Deloitte Advertising

Di Martino Giulia, Bozzi Martina, Di Razza Claudia, Pomanti Allegra

April 2022

Contents

1 Introduction 2

2 Methods 2

2.1 Exploratory Data Analysis 2

2.2 Data visualization 3

3 Experimental design 3

3.1 Univariate time series 4

3.2 Vector Autoregression Model 4

3.2.1 ADF test and differencing 4

3.2.2 Lag order identification and fit of the model with VAR 5

3.2.3 Granger Causality Test and Impulse response function 5

3.2.4 Variance Decomposition 6

3.2.5 Model validation 6

4 Code description of the multivariate time series 6

5 Results 8

6 Conclusions 8

6.0.1 Possible Future Strategies 8

7 Bibliography 9

8 Appendix 10

8.1 Exploratory Data Analysis 10

8.2 Data Visualization 10

8.3 Univariate time series 10

8.4 Vector Autoregression Model 10

8.5 ADF test and differencing 10

8.6 Granger Causality Test and Impulse Response Function 11

8.7 Variance Decomposition: graphs for Milan, Naples and Rome 11

8.8 Model validation: graphs for Milan, Naples and Rome 11

8.9 Peer self-evaluation 11

1 Introduction

Data regarding the behaviour of customers on the e-commerce website of the company Alpha has been collected across three different cities: Milan, Rome, Naples. The latter are considered as broadly representative of the Italian market in the three main geographical areas (respectively North, Center, South).

The objective of the analysis is to build a significant model which will determine, as a result, the impact (causal effect) of the advertising campaign launched by company Alpha on perfume's sales. In detail, the advertisement was conducted from December 3rd 2019 to December 24th of the same year.

2 Methods

2.1 Exploratory Data Analysis

The dataset used in the analysis has been obtained by merging the two assigned datasets: ”Sales”, and ”Web Traffic”. The resulting dataframe is composed of 1461 observations, and 7 variables (1 categorical, and 6 numeric, the target variable is included).

The variables are: **city** (describes the three cities in which the data of customers has been extracted. Those are: Milan, Rome, Naples) **day** (represents the date for each customer activity on the website. This variable ranges from the 15th of October 2018 to the 13th of February 2020); **visits** (number of visits to the ecommerce website); **conversion rate** (ecommerce conversion rate, computed as the fraction of website visitors that finalized an online purchase of the products); **average spend** (average expenditure (EUR) of visitors who finalized an online purchase); **sales** (identifies the sales finalized by customers and has been computed by multiplying visits with conversion rate); **total spending** (indicates the total spending (EUR) of customers, and has been derived by multiplying average spend with sales).

The second part of the analysis focused on data manipulation with the aim of maximising the performance of the data for later use. For this purpose, the two original datasets ”Sales” and ”Web Traffic” have been merged. The merged dataset was then used to compute two key features in the analysis : the variables ”sales” and ”total spending”. Specifically, the variable ”sales” has been computed by multiplying visits with conversion rate (since conversion rate is a ratio of visits that turn into a purchase). Whereas, the variable ”total spending” has been computed by multiplying average spend with the number of sales. Subsequently, in order to get a clearer insight of possible seasonality or trend in the data, a function to transform the daily data into weekly data was specified. This process resulted in a new dataframe containing the weekly data that ranges from the 1st of December 2018 to the 31st of January 2020 of the following variables: sales, conversion rate and website visits. A graph showing the sales (Appendix 8.1) in the three different cities, during the above cited period was then displayed: the graph shows that in December 2018 the sales were comparatively high in all the areas with a subsequent fall in January of 2019. Interestingly, December 2019 shows the same pattern, but with overall higher sales. Moreover, the norther region registered the highest number of sales, whereas the central region the lowest. Based on the EDA two models were employed throughout the analysis: a ”Univariate time series model” to detect seasonality, and a ”Vector Autoregression Model” to determine the impact of the advertising campaign on perfume's sales.

2.2 Data visualization

A Data visualization analysis has been performed in order to obtain deeper insights of the dataset. In fact, comparing the number of visits to the ecommerce website in the three regional areas during 2018 and 2019 (Appendix 8.2, graph 1-2) led to interesting considerations: the visits experienced an increase from 2018 to 2019 in all the cities. In detail, the city with the highest number of visits to the website both in 2018 and 2019 is Milan. Contrary, the city with the lowest number of visits throughout both the years is Rome.

Moreover, the graphical representation of the conversion rate pointed out some differences in the customers' purchase behaviour across the three cities throughout 2018 and 2019. In 2018 (Appendix 8.2, graph 3), the website visitors that finalized an higher number of online purchases are located in Rome; whereas the lower conversion rate is registered in Naples. In 2019 (Appendix 8.2, graph 4), the highest conversion rate is equal across the three cities; whereas the website visitors that finalized the lowest number of online purchases are located in Naples. Indeed, it can be deduced from the graphs that from 2018 to 2019 the purchase behaviour of the customers changed across the three cities, but Naples remained the one with the lower conversion rate throughout the two years.

Furthermore, the average spending habits of the customers also have mutated during the two years: in 2018 (Appendix 8.2, graph 5) the customers tended to spend less on average (around 20/50) whereas in 2019 (Appendix 8.2, graph 6) the customers slightly increased their average spending (around 30/60).

The previous trend is reflected also in the total spending of the customers throughout the two periods (Appendix 8.2, graph 7-8): the total spending increased from 2018 to 2019 across the three cities. Throughout 2018 and 2019, Rome is the city where customers totally have spent less; in Naples the spending pattern is similar to the one in Rome but customers tend to totally spend slightly more, whereas in Milan the total spending of the customers is much higher compared to the other two cities.

The heatmap (Appendix 8.2, graph 9) has been useful to display the correlation (that ranges from -1 to 1) between the variables of the dataset. It resulted that : the **sales** show a 0.6 correlation with the conversion rate, a 0.7 correlation with the total spending, and a 0.8 correlation with the number of website visits. The **average spending** is not correlated to the sales, but has a 0.7 correlation with the total spending. Interestingly, the **web site visits** show a correlation of 0.9 with the variable total spending, a slight uncorrelation with the conversion rate of -0.1. Lastly it emerged that the **total spending** and the conversion rate are slightly uncorrelated (-0.2).

3 Experimental design

This phase of the analysis aimed at detecting seasonality, and consequently at building a model to quantify the impact of the advertising campaign launched by company Alpha during December 2019, on sales by analyzing visits to its website and online purchases. Since the data refers to three different Italian cities which represent the three main geographical areas (North, Center, South) of the italian market; a model for each city was built with the objective of inspecting the trend of the perfume's sales. The procedure followed throughout the building of the model for each city has been the same.

3.1 Univariate time series

A time series is considered to be a sum or combination of the following components: trend, seasonality, and residual. The detection of the seasonal component, in this study, has been handled by employing an univariate time series (only the variable sales was under investigation) for each city. The decomposition graphs of the three cities (Appendix 8.3) outlined a clear weekly seasonal pattern in the sales: changes in data values are repeated regularly over the same time period (increases and decreases). Moreover, the ACF was graphically represented (Appendix 8.3). Analyzing the autocorrelation function enabled to identify seasonality in time series data. From the ACF, it is evident that the city with the highest number of lags (higher sales' seasonal component) is Naples, followed by Rome and Milan.

3.2 Vector Autoregression Model

The model employed for the following study is a Vector Autoregression model (VAR). The latter is a statistical model used to capture the relationship between multiple quantities as they change over time. The variables employed to build the model for each city are: sales, website visits and the conversion rate. In order to carry out a multivariate time series analysis, the dataset was splitted in two parts: the train set containing observations from December 2018 to December 2019, and the test set which was created including only the last part of the dataset (January 2020). At this point, a time series for each city was generated using the train set (Appendix 8.4). In the VAR(p) model each variable is a linear function of past lags of itself and past lags of the other variables. The VAR model is a multi-equation system where all the variables are treated as endogenous (dependent). There is one equation for each variable as dependent variable. In its reduced form, the right-hand side of each equation includes lagged values of all dependent variables in the system.

$$\mathrm{y}\ _t = \alpha_1 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + ... + \beta_p y_{t-p} + \epsilon_t$$

where:
yt: an (nx1) stationary vector of time series variables(sales, convrate, visits)
alpha: an (nx1) vector of intercepts
beta: (nxn) parameters matrix (coefficient matrices)
epsilon: an (nx1) vector of unobservable (zero mean error term, white noise)

In this case, the VAR model general equations (for each city) are as follows:

$$\begin{aligned} sales_{t,1} &= \alpha_1 + \beta_{11} convrate_{t-1,1} + \beta_{12} visits_{t-1,2} + \beta_{13} sales_{t-1,3} + \epsilon_{t,1} \\ convrate_{t,2} &= \alpha_2 + \beta_{21} convrate_{t-1,1} + \beta_{22} visits_{t-1,2} + \beta_{23} sales_{t-1,3} + \epsilon_{t,2} \\ visits_{t,3} &= \alpha_3 + \beta_{31} convrate_{t-1,1} + \beta_{32} visits_{t-1,2} + \beta_{33} sales_{t-1,3} + \epsilon_{t,3} \end{aligned}$$

3.2.1 ADF test and differencing

Subsequently, an Augumented Dickey Fuller Test was performed in order to check for stationarity of the time series. The statistical properties of a stationary time series such as mean, variance, and auto-correlation are constant over time, resulting into more accurate outcomes when performing predictions.

Since all the three models presented a p-value greater than 0.05 (implying non-stationarity), differencing was performed. The latter was employed in order to stabilise the mean of the time series, therefore eliminating trend and seasonality. One order of differencing was sufficient to reduce the p-values, making the series stationary (Appendix 8.5).

3.2.2 Lag order identification and fit of the model with VAR

In order to fit the three models, lag order identification was performed. According to the results, a max lag of 10 was considered the one to minimize the AIC for all the three models.
Based on these results, the Autoregressive Vectorial models were fitted.
Among all the metrics used to evaluate the accuracy of a fitted model, the MASE (Mean Absolute Scaled Error) was analysed in this case for each of the three models for mainly three reasons:

- MASE is independent of the scale of the forecast since it is defined using ratio of errors in the forecast. This means MASE values will be similar if we are forecasting high valued time series;
- MASE gives an indication of effectiveness of forecasting algorithm with respect to a naïve forecast. If MASE is less than 1, it means that the forecast is better than the naïve method on training data. If it is more than 1, it means that the forecast method is worse than the forecast using the naïve forecasting approach;
- Another advantage of MASE is that it can be used both on a single series, or as a tool to compare multiple series.

By comparing and analysing the accuracy of the three models it emerges that the MASE is less than 1, specifically: in Milan's model is 0.632, in Rome is 0.547 and in Naples is 0.542. It can be concluded that the forecast is better than the naïve method on training data.
For what concerns the adjusted R^2 for each equation of the three models, the values are the following: In Milan's model: adjusted R^2 of sales is equal to 0.66, of the conversion rate is 0.67 and of visits is 0.61. In Naples's model: adjusted R^2 of sales is equal to 0.71, of the conversion rate is 0.76 and of visits is 0.67. In Rome's model: adjusted R^2 of sales is equal to 0.72, of the conversion rate is 0.79 and of visits is 0.66.

3.2.3 Granger Causality Test and Impulse response function

At this point, causality has been checked through the Granger Causality Test. The latter is a crucial statistical hypothesis test used for determining whether a predictor in the time series is relevant, and indeed useful for forecasting another one.
In the Granger test the Null Hypothesis (H0) denies the existence of causality between the endogenous variables under analysis. In this case, since the p-values are all significant and below the threshold of 0.05 , indicating strong evidence against the null hypothesis, it can be deducted that there is, indeed, Grenger causality.
This means that one time series "Granger-causes" the other; suggesting that even tough it still can not be affirmed that one time series exactly causes the other, they can certainly be used to make accurate predictions of each other. This in practice can be translated in: if "visits" Granger-cause sales and convrate, then past values of visits should contain information to predict sales and conversion rate above and beyond the information contained in past values of sales and conversion rate alone.
The following step focused on the analysis of the Impulse Response Function's (IRF) graphs. The latter was used to describe the evolution of the models' variables in reaction to a shock in one or

more of the other variables. In this case, the three regions under analysis showed different results. Specifically:
the response of **sales** to a shock in website visits in Milan and Rome is significant at time 2,4 and 6; whereas in Naples it is significant only at time zero and six. As a matter of facts, Naples showed a slighlty higher value in the Grager-casuality test if compared to the other varibles.
The response of **sales** to a shock in the conversion rate in Milan showed significant results in a 95% confidence interval at time 3,4,7 and 8 with a similar trend recorded in Naples; whereas in Rome it is significant only at time 8 (Appendix 8.6).

3.2.4 Variance Decomposition

The forecast error variance decomposition is used to analyse the contribution of one variable to the h-step forecast error variance of the other endogenous variables present in the model.
This information is useful to understand in the forecast how much of the prediction of one variable depends on the others, and how much it depends on itself. For the current model, a step- forecast of five months was used. The result is a matrix of values in which:
In all the three cities (Appendix 8.7) the variable **sales** depends for the first month entirely on itself. This trend remains overall the same in the following months.
Among the three variables, the one showing most interesting results is **visits**. The latter in Milan and Naples shows an average composition of 50% on the variable sales in the first month. In Rome this trend is more evident: in the same period, the visits depend for 81% on the sales.
The variable **conversion rate** shows in Milan and Naples a high dependency on the sales (overall 80%). Whereas, in Rome this trend decreases with a dependency of 50% on the sales and 40% on itself.

3.2.5 Model validation

Finally, the three fitted models have been employed to carry out forecasting for the following 5 months. For each city, the forecast for variable sales, convrate and visits are represented in the following plot (Appendix 8.8). All the forecast values fall under a 95% confidence interval. Specifically, in Milan, all the three variables remain constant throughout the first period with a subsequent sharp drop in two intervals to then increase again in the last forecasted period.
For what concerns Naples, the variables tend to have a decreasing trend in the forecasted values. In detail, by comparing the the data regarding Milan and Naples results that (negative amount of values being the same in the five periods), the positive values in Naples are comparatively lower.
The forecasted results for Rome show overall a lower amount of negative values in the 5 steps; but also the lowest amounts of sales, conversion rate, and visits recorded when compared to the other two cities.

4 Code description of the multivariate time series

After importing the required libraries and the dataset, all the dates contained in the latter have been converted to date format. Since the dataset contained daily data, a specific lambda function named “day-week” was employed to convert the daily data into weekly data without losing any observation. The reason behind this choice was to easily identify seasonality later pn in the analysis for each week.

Weekly values for sales, conversion rate and visits have been computed by grouping the just mentioned features, and creating a new column with the sum of the sales, conversion rate and visits per week. Subsequently, a new dataset called “weekly data1” containing weekly observations for all the three cities under analysis has been created. In detail, the variables contained in it are:

- City: the city where the observations were taken from,
- Dayweek(day): the week number,
- Day: the date,
- Totsales: total of sales per week,
- Totrate: total conversion rate per week,
- Totvis: total of website visits per week.

Afterwards, the dataset has been splitted according to time-series requirements: the last part of the dataset (January 2020) has been used as test set, and the remaining part as train set.
In the following part of the code, data for each city has been analyzed and fitted employing a Vector Autoregression model. The code being the same for each city, and being the findings explained in an appropriate section, just one method will be explained here as metric to understand the others. Firstly, the city of interest has been specified by extracting it from the unique dataset. In order to get a visual representation of the three variables under analysis (totsales, totrate, totvis), a ggplot for each has been specified in relation to the time.
Later on, the data has been transformed to time series employing the “ts function” in R, which is included in the tseries package. The frequency is 61 weeks because one year and two months are taken in consideration during the analysis (which corresponds to 61 weeks). Subsequently a plot of the time series, and one of the three variables under analysis was graphically displayed.
The stationarity of the time series has been tested using the adf.test function; and since this resulted to be non-stationary, differencing was performed using the “diffM” function present in the lmtest library. Then, the differentiated time-series was plotted using “autoplot” function, by specifying the start/end parameters, and the frequency in order to get an accurate representation. Subsequently, in order to determine whether this passage was effective; the time-series was re-tested with another adf.test, and this time it proved to be stationary. Indeed, one order of differencing was enough. Lag order identification was performed using the “VARselect” function, of the differentiated time-series. The “type” parameter was set to ”none” since the time series was already made stationary using differencing, and a lag of ten was indicated in order to get a deep insight of how the AIC, HQ, SC, and FPE parameters behaved in this lag-range. A lag order of 10 was identified to be the one to minimize the most the AIC, and was used to fit the Vector Autoregression Model. The summary function has been used to get a summary of the model.
Consequently residual diagnostic was performed. The Granger causality test has been performed for each of the three variables inside the fitted model using the causality function. Subsequently, the IRF (impulse-response function) has been employed to understand how one variable responds to a shock in one of the others using the “irf” function present in R. Moreover, variance decomposition has been carried out using the “fedv” function contained in the var package. The latter shows the composition of each variable, and its dependency on the other two endogenous variables. A plot of the latter has been later deployed.
In conclusion, forecast on the fitted model has been performed in order to predict the trend of the variables for the following 5 months, and to detect a significant change in their pattern.

5 Results

At this point, the findings can be delineated based on the results obtained from the analysis. Firstly, the year 2018 was included in the analysis in order to analyse the counterfactual. The latter enables to estimate advertising effectiveness and indeed to get precise insights on what would have happened without the ad exposures. Consequently, weekly seasonality was removed from the training data (2018-2020). This step was performed because data at hand was generated in periods when the purchase probability is higher; making it harder to observe causal relationship between the ad campaigns and the variables sales, visits and conversion rate.

Moreover, Granger-causality test showed that the variables sales, visits and conversion rate for each city have a causal relationship: indeed, not only they depend on themselves, but also on the other variables.

Aside the analysis on the counterfactual, forecast on the three different models was conducted with the objective of investigating also the future trends of the variables. From that, a significant change in the future trend of the three variables was not detected.

6 Conclusions

In conclusion, by comparing the counterfactual, December 2018, (period without the ad campaign) and December 2019 (period in which the ad campaign was launched), it resulted that there has not been a significant increase in the sales, visits and conversion rate in all the three cities from December 2018 to December 2019. This outcome is reflected also in the variables' values obtained from the forecast: the future predicted values show a trend similar to the one of the periods under analysis. These findings lead to the conclusion that launching the ad campaign in December 2019 was not significantly effective.

The approach employed in this analysis for finding the impact of the advertising campaign on sales is mainly econometric. Therefore, as natural next step for a future work approach, what might be implemented is a numerical quantification of the impact of the advertising on the sales (the latter was found to be not significant, so a small percentage value would be expected as outcome).

6.0.1 Possible Future Strategies

Despite the ad campaign was found to not have a significant impact on the sales, different customers' behaviour emerged from the analysis. In detail, the Northern region presents overall higher values in the sales, visits and conversion rate with respect to the Central and Southern regions. Specifically, the region with the lowest scores recorded in all the features is the Center of Italy .

Indeed, from a business perspective the company Alpha might take the following potential actions:

- the company could exploit its already established presence in the Northern area, and the potential of the latter by concentrating future investments in this region;
- the company could decide to focus on the Central and Southern geographical areas of Italy, where all the variables scored overall lower values. This action might stimulate customers to make more purchases, increasing the presence of the company in this region.

7 Bibliography

Ahuja, A., (2021, January 11), Mean Absolute Scaled Error (MASE) in Forecasting, https://medium.com/@ashishdce/mean-absolute-scaled-error-mase-in-forecasting-8f3aecc21968

Eric, A, (2021, July 23), The Intuition Behind Impulse Response Functions and Forecast Error Variance Decomposition https://www.aptech.com/blog/the-intuition-behind-impulse-response-functions-and-forecast-error-variance-decomposition/: :text=Impulse%20response%20functions%20trace%20the,theoretical%20economic%20and%20finance%20models.

Gupta, A., (2021, September 5), Vector Auto-Regressive (VAR) Models for Multivariate Time Series Forecasting,https://medium.com/geekculture/vector-auto-regressive-var-models-for-multivariate-time-series-forecasting-106bb6f74add

Kumar, R., (2020, April 29), MAD over MAPE?, https://towardsdatascience.com/mad-over-mape-a86a8d831447

Mallick, D. (2020,November 25), Interpreting ACF or Auto-correlation plot, ://medium.com/analytics-vidhya/interpreting-acf-or-auto-correlation-plot-d12e9051cd14

Özen, A.(2021, February 27), Seasonality Analysis and Forecast in Time Series, Medium.com, https://medium.com/swlh/seasonality-analysis-and-forecast-in-time-series-b8fbba820327

Stephanie, A., (2019, July 29), Statistics How To, https://www.statisticshowto.com/mean-absolute-scaled-error

Zivot, E. (2013, May 29), Multivariate Time Series and Vector Autoregression

Zivot, E., (2003, September 3), Vector Autoregressive Models for Multivariate Time Series, http://www.ams.sunysb.edu/ zhu/ams586/VAR$_Lecture2.pdf$

8 Appendix

8.1 Exploratory Data Analysis

ggplot.jpeg

8.2 Data Visualization

visits_2018.png visits_2019.png

convrate_2018.png convrate_2019.png

avspend_2018.png avspend_2019.png

total_spending_2018_2019.png

correlazione_variabili.png

8.3 Univariate time series

DecompDsitcionDsMtiloanosNtagpileosRnoamgeof.srpne.png

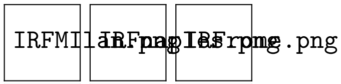
8.4 Vector Autoregression Model

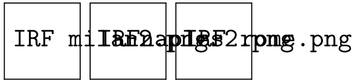
graph1ha2lpesgrapgh 11.png

8.5 ADF test and differencing

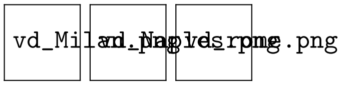
milan12nappo12rioma.pngpng

8.6 Granger Causality Test and Impulse Response Function





8.7 Variance Decomposition: graphs for Milan, Naples and Rome



8.8 Model validation: graphs for Milan, Naples and Rome

8.9 Peer self-evaluation

The following are the roles covered by the team members in this project:

- Martina Bozzi and Giulia Di Martino: Data Analysts.
  Martina and Giulia were in charge of the following tasks: exploratory data analysis, modeling (analysis and fit of both the models), coding, results, conclusions and report.

- Claudia Di Razza and Allegra Pomanti: Data Designers.
  Claudia and Allegra were in charge of the following tasks: exploratory data analysis, data visualization and Power Point preparation.