Advertising Response Measurement - Final Report

Team K - Zichen (Zoe) Huang, Jiaying (Claire) Wu, Chen (Cici) Chen

5/8/2019

Introduction

The ultimate goal of any marketing activities is to increase sales, either in short-term or long-term, and ideally each campaign or marketing channel should be evaluated based on the incremental profit, which is the additional sales we produce with advertising over what we would have sold without advertising, relative to its cost. While marketing efforts seem intangible as it is hard to land causal relationships between marketing events and transactions, advanced statistical methods could definitely shed the lights of tangible marketing impact on business growth, and guide the direction for future practice.

By analyzing the correlations between marketing activities and the transactions, weakness and strengths of various channels, seasonality of consumer's response, the marketing team could evaluate the cost efficiency and thus move forward to optimize the resources allocation for marketing spending, and strategically plan marketing events to improve the effectiveness of marketing efforts.

Among diverse marketing activities, advertising is a major mean for branding and informing. Both online (emails, social media, displays, etc.) and offline (direct mails) advertisements play an indispensable role in marketing. As placing ads is expensive in terms of the cost of time, recourses, and budget, one most frequently asked question from the management team is that: how does my advertising work? With this question in mind, our team leveraged four methods to tackle the advertising effectiveness measurement problem. The methods are last-click attribution analysis, holdout testing (experimentation), marketing mix models and model-based attribution analysis. The main goal of this project is to evaluate the **effectiveness of different advertising channels** from four different perspectives. Meanwhile, we would learn from the advantages and disadvantages of the approaches and gain a more comprehensive understanding of mix modeling methods for continuous study.

Sources of Data

The full dataset could be found in the submission files. As dataset related to advertisement responses and transaction records is very sensitive, we could hardly find real or original open data. The dataset we used for this project is a synthetic one simulated by the Elea McDonnell Feit, Marketing Professor of Drexel University, and is organized from three perspectives: customer, impressions, and transactions. Data generation method could be found here: https://github.com/eleafeit/ad_response_tutorial/blob/master/R%20code/AdResponseDataGeneration.R. The whole raw digital advertising dataset describes 10,000 customers as well as potential customers of a retailer, and the retailer uses four different advertising channels - display ads, social media ads, email ads, and direct mail ads for its marketing promotion.

Examination of the Data

We will be working with a simulated data set related to social media sites. The data are stored in following three files: **customer.csv**, **impressions.csv**, and **transactions.csv**. The first row of the data set includes the column names, and each subsequent row includes one observation of values. Here is a selection of 10 lines from each data file:

Inspect customer file

customer.csv: Information about the users with some fields from their profiles. Each row in the file represents a customer, 10,000 rows; the columns describe some basic information about each customer including id number, whether the customer has made a purchase prior to the observation period, and whether the customer is eligible to receive emails or direct mails.

Show 10 entries Search:

	id	past.purchase	email	direct
1	1	0	0	1
2	2	1	1	1
3	3	1	1	0
4	4	0	0	0
5	5	1	1	1
6	6	0	0	0
7	7	0	0	1
8	8	1	1	1
9	9	0	0	1
10	10	0	0	0

Showing 1 to 10 of 10 entries Previous 1 Next

- **past.purchase**: in the original dataset, the type of *past.purchase* is *int*, but this variable records whether the customer has made a purchase prior to the observation period so we decided to convert it into *factor*.
- **email**: in the original dataset, the type of *email* is *int*, but this variable records whether the customer is eligible to receive emails so we decided to convert it into *factor*.
- **direct**: in the original dataset, the type of *direct* is *int*, but this variable records whether the customer is eligible to receive direct mails so we decided to convert it into *factor*.

Inspect impression file

impressions.csv: Information about which users are connected to other users. Each row is an exposure of marketing communication to a specific customer, 501,336 rows; the columns describe the information about the customer's impression towards an advertisement including id number for the customer, date of impression, the channel of the ad exposure, and whether the customer clicked on the ad.

Show 10 entries Search:

	id	date	channel	click
1	1	2017-01-06	direct	0
2	1	2017-02-03	direct	0
3	1	2017-01-01	social	0
4	1	2017-01-02	social	0

	id	date	channel	click
5	1	2017-01-05	social	0
6	1	2017-01-06	social	0
7	1	2017-01-06	social	0
8	1	2017-01-07	social	0
9	1	2017-01-07	social	0
10	1	2017-01-07	social	0

Showing 1 to 10 of 10 entries

Previous 1 Next

- **date**: in the original dataset, the type of *date* is *chr*, and we decided to convert it into *date*.
- **channel**: in the original dataset, the type of *channel* is *chr*, and we decided to convert it into *factor*.
- **click**: in the original dataset, the type of *click* is *int*, but this variable records whether the customer clicked on the ad so we decided to convert it into *factor*.

Inspect transactions file

transactions.csv: Information about history of the user's account registrations (logins) over time. Each row is a transaction made by a customer; columns record the basic information of a transaction including customer id, date of the transaction, channel of the last ad impression the customer saw before the transaction, and channel of the last ad the customer clicked before the transaction.

Show 10 entries Search:

	V1	id	date	last.touch	last.click
1	1	2	2017-01-04	email	none
2	2	2	2017-02-12	email	none
3	3	3	2017-02-02	email	none
4	4	3	2017-02-14	email	none
5	5	5	2017-01-04	display	email
6	6	5	2017-01-13	display	email
7	7	5	2017-01-20	display	email
8	8	5	2017-01-26	display	email
9	9	5	2017-01-31	display	email
10	10	5	2017-02-06	display	display

Showing 1 to 10 of 10 entries

Previous 1 Next

- **V1**: in the original dataset, *V1* only represents the row number, so it is safe for us to remove this variable.
- **date**: in the original dataset, the type of *date* is *chr*, and we decided to convert it into *date*.
- **last.touch**: in the original dataset, the type of *last.touch* is *chr*, and we decided to convert it into *factor*.

- **last.touch**: in the original dataset, the type of *last.touch* is *chr*, and we decided to convert it into *factor*.
- **last.click**: in the original dataset, the type of *last.click* is *chr*, and we decided to convert it into *factor*.

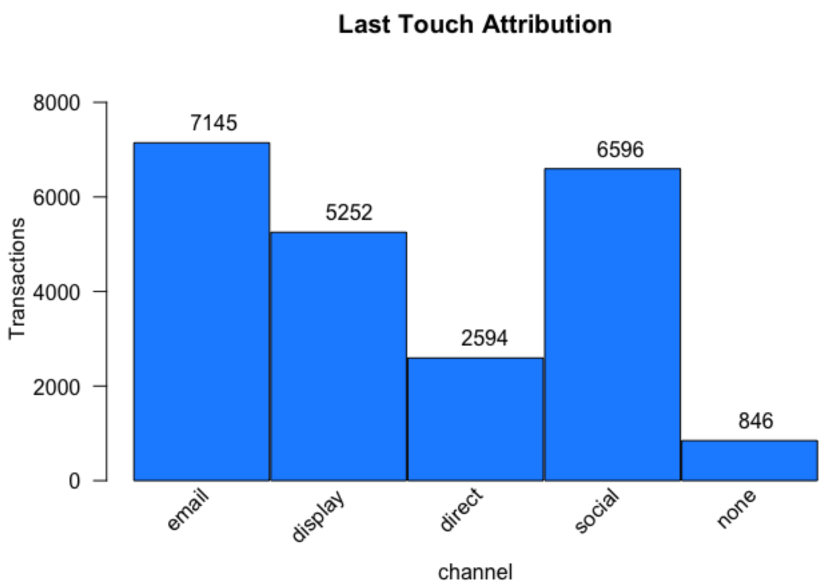
Method 1: Last Touch Analysis

Attribution rules with last-touch analysis

1. Investigation

Based on last-touch attribution, we are able to find the last ad the user clicked on prior to the conversion so that we can get the sales attributed to each channel. In this case, information about the last click and the last touch is stored in the transaction file. By doing a quick crosstab on the transaction table, we are able to calculate the number of transactions attached to each channel by last touch.

```
last.touch
 direct display   email    none  social 
   2594    5252    7145     846    6596 
```



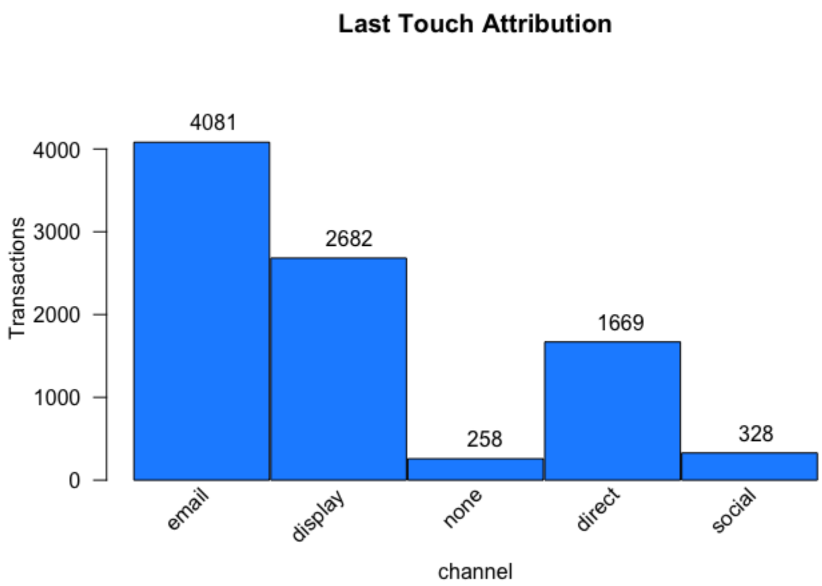
2. Results and Interpretation

Seen from the result, the incremental sales for social media ads are 6596.

Please refer to the dashboard for last touch analysis result for subgroups of transactions to see if some channels are not available, which channel of the rest is the best choice.

Also, we can see if the incremental sales for different ads would be different if we take time into consideration. For example, we can subset out the transactions based on date and then crosstab.

```
last.touch
 direct display   email    none  social 
   1669    2682    4081     258     328 
```



Seen from the result, in February, the incremental sales are far lower attached to social media ads. Probably, this is because that social media ads were ended at the end of January. Please refer to the dashboard to see the difference in different period of time.

3. Assumptions

When we do this, one assumption is that we ignore all the customers who didn't make a real transaction. Also, we assumed that all the sales are counted as incremental, which means that consumers who saw ads would not have bought if they hadn't seen the ads.

Method 2: Holdout Test Analysis

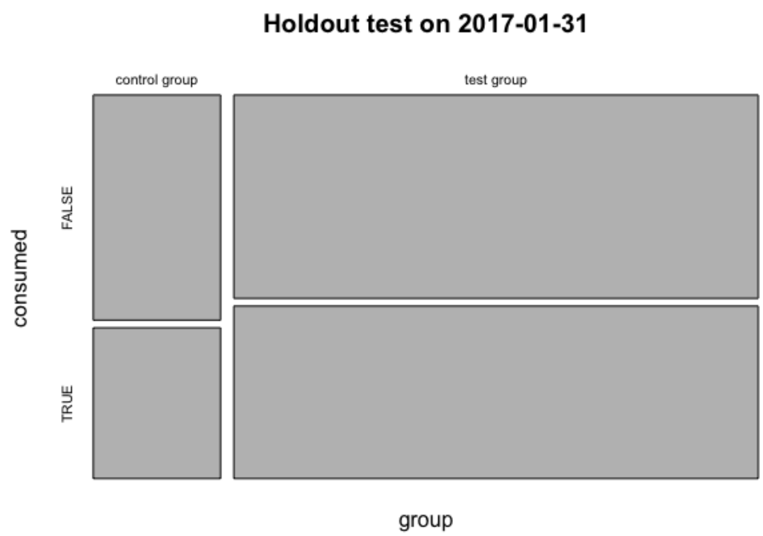
1. Investigation

Holdout testing, also called randomized controlled trial, is an experiment that randomly select the customers for the control group not exposed to an ad and see if there is any difference between people who received ads and those who didn't so that we can know if the ads is effective.

There was an email on 2017-01-03, 2017-01-17, 2017-01-24, 2017-01-31, 2017-02-07, 2017-02-14, and 2017-02-21 that included a holdout group by randomly selecting the customers for the control group to be not exposed to an email ad.

Here, we picked 2017-01-31 to analyze the result of holdout test for the first 10 days.

```
               consumed
group           FALSE TRUE
  control group   702  470
  test group     2613 2216
```



2. Results and Interpretation

Seen from the table as well as the plot, it is obvious that the proportion of people made actual consumption in test group who have received the email ads is higher than that in control group who didn't receive any email ads. Below is a proptest for more detailed information about the comparison of conversion rate between these two groups.

```
    2-sample test for equality of proportions with continuity
    correction

data:  ttable[, "TRUE"] out of xtabs(~group, data = total.tab)
X-squared = 12.541, df = 1, p-value = 0.0003981
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.08978269 -0.02595789
sample estimates:
   prop 1    prop 2 
0.4010239 0.4588942 
```

```
diff.prop 2         ci1         ci2 
 0.05787029  0.08978269  0.02595789 
```


```
    2-sample test for equality of proportions with continuity
    correction

data:  ttable[, "TRUE"] out of xtabs(~group, data = total.tab)
X-squared = 12.541, df = 1, p-value = 0.0003981
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.08978269 -0.02595789
sample estimates:
   prop 1    prop 2 
0.4010239 0.4588942 
```

```
diff.prop 2         ci1         ci2
 0.05787029  0.08978269  0.02595789
```

Seen from this example, the test group had a 45.89% conversion rate in the first 10 days after the email was sent, while the hold out group had a 40.1% conversion rate. The email on 2017-01-31 produced incremental sales. The incremental increase in conversion rate is between +2.60% and +8.98% (95% confidence interval). Please refer to the dashboard to see the difference about how much advertising increases sales with different experiment date and different window period. One inference we can get from the dashboard is that ad response is often greatest just after exposure and then falls off over time.

3. Assumptions

By randomly selecting the customers for the control group, we assumed that the two groups are the same on average. Both the treatment and control groups are assured to be similar in their propensity to transaction and response to ads, which is also called probabilistically equivalent.

Method 3: Marketing Mix Modeling

Based on our datasets, the data points are usually categorical response or logical response, but before we go deep into the logical regression, we want to do linear regression first to find the correlation between total sales to advertising spending on that same day/week/month.

1. Investigation

We did some research about marketing mix model before we started to build our models.

A "marketing mix model" or "MMM" is a regression relating advertising spending or total impressions to some response such as "sales (transactions)". Marketing mix modeling (MMM) is also a statistical analysis utilizing marketing time series data to estimate the impact of various marketing strategies on sales and then predict the impact of future sets of marketing mix or tactics. It is often used to optimize advertising mix and promotional tactics with respect to sales revenue or profit.

In order to investigate the relationships or correlations between total sales (transactions) and impression factors (four different advertising channels - display ads, social media ads, email ads, and direct mail ads) within a specific time period, we implemented linear regression model for marketing mix modeling.

A simple marketing mix model for our case could be represented as follows:

$$sales_t = \beta_0 + \beta_1 display_t + \beta_2 social_t + \beta_3 email_t + \beta_4 direct_t + \epsilon_t$$

the β's represent the unknown relationship between advertising spending and sales. For instance, β_1 means the increase in sales we will gain for each additional display impression unit while holding other impression factors fixed.

We have built five different linear models to see whether there are some thoughtful results which may help us understand the relationship between sales and advertising spending better.

- Model 1: Basic regression including email.holdout
- Model 2: Add in a day of week variable
- Model 3: Taking the advertising effect into consideration
- Model 4: Interactions terms ($Email \times Social$)
- Model 5: Interactions terms ($Direct \times Social$)

When we fit models, it will be more convenient to put the transactions and impressions data together in the same data frame. Also, it will be easy to do regression when we put the necessary variables together, so we did some data preparations before modeling.

so let's take a look at our summarized data, exploration plots, and the corresponding correlation matrix.

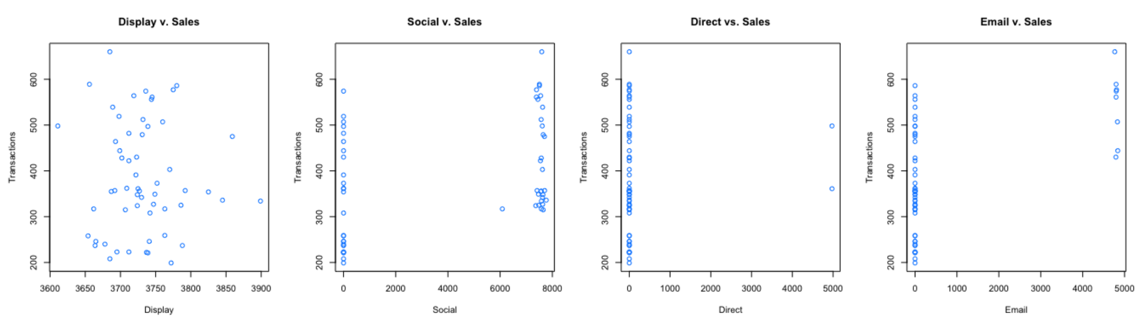
- Summarized data

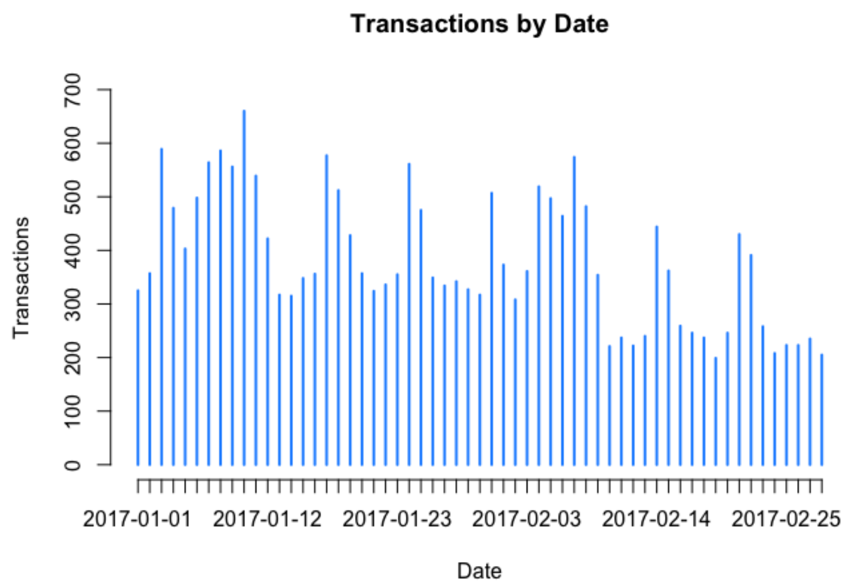
Show 10 entries Search:

	Sales	Direct	Display	Email	Email Holdout	Social	Day of Week	Email ad.effect	Disp ad.ef
2017-01-01	325	0	3786	0	0	7481	Sunday	0	
2017-01-02	357	0	3792	0	0	7416	Monday	0	4(
2017-01-03	589	0	3656	4798	1203	7505	Tuesday	4798	513
2017-01-04	479	0	3731	0	0	7648	Wednesday	2399	527
2017-01-05	403	0	3770	0	0	7620	Thursday	1199.5	5351.
2017-01-06	498	4974	3611	0	0	7614	Friday	599.75	5216.4
2017-01-07	564	0	3719	0	0	7552	Saturday	299.875	5283.92
2017-01-08	586	0	3780	0	0	7504	Sunday	149.9375	5365.177
2017-01-09	556	0	3744	0	0	7446	Monday	74.96875	5353.5532
2017-01-10	660	0	3685	4765	1236	7595	Tuesday	4802.484375	5291.06597

Showing 1 to 10 of 10 entries Previous 1 Next

- Exploration plots





Those plots gave us a general idea about how's the relationship between impression factors and sales, the Display channel tends to show a different trend compared to others, which will be explored more in the following model building part.

- Correlation matrix

Show 10 entries Search:

	Sales	Direct	Display	Email	Email Holdout	Social
Sales	1	0.07	-0.028	0.529	0.533	0.395
Direct	0.07	1	-0.235	-0.077	-0.077	-0.007
Display	-0.028	-0.235	1	-0.068	-0.071	0.142
Email	0.529	-0.077	-0.068	1	1	-0.025
Email Holdout	0.533	-0.077	-0.071	1	1	-0.02
Social	0.395	-0.007	0.142	-0.025	-0.02	1

Showing 1 to 6 of 6 entries Previous 1 Next

Based on the correlation matrix, it can be easily found that Email, Email Holdout and Social seems to have a big impact on Sales. They have positive correlations, which are 0.529, 0.533, and 0.395, and the other two factors seem to have a slight impact.

2. Results and Interpretation

By utilizing two datasets `impressions` and `transactions`, we can have observations of the sales and other advertising data for different impression channels, which will give us the data to estimate these unknown parameters of the model, and the process will be shown below.

Model 1: Basic regression including email.holdout relating transactions to impressions

The model 1 can be written as

$$Sales_t = \beta_0 + \beta_1 Direct_t + \beta_2 Display_t + \beta_3 Email_t + \beta_4 Email.holdout_t + \beta_5 Social_t + \epsilon_t$$

Show 10 entries Search:

	Variable	Estimate	Std. Error	t value	Pr(>|t|)	Coef.Lower.95	Coef.Upper.95
1	(Intercept)	438.501	940.623	0.466	0.643	-1405.085	2282.088
2	Direct	0.015	0.014	1.049	0.299	-0.013	0.042
3	Display	-0.035	0.252	-0.14	0.889	-0.53	0.459
4	Email	-0.162	0.322	-0.502	0.618	-0.792	0.469
5	\`Email Holdout\`	0.805	1.29	0.624	0.535	-1.723	3.334
6	Social	0.013	0.003	3.747	0	0.006	0.019

Showing 1 to 6 of 6 entries Previous 1 Next

- $Sales_t = 438.501 + 0.015Direct_t - 0.035Display_t - 0.162Email_t + 0.805Email.holdout_t + 0.013Social_t$
- From the above output, there is only one statistically significant effect, which is social impressions, and we gain 0.013 additional transactions for each social impression.
- The estimated effect of Email and Display is negative but they are not significant.

Model 2: Add the Day of Week variable

Besides advertising impressions, we introduced a new variable into the linear regression model - the Day of Week, which enable us to inspect the relationships between sales and marketing channels across the week rather than within the same day.

The Model 2 can be written as

$$Sales_t = \beta_0 + \beta_1 Direct_t + \beta_2 Display_t + \beta_3 Email_t + \beta_4 Social_t + \beta_i Day.of.Week(i\ is\ different\ weekdays) + \epsilon_t$$

Show 10 entries Search:

	Variable	Estimate	Std. Error	t value	Pr(>|t|)	Coef.Lower.95	Coef.Upper.95
1	(Intercept)	348.192	891.9491	0.3904	0.6981	-1399.996	2096.38
2	Direct	0.0295	0.0142	2.0852	0.0426	0.0018	0.0573
3	Display	-0.0311	0.2391	-0.1302	0.897	-0.4997	0.4374
4	Email	-0.9607	1.4635	-0.6564	0.5148	-3.8291	1.9077
5	Social	0.0128	0.0031	4.1445	0.0001	0.0067	0.0188
6	\`Day of Week\`Monday	71.4052	45.6163	1.5653	0.1244	-18.0012	160.8116
7	\`Day of Week\`Saturday	64.4455	45.6013	1.4132	0.1643	-24.9315	153.8224
8	\`Day of Week\`Sunday	55.2939	44.5984	1.2398	0.2213	-32.1173	142.7051
9	\`Day of Week\`Thursday	67.6637	45.341	1.4923	0.1424	-21.2031	156.5305

	Variable	Estimate	Std. Error	t value	Pr(>|t|)	Coef.Lower.95	Coef.Upper.95
10	`Day of Week`Tuesday	4877.2293	7029.9972	0.6938	0.4913	-8901.312	18655.7706

Showing 1 to 10 of 11 entries Previous 1 2 Next

When the new variable added into the model, it changed the coefficients of our model. Also, it can be noticed that there is no Friday, and it is just a coincidence that there is no Friday in the subdata which we merged from the original data that meets our assumptions in modeling for transactions and impressions.

In Model 2, we found a significant association between direct impressions and transactions, as well as social impressions and transactions.

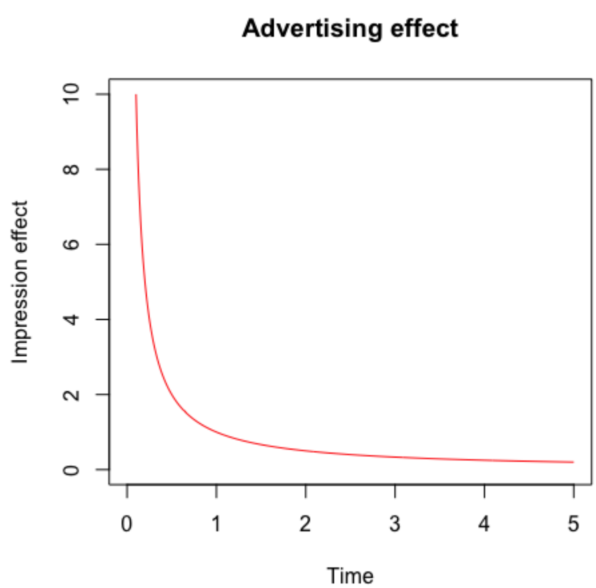
For instance, Direct impression has 29.5 additional transactions per thousand direct impressions, and Social social 12.8 additional transactions per thousand impressions.

Model 3: Taking the advertising effect into consideration

Now we have Model 1 and Model 2, and these two models assume that an impression on specific time t has an effect on the number of transactions on a specific time t and that those impressions have no effect on sales on other days.

However, this is not really happened in the marketing area, just like the results from the email holdout test, it suggests that email impression last about three days! It is easy for us to understand that an advertisement had its biggest advertising effect right after it is shown to the users and then the effect wears over time.

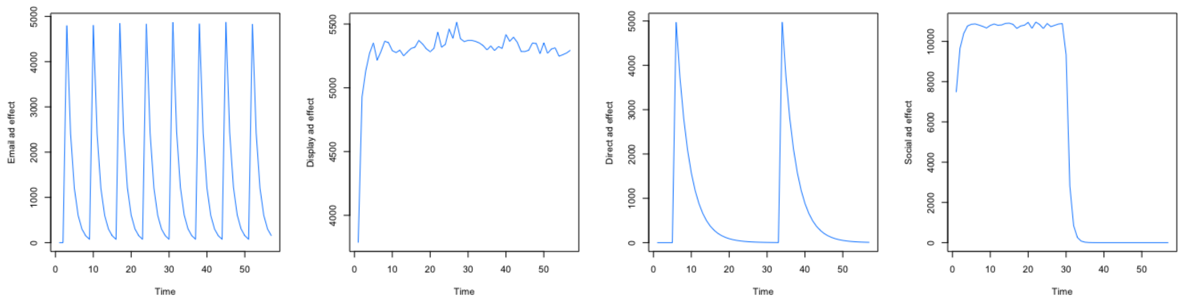
The following picture shows how's advertising effect looks like, and we usually use an exponential decay function to describe how the effect of the advertisements drops.



Additionally, markets defined a term named advertising adstock, which describes the prolonged or lagged effect of advertising on consumer purchase behavior. It is also known as 'advertising carry-over'. Adstock is an important component of marketing-mix models. An ad stock variable is created by computing the exponential decay of the impressions on each day and then sum up the total stock from impressions on previous days.

In order to make the report easy to understand, we will simply call this phenomenon as a advertising effect. In Model 3, we will count advertising effect into consideration.

By using R, we can transform the original impression factors into the new variables corresponding to four different channels, and their plots could be shown below:



The Model 3 can be written as

$$Sales_t = \beta_0 + \beta_1 Direct.ad.effect_t + \beta_2 Display.ad.effect_t + \beta_3 Email.ad.effect_t + \beta_4 Social.ad.effect_t + \epsilon_t$$

In order to make the results more precisely, we will remove first few observations to allow for the "Warmup" of the advertising effect based on the theory of advertising effect.

Show 10 entries Search:

	Variable	Estimate	Std. Error	t value	Pr(>|t|)	Coef.Lower.95	Coef.Upper.95
1	(Intercept)	199.476	704.798	0.283	0.779	-1181.902	1580.854
2	`Direct ad.effect`	0.056	0.007	7.777	0	0.042	0.071
3	`Display ad.effect`	0.004	0.132	0.028	0.978	-0.256	0.263
4	`Email ad.effect`	0.052	0.004	11.931	0	0.044	0.061
5	`Social ad.effect`	0.009	0.001	6.658	0	0.007	0.012

Showing 1 to 5 of 5 entries Previous 1 Next

- Positive effects for all forms of advertising.
- Email and Direct appears to have a similar influence on sales at about 0.056 and 0.052 additional sales.
- All effects are statistically significant except for Display. Display still has a high standard error after the adjustment of advertising effect and it indicates that we do not have a precise estimate of its effect. This happened because daily display impressions are pretty much the same every day.

Model 4: Interactions ($Email \times Social$)

Until now, we have three models, what's else that we need to consider? Interactions occured when there is an extra effect to have two advertising channels or impressions active meanwhile.

We model this by adding an extra interaction term into our basic model, and we take $Email \times Social$ as an example, which could be shown below:

$$Sales_t = \beta_0 + \beta_1 Direct.ad.effect_t + \beta_2 Display.ad.effect_t + \beta_3 Email.ad.effect_t + \beta_4 Social.ad.effect_t + \beta_5(Em$$

Show 10 entries Search:

	Variable	Estimate	Std. Error	t value	Pr(>|t|)	Coef.Lower.95	Coef.Upper.95
1	(Intercept)	204.787	710.729	0.288	0.775	-1188.217	1597.791
2	`Direct ad.effect`	0.056	0.007	7.494	0	0.041	0.07
3	`Display ad.effect`	0.003	0.133	0.025	0.98	-0.258	0.265
4	`Email ad.effect`	0.05	0.006	8.066	0	0.038	0.062
5	`Social ad.effect`	0.009	0.002	4.625	0	0.005	0.012
6	`Email&Social`	0	0	0.542	0.591	0	0

Showing 1 to 6 of 6 entries Previous 1 Next

There is no significant interaction effect between Email and Social at the significant level at 0.05 as its p-value is larger than 0.05.

Model 5: Interactions ($Direct \times Social$)

Similarly, we did the interaction factor in Model 5, which consider the interaction between Direct and Social impression factors.

Show 10 entries Search:

	Variable	Estimate	Std. Error	t value	Pr(>|t|)	Coef.Lower.95	Coef.Upper.95
1	(Intercept)	359.094	796.468	0.451	0.654	-1201.955	1920.142
2	`Email ad.effect`	0.053	0.005	11.448	0	0.044	0.062
3	`Display ad.effect`	-0.027	0.15	-0.178	0.86	-0.321	0.267
4	`Direct ad.effect`	0.058	0.008	7.441	0	0.042	0.073
5	`Social ad.effect`	0.01	0.002	5.423	0	0.006	0.013
6	`Direct&Social`	0	0	-0.446	0.658	0	0

Showing 1 to 6 of 6 entries Previous 1 Next

There is no significant interaction effect between Direct and Social at the significant level at 0.05 as its p-value is larger than 0.05.

There are different combinations in our case, and we will show the entire combinations in our reporting engine that you can choose whatever interaction terms you want eventually.

3. Assumptions

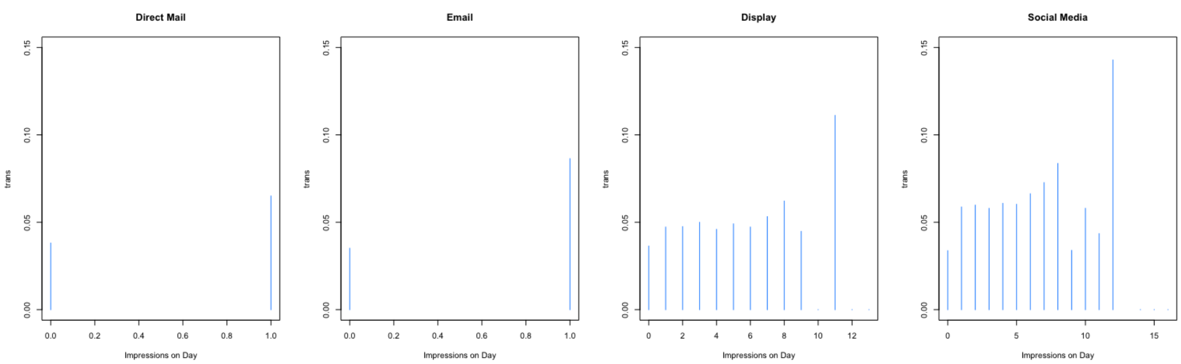
In our case, we assumed a decay rate for each channel when we created the advertising effect variables. We selected a small decay rate for display and social and a larger rate for email due to our intuition, which is reasonable as people normally have a short memory in display and social ads while having a slightly longer memory in email impressions.

We assume that there are no holiday effects in our modeling process. Take a holiday advertising as an example, the causality is actually reversed as the retailers are buying more ads as they know there will be a high demand than usual, so sales now have an influence on advertising impressions. As we want to know an ordinary result, it is reasonable to think that our simulation data does not include the holidays' factor.

Method 4: Model-based Attribution Analysis

1. Investigation

Different from the MMM method, model-based attribution analysis focuses on the **user level** by connecting the transactions of customers to their prior advertising impressions. By implementing this method, we could get a sense of how the advertising impressions could impact the user's purchasing behavior. To investigate this relationship, we firstly used the visualization to have a rough but quick idea.



The above four plots are telling the purchasing tendency in a very straightforward way. First, users of the online retailer convert more on days they get **emails** or **direct mails**. To be more specific, the channels of email and direct mail are more likely to bring more transactions to the store. (Email may bring even more transactions than direct mail.)

Additionally, we could also see that with more Display and Social Media advertisements, transactions may increase accordingly, and too many ads (around 10) on these two platforms would decrease transactions. However, there are fluctuations as reflected from the plots for Display and Social Media. We then implemented model-based analysis to further verify our findings.

2. Results and Interpretation

By implementing the logistic regression model on the dependent variable as transactions and the independent variables as the different advertising channels and whether or not the user has purchased before, we obtained the model summary as below:

Show 10 entries Search:

	rn	Estimate	Std. Error	z value	Pr(>|z|)	Odds.Ratio	OR.Lower.95	OR.Upper.95
1	(Intercept)	-4.023	0.014	-287.078	0	0.018	0.017	0.018
2	direct	0.413	0.042	9.896	0	1.511	1.392	1.639
3	display	0.101	0.007	15.291	0	1.106	1.092	1.12
4	email	0.762	0.02	38.099	0	2.143	2.061	2.229

	rn	Estimate	Std. Error	z value	Pr(>|z|)	Odds.Ratio	OR.Lower.95	OR.Upper.95
5	social	0.18	0.005	33.621	0	1.197	1.185	1.21
6	past.purchase1	0.957	0.016	61.079	0	2.604	2.526	2.686

Showing 1 to 6 of 6 entries

Previous 1 Next

From the statistical significance perspective, all the four channels and the binary past purchase record have significant impacts on the online retail transactions.

From the odds ratio perspective, we could tell that **email** has the highest positive impact on the transaction - each email could increase 114% of probability for the transaction. The second highest channel is **direct mail** - each direct mail could increase 51% of probability for the transaction. Another notable fact is that the customers who have purchased before will significantly and positively impact the transactions - they are 160% more likely to purchase.

Relating these findings to future marketing strategy, we would recommend the retailer to encourage the users with more than one purchase record to register for memberships, and design the email sending and tracking system specifically for those super users. In this way, we could match up the regular customer group with the most effective advertising channel and would likely to maximize the efficiency of the marketing efforts.

3. Assumptions

In this analysis method, we assumed that direct mail, email, social media, and display are the only four channels that the retailer applied for its marketing promotion, such that the effectivenesses are only compared between these four channels. Furthermore, there are no special events (holiday, big discounts, etc.) happening in the same time period (as the date in the dataset) so that the impact of the advertisement would not be influenced by other factors.

Limitations and Uncertainties

For these four methods, each of them has its own advantages and disadvantages:

Method	Advantages	Disadvantages
Attribution rules	Easy to compute and understand	Ignores non-transaction customers
Holdout testing	Simple analysis	requires planning in advance
Marketing-mix modeling	Radically reduces data size	sensitive to model assumptions
Model-based attribution	More flexible	requires large dataset

Limitations and uncertainties of this case mainly come from four aspects:

Firstly, this dataset is not a real one but simulated for training only, which means the result may not be a good reveal of reality;

Secondly, this case is limited by marketing channels provided in the original dataset so that we cannot exclude the impact of other marketing channels, such as friends' recommendation; also in our analysis, we count all the sales as incremental, assuming that customers who saw ads would not have made a transaction if they hadn't seen these ads in the first place;

Thirdly, each holdout test was conducted on Tuesday. Therefore, it is hard to see whether there would be any difference if the holdout test was conducted on other days during a week.

Last, the dataset is not large enough and especially after combining the data, the informative data is not enough.

Areas of Future Investigation

By using last-click attribution, experiments(holdout testing), marketing mix models and attribution models, we are likely to find out the reasons behind the increasing or decreasing sales as mentioned above.

However, there are still rooms for us to make additional contributions based on this theory when applied to reality in the future. Correlation is a complicated magic that people may spend a long time on it. Also, it will be interesting to focus on the trends on different channels besides the impression factors we mentioned before.

We may be able to explore more about the connections among different channels. With the rapid development of the social media and the internet, the traditional way of advertising has been changed a lot. What we studied in our project is just a small projection about what is really happened every single day in our real life. The simple theories used here is still useful in many areas of today's world.

Moreover, breaking down each advertising channel would give us a more explicit and detailed overview of the effectiveness across different promotion platforms. For example, having data aggregated from social media channels - Facebook, Instagram, Snapchat, LinkedIn, etc., we could investigate the advertising platform with the highest cost efficiency, or ARPPU (average revenue per paid user), and the time slots with highest response rate.

Additionally, adding back the users' profile data would also largely increase the potential of learning the user response behaviors at a granular level. Customer information such as age, gender, education level, and personal income would help us do customer segmentation before implementing the analyses, which would distill more insights of advertisement response from different user groups.

There are still a lot of potential areas of future investigation, such as Marketing optimization. Marketing Optimization services and tools can help improve efficiency and profitability by pinpointing the level of marketing activity required to maximize your sales and profit and minimize costs. Methods we utilized before could be excellently generalized in this service in reality. It helps the retailers plan budgets, manage multi-channels, create pricing strategies and produce optimal revenue for a single brand.

Marketing data could be quantified by using these different methods and thanks to these methods, we will be able to understand that why every dollar spent in advertising is actually working for your business.

References

- Marketing mix modeling. (2019, April 24). Retrieved from https://en.wikipedia.org/wiki/Marketing_mix_modeling
- Lohse, A. (n.d.). TV Effectiveness Analysis is Incomplete Without Ad-stock Consideration. Retrieved from https://www.martechadvisor.com/articles/ads/tv-effectiveness-analysis-is-incomplete-without-adstock-consideration/
- Marketing Optimization. (n.d.). Retrieved from https://analyticpartners.com/marketing-optimization/?gclid=CjwKCAjwk7rmBRAaEiwAhDGhxMBUlY0olF0IYV-oplCVBS1f4HbT8dV4p2aRmEaUghp50RSXM1-sExoCXLIQAvD_BwE