Machine Learning for counterfactual estimation and forecasting

Carlo Ardito, Martina Manno, Olimpia Sannucci

May 2022

Contents

1 Introduction

The following business case was assigned to Data Science and Management students for the Machine Learning project by Deloitte, as the leading global provider of audit and assurance, consulting, financial advisory, risk advisory, tax, and related services. The challenge presented is a real-life challenge that a data scientist may face every day. Therefore, the approach adopted to achieve the goal is realistic and practical and will take into account the economic theory, statistical models but also more recent algorithms used by the 'BIG' such as Amazon and Google.

1.1 Business case

The scope of this project is to detect the effect of an advertising campaign. In particular, the case refers to an e-commerce of perfumes that after running an advertising campaign wants to analyse the results. The advertising campaign started on 2019 December 3rd and ended on 2019 December 24th, so it was run only for few days, but it is fundamental to quantify the impact it has had in order to be able to launch a massive advertising campaign later on.
The starting point of case is a fact that is immediately apparent: there is a significant peak in website visits in the December 2019. Before jump to any conclusions, it is necessary to analyse the elements that are not so obvious making a more in-depth analysis focused on the historical trends and the company's behaviour in the last years. Hence, the main request of this project is to quantify the impact of the advertising campaign on sales by analyzing the website and online purchases performances.

1.2 Economic theory

To achieve the business case goal, it is not enough to consider that, ceteris paribus, if the advertising campaign has an effect, the outcome should increase. This approach would restrict and simplify the analysis not considering other influencing factors. Actually, the relationship between the advertising campaign and the sales could be useful to make prediction about the historical data but not about the causal impact of advertising resources on the outcome. Therefore, oversimplifying the case could wrongly attribute a causal-effect relationship between the advertising campaign and sales basing on an observed association or correlation between them. In economic classes, the approach to this problem is always supported by the logical statement: "correlation does not imply causation". It serves as a useful reminder of how to think about the relationship between two variables X and Y. In the specific case, if the advertising campaign and sales increasing seem to be linked, it's possible but not certain that the campaign caused the sales increasing. It's also possible that some third variable omitted that affected both at the same time, making the two variables move together.
Therefore, in order to conduct a valid analysis, the statistical causal inference model must be take into account. The critical step in any causal analysis is estimating the counterfactual, a prediction of what would have happened in the absence of the treatment. The basic approach consists of applies a treatment to some set of subjects and observes some outcomes. The outcomes for the treated subjects can be compared with the outcomes for the untreated subjects (the control group) to determine the causal effect of the treatment on the subjects. So the basic identity for causal inference is as follows:
Outcome for treated — Outcome for untreated = [Outcome for treated — Outcome for treated if not treated]+[Outcome for treated if not treated — Outcome for untreated] = Impact of treatment on treated + selection bias.
This shows that the critical concept for understanding causality is the comparison of the actual outcome (what happens to the treated) compared with the counterfactual (what would have happened if they had not been treated).

Another approach is to compare the outcome after running the advertising campaign to an estimate of the counterfactual, so what would have happened during the limited period without that increase in spend for the advertising campaign. The counterfactual estimation comes from a predictive model developed using data from before the experiment was run. The powerful techniques used in machine learning may be useful for developing better estimates of the counterfactual, potentially improving causal inference.
When interpreting causal analysis, the possible presence of confounding factors in the data must always be considered. In general, analyzing the success of an advertising campaign it is possible to consider two types of confounding factors, which are the participant variability and the seasonality. Seasonality is a characteristic of a time series in which the data experiences regular and predictable changes that recur every calendar year. Any predictable fluctuation or pattern that recurs or repeats over a one-year period is said to be seasonal. And this is the main difference with the cyclical effects, as seasonal cycles are observed within one calendar year, while cyclical effects, can span time periods shorter or longer than one calendar year. How to understand seasonality will be discussed more in detail in the next section.

2 Data Management

2.1 Data understanding

The datasets provided by Deloitte are 3. The first dataset, called *Web_traffic*, contains information about the daily website traffic divided by city. The cities in which the advertising campaign was run are Milan, Rome and Naples. Hence, for each city the dataset provides daily data about the number of visits to the e-commerce website.
The second dataset, called *Sales*, contains information about the daily sales divided by city. Then, for each city it provides daily data about two important indicators:
– the conversion rate which is computed as the fraction of website visitors that finalized an online purchase of the products,
– the average spend which refers to the average expenditure in EUR of visitors who finalized an online purchase.
The last dataset provided is called *Gsearchdata* and it gives information about the daily number of Google queries on company's product divided by city.
In order to understand the data available, plotting data can be useful.Analyzing the graphs below, the website visits result to have a positive trend over this one and a half years with two important peaks in the two period corresponding to December 2018 and December 2019. The first question to be answered is related to how is it possible that in December 2018, the e-commerce perceived an increase in website visits if they didn't run the advertising campaign in that year? And so the second question arises is about understand if the peak in December 2019 would have been there regardless of the advertising campaign. If so, whether the peak would be of equal or lesser magnitude? These are the main questions that will be answered in this report.

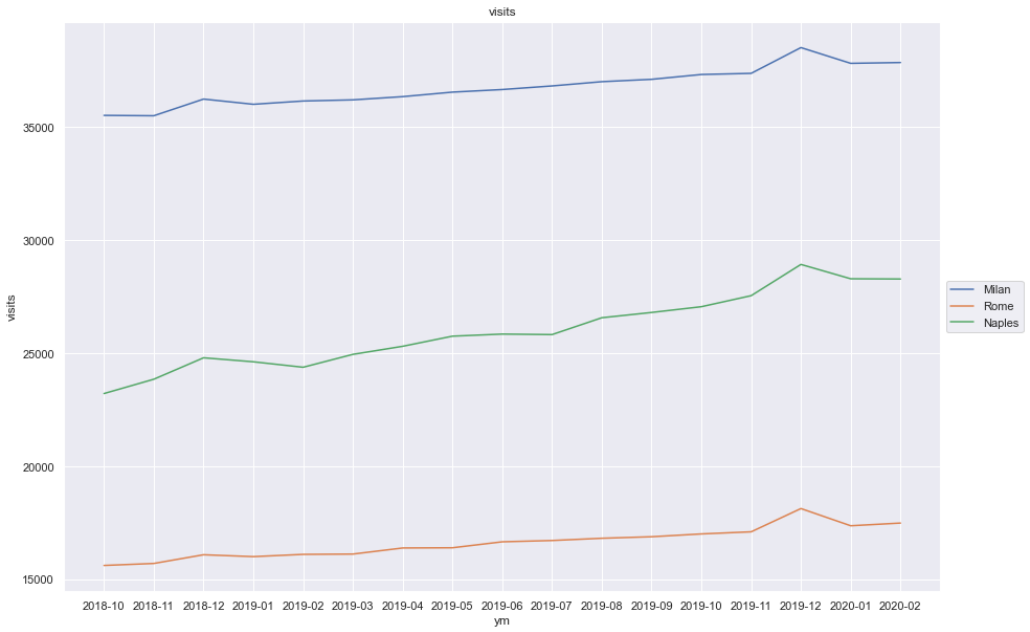

Figure 1: Website visits for each city

2.2 Data preparation

The first step that has been taken is to merge the different datasets basing on the date variable to have an overview of all data. Then, it is fundamental to obtain data about sales and online purchases hidden in the available data. Indeed, it is possible to get data about online purchases by multiply the conversion rate with the numbers of visits for each day and for each city. The other important missing data is the number of sales, that can be retrieve by multiply the number of online purchase with the average expenditure for each day and for each city. After these manipulation, the final dataset consists of 8 variables:

– daily date from 15 October 2018 to 13 February 2020, for a total of 487 days.
– city name which can be one of these: Milan, Rome or Naples.
– conversion rate
– average expenditure
– website visits
– sales
– online purchases
– Google queries

By looking graphically at the data, one can see that for each city both website visits, Google queries and sales have a positive trend with peaks in December in both years, although obviously the sales numbers are much larger than the visits numbers, which in turn are larger than the google queries numbers.

Figure 2: Sales for each city

It can be also useful to have a look at the correlation matrix which is a table showing correlation coefficients between variables. As expected, the matrix turns out that there is a strong positive correlation (0.92) between the visits and the sales because obviously many of the website visitors then finalise the purchase. And the same reasoning also applies to the Google query, which have a positive correlation (0.82) with sales. The matrix also shows the positive correlation (0.81) between the online purchases and the visits. This is probably related to the fact that the variable created of online purchases depends on visits. On the other hand, there is a negative correlations, between the average expenditure and the conversion rate (-0.63).

Figure 3: Correlation Matrix

At this stage, we decide to aggregate the data by date through the sum and therefore not considering the city of reference. This choice is motivated by the fact that the behaviour of the variables is the same in all cities and that the effect that actually needs to be detected is that of the advertising campaign in general. In addition, no specific data have been provided about the type of advertising campaign that was sent. Therefore, it is possible to assume that the resources spent, the target and the transmission channels are the same for all three cities.

2.3 Seasonal Decompose

In order to start the approach on a time series, the first step was to check if there was any seasonality or a specific trend for each of the given variables. The command *seasonal_decompose* has been used to achieve this task. The model was considered 'additive' because data can be considered as as the result of adding numbers, and this type of data tends to show a linear trend.

```
result = seasonal_decompose(df.tot_sales, model='additive', period = 12)
```

The output highlighted that all of the three variables considered (sales, visits and online purchases) follow a seasonality pattern. The figure below represents the case of sales. The command also shows if there is any type of trend in the series: both visits and sales show an increasing linear trend, while that does not apply to online purchases. Detecting seasonality is fundamental in order to proceed with the predictions.

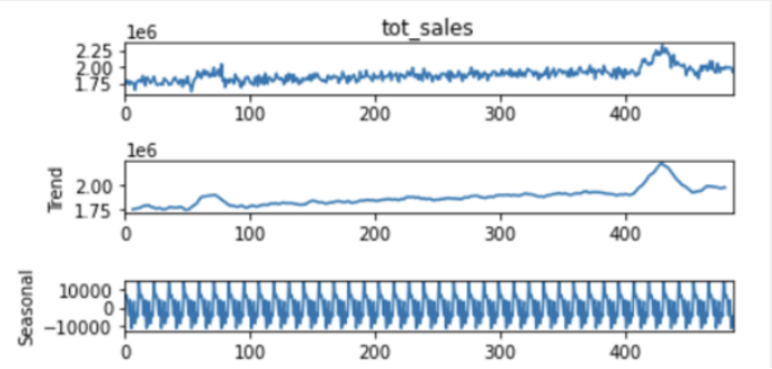

Figure 4: Seasonal Decompose on sales

2.4 Augmented Dickey-Fuller

Augmented Dickey Fuller test (ADF Test) is a common statistical test used to test whether a given Time series is stationary or not. It is one of the most commonly used statistical test when it comes to analyzing the stationary of a series.In time series forecasting, the first step is to determine the number of differentiating required to make the series stationary.The aim is to test the null hypothesis that states that the series is stationary.
The results show that, again, both visits and sales are non stationary (since we rejected the null hypothesis) while the online purchases are stationary (we accepted the null hypothesis).

```
from statsmodels.tsa.stattools import adfuller
test_result=adfuller(df1)
def adfuller_test(df1):
    result=adfuller(df1)
    labels = ['ADF Test Statistic','p-value','#Lags Used','Number of Observations Used']
    for value,label in zip(result,labels):
        print(label+' : '+str(value) )
```

```
    if result[1] <= 0.05:
        print("strong evidence against the null hypothesis(Ho), reject the null hypothesis.
    Data has no unit root and is stationary")
    else:
        print("weak evidence against null hypothesis, time series has a unit root, indicating
    it is non-stationary ")

adfuller_test(df1)
```

2.5 GARCH

At this point, we wanted to check if there were any significant months or days of the week among the time series. The model that used to achieve this task is the Garch model, which is a model used in time-series data when the variance error is believed to be serially correlated. This model also takes into account the moving average component.
Dummy variables for months and the days of the week were created to perform the analysis. Also, since a trend was detected by the AD Fuller test, differentiating was performed in order to stabilize the mean of the time series by removing changes in the level of a time series, and so eliminating (or reducing) trend and seasonality. Differentiating is performed by subtracting the previous observation from the current observation. In this way, a series of differences can be calculated.We ran this model on three variables: sales, visits and online purchases. The Garch analysis was performed on Gretl.
For the visits the significant month resulted to be December while for the days Monday, Tuesday, Thursday and Friday were negatively significant while there is a positive significant coefficient for Saturday For online purchases there was no significant month while it resulted significant in a negative way Monday; Saturday was not significant but had the most positive coefficient. For the variables sales, as you can see below, for what concerns the month, it did not give any significant values, but the greatest positive coefficient is December, while for the days of the week Monday as a significant.

	coefficiente	errore std.	z	p-value	
day_of_week_Fr~	-83,0387	35,1751	-2,361	0,0182	**
day_of_week_Mo~	-161,720	38,6292	-4,186	2,83e-05	***
day_of_week_Sa~	69,7873	36,7439	1,899	0,0575	*
day_of_week_Su~	552,668	37,0212	14,93	2,15e-050	***
day_of_week_Th~	-22,8286	40,1989	-0,5679	0,5701	
day_of_week_Tu~	-154,545	38,8788	-3,975	7,04e-05	***
day_of_week_We~	-48,2477	39,1856	-1,231	0,2182	

Figure 5: Garch results on sales

3 Modelling Approach

The models used to achieve the case goal are listed in this section. The approach for each model focuses on how it is designed for the specific case and what is its result.

3.1 Linear regression

The Multiple Linear Regression is the first model used to get an initial overview of the challenge. It is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables

and response (dependent) variables. Generally, it is always the first approach to this kind of problems because of it simple interpretation, but the trade-off between interpretation and accuracy must take into consideration. For this reason, this model will be used only to highlight the significant variables to be taken into account. Before performing the model, an additional feature is created. This is a dummy variable which indicates with 0 the period non affected by the campaign, and with 1 the opposite. The Linear model performing has sales as dependent variable and contains many regressors:
- visits,
- online purchases,
- Google queries,
- presence or not of the advertising campaign
- dummy variables about the days of the week and the months of the year. Then, applying the feature selection, the model was reduced with only the statistically significant variables:
- visits,
- presence or not of the campaign
- day of week = Sunday
- month of the year = December

```
Call:
lm(formula = sales ~ visits + Si.no + day_of_week_Sunday + month_December,
    data = df_final)

Residuals:
    Min      1Q  Median      3Q     Max
-610113 -182373  -32684  166064 1032498

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.025e+06  3.276e+05  -3.130 0.001857 **
visits             8.189e+01  4.141e+00  19.778  < 2e-16 ***
Si.no              2.513e+05  7.058e+04   3.561 0.000406 ***
day_of_week_Sunday -6.873e+04  3.199e+04  -2.149 0.032161 *
month_December      1.430e+05  4.065e+04   3.519 0.000475 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 245700 on 482 degrees of freedom
Multiple R-squared:  0.5891,    Adjusted R-squared:  0.5857
F-statistic: 172.8 on 4 and 482 DF,  p-value: < 2.2e-16
```

Figure 6: Linear Model results

As already state, the model has not a good accuracy and this is the reason why it is not used to make prediction but it reveals interesting insights into the variables and it will therefore be as a baseline for the next considerations.

3.2 Prophet Algorithm

In order to quantify the impact of the campaign, what would have happened without the campaign had to be predicted. To accomplish this task, the Prophet algorithm was used. The Prophet algorithm is an additive model,developed by Facebook, which means that it detects the following trend and seasonality from the data first, then combine them together to get the forecasted values. The model requires two variables: ds, which represent the data values, and Y, which the dependent variable. For the analysis, data regarding the period before the campaign were considered and for more accurate predictions the model was implemented with all the Italian festivities like Christmas, New Years, Easter...

```
model = Prophet(seasonality_mode= 'additive',interval_width= 0.95, weekly_seasonality= 'auto',
    yearly_seasonality=True,daily_seasonality='auto')
model.add_country_holidays(country_name='IT')
model.fit(dfp)
```

To quantify the impact of the campaign, the difference between the actual values and the forecasted ones was computed and as a result there is a difference of 3.8%, representing the impact. The same model was applied to the variable visits.

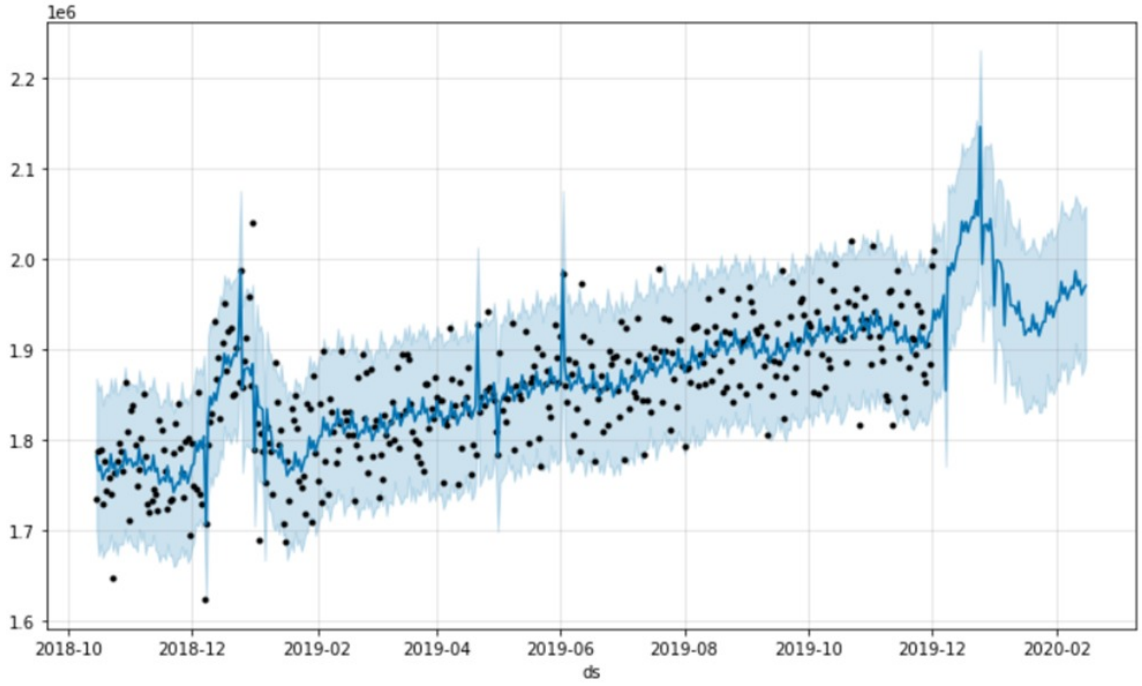

Figure 7: Prophet results

The additional feature of this model is the weekly decomposition. The results were very similar from what obtain during the Garch analysis: for what concerns the visits, the negative coefficient were Monday, Tuesday and Friday, and as a positive coefficient Saturday; in the image shown it is visible a decreasing trend from Monday through Friday. While for the sales in the Garch analysis the negative and positive coefficients were respectively Monday and Saturday, and looking at the Prophet results there is an increase between Saturday and Sunday and a decrease in the sales from Monday. The figure below shows the predicted values (blue line) and the actual values (orange line).

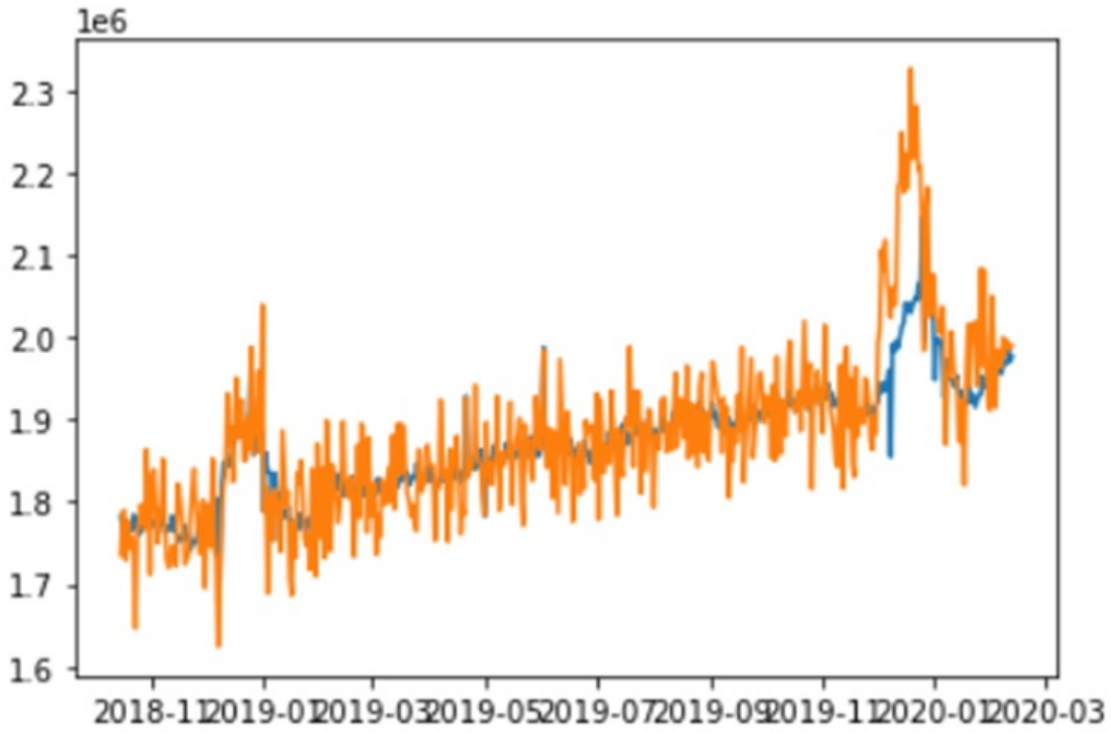

Figure 8: Predicted vs Actual

3.3 Causal Impact Algorithm

The Causal Impact is an algorithm built by Google for time series data in order to create a Bayesian structural time series model based on multiple control groups to estimate a series of baseline values for the time post-intervention. The causal impact analysis explains if the action taken had effects on an outcome metric. In our scenario, the action taken was running an advertising campaign and the goal is to detect if it had a statistically significant effect on visits, sales and online purchases. Basically, the scope is to understand by estimating the counterfactual if these three metrics had increased after the start of the campaign.
In the specific case, the model doesn't fit very well to the data because there are not control groups and there is no possibility to create control group with the Propensity Score Matching.
So, the first step was to obtain daily data from Google Trends about any other e-commerce of perfumes in order to use these information as control group. The e-commerce take into account are: Sephora and Douglas.
The second step is to split the data in periods: before and after the advertising campaign. The model takes into consideration the visits and sales variables as variables affected by the advertising campaign while the Google Trends variables as non affected. Another variable is enter in the model: the Google queries about our e-commerce. About this variable, there is not much information butt could be possible to make some assumptions. Since when looking at the data the google queries are significantly lower than the website visit, this could be due to the fact that if the advertising campaign was broadcast via social media with a direct link to the site, many visitors probably did not go through Google searches. Therefore, one could assume that this variable is not directly affected by the campaign. Indeed, by performing the Causal Impact the result is the same if the model contains or not the Google queries variable.As shown in the figure below, these are the results of the Causal Impact:
- In absolute numbers, during the post-intervention period, the response variable had an average value of

approx. 2.04M. By contrast, in the absence of an intervention, we would have expected an average response of 1.99M. The 95% interval of this counterfactual prediction is [1.97M, 2.01M]. Subtracting this prediction from the observed response yields an estimate of the causal effect the intervention had on the response variable. This effect is 0.05M with a 95% interval of [0.03M, 0.07M]. Summing up the individual data points during the post-intervention period, the response variable had an overall value of 148.86M. By contrast, had the intervention not taken place, we would have expected a sum of 145.05M. The 95% interval of this prediction is [143.46M, 146.58M].
- In relative terms, the response variable showed an increase of +3%. The 95% interval of this percentage is [+2%, +4%].

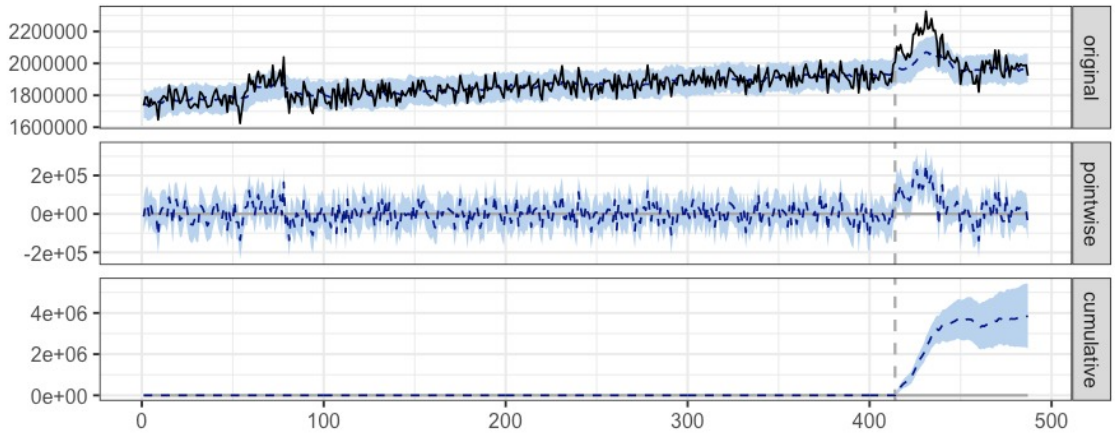

Figure 9: Causal Impact results

This means that the positive effect observed during the intervention period is statistically significant and unlikely to be due to random fluctuations. It should be noted, however, that the question of whether this increase also bears substantive significance can only be answered by comparing the absolute effect (0.05M) to the original goal of the underlying intervention. However, the probability of obtaining this effect by chance is very small (Bayesian one-sided tail-area probability p = 0.001). This means the causal effect can be considered statistically significant.

4 Results

From each model, some interesting results were extrapolated. Starting with the Linear Model, this indicated which variables were statistically significant: visits, the presence or not of the advertising campaign, and the month of December. While the two models, Prophet and Causal Impact, confirmed the same result: the campaign had a positive effect on sales. As far as Prophet is concerned, it seems that the advertising campaign led to a 3.8% increase in sales. On the other hand, the Causal Impact Algorithm shows a 3% increase on sales with a confidence interval [2%,4%].

5 Conclusions

The model's results are precise but not enough to make a better and conscious decision for future advertising investments.
Indeed we asked ourselves how to enrich this analysis to provide more insightful pieces of information for future investment analysis, and we came up with these conclusions.
Understanding the amount of capital invested in the campaign could be an opportunity to calculate the Return On Investment (ROI) and improve other advertising decisions.
Both prediction models, as explained, are training on data not affected by the campaign, so acquiring more historical data will improve the precision of these models.
Beyond those calculations, we thought that acquiring qualitative data about the advertising campaign (channels, buyer persona, tone of voice, etc.) could give us a bigger picture to make future decisions.

The project was carried out on GitHub, so that each team member could share their progress with the others in real time. This is the link to access all our work: https://github.com/martinamanno/DeloitteXLuiss .

6 References

– Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani "An Introduction to Statistical Learning with Applications in R", Springer.
– Hal R. Varian, "Causal inference in economics and marketing".
– James H. Stock, Mark W. Watson, " Introduzione all'econometria",Pearson.
– Roger A. Kerin, Steven W. Hartley, Luca Pellegrini, Francesco Massara, Daniela Corsaro "Marketing", Mc Graw Hill.
– Prophet Algorithm: https://facebook.github.io/prophet/
– Causal Impact Algorithm: https://google.github.io/CausalImpact/CausalImpact.html

7 Appendix

Each team member has equally contributed in assessing the problem and evaluating the possible solutions.The workload was evenly divided among team members and personal skills have been leveraged in order to reach the objective. Specifically, Olimpia Sannucci has contributed in the analysis of the seasonality (implemented with the Garch Model) and the application of the Prophet Model; Martina Manno worked on the data visualization and pre-processing analysis. Within the R environment, Martina Manno performed the Causal Impact model with the support of Carlo Ardito, who helped transform the Google Trends data. Overall, Carlo Ardito has contributed to the general assessment of the problem and the implementation of the solution within a business model.