Data-Driven Multi Touch Attribution

Daffney Viswanath
New York University
NY, USA
ddv246@nyu.edu

Gunjan Desai
New York University
NY, USA
grd285@nyu.edu

Priyal Nile
New York University
NY, USA
pan303@nyu.edu

Sonal Sharma
New York University
NY, USA
ss13449@nyu.edu

ABSTRACT

One of the toughest challenges in Digital Marketing is to gauge the impact of Digital Marketing Channels' contribution to customer conversion down the sales funnel. Several heuristics are used for attribution in practice; however, there is no formal justification for them and many of these fail in simple canonical settings and aren't correct representations of the customer's journey. Data-Driven Multi-Touch Attribution is the solution for analyzing the value of customer touchpoint/marketing channel that leads to a conversion. In this report, we discuss Shapley Values and Markov Chain Models along with Heuristics models. As per analysis, Shapley Model would be a good move towards the implementation of the first Multi-Touch Attribution at Policygenius Inc. Shapley values method is a straightforward approach to the attribution problem in which the insignificance of sequence makes it easier to implement. Its results are usually more stable in practice in a manner that given for a longer user journey, it always takes into account the unique channels and removes repetitive channels in finding attribution. Therefore, the results of Shapley values are in general less sensitive to the input data and would be good for analyzing longer user journeys.

KEYWORDS

Attribution Modelling, Heuristics Model, Markov Chain, Shapley Values, Google Cloud Platform, Big Query, Policygenius as PG, Channel Attribution, User Journey (Marketing Channel Subset)

1. INTRODUCTION

Attribution modeling is a strategy that allows marketers to analyze and assign credit to marketing touchpoints that occur at the specific steps of the customer journey, from searching for a product online to making a purchase, and every action in between. Using attribution models helps marketers better understand which parts of their marketing effort are driving the most leads to that part of the sales funnel. As a marketer running multiple campaigns on multiple platforms, it can be difficult to determine which mix of PPC keywords, display ads, landing pages, and SEO are generating leads that move efficiently through the sales funnel and down the conversion path. With attribution modeling, marketing teams get a bird's eye view of each customer journey from the starting point to the end purchase.

a. Company Introduction

Policygenius is America's leading online **insurance marketplace** with HQ in New York City & Durham, North Carolina. It is a one-stop-shop for all different insurances and their comparisons. Its mission is to help people get insurance right by making it easy for them to understand their opinions, compare quotes and buy a policy in one place. The objective of this project is to increase the effectiveness measurement of the marketing process at Policygenius Inc and this Project belongs to the Structured Marketing Team sector at Policygenius Inc.

b. Why Multi-Touch Attribution?

PG is already using First touch Attribution Modelling however, there are a number of shortcomings of this approach. Entire credit is given to the first digital marketing touchpoint whether or not it even drove the customer into the sales funnel. Also, the customer journey is generally through multiple channels so using a data-driven multi-touch attribution modeling is a better solution to address and answer many questions like How to spend Effectively for marketing channels? Which is the more efficient channel? Are there any trends we are observing? Are we somehow disregarding any channel that is giving us the exposure? How to increase the revenue by minimizing the waste of expenditures in marketing efforts and also to correctly penalize the channels which used first touch methods.

Thus, MTA is helpful for In-Channel Performance Optimization, Cross-Channel Performance Optimization and budget planning. We touch on each aspect of the Project in greater detail in the following subsections of this project paper.

2. PROBLEM DEFINITION AND ALGORITHM

a. Task

The goal of this project is to build a data driven MTA model to identify/prioritize the marketing channel that has the most impact on customer conversion thereby aid in a well - informed budget allocation and improve an in-flight campaign. This MTA model is for the life vertical within the first stage of funnel conversion.

b. Algorithm

Policygenius uses a single-touch attribution model. The major drawbacks of a single touch attribution model is that only the first touchpoint is taken into consideration. Moving from single-touch to multi-touch attribution modeling will allow all touchpoints of a customer's journey to be taken into account and assign fractional credit to each touchpoint so as to see how much influence each channel has on sales at PG.

3. EXPERIMENTAL EVALUATION

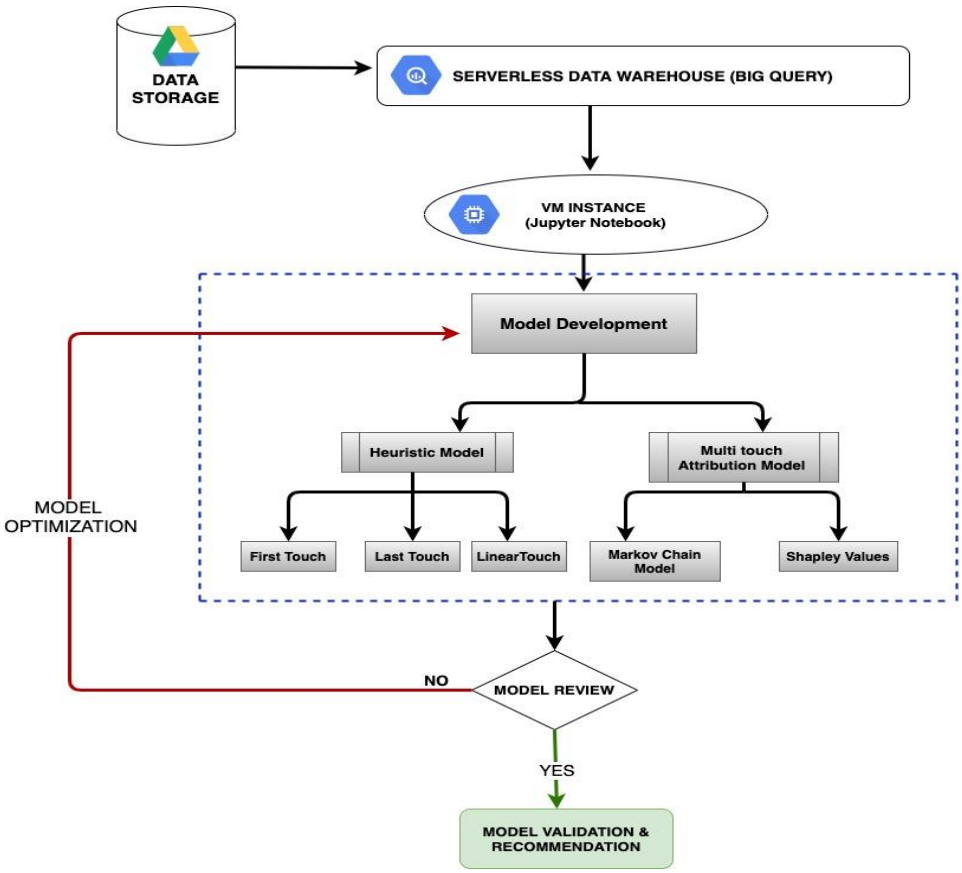
a. Data

The data was provided by the Policygenius Team. The dataset contains 27 million rows and 12 columns. The Dataset was ~ 6GB in size and provided valuable information necessary on life vertical within the first stage of funnel conversion. The Dataset was taken from the period of March 2020 to March 2021.

Sr No.	Column Name	Description	Pandas Data Type	Examples	Unique Entries
1	user_analytics_id	customer id	object	7e9e34533f6b42cd9f44fa41750d3e10	2302725
2	user_agent	device, os system, etc. used by customer to access PG Site	object	'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/85.0.4183.121 Safari/537.36'	105927
3	referred_to_url	site for touchpoint (the Marketing touchpoint/site from which user entered to PG)	object	'/life-insurance/signatures/basics-confirmation/630123e5-8b03-42e1-a9be-13da9a12a616'	98527
4	created_at	reference date	datetime64	2020-10-08 18:39:07 UTC	2465676
5	visitor_source	channel source	object	'Organic - SEO'; Partnerships; Email; Earned Media; Facebook; Referral; Paid Email; YouTube	24
6	conv_#_life	# stage of funnel (#1 is PRIMARY RESPONSE LABEL)	float64	1.0 (always this value)	1
7	conv_#_life_date	date of conversion	datetime64	2020-11-25 17:21:11.70; 2020-12-10 09:09:54.688	
8	exit_survey_source_category	exit survey response for source	object	google_or_other_search_engine; online_article; podcast; facebook; youtube; email; quora; instagram	34
9	exit_survey_submitted_product	exit survey product type	object	life; Home and Auto; disability; renters	4
10	exit_survey_source	exit source detailed response from source_category	object	other; ben_shapiro; policygenius_blog; youtube_video_ad; travel; the_classicist; nurses_on_fire	342
11	product_visited	product associated with referral site	object	other; life; home; auto	4

b. Project Workflow

Project Architecture:



i) Data Loading

Data was provided in Google Drive by the Policygenius team so this script was used to load the data from Google Drive to BigQuery. This creates a table named *mta_data*.

ii) Data Understanding and Data Cleaning

The Data has 26615013 rows and 16 columns. Conv_1_life, conv_2_life and conv_3_life columns were of type float while the id column was of type Int. The rest of the columns were of the object type. By checking for the percentage of distinct values in each column, we found that columns such as user_analytics_id, created_at and id had more uniqueness. This, in turn, proved that the columns we look more into such as visitor_source, product_visited had less unique values, meaning that the categories were repeated.

Basic Data Cleaning :

1. Filling null values: we replaced null values in exit survey(exit_survey_source_category, exit_survey_submitted_product, exit_survey_source), user_agent and referred_to_url columns with "unknown". Similarly for conv_1_life, the null values were replaced with 0. We replaced null values in referred_to_url with '/' For entries with visitor_source="Direct".

2. Convert to datetime: Columns that included date and time such as conv_1_life_date, conv_2_life_date, conv_3_life_date were converted to datetime format for future reference.

3. Identify the referred_to_url that contains "-life%" and if the product_visited for such URL is not already "auto" or "home" then we can safely categorize them as "Life" in product_visited

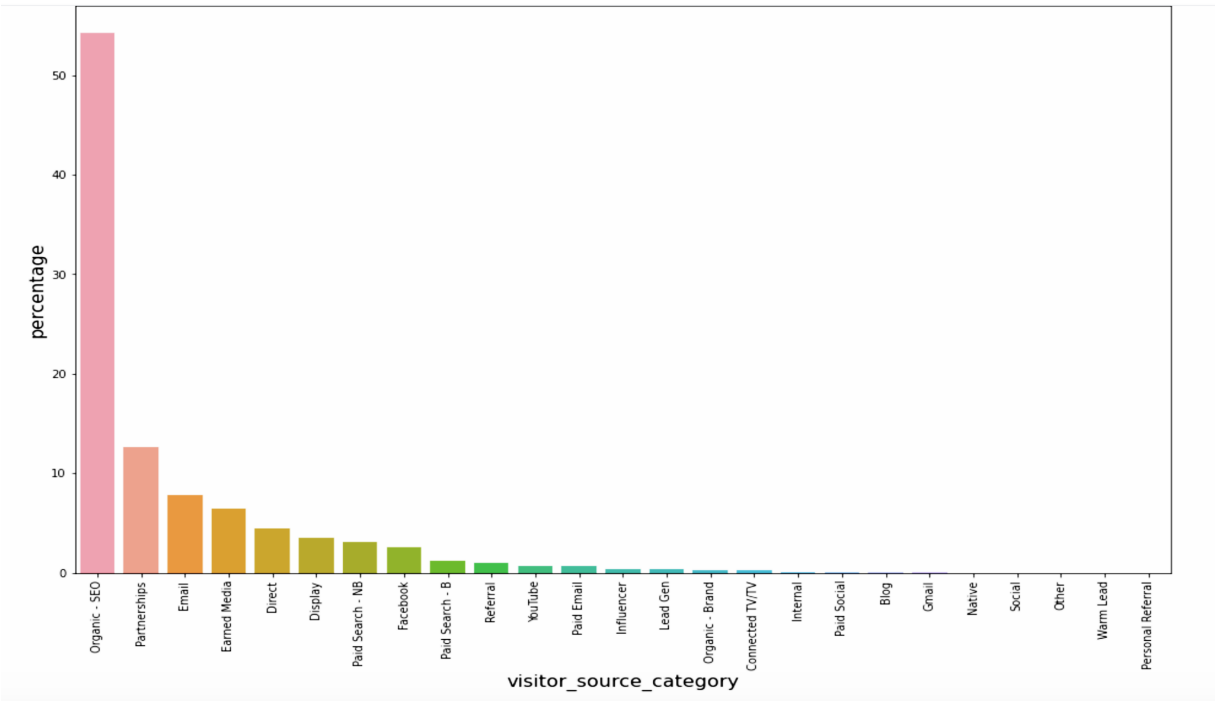
4. Calculate the time difference between the consecutive interactions for each user in time_diff_sec

For the Data cleaning step, this script was used to perform all data cleaning operations in one go. The script would create a BigQuery table named *mta_data_clean*.

iii) Univariate Analysis

In Univariate analysis, the number of unique categories or the cardinality of every column was checked. The level of cardinality depends on the percentage of distinct values in each column. As we expected, columns such as User_agent, referred_to_url had lower cardinality while user_analytics_id, created_at had higher cardinality.

Visitor_Source: It is one of the most important columns in the data. It has 25 different types of data. Organic-SEO has the highest percentage of data followed by the partnership(12%) and email(7%).



The Exit_Survey columns : There are three columns in exit_survey namely exit_survey_source_category, exit_survey_submitted_product, exit_survey_source. As mentioned in the data dictionary, Exit_survey_source_category is simply the category in which each exit_survey_source value falls into. Also the exit_survey_columns had almost 95% of null values which were replaced with unknown during data cleaning.

Product_Visited: This column has only 4 different types of data, life, home, auto, and other. About 50% of the data were of type other and 35% of type life. We are more concerned about Life vertical only.

iv) Hypothesis Testing

The categorical columns, visitor_source, exit_survey_submitted_product and Product_visited, require correlation check. This is performed to analyze if one column is dependent on another so as to be useful for predictive modelling, if used.

Chi-square Test: The Chi-square test of independence is a statistical hypothesis test used to determine whether two categorical or nominal variables are likely to be related or not.

Null hypothesis: There is no association between the two variables.

Alternative hypothesis: There is an association between the two variables.

Chi-square Statistic: The formula for calculating the Chi-square statistic (X^2) is shown as follows: X^2 = sum of [(observed-expected)^2 / expected]

The term ‘observed’ refers to the numbers we have seen in the contingency table, and the term ‘expected’ refers to the expected numbers when the null hypothesis is true.

we know that the 'observed' should be close to 'expected' under the null hypothesis which means X^2 should be reasonably small. When X^2 is larger than a threshold, we know the p-value (probability of having such a large X^2 given the null hypothesis) is extremely low, and we would reject the null hypothesis.

Visitor_Source vs Product Visited: We found that the **alternative hypothesis** is true, i.e there is an association between visitor_source and product_visited

Exit_survey_submitted_product vs Product_visited: We found that the **alternative hypothesis** is true, i.e there is an association between Exit_survey_submitted_product vs Product_visited.

v) Creating final data

Step 1: Use this script *(Create_User_Session.sql)*
For each user, we have created separate sessions based on their inactivity on the page. For example, if a user is inactive for more than 30 mins, we tag his next interaction as a new session. The above script assigns a session id in an incremental format for each user based on the time difference between their consecutive interactions. Also, we assume that a new session starts with a new day. The result of the script would be stored in a BigQuery table named *mta_data_session.*

Step 2: Use this script *(Aggregate_User_Sessions.sql)*
Once we have created unique sessions, we identified multiple interactions of every user in every session. In Aggregation, we have aggregated the interactions each user had within each session. This could also be thought of as removing duplicates. In the above code, we now have a unique entry for each user based on *user_analytics_id, Date, session_id, user_agent, referred_to_url, product_visited,* and *visitor_source* columns. The result of the script would be stored in a BigQuery table named *mta_data_agg_0.*

Step 3: Use this script (*Group_consecutive_channel.sql)*
Since we already have session ids for each customer session in place which indicates a changed session. Now, we create the running sum (creating a session_cumsum column) for the user session ids in order to get the distinctive sessions for the whole marketing journey of the user in an incremental manner. The result of the script would be stored in a BigQuery table named *mta_data_agg_1*.

user_analytics_id	session_cumsum	conv_1_life	visitor_source
59d97098bc4f482f9ccfbb93c6c7b021	1	0	Organic - SEO
59d97098bc4f482f9ccfbb93c6c7b021	2	0	Email
59d97098bc4f482f9ccfbb93c6c7b021	3	0	Organic - SEO

Step 4 (optional): Use this script *(Aggregate_User_Path.sql)* would create the user journey which comprises all marketing channels that a user has been through and whether or not that user has converted. Then, for each marketing channel subset, we'll find out how many conversions (conv_1_life column) and non-conversion (total_null column) have been derived from that channel. The result of the script would be stored in a BigQuery table named mta_data_agg_2. This step is optional since we have done this aggregation in Jupyter notebooks also for quick analysis through pandas.

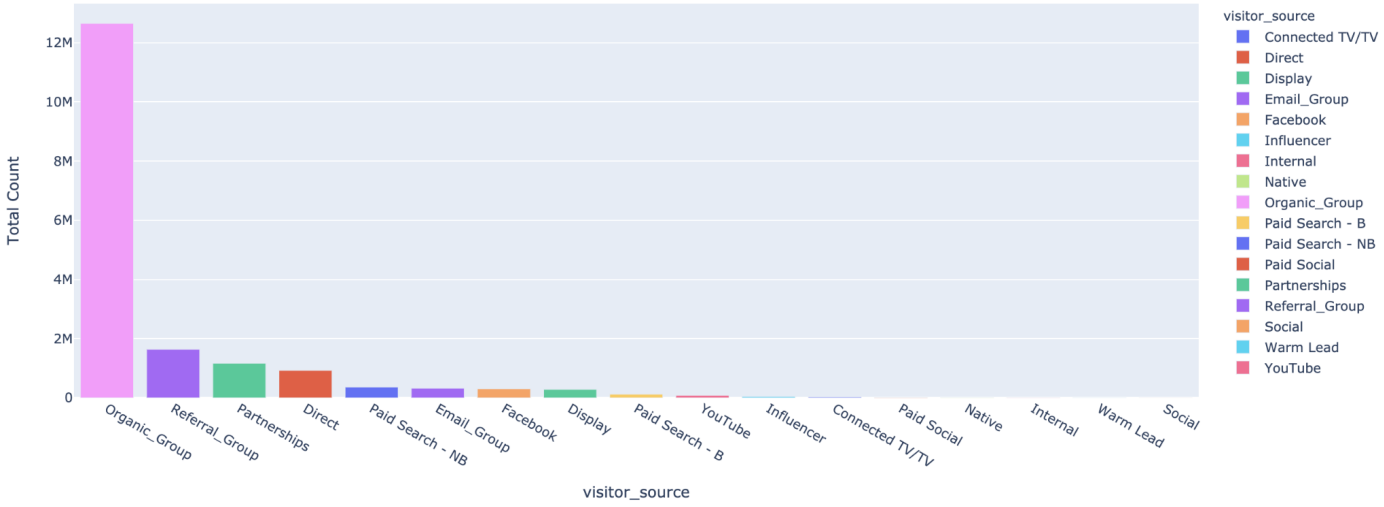
user_analytics_id	marketing_channel_subset	conv_1_life	total_null
4c031f666233482eaf8a8057e1920d87	Paid Search - B > Email > Referral > Organic - SEO	1	0
69da54250063421ba8c1ccae3a932d57	Organic - SEO > Paid Search - NB > Paid Search - NB	1	0

Then, we would analyse conversions and non-conversions driven through a particular marketing channel subset.

marketing_channel_subset	conv_1_life	total_null
Referral > Organic - SEO > Email > Paid Search - B > Earned Media	0	257
Facebook > Influencer	0	257
Earned Media > Internal > Organic - SEO > Direct	0	257

Because of computational issues, it was recommended by PG Team to group some similar channels. Therefore, the grouping has been done in the following manner as per PG Team Suggestions:

- Group Blog/ Organic SEO/ Organic - Brand => ***Organic_Group***
- Group Earned Media/ Lead Gen/ Personal Referral/ Referral/ Other/ Null => ***Referral_Group***
- Group Email/ Gmail/ Paid Email => ***Email_Group***



vi) Model

We have implemented the following models (Jupyter notebook can be found here):

1. Heuristics Model (Jupyter notebook can be found here)
    a. ***First Touch***

        This model gives all of the credit to the very first interaction the business had with a customer before they convert i.e. make a purchase. This model is beneficial to see what catches new customers' attention at the top of the funnel.

    b. ***Last Touch***

        This model assigns 100% of the credit to the last marketing touchpoint.

2. Multi-Touch Attribution Model (Jupyter notebook can be found here)

a. ***Linear Touch***

This model involves dividing the credit equally amongst all of the touchpoints in a conversion path or customer journey. For example, if there were 3 touchpoints in the consumer journey, each of those points would get 33.33% of the overall credit for that sale.

b. ***Markov Model***

It assigns the weights according to the probability of each event depending on the state attained in the previous event. It works on the principle of removal effect where it examines what would happen to the overall conversions if we took one of those channels out. The advantage of this model is that it allows you to evaluate the mutual influence of channels on conversions and find out which channel is the most significant. Therefore, it's suitable for businesses that have all their data collected in a single system. The disadvantage is that it underestimates the first channel in the chain and requires programming skills.

c. ***Shapley Model*** (Jupyter notebook can be found here)

Based on the Cooperative Game Theory, the Shapley Value is a solution concept of fairly distributing both gains and costs to several actors (here, marketing channels) working in coalitions (various ways in which channels interact with accounts throughout the buyer journey). The Shapley values apply primarily in situations when the contribution of each actor is unequal, but they work in cooperation with each other to obtain a payoff here, conversion of a customer into sales funnel).
Thus, Shapley Values provide a stable way to measure channel influence and fairly divide the credit for sales conversions between channels based on their individual contribution to total payoffs.

There is a slight tradeoff while building a Shapley Values Model (One of the Multi-Touch Model) as it does well up to 10-15 channels. After that payoffs per channel get exponentially complicated to compute. In order to address this issue, we worked with the PG Team to identify and group some channels as one and reduce the count from 20+ channels to 17.

vii) Insights / Key Takeaways

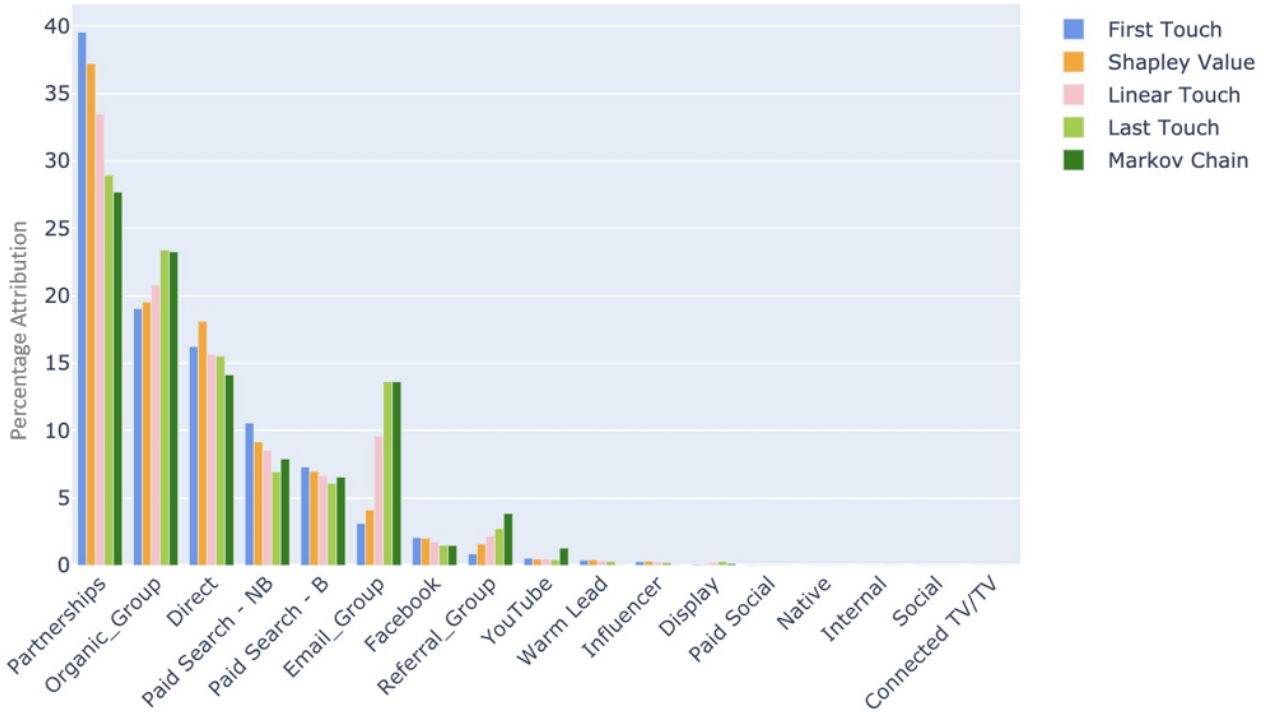
Final Data Frame :

After data cleaning, grouping and performing the data aggregation on the user level, below are the key findings for the user journey:

- Total conversions for the dataset: 47767
- Total data points in the dataset: 17200179
- Total conversion rate: 0.278%
- Total path before grouping: 14527
- Total path after grouping: 12299
- Total unique path after grouping for Shapley Model(unique channels): 3716
- Min length of a path (Min channels present in a single path): 1
- Max length of a path (Max channels present in a single path): 266
- Top Marketing Channel Subsets w.r.t Max Total Conversions are as follows:

Marketing Channel Subset	Total Conversions	Total Non-Conversions

Partnerships	12,313	2,003,632
Organic - SEO	5,433	12,826,015
Direct	4,622	569,543
Paid Search - NB	2,906	371,710
Paid Search - B	2,257	90,638
Organic - SEO > Partnerships	1,080	123,195
Email > Partnerships	1,070	64,375
Partnerships > Email	1,042	60,235
Partnerships > Organic - SEO	1,027	112,114
Direct > Organic - SEO	685	385,184



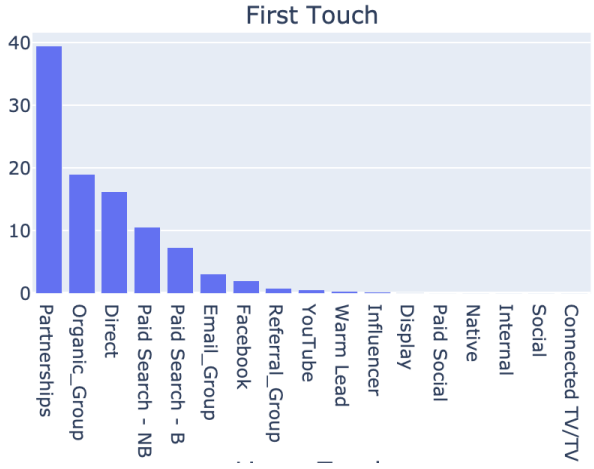
Channels	First Touch	Last Touch	Linear Touch	Shapley Value	Markov Chain
Partnerships	39.52728871	28.92582745	33.47905831	37.20379647	27.68143866
Organic_Group	19.03615467	23.37806435	20.7978273	19.52563291	23.24983943
Direct	16.23505768	15.52117571	15.66315104	18.10843362	14.12973667
Paid Search - NB	10.55749785	6.933657127	8.542126408	9.161868294	7.899807322
Paid Search - B	7.312579815	6.089978437	6.648782008	6.969731867	6.55105973
Email_Group	3.115121318	13.62446878	9.605998426	4.118443024	13.61592807
Facebook	2.059999581	1.492662298	1.741034998	2.023473638	1.477199743

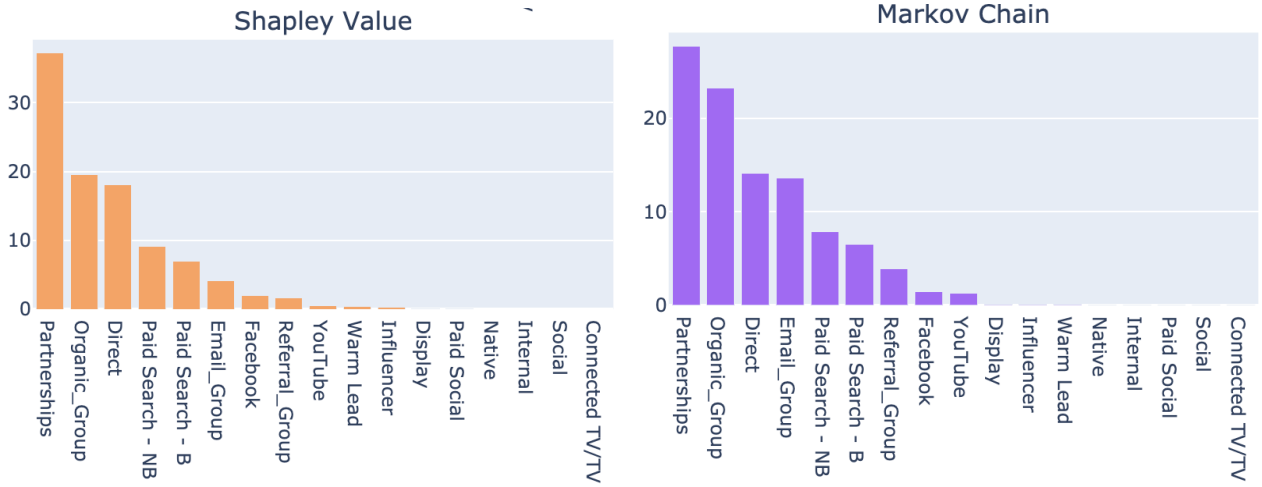
Referral_Group	0.860426654	2.73201164	2.183353042	1.600976949	3.853564547
YouTube	0.540121841	0.43335357	0.487400646	0.465442122	1.284521516
Warm Lead	0.389390165	0.309837335	0.335255029	0.41889791	0.064226076
Influencer	0.276341407	0.226097515	0.252370884	0.314773229	0.064226076
Display	0.073272343	0.303556849	0.240822816	0.074134853	0.128452152
Paid Social	0.016747964	0.014654469	0.017376013	0.014395117	0
Connected TV/TV	0	0.004186991	0.002337737	0	0
Internal	0	0.010467478	0.003105352	0	0
Native	0	0	0	0	0
Social	0	0	0	0	0

The table above summarizes the contribution of each marketing channel assigned by several models we implemented. The values in the table above are in terms of percentage %. We can see that "Partnership" was assigned the highest score by all the models.

Markov Chain:
Max attributed channels are as follows: Partnership (27.68%), Organic_group (23.24%), Direct (14.12%), Email_group (13.61%)

Shapley Values:
Max attributed channels are as follows: Partnership (37.20%), Organic_group (19.52%), Direct (18.01%), Paid Search - NB (4.11%).





Model Comparison:

Markov and Shapely models have their own pros and cons. However, after comparing the first model and Shapley values, the attribution for almost all channels (Except Direct) is comparably the same. Similarly, the Last Touch model and Markov Chain model are giving similar attribution for all channels. When it comes to a longer period of attribution window, shapely works better than Markov.

4. CONCLUSION

In conclusion, multi-touch attribution modeling is undoubtedly better than heuristic models as it removes the guesswork. Comparing Shapley value with Markov, Shapley Values works better than Markov for a longer period of attribution window. Also, Shapley Value model results are pretty much the same as the First Touch Attribution Model. Thus, Shapley Model can be a potentially good move for migrating PG's Marketing Business model towards Multi-Touch Attribution. Of course, there are limitations of this approach, because of the computational complexity (2^n) of Shapley Value Models above 15 Marketing Channels (n is the number of Marketing Channels).

Additionally, Shapley values method is a slightly more straightforward approach than Markov Model to the attribution problem in which sequence doesn't matter which makes it easier to implement. Its results are usually more stable in practice in a manner that given for a longer user journey, it always takes into account the unique channels and removes repetitive channels in finding attribution. Therefore, the results of Shapley values are in general less sensitive to the input data.

5. RECOMMENDATIONS

- Incorporation of cost and revenue to the attribution model would give correct insight on which part of revenue - channel attribution, incrementality of traffic source and better KPIs to compare different models attribution in terms of cost in addition to the weights being assigned.
- Multiple Marketing Event Data (Impressions, Clicks along with conversions), as well as cost data, **ROI** can be integrated to generate more insights. In the case of PG's in-house analytics, clicks can be incorporated into the data. The consistent UTM Tracking of online channels can be used to get the clicks.

- Additionally, incorporating bounce rates (soft bounce and hard bounce) will give insights into the true impressions on the website.
- With more growth of the PG in the long-term future in terms of revenue, size, and market share, the impressions (one of the costliest event data) can also be incorporated into the modeling research efforts.

6. ACKNOWLEDGMENT

We want to thank the several helping hands who contributed to the successful completion of this project. Professor Evan Korth, Teaching Assistant Nolan Filter for the continuous support and this opportunity through the ITP Course.

Special thanks to the Policygenius Team for helping us in every step of the way through their technical and business expertise and to help us to complete this project.

7. REFERENCES:

[1] Martijn van Otterlo, Grzegorz Chrupala, Peter van Eck, Evaluating attribution models on predictive accuracy, interpretability, and robustness,
[2] Kyra Singh, Jon Vaver, Richard Little, Rachel Fan, Attribution Model Evaluation
[3] Xuhui Shao, Lexin Li, Data-driven Multi-touch Attribution Models
[4] Ya Zhang, Yi Wei, Jianbiao Ren, Multi-Touch Attribution in Online Advertising with Survival Theory
[5] C. B. Yuvaraj1 , B. R. Chandavarkar2 , V. Santhosh Kumar3 , B. Satya Sandeep, Enhanced Last-Touch Interaction Attribution Model in Online Advertising
[6] https://github.com/jmwoloso/pychattr => Library used for Heuristic Models and Markov Chain model
[7] https://blog.dataiku.com/step-up-your-marketing-attribution-with-game-theory => Shapley Values Modelling
[8] https://medium.com/analytics-vidhya/the-shapley-value-approach-to-multi-touch-attribution-marketing-model-e345b35f3359 => Shapley Values Modelling