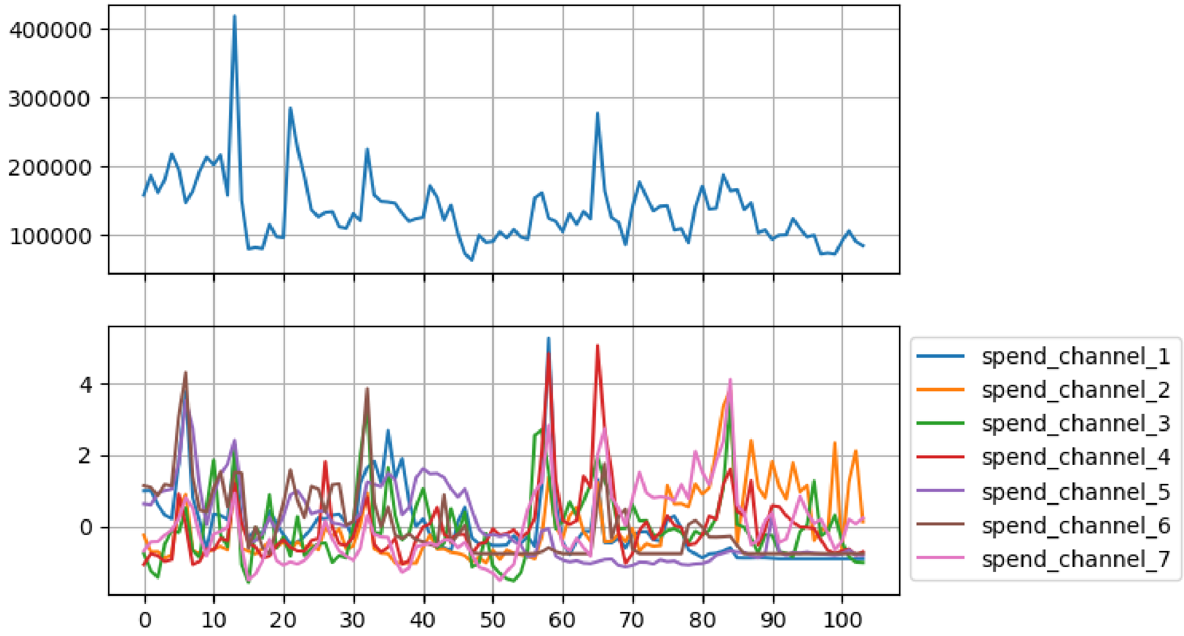
In order to maximize profit and generate as much revenue as possible with as few resources as needed, Company X has turned to implementing a Mixed Media Model. There is one variable for revenue which we will try to model using information on the date and our 7 spend channels.

Exploratory Data Analysis

Beginning with EDA, I plot the revenue and spend channels in the below graph along the time axis. This can give me some insight into how which variables could affect revenue. The revenue channels have been standardized by demeaning and rescaling according to their standard deviation to make the analysis easier.



Based on the above graph I believe that Purple (5) will have some effect on the revenue. Between the 0th and 10th observation we have a spike in both revenue and spend. Additionally , I believe that Brown(6) will have an effect on revenue. Around the 30th observation there is a sudden increase in both revenue and spend. Furthermore I believe that Red(4), green(3) probably won't have an effect because there is spike between the 50th and 60th observation which results in no change in revenue. I also believe that Blue(1) won't have an effect since there is a spike between 30 and 40 which makes no change to the revenue. I don't believe that Yellow(2) and Pink(7) will have an effect. The increase in spend between the 80th and 100th results in no major change in revenue.

From the graph of revenue we can also see that it has a slight downwards trend. Additionally, I believe that it has a weak autoregressive (AR) process since the values are not distributed around the trend, but instead depend on its latest values. From an economic point of view, it also makes that it should have an AR process, as consumers may take notice of the product if sales were high in the last week. The outliers found around observation 10, 20 does not seem to be well explained based on this EDA and might introduce issues further in the modelling. For the sake of simplicity I chose to ignore them but would want to revisit what they are caused by and if they are true outliers in further analysis.

For the seasonal component we can see that November is the month which sees the highest revenue. Something surprising is that December is on average below the revenue earned during the sample period. We would expect that Christmas would have a big positive effect during December because of this.

Table of deviation from average revenue per month

Month	Deviation from average revenue
January	20454
February	-7950
March	-2141
April	26842
May	-14525
June	-10984
July	-48706
August	-34480
September	13332
October	12458
November	69921
December	-25602

The priors I chose for modelling are that we have a seasonal component in our revenue, which will be captured by a monthly dummy, a trend variable, and spend channels with various lags. I believe that the coefficients for these variables and for revenue are normally distributed. I also believe that the AR process is bounded between 0 and 1, however due to ease of implementing it into the model, I let it have the same distribution as all the other coefficients.

While the revenue should be truncated at 0 since we can't have negative sales, I was not able to implement this.

In addition to the above transformations, I also chose to have all values in their level format, as opposed to taking the natural logarithm. I did this to estimate the effect that 1 euro/dollar/yen of increased spending has on each channel and because some spend channels had values of 0 at certain observations.

I split the data into train and validation data into a 90/10 split and will use the validation data to select the final model.

Modelling

I try out two different ways of modelling spend carry over. The first way is by having additional lags of each spend column. I select to have 4 lags since then the previous month's spend will affect the current revenue however 4 lags is kind of arbitrary and could be extended

The other model is by introducing an autoregressive component. I lag revenue by 1 which captures the effect of previously high revenue. An AR(1) component can be reasoned to capture the effect of previous increases in spending. Other benefits of the AR(1) is the data lost due to lags is reduced and we can assume that it would affect based on the economic arguments outlined above.

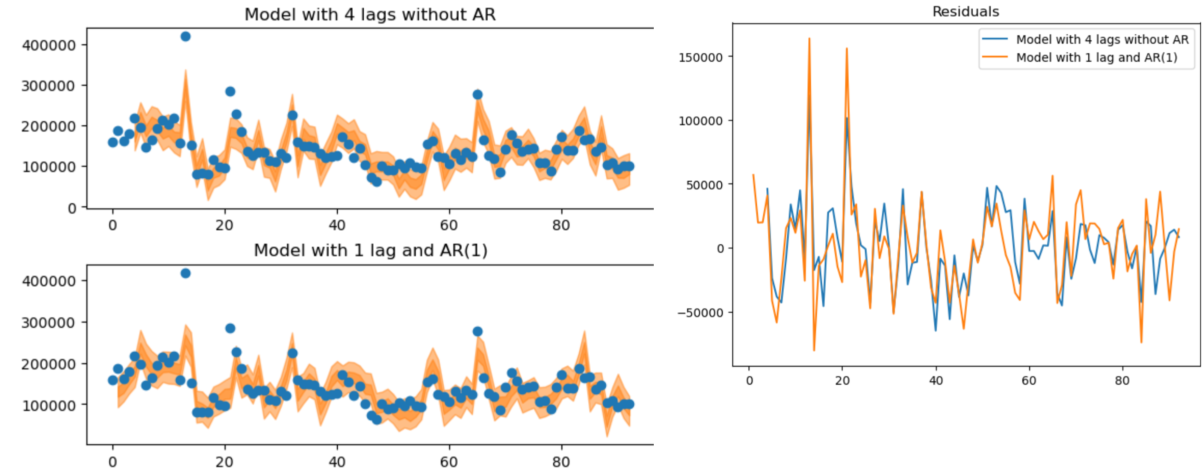
The monthly dummies, trend and contemporary spend columns is included for both models.

Posterior sampling

Introducing our model to observed data, we can create a pretty good fit. On the left hand side of Table 2 we can see that both models fit the data OK. While some of the outliers like in the beginning of the time series is not fully captured, the other data points and the smaller spikes later on seems to be captured within our High Density Interval (HDI).

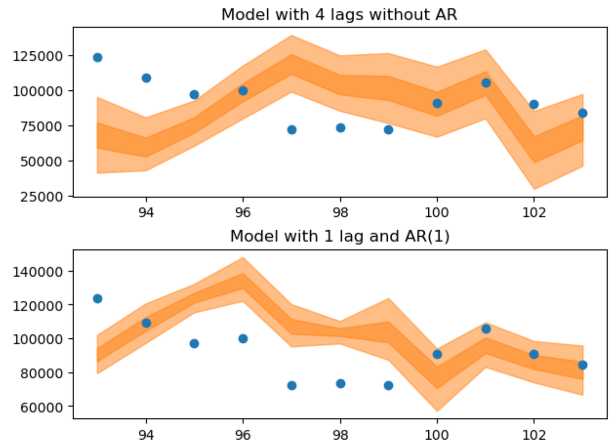
Examining the residuals for any autocorrelation in the errors we see that it is likely, but does not have a big effect. While the residuals are not completely white noise, as we would want, there is not too much information gained from the previous value of the residual. This does not diminish the fact that our residuals are dependent of each other, which biases our coefficients. Based on the residuals I would prefer the model with 4 spend lags, since it has lower variance and exhibits a more stochastic walk than the model with 1 lag and AR(1) component.

Table 2 – results of models



As a more objective measure of our model, I examine the Mean Absolute Percentage Error (MAPE) for both models. The model with 4 spend lags has a MAPE of 16.92% and the model with 1 lag spend lag and AR(1) has a MAPE of 18.32%. Based on this I still believe that the model with 4 spend lags is the best out of the two. While 16.92% is still too high for my liking, I consider it good for a first run.

Out of sample results



For the out of sample (OOS) predictions, we can see that it does a poor job at predicting the dip in the middle of our OOS period between the 93rd and 95th observation. Our predictions follow the same path with a downwards trend and the small changes after observation 96.

I examine the MAPE for the out of sample (OOS) period. The preferred model with 4 spend lags has an OOS MAPE of 28.54% and the model with 1 spend lag and AR(1) has an OOS MAPE of 20.19%. Based on this I believe that the model with 1 lag and AR(1) best describes the data and is my preferred model. I will continue to comment on both for the sake of exposition.

Return on investment (ROI)

In order to produce a recommendation for the client I calcualte the return of investment from spending one euro, dollar, or other currency on each of the spend channels. I do this by first predicting the values for a baseline where no change happens. In this case I set all variables to zero, including the trend and monthly dummies. For each spend channel, and its lags, I set one observation equal to 1 where appropriate. By taking the difference between the baseline, and the prediction I can estimate the effect of one additional spend. This produces the ROI.

For the model with the AR(1) component I divide it by the coefficent of it’s lagged value. This gives me the long run effect of the impulse.

Based on the column “ROI Model with 1 lag with AR(1)”, we can see that channel 1 and 2 are in fact hurting revenue. Given the magnitude of the two channels, it seems like the model is capturing some variance and attributing to these channels as it seems unlikely that they would damage revenue this much. Channel 5 is the highest ROI channel, followed by channel 6 and channel 3. This is partially in line with my belief as outlined in the EDA, where I believed that channel 5 and 6 would have an effect. Contrary to my expectations channel 3 had a positive impact on the revenue. The HDI at 3% and 97% for the contemporary coefficients for channels 5, 6, and 3 does not include 0. On the other side, the lagged coefficients for channels 6 and 3 do include 0 in their HDI. As such, focusing on channel 5 is strongly recommended as it is the most economically and statistically significant.

Channel	ROI Model with 4 lags without AR	ROI Model with 1 lag and AR(1)
Spend channel 1	-14,91	-12,01
Spend channel 2	-65,52	-16,93
Spend channel 3	3,65	2,15
Spend channel 4	0,24	2,11
Spend channel 5	1,8	4,1
Spend channel 6	4,5	2,69
Spend channel 7	2,25	1,76

Summary

To summarize, Company X should focus their investments on channels 5, 6, and 3 as these are most likely the ones to give a good return on investment. In particular channel 5 should be focused on. This should be cautioned with the fact that the models are not good at explaining out of sample data, may have autocorrelated residuals, and produces strange ROI for spend channels 1 and 2.

As a final point I provide questions to the client that could help with this analysis and points on how the model can be improved upon.

Questions to client

1. What does each of the channel represent and can they be aggregated in some way?
2. Is there other factors which affects the product such as competitor entry/exit, surge in media attention, or other similar factors?

Points of improvement

1. Introducing saturation and shape effect of the adstock.
2. Possibility of aggregating spend channels.
3. Introducing a moving average (MA(1)) component to the model.
4. Exploring the outliers in the beginning of our sample period.