in the previous week.

Assuming an equal amount spent on each channel for the two weeks the following is obtained.

Normal coefficients:

Revenue = 164k * sTv_C + 64k * sOoh_C + 66k * sPrint_C + 79k * sFacebook_C + 530k * sCompetitor_P + 201418.35

Informative coefficients:

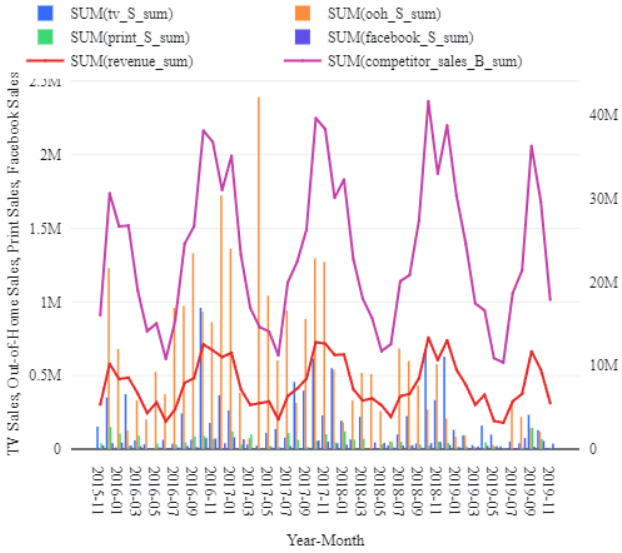
Revenue = 1.87 * Tv_C + 0.26 * Ooh_C + 1.08* Print_C + 9.59 * Facebook_C + 0.26 * Competitor_P + 201418.35

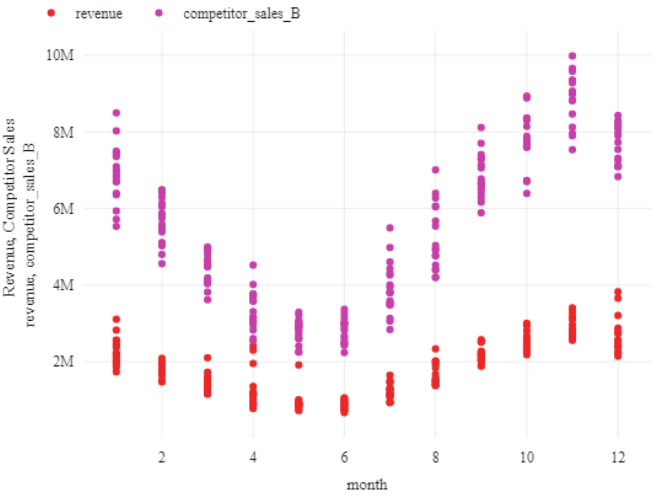
Which provide a clearer picture of each channels influence on revenue and their efficacy. It results that while Tv has been more influent, investing in Facebook advertising is more efficient. Furthermore, the Out of home channel, despite being the one in which the most investments have been done, is also the one with the worst efficacy.

However, it is critical to remember that when making new investments, if they are too different from the ones in the training data then the model might lose accuracy and its evaluations might become meaningless. Therefore, it is always good practice keeping in mind the characteristics of the training data for a correct use of the machine learning models.

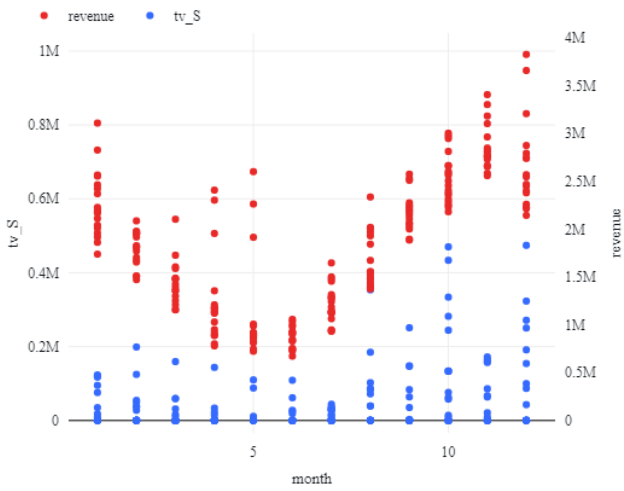
Seasonal trend

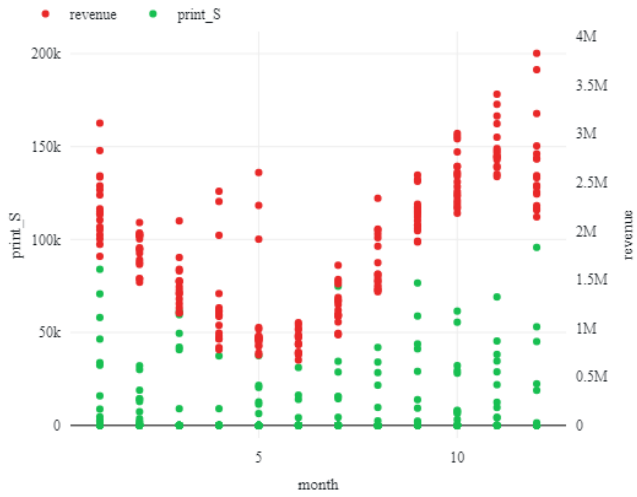
The seasonal trend has been observed in the early stages of the analysis and is clearly visible in the following plots.





As mentioned above, the trend interests both revenue and competitor sales. Although it can also be seen as a consequence of the distribution of advertising spends across the year, as demonstrated by the following plots.





In order to train an unbiased model, it has been made the decision to not include months as a feature.

External influences

It has been observed that there might be some kind of market trend influencing the revenue. This deduction comes from the analysis of the performances of two models: one trained using previous week competitor sales data, and the other one using the previous week revenue data (alongside the advertising spends in both cases). The results show that the former achieves better predictive performance with a margin of 0.12 in the R-squared score over the latter. This significant drop indicates that the competitor sales feature, containing information less correlated to our spends, can instead be useful in providing clues related to external factors.

Synergy Effect

To evaluate the presence of a synergy effect between couples of features, three different models for each combination have been trained:

1. The first has 3 features: two advertising channels and their product.
2. The second has 2 features: two advertising channels.
3. The third has 1 feature: the product of two advertising channels.

The evaluation will follow this logic: if the models with 2 and 3 features have the same performance (measured with R2 score) then it means that there is no performance gain in adding a synergy feature and so it will be concluded that there is no synergy effect with the specific combination of advertising channels. The third model has been used to further validate the results. Note that for this analysis features not related to the advertising channel spends have not been used (e.g. 'competitor sales').

Here the results produced are reported in a table.

Channel couples	R-Squared score		
	3 Feat. M.	2 Feat. M.	1 Feat. M.
Tv / Print	0.15	0.16	0.05
Tv / Facebook	0.16	0.18	0.11
Tv / Out of home	0.09	0.10	0.02
Print / Facebook	0.11	0.12	0.09
Print / Out of home	0.02	0.05	0.01
Facebook / Out of home	0.09	0.07	0.04

By looking at the R2 scores it can be deduced that no combination of advertising channels results in a synergy effect (the models with 2 features and their respective models with 3 features counterparts have the similar performance). The 1 feature models performance highlight that the synergy feature alone is not enough to achieve the same effectiveness of the other models, further validating the first observation.

Saturation Effect

The analysis of the saturation effect followed a logic similar to the one used for the synergy effect. Two models has been built:

1. The first one with 2 features for each adv channel: ch_S and sqrt(ch_S), where ch_S is the advertising channel spend.
2. The second one with 1 feature for each adv channel: sqrt(ch_S).

Furthermore, previously built model (baseline, R-squared: 0.23) with only 'ch_S' features has also been used for comparisons. Also for this analysis features not related to the adv channel spends have not been used (e.g. 'previous revenues').

The results show that the first model achieves an R-squared score of 0.20, worse than the baseline by a margin of 0.03. While the second one achieves a score of 0.24, improving over the baseline by a small margin of 0.01. The drop in performance of the first model might be due to collinearity. The improvement of the latter, instead, can be seen as a sign of small saturation effects, but not enough to be considered relevant.

Carryover Effect

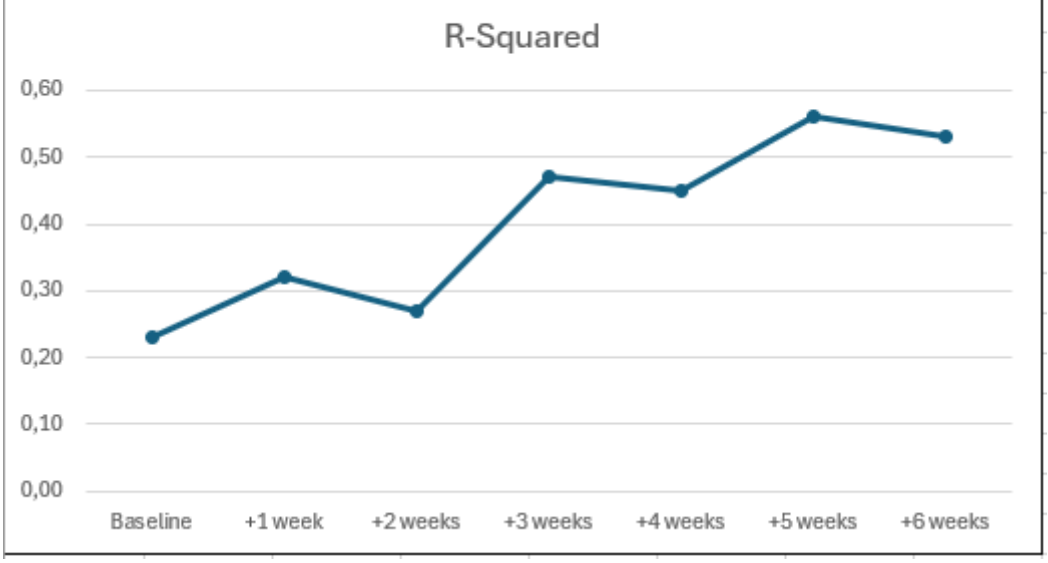
The analysis of the carryover effect also followed a logic similar to the one previously used. Multiple models have been built. Each model will have features from increasingly more weeks before the current one:

- The first one with 2 features for each adv channel: ch_S and 'ch_S_shifted', where ch_S is the advertising channel current spend and ch_S_shifted is the advertising channel previous week spend.
- The second one with 3 features for each advertising channel, and so on until no improvements are observed anymore.

Furthermore, like in the saturation case, the baseline model (R-squared: 0.23) with only 'ch_S' features has also been used for comparisons. Features not related to the adv channel spends have not been used (e.g. 'previous competitor sales).

The results are shown in this plot.

	Baseline	+1 week	+2 weeks	+3 weeks	+4 weeks	+5 weeks	+6 weeks
R-Squared	0,23	0,32	0,27	0,47	0,45	0,56	0,53



The R-squared scores increase as more weeks are added, clearly signalling the presence of carryover effects. However, when training a complete model, including the competitor sales variable among its features, only the latest two weeks result to be relevant (the R-squared score is 0.78 with two weeks data and 0.72 with 3 weeks data). Therefore, it can be concluded that while the carryover effect is present, it is mostly relevant in the last 2 weeks. Note that this analysis did not evaluate channels individually.

Further details on the methodology

For the sake of transparency, in this section a more detailed description of the methodology applied is reported.

Data Loading

In this initial step, the data provided in csv files is imported in Databricks. This phase ensures that all data is consolidated, accessible, and undergoes preliminary checks for integrity and completeness, setting a solid foundation for subsequent analyses.

Data visualization

Data exploration begins with data visualization and understanding, serving as the foundation for deriving insights. The visualization of the dataset helps in guiding the first steps of the data exploration by providing initial clues of what needs to be analysed further.

Some early observations have produced what follows. The dataset contains 208 total rows. The samples consist in weekly updates starting from '23-11-2015' and ending with '11-11-2019'. From a first look at the data there seemed to be many zero values in the columns relative to the spends. There is a boolean event column which also contains many zeros and it seems that no scheme has been used in the past to coordinate spends across the channels, this is helpful for building less biased models.

Checking data types

Making sure that data types are consistent with the variables ensures robustness and reliability for the following analysis. No issues with data types where found.

Summary statistics

Similarly to the first step of dataset visualization, few simple metrics like means, standard deviations, min and max values, number of missing values and values distribution help in finding what needs to be analysed further or cleaned during the data preparation.

summary	_c0	revenue	tv_S	ooh_S	print_S	facebook_I	competitor_sales_B	facebook_S	events_event2
count	208	208	208	208	208	208	208	208	208
mean	103,5	1822142,772	44531,07365	129653,8231	11185,89679	24460244,99	5538024,904	6436,973479	0,004807692
stddev	60,1885925	716228,6067	85675,09964	251974,28	19449,2107	35097382,35	2077191,971	9481,09141	0,069337525
min	0	672250	0	0	0	0	2240235	0	0
25%	51	1160205	0	0	0	0	3588301	0	0
50%	103	1862458,333	0	0	0	0	5535394	0	0
75%	155	2377706,667	54362,88	149873,6	14290,66667	41147695,82	7311071	10869,0367	0
max	207	3827520	474139,72	1501084,8	95766,93333	178298272,9	9984742	46201,17422	1

In this phase the following has been observed. The mean and standard deviation values seemed to be reasonable for all variables except for the 'event' one. Although, the first and second quartile underline a potential issue with zero values. Overall, the min and max values indicate that there are no out-of-range values. It could also be observed that a lot more money has been spent on average on 'Out of Home' advertising, while 'TV', 'Print' and 'Facebook' follow with lower amounts.

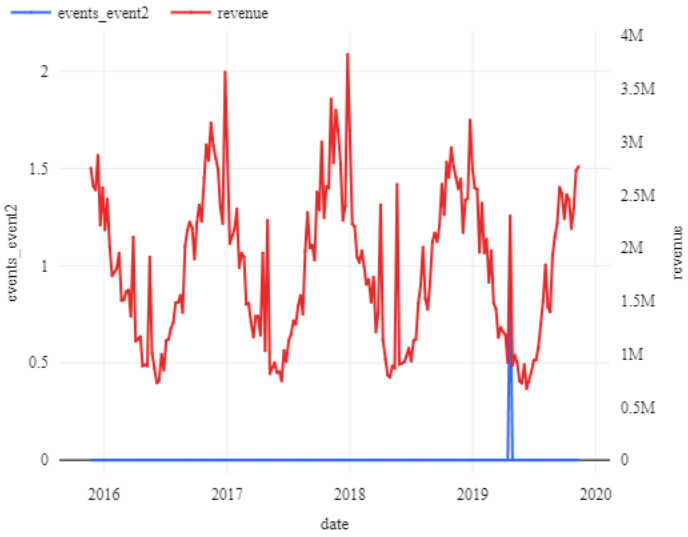
There were no missing values but high percentages of zero values in many columns. The amount of zero values seemed to be similar across the 'Spend' columns and the 'Impressions' one. Since in some weeks the spends for a certain type of adverting could be

zero these values seemed reasonable and have not been considered as missing values. The 'event' column was highly anomalous. Another detail, that needed further investigation, involves the 'Facebook' columns not sharing the same amount of zero values.

Feature operations

Before continuing with further visualization some operations could already be carried out, simplifying the following steps and eliminating unnecessary data.

The column '_c0' simply contained the row indexes of the dataset, therefore was dropped early. While a closer look at the ‘event’ column highlighted the presence of only one non-zero value which was considered irrelevant.



The only non-zero value in the 'events_event2' column coincides with a peak in the revenues. However, there are several other peaks consistently repeated throughout the years that do not have a corresponding non-zero value in the column.

Suspicious zero value

Inspecting the relation between the 'Facebook_S' and the 'Facebook_I' columns has highlighted an inconsistency. Since the Facebook Impressions probably come from a secondary source w.r.t. the sales amount, it can be assumed that, only in this row, the zero value in the 'Facebook_S' column is actually a missing value (Missing Completely At Random).

Even if there is only one such case in the whole dataset, since the amount of available data is relatively small this value was later be replaced appropriately. Furthermore, in this specific case, keeping all the samples is important and is beneficial during the analysis because of the precise weekly cadence of the data.

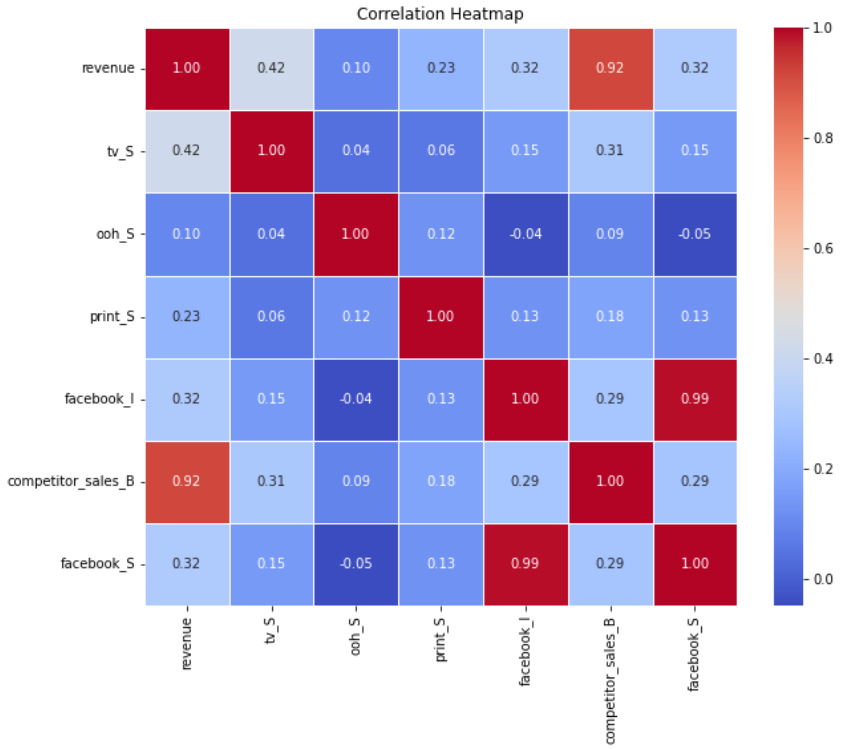
Advanced visualizations

More advanced visualizations allow for the representation of complex datasets in a comprehensible manner, facilitating pattern recognition and anomaly detection. Various

techniques, including scatter plots, histograms, and heatmaps, offer different perspectives on the data, enabling analysts to explore relationships, distributions, and trends effectively. Understanding the data through visualization involves uncovering hidden patterns, outliers, and correlations, which may not be apparent in raw datasets.

Correlations

The correlation matrix is a fundamental tool in data analysis, providing insight into the relationships between variables within a dataset. Its importance lies in its ability to quantify the strength and direction of associations between variables, aiding in identifying patterns, dependencies, and potential causal relationships. This knowledge is invaluable for decision-making processes, such as feature selection in machine learning models, identifying multicollinearity in regression analysis, or understanding the interplay between different factors in complex systems.



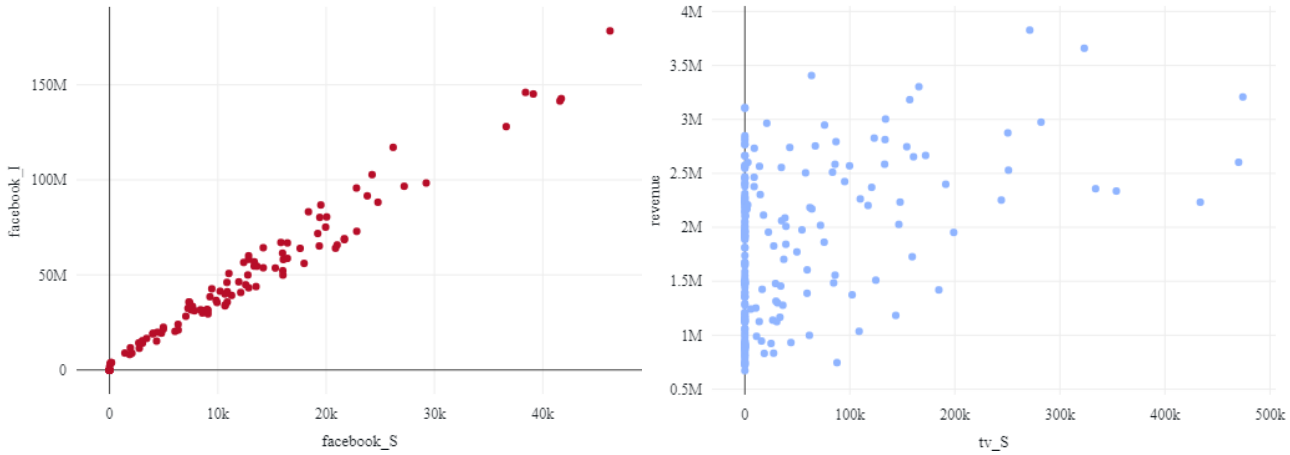
The correlation matrix underlines two strong linear correlations:

- the first one is between the two 'facebook' columns.
- the second one is between the 'competitor sales' and the 'revenue'.

The first is a clear indication of the fact that the facebook impressions directly depend on the facebook spend. This relation can be exploited to replace the missing value mentioned earlier. The second one might indicate that the revenues follow some general market trend and thus oscillate together with the competitor sales. When selecting the features to eventually build the models these strong correlations must be considered. Regarding the correlation strenghts of the spends with the revenue, they range from intermediate to low in the following order: 'TV', 'Facebook', 'Print' and 'Out of Home'.

Scatterplots

To find possible non-linear relations between the variables, scatterplots are also needed. They provide a clear and intuitive representation of how changes in one variable correspond to changes in another. Two examples are reported here.



No non-linear correlations were highlighted by these graphs.

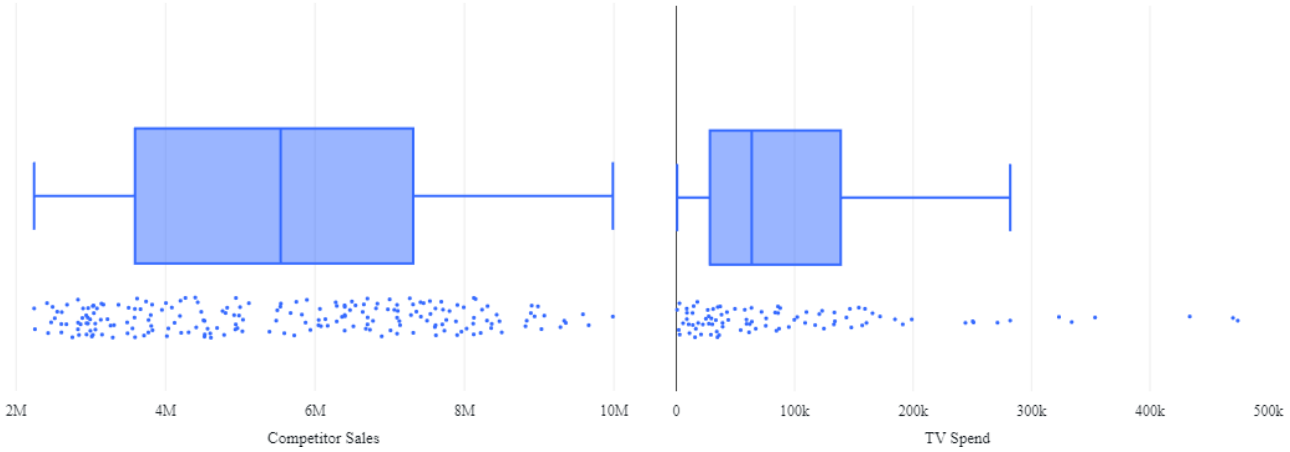
Missing value replacement

Since a strong linear correlation between the two 'Facebook' variables has been observed, using a linear regressor trained with feature 'Facebook_I' can be good to predict the missing value in 'Facebook_S' earlier discussed. The precision of the trained model is obviously high, and it can be used to predict a new value needed to replace the suspicious zero value.

Box-plots and outliers removal

Box plots are essential tools in data analysis for visualizing the distribution and variability of a dataset. They provide a concise summary of key statistical measures, including the median, quartiles, and potential outliers. Box plots offer a clear representation of the central tendency and spread of the data, enabling quick comparisons between different groups or categories. By displaying the range and variability of the data in a visually intuitive manner, box plots facilitate the identification of skewness, asymmetry, and the presence of extreme values. This makes them invaluable for identifying potential discrepancies within the data.

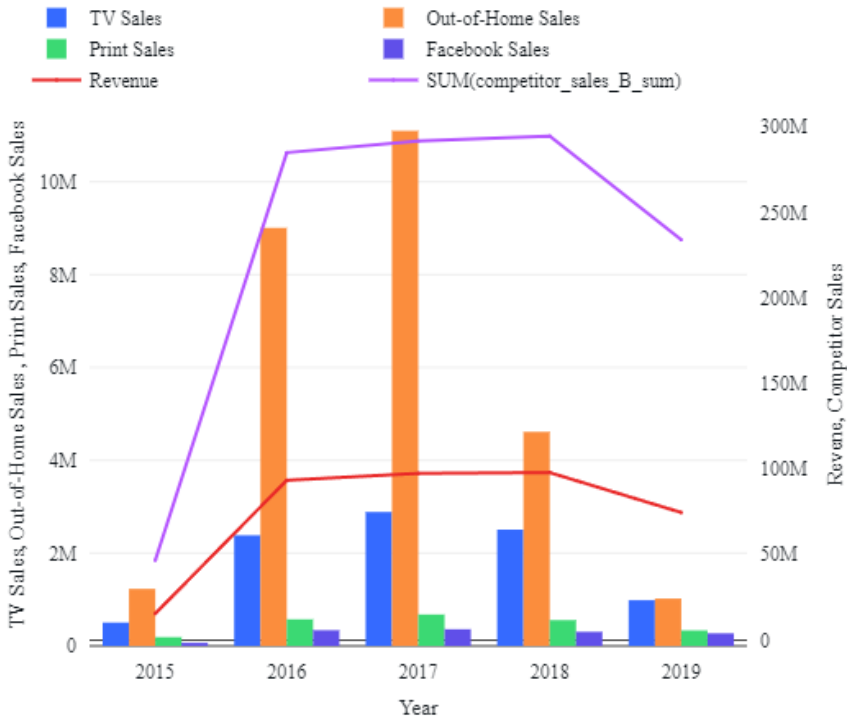
The presence of many zero values in the data could negatively impact this part of the analysis, resulting in too many data points being signalled as outliers. To address this issue all samples containing zero values have been systematically removed during this phase. Here some examples are reported.



For all the 'Spend' variables very few values fall in the outlier range (and even those data points appear reasonable). Consistently with the policy used earlier in this analysis all values were kept (the temporal regularity of data is highly valued). This also prevents from further reducing the amount of available data to train the models. In the 'Competitor sales' and 'Revenue' columns no outliers were observed.

Yearly based visualization

Bar plots are essential visualizations in data analysis, providing a straightforward way to compare group-level summaries. They make it easy to visualize and interpret differences between categories or groups. Bar plots are particularly useful for displaying frequency distributions, proportions, or aggregated metrics across different categories. They offer a clear and intuitive way to identify patterns, trends, and disparities within the data. In this case, Bar plots have been used to visualize temporal aspects of data. In the first plot data is aggregated year by year (aggregation is performed by means of a sum).

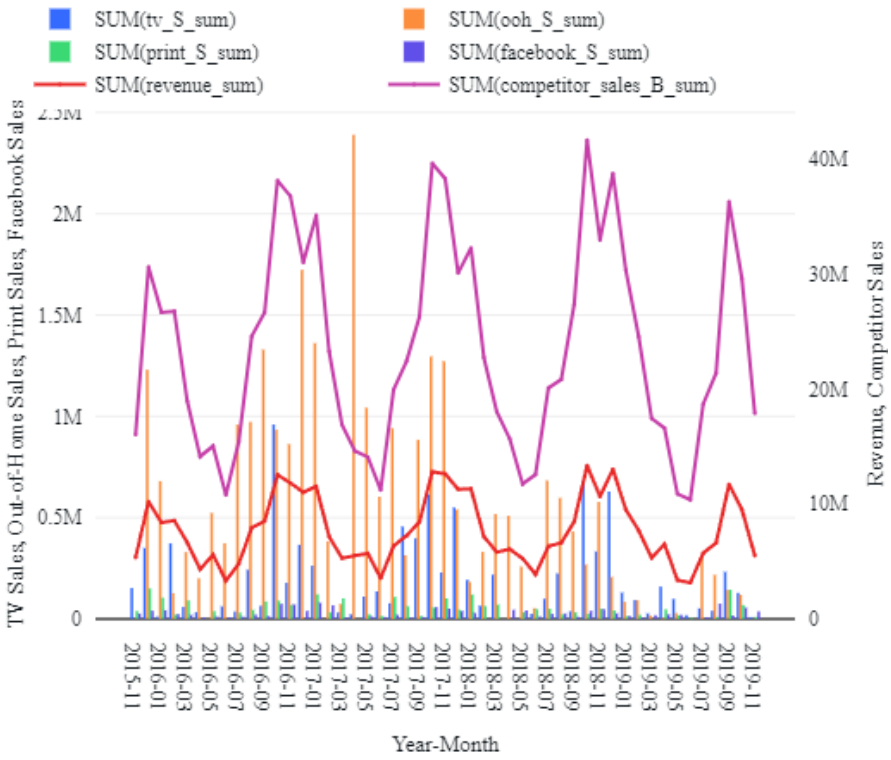


A few insights emerge:

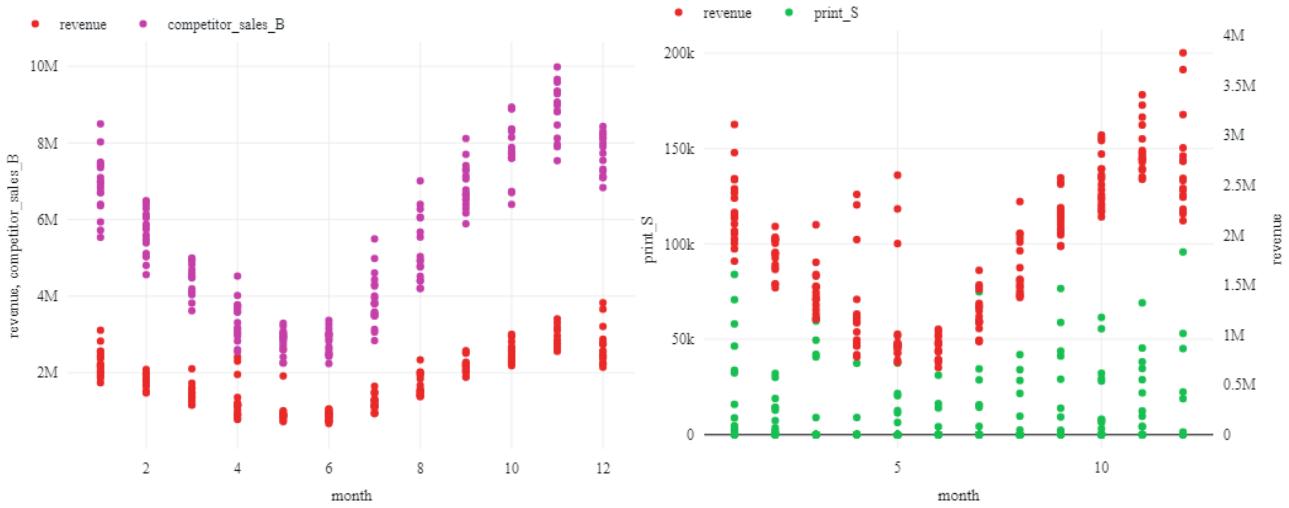
- the year 2015 has less expenses but this is due to the fact that only two months are present in the data.
- the year 2019, on average has less expenses (only one missing month isn't enough to justify this decrease).
- in the last year revenues decrease and this trend is also observed for the competitor sales.

Monthly based visualization

In the second plot data is aggregated month by month (aggregation is performed by means of a sum).



From this visualization, a seasonal trend emerged both for revenues and competitor sales. From December to June the revenues decrease and from June to December they increase. The trend holds true for all the years considered. To explore further this phenomenon more plots have been traced. Here some examples are reported.



It seemed that the revenue trend might be caused by an uneven distribution of the spends across the different periods of the year. Therefore, months have not been used as a feature to avoid bias in the models.

Feature engineering

Feature engineering is a pivotal process in machine learning and data analysis, involving the creation or transformation of variables (features) to enhance model performance or extract meaningful insights from the data. It plays a crucial role in improving model accuracy, interpretability, and generalization by capturing relevant information and reducing noise in the dataset. Feature engineering encompasses a range of techniques, including encoding categorical variables, scaling numerical features, creating interaction terms, handling missing values, and generating new features through mathematical transformations or domain knowledge. Through careful feature selection and transformation, hidden patterns can be uncovered, overfitting reduced, and model robustness improved.

Feature selection

At this point two columns of the original dataset (Index column and Event column) had already been dropped. Since the two 'Facebook' features are strongly correlated, 'facebook_I' has been dropped too. In this case keeping 'facebook_S' is undeniably better because it is directly comparable with the other 'spend' features, therefore more useful to the Marketing team.

2 weeks granularity and 1 month granularity

In addition to the weekly dataset, a second aggregation (by means of a sum) with two weeks long intervals and a third aggregation with one-month long intervals have been introduced for the experiments.

Addressing zero values issue

The high amount of zero values in the data can negatively impact the models' performance. One way experimented to address this issue involved using additional columns containing boolean values that match the presence of zeros in the spends.

Model Development and Evaluation

The focus of the analysis is to understand the contribution of each individual advertising channel to sales, particularly when expenditures are made across all channels. Specifically, the analysis should quantify the increase in sales for every dollar spent on advertising within each channel. Additionally, it is crucial to determine the accuracy with which these sales increases can be predicted. To meet the analysis objective, the predictive model chosen is the linear regressor.

Linear regression models

A linear regressor is a statistical model used to predict a dependent variable based on one or more independent variables, assuming a linear relationship between them. Characteristics of linear regressors include simplicity, interpretability and quick computation. They are commonly used for forecasting, trend analysis, and determining the strength of predictors. However, they are limited and prone to underfitting.

Preliminary operations

A test set has been extracted from the data and 5-fold cross-validations have been used to validate the feature choices and to carry out a less biased analysis. Important notes:

- The first 4 rows of the dataframe have always been removed to make the all the models directly comparable.
- In the Model development and evaluation and Model tuning phases, models are evaluated on the train + validation set using 5-fold cross-validation. This avoids biases and ensures a correct test evaluation for the final model.

Baseline

A first model has been built on the 'weekly' dataframe using 'tv_S', 'ooh_S', 'print_S' and 'facebook_S' as features. This model will be considered the baseline for most of the following tests. The R2 score achieved is 0.23. While the accuracy is too low for a marketing analysis, this model can still be valuable for comparisons with other models.

Competitor sales

The 'competitor_sales' feature has also been used. Using its values as-is wouldn't make sense because the sales are only known at the end of the week, therefore, they cannot be used to predict the same weeks' revenue. For this reason, this column has been shifted by one week, meaning that the competitor sales from the previous week has been used for predicting the revenue of the current week.

The newly added feature greatly improves the model performance (R2 increases from 0.23 to 0.76). The high correlation with the target feature is certainly helping. This results in a more reliable model that better distributes importance to the different features.

Past revenues

Given the high correlation between the revenue and the competitor sales, it can be expected a similar result by using a 'shifted revenue' feature instead of the 'shifted competitor sales' one. Note that this means using the information about the past week revenue to predict the current week revenue. As expected also in the case there's a performance improvement w.r.t. the baseline but worse performance w.r.t. the 'competitor sales' model: R2 decreases from 0.76 to 0.64.

Boolean values

Adding these features seems to not improve the performance, on the contrary, it slightly reduces it (R2 decreases from 0.76 to 0.74).

2 weeks granularity (with competitor sales)

To better capture the past data information an aggregation of data in 2 weeks long intervals have been taken into consideration. The features are the sum of each channel’s last two weeks spends. Note that the label is the revenue of only the current week. This new model slightly improves the performance (R2 increases from 0.76 to 0.77) over the weekly one, indicating that past data might be relevant.

Monthly granularity (with competitor sales)

Another experiment has been carryed out using monthly aggregated features. This model has worse performance (R2 decreases from 0.76 to 0.70), in contrast with the 2-weeks aggregation findings. This might indicate that aggregation by sum is not effective on longer time periods.

Synergy, Saturation and carryover Effects

These effects have already been discussed sufficiently in previous sections.

Model Tuning

Model tuning is a critical step in machine learning model development aimed at fine-tuning model parameters to improve performance and generalization. It involves systematically searching through a defined hyperparameter space to identify the combination that achieves the best trade-off between bias and variance. Through iterative experimentation and cross-validation, hyperparameters that yield the best performance on unseen data can be identified. Overall, model tuning impacts model accuracy, robustness, and scalability, ultimately leading to more reliable and effective predictive models.

First model

The first parameter selection included data from the current week using 'sqrt features' and data from the previous week using 'normal features'. 'Competitor sales' has also been used. This model achieved an R-squared score of 0.78.

Second model

The second model is similar but only uses 'normal features' and its performance are slightly lower than the previous case (R-squared decreases from 0.78 to 0.77).

Third model

The third model only uses 'sqrt features' and also in this case the performance are slightly lower than the best case (R-squared decreases from 0.78 to 0.77).

Fourth model

In the fourth model adding more weeks has been experimented but it seems to decrease the performance (R2 from 0.78 to 0.72).

Final model

The last phase of feature selection and engineering has been guided by all the information acquired during the analysis. Given all these observations, a final model, with a structure like the one used for the first model of the tuning phase, has been trained on the training + validation sets and tested on the test set (parameter selection included data from the current week using 'sqrt features' and data from the previous week using 'normal features'. 'Competitor sales' has also been used).

The model achieves a R-squared score of 0.94 and a Root Mean Squared Error of 173289. These performances guarantee the reliability of the coefficients in providing insights useful for the marketing team.

Also, a second model, slightly less performant has been trained to obtain an even more interpretable result (using ‘normal features’ for both current and past data). The performance decrease is negligible (R2 from 0.942 to 0.935).