Email Campaign Lift Analysis
Bluecore

Given a set of data of email and purchase records explore and analyze the data for patterns and results.

DATA IMPORT AND EXPLORATION

The first step in this analysis is to import and view the data to determine how it is organized, labeled and typed.

I chose the Pandas library, an open source data analysis library, in Python to view and manipulate the data. After importing the data, I looked closely at the initial 5 rows to determine that the data file matches the description in the email.

df.head()

	email	event_date	event_type	action	total
0	4867784685125632	2015-10-27	delivered	window_shopping	-1
1	5352432066363392	2015-10-27	delivered	window_shopping	-1
2	6024938649550848	2015-10-27	delivered	window_shopping	-1
3	6191500064980992	2015-10-27	delivered	window_shopping	-1
4	5786846443339776	2015-10-27	delivered	window_shopping	-1

This data matches with the initial brief about how the data is intended to be delivered.

Next, I use other methods in pandas to check the size and type of data.

df.info()

```
&lt;class 'pandas.core.frame.DataFrame'&gt;
RangeIndex: 1244308 entries, 0 to 1244307
Data columns (total 5 columns):
email         1244308 non-null int64
event_date    1244308 non-null object
event_type    1244308 non-null object
action        1244308 non-null object
total         1244308 non-null object
dtypes: int64(1), object(4)
memory usage: 47.5+ MB
```

First thing to notice is that there are sizable amount of rows, but not unmanageable. Consideration for time should be made on certain methods, but nearly all Python operations on this data will be reasonable.

Next, The data types for 4 of the 5 columns are "non-null objects". The "event_date" and "total" will need to be further explored to be certain that the can be manipulated into Datetime and Float types respectively.

> *event_date* - Doing a sort on this column shows that all values are indeed timestamp like strings, which can be manipulated. Additionally, there are no NaN or null values listed here and thus, this column can easily be converted into a Pandas Datetime object for further manipulation.
>
> *total* - This column has both floating point values and strings. In sorting the data to view it, it looks as though there are a limited amount of values labeled "MISSING". Further, there are negative values and zero values.
>
>> Decision **-** The brief asked for several metrics concerning whether "*Conversions"* were made. I will interpret the term "MISSING" to mean only that dollar value wasn't available, but a purchase was still made. Thus, I will leave all values in the total column alone until I need to look at attributed revenue, then make a decision to remove, replace, or keep.

Next in the exploration process, I looked at the number of unique values across all columns, this gave me a sense of the data as a whole and confirmed for certain columns that there were *only* the intended choices, such as "Delivered, halted, or Purchased" in the event type column.

CONVERSION CALCULATION PROCESS

Knowing that the data was clean and as intended, I moved on to the process of calculating each conversion metric.

The data is grouped in three events and are flat. In order to run further analysis, I decided to temporary separate the data and prepare for calculations. I determined the next steps to be:

- Filter out two lists of delivered actions and a list of halted actions.
- Take a count of each list for unique emails and unique users
- Gather unique users from each list and use it to match users to purchased actions
- Build a python function to filter purchases from within 5 days of email and label as converted.
- Use these four lists to calculate metrics for table and sense check for accuracy

To make these comparisons in date across each table, the event_date column must be manipulated to become a pandas DateTime object. Pandas will quickly find a "Time delta" between dates for our boolean comparison.

BUILD "CONVERTED" FUNCTION

From here we need to see which purchases were made as a result of a previous action.

I chose to write a quick function to validate an email was sent within 5 days of a purchase (Quick to write. NOT quick to run… yet ;)).

The function starts with our purchased lists. It validates a date and email hash of a purchased row. It then uses that same email hash to parse the larger list of all email actions. It iterates each date of the emails until it meets the date criteria. At this point the function appends the purchased list with the information of the purchase being the result of a conversion.
The function is run twice. Once on the list of purchases from the BC cohort, the second time on the smaller control group.

The BC and Control purchased list are now easily filtered by conversions. This will allow for quick calculations to look for lift from the email campaign.

CONVERSION METRICS

Purchase Conversion Rate and Lift -

The first metric to produce is the purchase conversion for both cohorts and the % lift from the email action.

The method for this was simply to find the number of purchases in each cohort and divide by the amount of email or halted actions total in that cohort.

Considerations -
1. If more than one email was sent within the 5 day period before a single purchase was made.
2. If more than one purchase were made on the same day, each purchase is counted as a conversion.

Response -
I chose to let these two considerations as is. Potentially they may offset each other to a negligible response.

Further -
An interesting question to determine whether this is worthwhile in the future would be to look at how 2 emails inside of 5 days effect order value. If order value is higher, we might assume each email aided in the purchase.

Purchase Conversion Rates within 5 Days

		Converted	Total	Conversion Rate	% Lift
Purchase Conversion Rate	BC Cohort	14,459	38,298	37.75%	-38.00%
	Control Cohort	1,258	2,066	60.89%	

Customer Conversion Rate and Lift -

This metric is similar to the first on the brief, however here we are looking specifically at the unique number of customers which which were emailed/halted versus the unique number of customers that our email campaign helped to convert to a purchase and the subsequent lift created from the campaign.

I first filtered the purchased lists to show only the *converted* purchases. From there I made a list of the unique email addresses of each list and divided by the number of unique email addresses on the email/halted lists.

Considerations -

Customer Conversion Rates within 5 Days

		Converted	Total	Conversion Rate	% Lift
Customer Conversion Rate	BC Cohort	10,060	506,406	1.99%	-29.52%
	Control Cohort	806	28,596	2.82%	

Email Conversion Rate and Lift -

This metric is again similar to the purchase conversion rate but in reverse. It checks to see if more than one email happened inside the 5 day window of a purchase. In this metric each email counts toward the conversion of a purchase.

I will adjust the previous converted function to add a converted marker to each email. Then take a count of emails that converted divided by the total emails.

Considerations -
1. Multiple emails could be counted toward the same purchase.
2. Multiple purchases might happen in the 5 day window after an email

Response -
1. Multiple emails effecting purchases could be validated as important if we can also look at order volume when two emails are responsible for a single purchase. Additionally if we can see a pattern that a combination of emails results in a higher lift.
2. Our results show that this second scenario is more prevalent in the current data. There are more purchase conversions (14153) than there are email conversions (7,119). In this case I would like to question the data source about how purchases are recorded.

Single user journey with multiple purchases

	email	event_date	event_type	action	total
67529	5312915884212224	2015-10-01	purchase	-1	59.98
911850	5312915884212224	2015-10-05	delivered	window_shopping	-1

	email	event_date	event_type	action	total
1240953	5312915884212224	2015-10-12	delivered	abandoned_search	-1
470065	5312915884212224	2015-10-16	delivered	window_shopping	-1
120934	5312915884212224	2015-10-20	purchase	-1	179.98
120995	5312915884212224	2015-10-20	purchase	-1	59.99
683396	5312915884212224	2015-10-21	delivered	abandoned_search	-1

I looked at several users journeys in the data to see multiple purchases on the same day. At this volume it is unlikely this many people are making more than one purchase in a single day (shipping costs?). I might assume this is retail record keeping where an order shipped from two warehouses appears separately. Thus, I will leave the result and count the email as having converted only once until I learn more about the record keeping process.

Email Conversion Rates within 5 Days

		Converted	Total	Conversion Rate	% Lift
Email Conversion Rate	BC Group	7,119	1,062,584	0.67%	-48.27%
	Control group	1,384	106,866	1.30%	

Attributed Revenue and Lift -

This metric is looking at the total volume of purchases from within 5 days of an email or halted action.

Considerations -
Earlier in the process I saw that this “total” column was not a floating point number and additionally it had many values labeled “MISSING” or were less than zero, which should not be calculated as a purchase.

Response -
I have decided to remove all of these data points. I first looked at the volume of missing or zero values and see that in each list the number is less than 10% of the total number of values. The impact on the total rate should not be greatly effected.

Additionally, negative values in the total column (ie Returns) are generally consider faulty product and not bad marketing. Most retail firms don’t include these in AOV.

Further -
I could ask someone from the data source more about these values, if they can be corrected or if more information is known about how they came to be.

Attributed Revenue within 5 days

		Total Revenue	Total orders	Average Revenue per order	Lift
Attributed Revenue per Email	BC Group	$2,021,041.00	13,475	$149.98	-5.68%
	Control group	$183,340.00	1,153	$159.01	

NEXT STEPS

THOUGHTS

As mentioned in the brief, the data has produced some very strange outcomes. Outcomes that DO NOT pass the smell test. It seems that in all cases of my study, consumers getting an email have produced worse results!

The first response to this was to check and recheck my data collection, code, and maths. I have added more print statements in each process and made certain filters were pulling and comparing the correct groups of rows.

From here I would generally work with a teammate to ask for a fresh consideration to potentially see any errors I might be missing.

FURTHER EXPLORATION

1. The first question I would ask of the data to further analyze would be to understand if we have the correct response time. Is a window of 5 days too short or long to be compared to the control.

1 DAY
I tested this first. I had a hypothesis that the closer the window to the email event would prove greater difference in results. Humans truly influenced by the marketing are less likely to respond several days after a message.

Purchase Conversion Rates with 1 Day

		Converted	Total	Conversion Rate	% Lift
Purchase Conversion Rate	BC Cohort	7,418	38,298	19.37%	-56.31%
	Control Cohort	916	2,066	44.34%	
Customer Conversion Rate	BC Cohort	5,843	506,406	1.15%	-46.09%
	Control Cohort	612	28,596	2.14%	
Email Conversion Rate	BC Group	3,630	1,062,584	0.34%	-47.99%
	Control group	702	106,866	0.66%	
		Total Revenue	**Total orders**	**Average Revenue per order**	**Lift**

		Converted	Total	Conversion Rate	% Lift
Attributed Revenue per Email	BC Group	$1,002,090.00	6,915	$144.92	-3.79%
	Control group	$127,434.00	846	$150.63	
		2,477 records removed from BC and 151 records removed from Control			

10 DAYS

As the previous trial produced strange results I also ran a trial of the conversion metrics with an extended window. This also did not increase the lift to a position that I would hav assumed.

Purchase Conversion Rates with 10 Days

		Converted	Total	Conversion Rate	% Lift
Purchase Conversion Rate	BC Cohort	20,625	38,298	53.85%	-24.21%
	Control Cohort	1,468	2,066	71.06%	
Customer Conversion Rate	BC Cohort	13,392	506,406	2.64%	-18.86%
	Control Cohort	932	28,596	3.26%	
Email Conversion Rate	BC Group	10,711	1,062,584	1.01%	-48.83%
	Control group	2,105	106,866	1.97%	
		Total Revenue	**Total orders**	**Average Revenue per order**	**Lift**
Attributed Revenue per Email	BC Group	$2,836,717.00	19,207	$147.69	-4.90%
	Control group	$209,660.00	1,350	$155.30	
		2,477 records removed from BC and 151 records removed from Control			

2. Compare conversion metrics by action. Did our email events drastically change conversion for distinct actions? Run each of these metrics filtering for one of the four action types.

3. Potentially, we may need a larger Control group. The values don't seem to be in a statistical confidence interval to reject the null hypothesis.

4. Start with the WELCOME email. I would like to clean the data in a manner to view the customer journey from the welcome email. Perhaps the influence of the email marketing is more effective at creating customers than maintaining them.

Single user journey with purchase start

	email	event_date	event_type	action	total
75151	5994933019213824	2015-10-04	purchase	-1	99.99
937082	5994933019213824	2015-10-05	delivered	abandoned_search	-1
81259	5994933019213824	2015-10-06	purchase	-1	119.99
1070970	5994933019213824	2015-10-06	delivered	window_shopping	-1