Supervised Learning for Attribution Analysis

Final Report: NBA Lead Reasons Attribution Analysis

Submitted to the Prof. Yuri Lawryshyhn as the Final Project Report as a part of the ESC490 Independent Study Project Evaluations

University of Toronto
Center for Management of Technology & Entrepreneurship

By Xuanze (Charlie) Li (1005854309)

Total Length: 37 Pages (Excluding the Title Page, Reference List and the Appendix)

Acknowledgement

First off, I want to thank Prof. Lawryshyn for providing me this valuable opportunity to learn more about the field of machine learning and data science. I have learnt a lot in terms of both technical skills and interpersonal communication skills by doing this project with the company sponsor. This project wouldn’t be possible without his support during our weekly meetings and therefore, Thank you dearly professor for providing me this opportunity to learn and grow as one of your students.

I would also like to thank Francis for helping me through this project. Tolerating my ignorance and himself under an extremely busy daily working schedule, he took countless hours out of his time explaining concepts to me, defining goals for each step of our project, providing essential feedback on my work and guiding me on our next steps. Besides, he is also a great friend who is constantly supporting me during my learning process. I started the project with only little knowledge about the subject area and ended with a strong grasp of many important attribution modeling techniques. Thank you Francis for helping me this much so far.

Table of Contents

I. Introduction

1.1 Background Information

Recently, one of CMTE's company sponsors asked us to provide an attribution analysis on its customer's come-back purchase reasons. Each buyer, also known as a customer of the company, comes to the company's advising teams who conduct thousands of tickets per year and we want to study the main reasons that led those customers to come back to purchase company products after they made an appointment with the advising team. However, the company is only interested in those advisors whose total gross purchase is greater than 10K monetary value to eliminate potential biases in the dataset. Therefore, the goal of this project is *to generate each lead reason's contribution towards the advisor's big tickets (>$10K)*. The study subject of this project is the closed Next Best Action (NBA) leads, which stands for the reasons/ behaviors which customers have accepted and made contact with later with the company, following the NBA lead outputs. The study uses all the data from 2021-01-01 up to 2021-10-31.

1.2 Data Pre-Processing & Model Selection

From the information we got directly from the sales team, there are 11 total NBA Lead Reasons, namely:

A. Short-Term -ve GS Momentum
B. RRSP/SPRING/SUMMER/FALL PROSPECT
C. Fund Reco
D. Service Standard / BDM Service Standard
E. Web Leads/ETF Web Leads/MM Web Leads/Liq Alt Web Leads
F. Webcast Reco
G. Trending Low in RRSP/SPRING/SUMMER/FALL GS
H. Redemption Risk
I. Short Term +ve GS Momentum
J. Webcast Follow Up / First Webcast Follow Up
K. High Fund/PM Concentration
L. Rec List

These 11 reasons are the ones contributed the most and occur the most frequent in the dataset. Therefore, we plan to evaluate each reason's contribution to the customer's final purchase through a rigorous mathematical approach. The models I selected and implemented in detail in this report are **1) First-Touch, 2) Last-Touch, 3) Linear, 4) Position-Based, 5) Time-Decay 6) Shapley Value, 7) Markov Chain, 8) Simple Probabilistic Model, 9) Bagged Logistic Regression Model & 10) Additive Hazard Model.**

The first five models are all heuristic models which determine each channel's contribution in a deterministic way (rule-based models). The last five models are all data-driven machine learning models which we used to extract more insights from the company's dataset.

1.3 My Contribution

In this project, my contribution primarily lies in researching and implementing different marketing attribution channels including the Shapley Value Model, Markov Chain Attribution Model, the simple Probabilistic Model, Bagged Logistic Regression Model & the Additive Hazard (Survival Theory) Model. Previously, 2 thesis students at the company have already implemented First-Touch and Last-Touch as the heuristic models to compare with the Shapley Value and Markov Chain models, I went deeper by also implementing several other well-known heuristic models (Linear model, Position-Based model and time-decay model) along with 3 more data-driven models (Simple Probabilistic Model, Bagged Logistic Regression Model and the Additive Hazard Model) to help the company better generalize the result. I also "ensembled" these models' results into one final model and made recommendations accordingly.

1.3 Objectives & Deliverables

1.3.1 Higher-Level Objectives (HO)

- Research and Develop more data-driven attribution models other than Shapley-Value & Markov Chains.
- Improve upon the Shapley Value & Markov Chain Model.
- Generate analysis for NBA Lead and make recommendations accordingly.
- Construct an Ensembled Model which can be used by the company in all attribution analysis problems, not just restricted to \marketing channel analysis and next best action analysis.

1.3.2 Lower-Level Objectives (LO)

- Run an Exploratory Data Analysis on the dataset
- Construct and Finish the testing/evaluation pipeline

- Enhance model development and compare with the old approach
- Research and implement more data-driven models

1.3.3 Deliverables

- Python code for the 10 models implemented (Including all datasets and models)
- Final presentation with the company's Business Analytics Team
- Final report detailing all results and implementations of the models.

II. Literature Review

I will begin the literature review section by providing a general overview of the marketing channel attribution research over the past years.

Over the past two decades, digital advertising has played an increasingly important role in each company's business analytics team. [1] According to some past research [1, 2, 3], Internet ad expenditure surpassed broadcast TV for the first time in 2013. The following two factors are the driving forces behind digital advertising's rapid rise. First, increased web usage allows the Internet to have a stronger influence on consumers' purchasing decisions. Second, digital advertising allows for better focused delivery and hence more successful advertising due to its capacity to track consumers' interactions with commercials. Textual ads, banner ads, rich media ads, and social network sites are just a few of the types accessible for digital advertising. Advertisements are provided through a variety of media channels, including search, display, social, mobile, and video. An advertising channel is usually made up of a series of linked ads that come in a variety of forms and delivery methods. How to allocate credits among multiple ad formats and ad channels to improve overall return on ad advertising has become one of the most important research questions in digital advertising. Many past researchers have tried to evaluate the effectiveness of advertising from some rule-based methods[1, 2, 3, 4]. However, the majority of available research relies on small-scale user surveys and ignores the interaction between different advertising channels.

In most cases, a single user is exposed to many advertising impressions supplied via various advertising channels. Understanding the contribution of various ads in leading consumers to desired behaviors such as clicking or making a purchase is known as attribution. Attribution has been identified as one of the most essential difficulties in digital advertising in order to quantify the impact of different adverts and hence optimize advertising channels. One of the early attribution models, Last Touch Attribution, assigns full credit to the last advertising a user sees before converting. Last touch attribution is widely used in practice and is included in most online analytics programmes as a standard attribution model. Despite its simplicity, one important downside of last touch attribution is that it only recognises the contribution of a single marketing impression for every conversion and does not credit the advertisement displayed before to the 'last touch.' In fact, conversion is frequently the result of a cascade of advertisements. Simply putting the credit on the last touch may overestimate the contribution of some types of advertisements, such as Search ads, which are triggered by user queries. In truth, the inquiry is frequently caused by previously watched adverts.

Multi-touch attribution, which distributes credit to all associated commercials based on their respective contributions, was launched lately and has quickly become a hot topic in digital advertising research. Last-touch model, first-touch model, linear model, and time decay model are some of the rule-based attribution models that have been presented, with last-touch and first-touch models being special examples of multi-touch attribution models [5, 6].

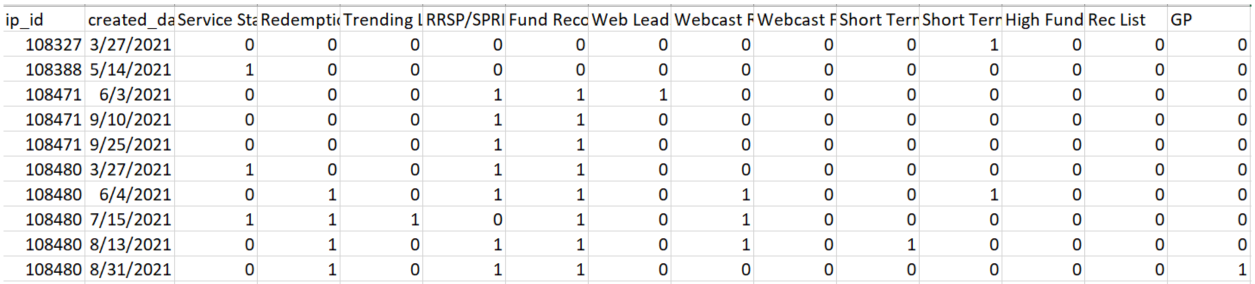
However, the disadvantage of the above rule-based models is that the rules are generated from simple intuition and may not be well suited to reality. With the capacity to track advertisement placement and users' interactions with advertisements improving all the time, a few data-driven multi-touch attribution models have recently been presented, which aim to infer the contribution from real user interaction data. Therefore, in this report, I will discuss 5 state-of-the-art data-driven attribution models along with their pros and cons.

Note that although this project is about NBA lead attribution analysis, we treat the problem the same as a general marketing channel attribution analysis as the main idea behind these problems are the same. Therefore, sometimes in the report I might accidentally use the word “channel” which actually stands for each NBA lead reason in the context.

III. Data Preprocessing

In this section I will describe my approach to clean the dataset and present my EDA results.

The data I extracted using SQL from the company databases are in the following format:

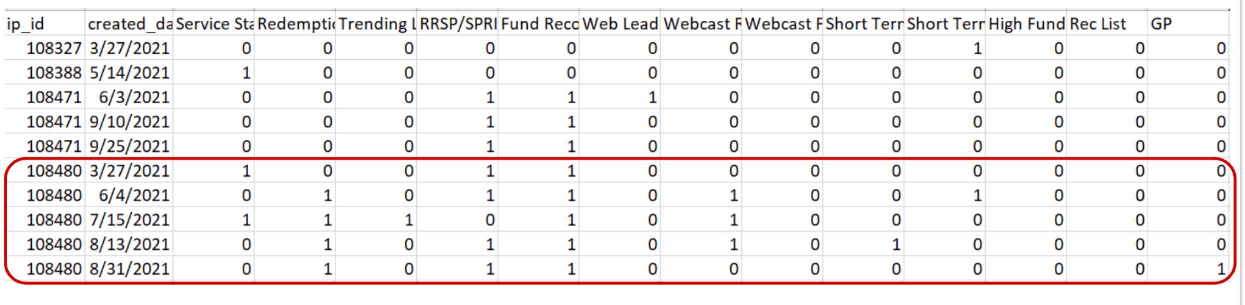


ip_id	created_da	Service Sta	Redempti	Trending L	RRSP/SPRI	Fund Reco	Web Lead	Webcast F	Webcast F	Short Terr	Short Terr	High Fund	Rec List	GP
108327	3/27/2021	0	0	0	0	0	0	0	0	0	1	0	0	0
108388	5/14/2021	1	0	0	0	0	0	0	0	0	0	0	0	0
108471	6/3/2021	0	0	0	1	1	1	0	0	0	0	0	0	0
108471	9/10/2021	0	0	0	1	1	0	0	0	0	0	0	0	0
108471	9/25/2021	0	0	0	1	1	0	0	0	0	0	0	0	0
108480	3/27/2021	1	0	0	1	1	0	0	0	0	0	0	0	0
108480	6/4/2021	0	1	0	1	1	0	1	0	0	1	0	0	0
108480	7/15/2021	1	1	1	0	1	0	1	0	0	0	0	0	0
108480	8/13/2021	0	1	0	1	1	0	1	0	1	0	0	0	0
108480	8/31/2021	0	1	0	1	1	0	0	0	0	0	0	0	1

Figure 1. Raw datasets for lead reasons and customers gross purchases

Here, as shown above, the first column is the customer's IP id, the second column being the contacting date when a company's sales representative has made a contact with the customers, the third to the thirteenth column being the name of the lead reasons and the last column being the gross purchase done by the customer. The data is in binary form where 0- means no lead reasons/no purchase, and 1 means a lead reason/a purchase is made.

After this step, I took the raw dataset and reformat them using a data_reformmater.py file to form the purchasing path of each IP by considering all possible permutations of lead reasons which could lead to a GP (gross purchase) in the end. Here, taking the 108480 as an example:



ip_id	created_da	Service Sta	Redempti	Trending L	RRSP/SPRI	Fund Reco	Web Lead	Webcast F	Webcast F	Short Terr	Short Terr	High Fund	Rec List	GP
108327	3/27/2021	0	0	0	0	0	0	0	0	0	1	0	0	0
108388	5/14/2021	1	0	0	0	0	0	0	0	0	0	0	0	0
108471	6/3/2021	0	0	0	1	1	1	0	0	0	0	0	0	0
108471	9/10/2021	0	0	0	1	1	0	0	0	0	0	0	0	0
108471	9/25/2021	0	0	0	1	1	0	0	0	0	0	0	0	0
108480	3/27/2021	1	0	0	1	1	0	0	0	0	0	0	0	0
108480	6/4/2021	0	1	0	1	1	0	1	0	0	1	0	0	0
108480	7/15/2021	1	1	1	0	1	0	1	0	0	0	0	0	0
108480	8/13/2021	0	1	0	1	1	0	1	0	1	0	0	0	0
108480	8/31/2021	0	1	0	1	1	0	0	0	0	0	0	0	1

Figure 2. Sample Path Formation raw Dataset

A potential path here for this person would be Service Standard -> Redemption Risk -> Seasons Prospect -> Webcast Reco -> Fund Reco -> GP or Service Standard -> Redemption Risk -> Seasons Prospect -> Webcast Reco -> NULL. I considered the path formation in both cases either lead to a conversion or not. Therefore, for this particular simple example, there could be potentially more than 182250 paths in total. (Details please refer to presentation slides 20).

After forming the paths for the customers, I did exploratory data analysis on the raw dataset where I first looked at the percentage of each lead reason which could lead to a conversion in the end.

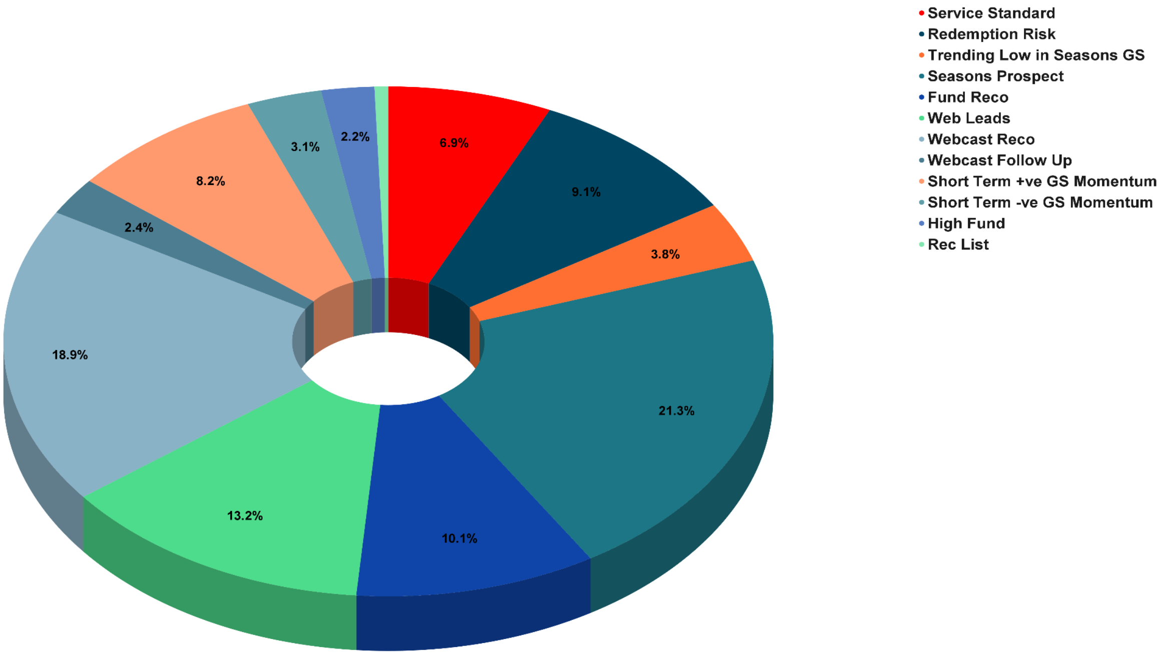


Figure 3. Percentage Pie Chart of all lead reasons which lead to a conversion

Here, as shown by the pie chart above, it is evident that the top three lead reasons appear in the path/raw dataset are **Webcast Follow up, Webcast Recommendation and Web Leads**, which are all website related lead reasons. The significance of this step is that after we get all raw lead reasons results using the 10 models we developed, it is essential that we normalize the frequency of the results for each model using the frequency of each lead reason appearing in the raw dataset. Because some lead reason's contribution is reduced by their relative low frequencies in the raw dataset. Therefore, we want to normalize the results using the frequencies of the lead reasons that appear in the dataset.

In addition, I also studied journal length distributions. I plotted the typical journey length for a customer which lead to a conversion as below:

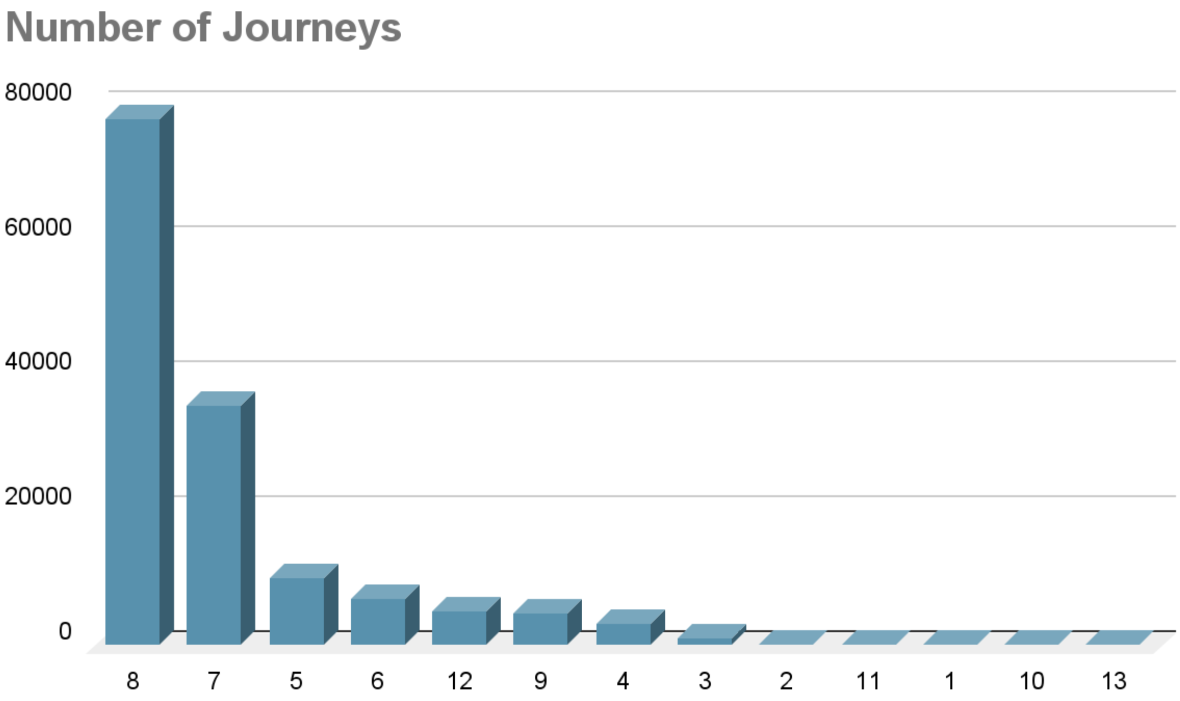


Figure 4. Journey Length Frequencies Chart

As shown above, a typical journey length which leads to a conversion in the end is 8. This means that on average, the company needs the marketing team to contact the customers 8 times to make a purchase, which is in line with what we get from the company.

IV. Heuristic Models for Attribution Analysis

4.1 Introduction

In this section I will introduce the five heuristic models I implemented as the base-line models for the NBA weighting analysis.

4.2 First-Touch Model

The first-touch model assigns all credit to the first touched points in the marketing channel analysis. In our case, we assign 100% credit to the first lead reason which leads to a conversion in the end.

4.3 Last-Touch Model

The last-touch model assigns all credit to the last-touch point in the marketing path. In our case, we assign 100% credit to the last lead reason in the path which leads to a conversion.

4.4 Linear Model

The linear model assigns all credit equally in a purchasing path to all lead reasons. In our case, we assign equal percentage credit to all lead reasons which lead to a conversion in a path.

4.5 Position-Based Model

The position based model assigns 40% credit to both the first and last lead reason in a path, and equally divided the remaining credits among the remaining lead reasons in a purchasing path.

4.6 Time-Decay Model

The time-decay model considers the effect of the time-decay in a consumer's purchasing path. For example, we would like to assign more credits to the last lead reasons than the second last lead reasons in a purchasing path. I achieved this criteria by assigning a gradual time-decay rate (10% decay per previous lead reason) and assigning weights accordingly in each consumer's purchasing path.

4.7 Implementations (Code)

Please refer to the Code **MLA_Lead_Attribution_with_Normalization_Charlie.ipynb** for implementation details. The code is double checked by Francis and no errors are found.

4.8 Results & Discussion for each individual models

Here I documented all the results for each model and provided detailed discussions on the results and the models' pros and cons.

4.8.1 First Touch Result

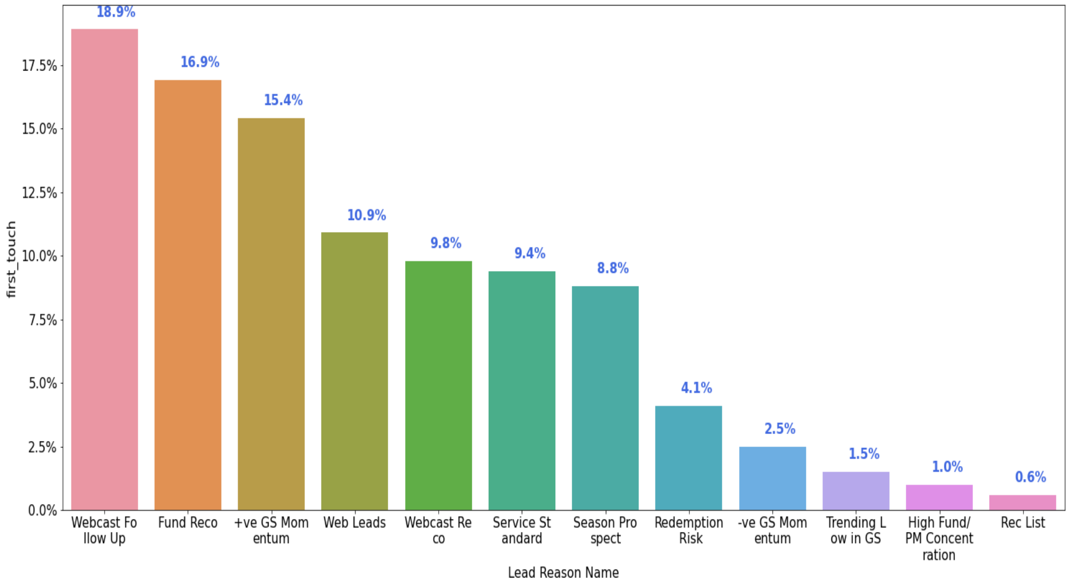


Figure 5. First-Touch Model Results

Discussions: According to the first-touch model, the top 3 lead reasons are Webcast Follow up, Fund Recommendation and +ve GS momentum (18.9%, 16.9% and 15.4% each). It intuitively makes sense as the company should focus more fund recommendations, follow up more on the website after customers made an appointment with the sales teams and keep increasing the momentum for gross sales. However, the first-touch model fails to consider the purchasing path of each individual customer and therefore, it should be treated carefully and more insights should be added to the model.

4.8.2 Last-Touch Result

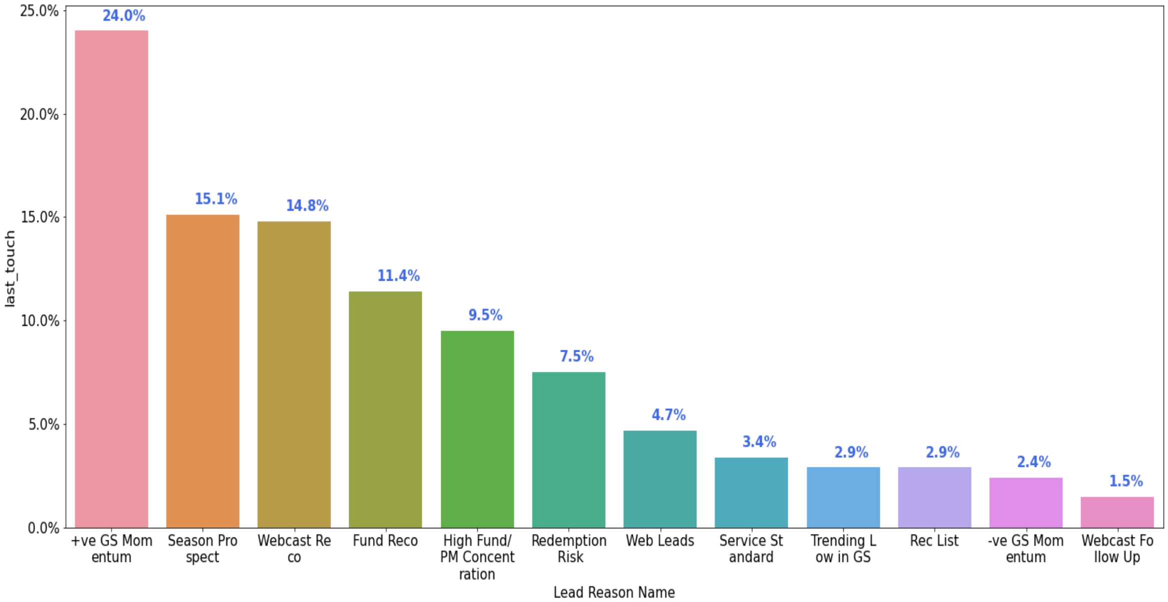


Figure 6. Last-Touch Model Results

<u>Discussions:</u> Based on the results above, the last-touch approach recommends that the marketing team should focus more on increasing the +ve GS momentum (increase gross sales), keep the season's prospects and continue to do more webcast recommendations. This result also makes sense as recently the company has decided to focus more on increasing +ve GS momentum and do more fund recommendations in general. However, similar to the first-touch model, this model aslo only considers one touchpoint and fails to consider all paths in a customer's journal. Therefore, a more complex model such as linear and time-decay model should be used as a baseline for this project.

4.8.3 Linear Model Results

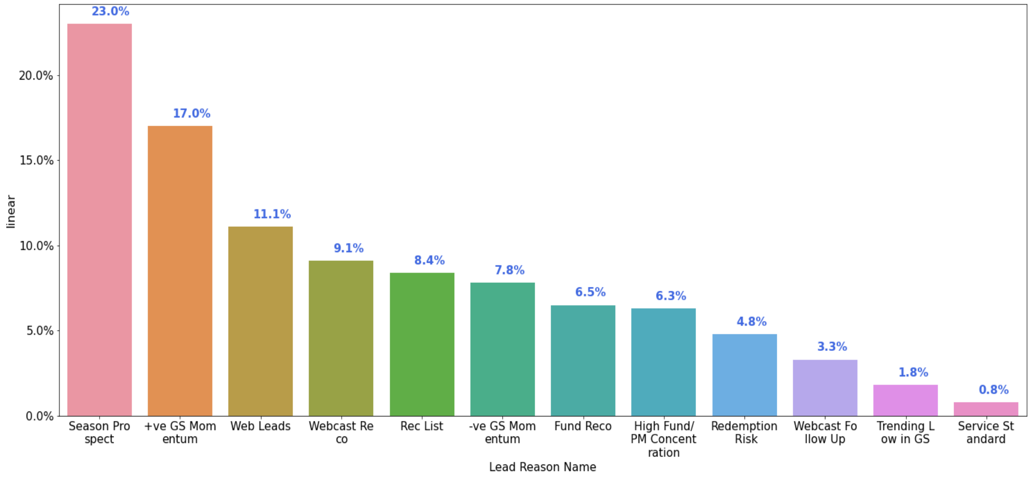


Figure 7. Linear Model Results

<u>**Discussions:**</u> The linear model recommends that it is best to focus on the season's prospect side and continue to increase the gross sales momentum. In addition, it provides a new lead reason that might be worth investigating in – the web leads. Web leads are those leading advertisements published online which would generally lead to a customer's come-back purchase. However, this model fails to consider how the purchase time would play a role in the customer's journey and therefore a more complex model should be used as well.

4.8.4 Position-based Model Results

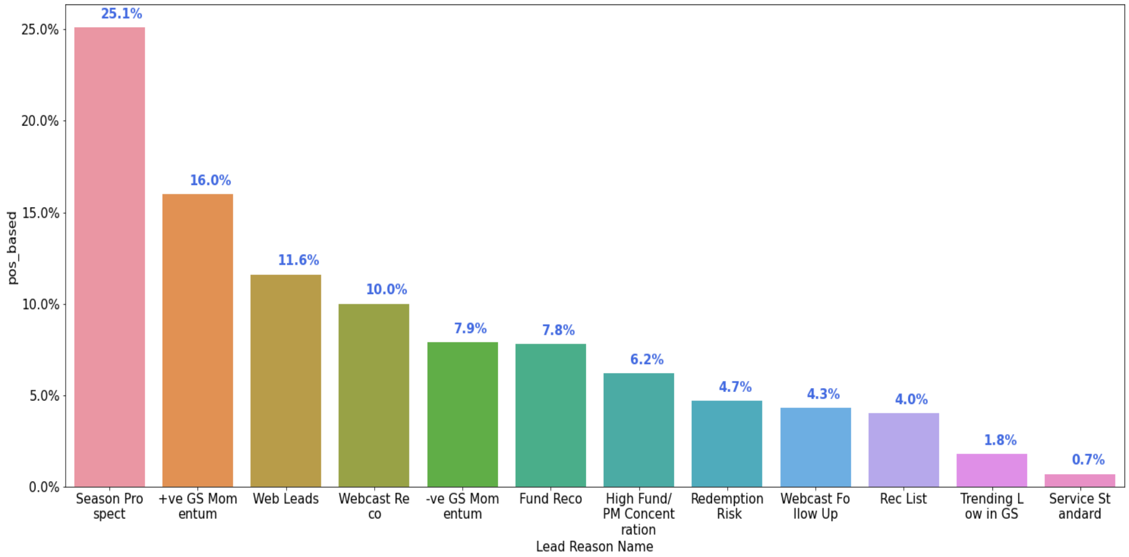


Figure 8. Position-based Model Results

<u>Discussions:</u> The position-based model produces really similar results compared with the linear model as it uses a combination of linear model and first-last touch. This approach is apparently better than the previous ones as it assigns different weightings to different channels based on their position in a purchasing path. It has several practical implications: first off, considering a customer visits the company several times during a year. Even though he might make a purchase at the end of the year, it is very hard to tell if it is the last interaction that helps him make the decision or it is because the series of interactions with the marketing team helped him make the decision. Thus, we would prefer a model that can emphasize these types of information in the model. Although it is unclear if the first and last channel are really that powerful in driving a conversion, we still accept this approach as it gives a better understanding for the customer's journey than the previous ones. Overall, we decide to accept this model as it produces valuable insights as a baseline model.

4.8.5 Time-Decay Model Results

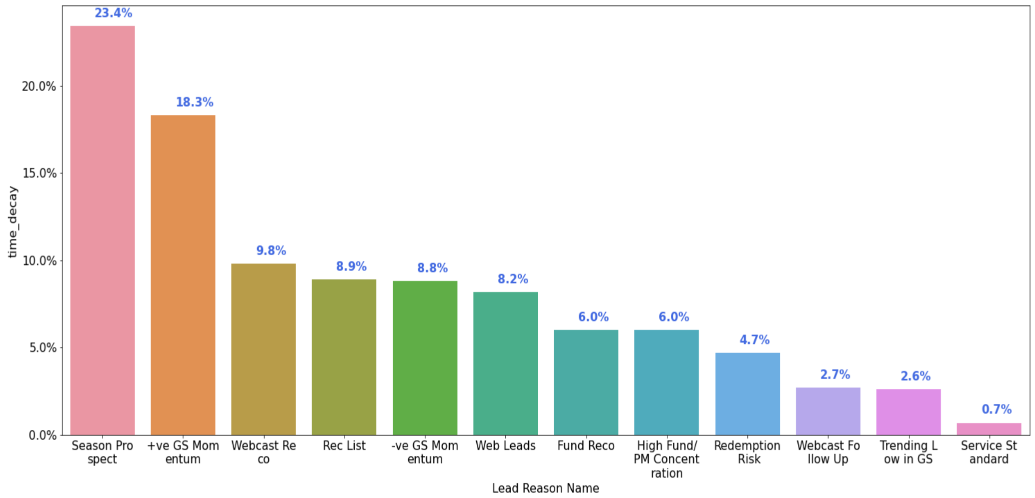


Figure 9. Time-Decay Model Results

Discussions: This time-decay model also produces very similar results compared to the above baseline models. It recommends the company to focus more on the season's prospect, +ve GS momentum and Webcast Recommendation which inline with what the company is asking for. This model is by far the best heuristic model that we can find online as it considers the time influence on the channels on a gradual basis. It takes the journey's length as an input and assigns credits by gradually decreasing each channel's contribution on an arithmetic series basis. Overall, it is the best heuristic model and this should be used as a main reference when comparing results with the data-driven models. In addition, the company also accepts this model as its baseline model and the results are inline with what they expected.

4.9 Overall Discussion on these Baseline Models

In summary, as demonstrated by the above results, the five heuristic based-line models all produce very good grounding results which is in line with what the company is currently expecting. I picked these models based on their popularity among the current research direction in the marketing channel attribution field and verified their effectiveness with the company dataset. However, none of these models consider the attribution problem from a dynamic point of view and therefore, they should only serve as baseline models in this project to compare with more complex data-driven models.

V. The Shapley Value Approach for Attribution Analysis

5.1 Introduction

Developed by the economics Nobel Laureate Lloyd S. Shapley, the Shapley Value is now commonly used as an approach to fairly distributing the output of a team among the constituent team members. In game theory, the Shapley value is a solution concept of fairly distributing both gains and costs to several actors working in coalition. The Shapley value applies primarily in situations when the contributions of each actor are unequal, but they work in cooperation with each other to obtain the payoff. [7] Nowadays, this approach is usually used in marketing attribution problems, where it is used to determine a value attributed to each "channel" which later leads to a sale in the sales team. The basic idea behind it is that it takes the weighted average of its marginal contribution over all possible coalitions for each channel. For example, imagine a user has touched multiple channels before making a purchase at a company. He might search the company's website, click the promotion email sent by the saling company or even just click the advertisement from instagram which later directs him to the purchasing page. Each of these emails, websites and advertisements are considered as a single channel in the problem and the Shapley value approach basically uses a complex mathematical approach to get the average of its marginal contribution for all possible combinations for each channel.

Following the exact same idea, we propose to use this method to also study the Next Best Action Lead reasons after the customer has made an appointment with one of our company advisors and made the purchase as the problem is the same just with a different context.

5.2 Methodology & Mathematical Derivation of the Shapley Value

Shapley Value Approach (SVA) treats each NBA lead reason as a player in a cooperative game. It directly measures each reason's contribution to the sales amount, and takes the mean of all added contributions as attribution values. Specifically, given a grand coalition P consisting of lead reasons $\{x_1, x_2, \dots, x_p\}$, we use a utility function $v(S)$ to describe the contribution of S which denotes a coalition of reasons. The Shapley value can be calculated through the following formula,

$$\phi_j = \sum_{S \subseteq P \setminus \{x_j\}} \frac{|S|!\,(p-|S|-1)!}{p!} \Big(v\big(S \cup \{x_j\}\big) - v(S) \Big),\ j = 1, \dots, p, \qquad (1)$$

Equation 1. Formula used to calculate the Shapley Values for each channel. [1]

where $|S|$ is the cardinality of coalition S and the sum extends over all subsets S of P not containing lead reason x_j. $v(S \cup \{x_{\ j}\}) - v(S)$, which is also denoted as $M(j, S)$, is the marginal contribution of reason x_j to the coalition S.

As noticed in Equation (1), Shapley value method takes the weighted average of its marginal contribution over all possible coalitions for each lead reason. The coalition's contribution and marginal contribution can be measured in multiple ways to gain insight from different aspects. In the following sections, we will discuss the definition of contribution and marginal contribution in the Shapley value approach.

5.2.1 Contribution

The utility function $v(\cdot)$ measures the contribution of a lead reason's coalition to the total sales in the absence of the rest. For the grand coalition P that includes all the reasons in the channel, $v(P)$ should equal to the total gross product created by the entire channel. The channel value is generated by all the converted customers that have made a contract with the company. In reality, most converted customers make decisions through a particular subset of lead reasons even though they might have all available reasons in P. If we define u_j to be the set of customers who have lead reason x_j before making a purchase (they may or may not have other lead reasons), the total channel value is generated by the users is

$$U_P = \bigcup_{j \in P} u_j, \qquad (2)$$

which contains all the converted customers. These customers can be further grouped into multiple customer types according to the lead reasons they have. For example, user type {1} contains all the customers who have made purchase with lead reasons x_1 only, user type {2} contains all the customers who have made purchase with lead reasons x_2 only, user types {1,2} contain all of the users who have made purchase with both lead reasons x_1 and x_2, and so on. Consequently, we can define the contribution of P as

$$v(P) = \sum_{S \subseteq P} R(S), \qquad (3)$$

where *R*(*S*) is the total value created by the customers who have made purchase with all lead reasons in *S*. We then name it as the individual contribution of coalition *S*. Therefore, *v*(*P*) contains the contributions from all of the converted customers. It is also the total channel value and the total credits to be allocated in our lead reasons attribution problem. Similarly, for each coalition *S* ⊆ *P*, the contribution of *S* made by the users is

$$U_S = \bigcup_{j \in S} u_j \setminus \bigcup_{k \notin S} u_k. \qquad (4)$$

This means that the contribution of any coalition *S* is measured as the value created by the customers who have converted with some reasons only in *S*. These customers shall not have any other reasons outside of *S*. Therefore, for any two coalitions *S*1 and *S*2 satisfying *S*1 ⊆ *S*2, we have,

$$U_{S_1} \subseteq U_{S_2}. \qquad (5)$$

Similar to (3), we define,

$$\nu(S) = \sum_{T \subseteq S} R(T). \qquad (6)$$

In practice, *R*(·) could be the total number of reaches, total number of conversions or revenue/profit. We hereafter use revenue as our total sales amount. Our analysis can be extended to other cases as well.

5.2.2 Marginal Contribution

We quantify the additional contribution of a reason (*x*<j) when added to a coalition (*S*) as its marginal contribution to *S*.

$$M(j, S) = \nu(S \cup \{x_j\}) - \nu(S). \qquad (7)$$

In practice, this means we add a new channel to an existing channel that contains all the reasons in *S*. This new reason may likely bring new customers who are not aware of the company products. It may also further influence the existing customers who have been exposed to some of the lead reasons in *S* or convert the customers directly regardless of their prior experience with the company. On the other hand, availability of additional reasons does not necessarily mean that every member of the population will continue to be targeted by all existing reasons plus this new reason. Some customers will convert right after having these new reasons before they have those previous reasons. A bulk of the customers will not get an impression from this new reason because they convert before this new reason comes to their mind.

The contribution $v(S \cup \{x_{\text{j}}\})$ is made by the following three types of customers.

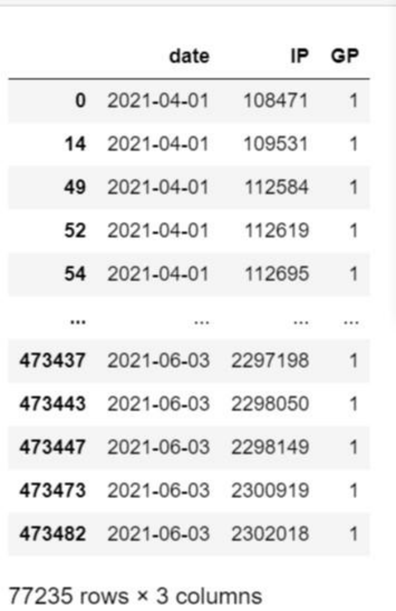
- Type 1: Customers who have some reasons in S only.
- Type 2: Customers who have reason x_{j} only.
- Type 3: Customers who have reason x_{j} AND some other reasons in S.

Type 1 customers contribute to $v(S)$, which is part of $v(S \cup \{xj\})$ and eventually part of $v(P)$. Type 2 customers contribute to $M(j, S)$, i.e., the individual contribution from reason x_{j} . Lastly, type 3 customers contribute to $M(j, S)$ as well, which is regarded as the advertising synergy among x_{j} and the reasons in S.

*Note that the marginal contribution is generated by additional customers (types 2 and 3). Therefore, $M(j, S)$ should always be non-negative.

5.3 Data Processing

First, we use SQL to select and extract the data from the database and re-ordered them into a python dataframe as follows:



```
              date       IP  GP
     0  2021-04-01   108471   1
    14  2021-04-01   109531   1
    49  2021-04-01   112584   1
    52  2021-04-01   112619   1
    54  2021-04-01   112695   1
   ...         ...      ...  ...
473437  2021-06-03  2297198   1
473443  2021-06-03  2298050   1
473447  2021-06-03  2298149   1
473473  2021-06-03  2300919   1
473482  2021-06-03  2302018   1

77235 rows × 3 columns
```

Figure 10. Raw data from the company's database

Here, “Date” represents the usable data from 2021-04-01 to 2021-06-03 for all the people who made a purchase and those advisor’s tickets are greater than 10k. “IP” represents each customer’s personal ID and “GP” represents their gross purchase expressed in a binary form, where 1 stands for a purchase and 0 means no purchase.

Then, we sorted the data based on the customer's ID and the ticket’s created date, and later we summed up the reasons for each customer on each day and reordered the dataframe. Finally, we re-ordered the data frame to get the coalition of reasons for each IP and their corresponding Gross Purchase value, which was later used to calculate the conversion sum and the shapley value for each reason using the approach defined above.

5.4 Model Implementation

Please refer to the Code **MLA_Lead_Attribution_with_Normalization_Charlie.ipynb** for implementation details. The code for the Shapley Value Model is double checked by Francis and with many online resources and no errors are found.

5.4 Results & Discussions

The reason behind using the Shapley Value Approach (SVA) is that it applies primarily in situations when the contributions of each actor are unequal, but they work in cooperation with each other to obtain the payoff. In our scenario, the actors in the SVA are the generated lead reasons and it fulfills the situation where SVA should be applied as all these reasons work in cooperation to the customer's final purchase but their individual contribution is unknown.

Based on the method defined above, with the dataset range from 2021-01-01 to 2021-10-31, we get the lead reason’s contribution distribution below:

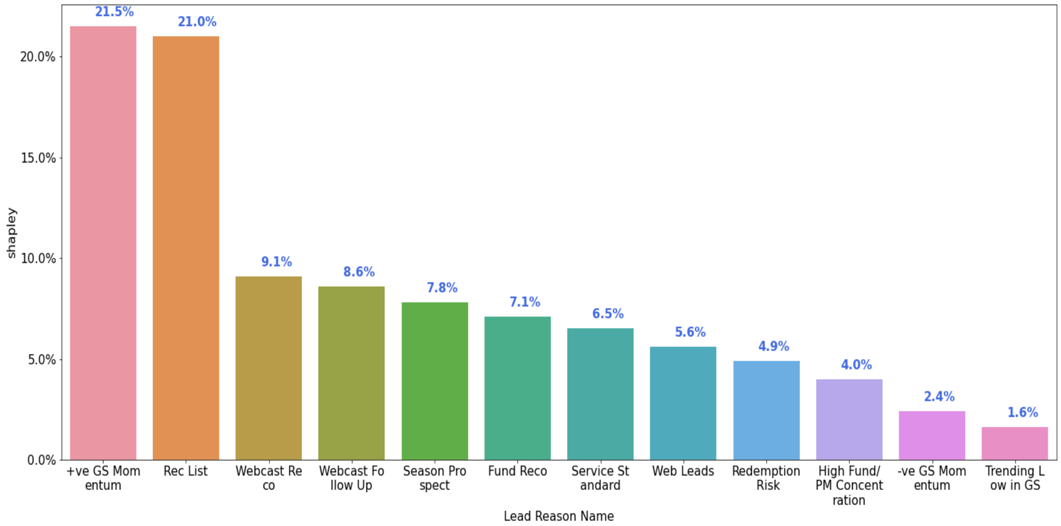


Figure 11. Results of Lead Reason Analysis using the Shapley Value Approach

Here, as shown above, the lead reason which contributes the most to customer's come-back purchase is +ve GS Momentum - 21.5%. Followed by the second and third leading reasons being Rec List - 21.0% and Webcast Reco - 9.1%. Compared with the baseline model, this result meets our expectation in a way that the company is indeed currently focusing on increasing the positive saling momentum and the fund recommendation list. This method has its advantage in it considers how each channel works together to generate a conversion rather than simply assigning a fixed rule to the model. The results are therefore highly data-dependent and it produces a new perspective for us to focus on improving the gross sales.

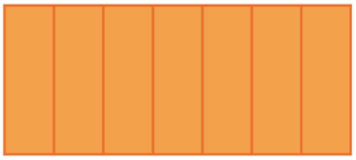
5.5 Limitations

Despite its evident advantages, the Shapley Value approach also has some drawbacks. First off, the model defined above does not consider the sequence or frequency of the input data. Results derived from a method like this will not give useful insights on how multiple contacts made by a single customer on a single day will lead to a purchase or vice versa. Therefore, further investigations in the sequence and frequency of input datasets should be studied.

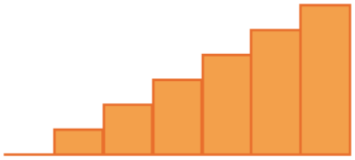
VI. The Markov Chain Attribution Model for Attribution Analysis

6.1 Introduction

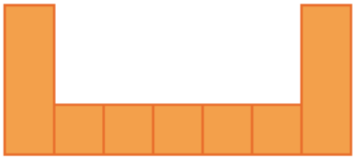
As discussed above, all attribution models have their pros and cons, but one drawback the heuristic models have in common is that they are rules based. The user needs to decide up front how they want the credit for sales events to be divided between the touchpoints. As stated above, some heuristic models we used in this project include:



Linear: credits an equal share of the value between all touchpoints



Time-decay: credits a decreasing percentage of value the further away in time a touchpoint is from the sales event

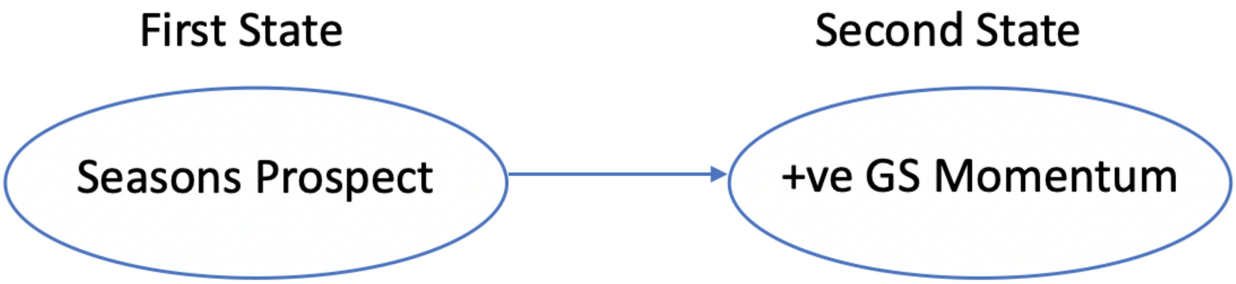


Positional: credits 40% to the first and last touches, and the remaining 20% is evenly distributed to the touches in between

Luckily apart from the shapley value model, there are more sophisticated data-driven approaches that are able to capture the intricacies of customer journeys by modeling how different channels actually interact with customers, and each other, to influence a desired conversion. A data-driven model such as Markov Chain Attribution Model provides marketers with deeper insight into the importance of channels and channels, driving better marketing accountability and efficiency. Therefore, in this section, I will present a data-driven channel attribution method using the Markov Chain.

6.2 Methodology & Theory

Markov Chain is a probabilistic model that represents all customer's journeys as a graph, with the graph's nodes being the "markov states", and the graph's connecting edges being in our case, a customer first demonstrated a seasons prospect as its first lead reason (first state) then transit to +ve GS momentum (transition) where they ended up booking another appointment with the company sometime in the future (second state). [6]



The transition probabilities are the model's most important component (the likelihood of moving between states). The number of times a customer has changed states is translated into a probability, and the entire graph can be used to assess the value of each stage and the most likely paths to success (conversion).

A Markov chain model can be used to measure the importance of each channel by calculating the **Removal Effect**. A channel's effectiveness is determined by removing it from the markov graph and simulating customer journeys to measure the change in conversion rate without it in place. Removal Effect is a proxy for weight, and it's calculated for each channel in the Markov graph.

Therefore, to calculate each channel's attribution value we simply use the following formula: ***V = T * (Rt / Rv)***

- ***V*** = Channel's attribution value
- ***T*** = Total value to divide. For example, the total revenue of all customer's journey which lead to a GP as input to the Markov model
- ***Rt*** = Channel's Removal Effect
- ***Rv*** = Sum of all Removal Effect values

To construct the model, the input data is first grouped by an IP ID in the order of visits; each IP ID is then associated with a list of the channels they interacted with in the order of their interaction. The last element of each list is either "Conversion" or "Null", depending on whether the user concluded their customer journey by one of the valid conversion types or did not convert. An example of what this journey might look like is the list ["Seasons Prospect', "+ve GS Momentum", "NULL"].

After that, I formed the pairs between each two possible channels in the format of ("Seasons Prospect", "+ve GS Momentum") and the complete universe of journeys is iterated over when the set of pairs is generated, and the number of times a certain state transition occurs is counted. The counters for transitions states like ["Seasons Prospect" -> "+ve GS Momentum"] and ["Web Reco" -> "Conversion"] would all increment by one using the same example customer journey given in the previous paragraph. Each count of state transitions is divided by the total count of state transitions with the same source state at the end of this phase to determine the probabilities associated with each edge of the Markov Chain. The transition matrix, M, is therefore in the form

$$M = \begin{bmatrix} p_{a,a} & p_{a,b} & p_{a,c} & \\ p_{b,a} & p_{b,b} & p_{b,c} & \cdots \\ p_{c,a} & p_{c,b} & p_{c,c} & \\ & \vdots & & \ddots \end{bmatrix},$$

where each entry $p_{i,j}$ represents the calibrated probability of transition occurring between source state i and destination state j. Once this transition matrix is constructed, we use the removal formula defined above to determine the relative importance of each channel to driving conversions. After this step, bhy doing a series of matrix manipulations and found the eigenvalues, we can finally calculate the attribution weight for each channel as

$$\mathrm{w_i} = \frac{r_i}{\sum_j r_j}.$$

Each attribution weight $\mathrm{w_i}$ denotes the credit of the channel I in driving conversions in the existing universe of customer journeys, as well as the percentage of conversions that can be attributed to that channel. And these weights are identical in nature to those of the shapley value & heuristic models.

6.3 Model Implementation

Please refer to the Code **MLA_Lead_Attribution_with_Normalization_Charlie.ipynb** for implementation details. The code for the Markov Chain Model is also double checked by Francis and cross-validated with many existing code online and no errors are found.

6.4 Results & Discussions

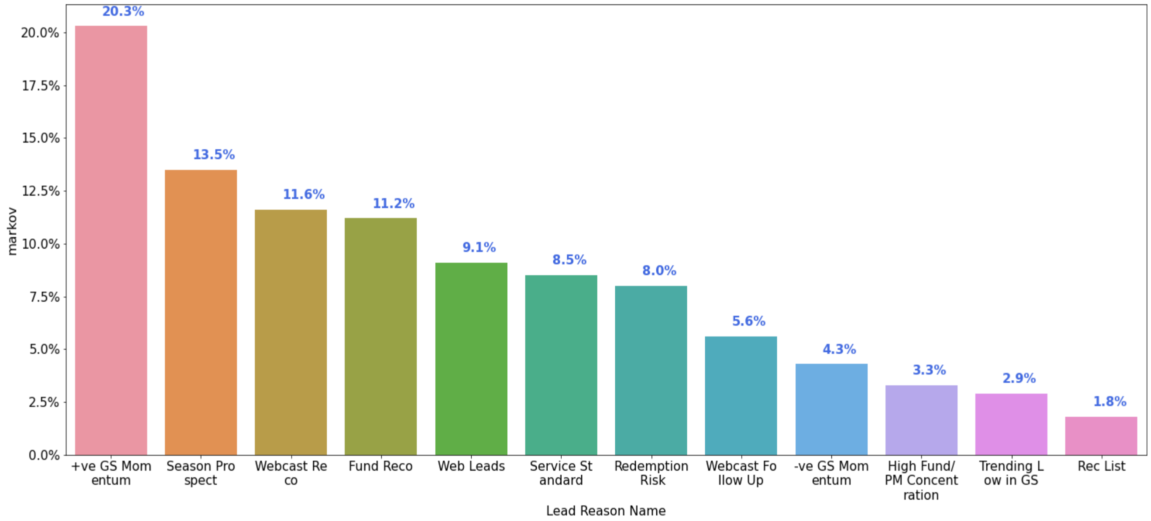


Figure 12. Markov Chain Attribution Model Results

As it can be shown above, the markov chain model also recognizes the +ve GS momentum, Season's Prospect and Webcast Recommendation as its top 3 NBA leads. This result is cross-validated through all of our previously implemented models and the results are evenly distributed and no skewed results produced. Therefore, we accepted the model results and included it into our final ensembled results.

VII. Simple Probabilistic Model for Attribution Analysis

7.1 Introduction

According to a paper published by Shao and Li in 2011 [6], they developed a simple probabilistic model based on a combination of first and second-order conditional probabilities to calculate each channel's attributions. Simplicity of model translates to minimal estimation variability and ease of understanding, although accuracy is sacrificed. The probabilistic model is generated using the following steps:

Step 1: For a given data set, compute the empirical probability of the main factors,

$$P(y|x_i) = \frac{N_{positive}(x_i)}{N_{positive}(x_i) + N_{negative}(x_i)} \qquad (1)$$

and the pair-wise conditional probabilities

$$P(y|x_i, x_j) = \frac{N_{positive}(x_i, x_j)}{N_{positive}(x_i, x_j) + N_{negative}(x_i, x_j)}, \quad (2)$$

for i ≠ j. Here y is a binary outcome variable denoting a conversion event (GP or non-GP), and x_i i = 1, . . . , p, denote p different advertising channels. $N_{positive}$ (xi) and $N_{negaitive}$ (xi) denote the number of positive or negative users exposed to channel i, respectively, and $N_{positive}$(xi, xj) and $N_{negaitive}$(xi, xj) denote the number of positive or negative users exposed to both channels i and j.

Step 2: The contribution of channel i is then computed at each positive user level as:

$$\begin{aligned} C(x_i) \quad &= \quad p(y|x_i) + \frac{1}{2N_{j \neq i}} \sum_{j \neq i} \Big\{ p(y|x_i, x_j) \\ &\qquad - p(y|x_i) - p(y|x_j) \Big\}, \qquad (3) \end{aligned}$$

where $N_{j \neq i}$denotes the total number of j's not equal to i. In this case it equals to N-1, or the total number of channels minus one (the channel i itself) for a particular user.

The model is essentially a second-order probability estimation. There is a substantial amount of overlap between the influences of different channels due to similar lead reasons such as Web leads and Webcast Reco and customer's exposure to multiple NBA leads. As a result, including the second-order interaction variables in the probability model is vital. Theoretically, as the paper suggests, we can get to third-order, fourth-order, or higher order terms. But the model will crash as we increase the order of probability estimations. [6] As a result, attempting to estimate the empirical probability with the third or higher order is of little practical benefit. In addition, we make a key assumption in the probability model: the net effect of the second-order interaction is distributed evenly across the two factors. We believe this is the minimum assumption we need to make based on the Occam's Razor concept [6], as there is no data evidence to imply otherwise. Focusing on the first and second-order terms also helps to keep any assumptions to a bare minimum — splitting the effect in third-order interactions, for example, can be more dangerous than splitting the effect in second-order interactions.

7.2 Model Implementations

Please refer to the Code **MLA_Lead_Attribution_with_Normalization_Charlie.ipynb** for implementation details. I manually code this part by myself and checked with Francis. The code should function properly without error, although there is no resource online to compare as no one attempts to replicate this method before.

7.3 Results & Discussions

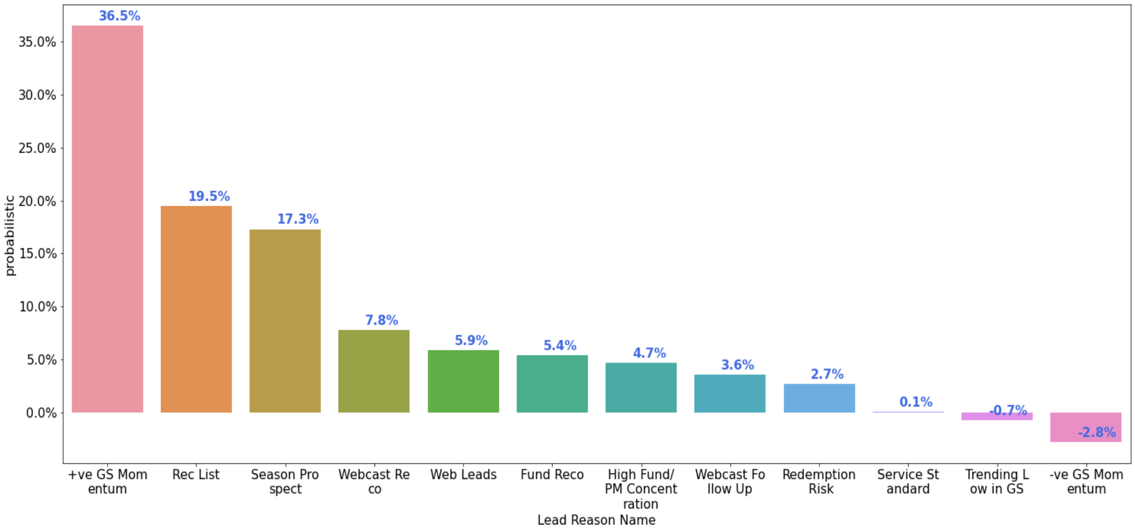


Figure 13. Simple Probabilistic Model Results

As shown above, the simple probabilistic model shows a different perspective than the other models. It gives the Rec_List 19.5% attribution credit which is more than doubled than its results from previous models. Furthermore, the result is very skewed as it assigned nearly 75% to the top three lead reasons while only giving 25% to the remaining channels. In addition, the model gives a -2.8% attribution to the -ve GS momentum. A negative number in an attribution model analysis problem means that the channel does not contribute positively to the overall gross purchase. This intuitively makes sense as a -ve GS momentum means that the marketing team fails to keep up their saling momentum in the recent month and thus they need to increase their sales in the future. The results obtained from this model are very close to what the company Business Analytics People presumes and therefore, we decide to keep the model as a potential candidate for the ensembled model.

VIII. Bagged Logistic Regression Model for Attribution Analysis

8.1 Introduction

Also Proposed by Shao & Li in the paper published in 2011 [6], I manually implemented the bagged logistic regression marketing attribution model specified as below. The bagged logistic regression is fitted using the following steps:

- **Step 1:** For a given data set, sample a proportion p_s of all the sample observations and a proportion p_c of all the covariates. Fit a logistic regression model on the sampled covariates and the sampled data. Record the estimated coefficients.
- **Step 2:** Repeat Step 1 for 2000 iterations, and the final coefficient estimate for each covariate is taken as the average of estimated coefficients in 2000 iterations.

The sample proportion p_s, the covariate proportion p_c, and the number of iterations are the parameters of the bagged logistic regression. According to the paper, it is observed that, for a range of values of p_s and p_c that are not close to either 0 or 1, the bagged logistic regression yields similar results. Besides, the results are not overly sensitive to the choice of number of iterations.

The reason behind selecting the logistic regression model is that since our customer purchase data and each generated lead reasons data are binary, we want to explore the part where we also consider the sequence of purchase instead of blindly doing statistical weighting analysis. Logistic Regression Models are used extensively in classification problems and therefore we want to use it as a baseline model to compare with the results derived from the Shapley Value , Markov Chain and other models.

8.2 Model Implementation

Please refer to the Code **MLA_Lead_Attribution_with_Normalization_Charlie.ipynb** for implementation details. I manually code this part by myself and checked with Francis. The code should function properly without error, although there is no resource online to compare as no one attempts to replicate this method before.

8.3 Results & Discussions

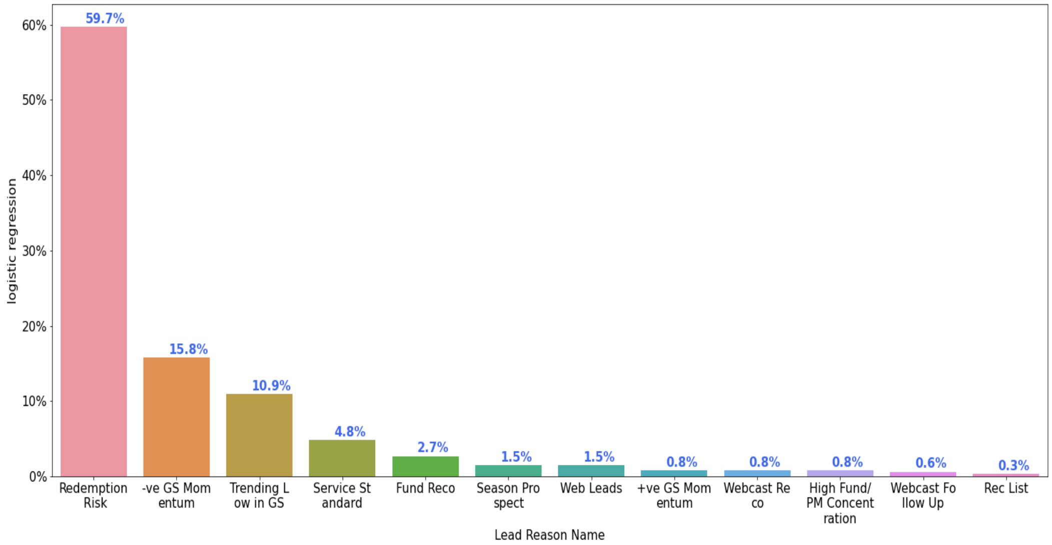


Figure 14. Logistic Regression Model Results

As shown above in the graph, the logistic regression does not yield a similar result compared with other models. It recommends the marketing team to focus more on redemption risk, -ve GS momentum and Trending Low in GS, which intuitively does not make any sense. The -ve GS momentum stands for a negative trend in recent gross sales which shouldn't be something that the company tries to improve upon. In addition, the result is highly skewed as it assigns almost 60% to redemption risk, a lead reason which no other model ranked as top 3. However, I still include it as part of the final model as it gives a different perspective on the NBA lead analysis.

IX. Additive Hazard Model for Attribution Analysis

9.1 Introduction

In 2014, Zhang, Wei and Ren [8] showed a promising result in combining attribution analysis with survival theory models. In their model, they demonstrated that by adopting a probabilistic framework, the proposed model is able to remove the presentation biases inherent to most of the other attribution models such as the simple probabilistic model and the bagged logistic regression model. In addition to modeling the attribution, the proposed model is also able to predict the user's 'conversion' probability.

9.2 Methodology & Theory

9.2.1 Survival Theory

A. The Survival Function

The survival function is defined as:

$$S(t) = Pr(T > t),$$

where t is some time and P r stands for probability. The survival function is the probability that the time of 'death' is later than some specified time t. And then the cumulative distribution function F(t) is defined as:

$$F(t) = Pr(T \leq t) = 1 - S(t).$$

If F(t) is differentiable, the probability density function (p.d.f) f(t) is :

$$f(t) = \frac{d}{dt}F(t).$$

B. The Hazard Function

Formally, we define the hazard function (or instantaneous rate of occurrence of the event) as:

$$\lambda(t) \quad = \quad \lim_{dt \to 0} \frac{Pr(t \leq T \leq t + dt | T > t)}{dt}$$

Furthermore, we have

$$\begin{aligned}
\lambda(t) &= \lim_{dt \to 0} \frac{Pr(t \le T \le t + dt | T > t)}{dt} \\
&= \lim_{dt \to 0} \frac{Pr(t \le T \le t + dt) / Pr(T > t)}{dt} \\
&= \lim_{dt \to 0} \frac{(F(t + dt) - F(t)) / S(t)}{dt} \\
&= \frac{f(t)}{S(t)} \\
&= -\frac{S'(t)}{S(t)}.
\end{aligned}$$

And by doing some substitution of variables, we can finally get:

$$S(t) = \exp(-\int_0^t \lambda(u) du).$$

9.2.2 Additive Hazard Model for Attribution Analysis

Now using a similar approach defined above, I followed the theory presented in the paper published by Zhang, Wei and Ren, I manually implemented the solution for additive hazard model and compared it with some online existing code to make corrections. I was able to fully implement the model and get the result. I will not show the full derivation of the weight parameters here as it is a very long and tedious linear programming problem and requires optimization using simplex methods. For detailed derivation, please refer to the original paper.

9.3 Results & Discussions

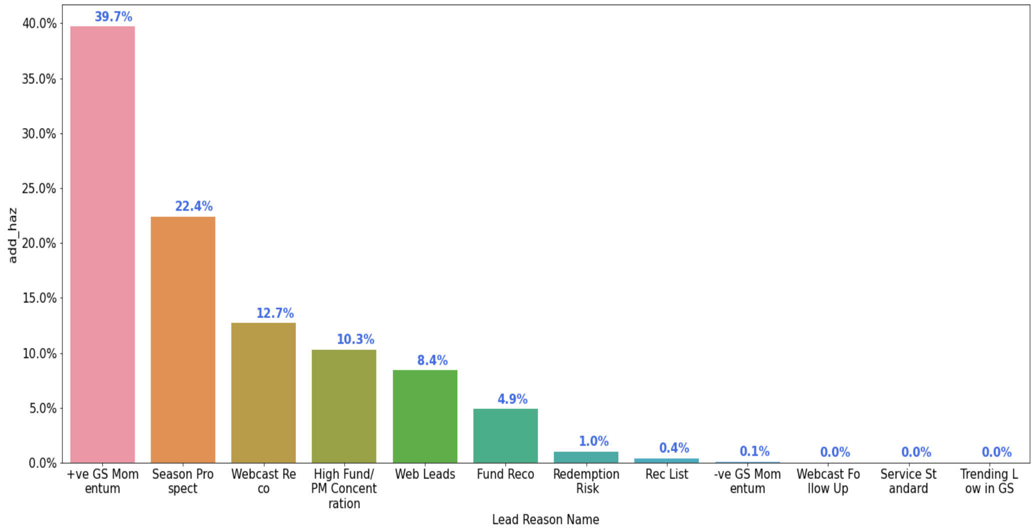


Figure 15. Additive Hazard Model Results

As shown above, the additive hazard model predicts the top 3 lead reasons to be +ve GS momentum, Seasons Prospect and Webcast Reco, which is in line with what we have for the majority of previous models. One downside of this approach is that the result is highly skewed, it tends to assign most credit to the top channels and leaves everyone else 0.0%. However, I still include it into the final ensembled result as it strengthened the company's current focus and is in line with the baseline models.

X. Ensembled Results

10.1 Introduction

The reason that we consider more than one model is the following: Digital advertising relies on a fair amount of subjectivity. Having 10 different modeling approaches give the company marketing team the flexibility to choose and combine the results if necessary. Heuristic Models are widely used in industry and should be generally considered into the ensembled results to provide a baseline estimation. The Shapley Value and Markov Chain models are the most widely used data-driven attribution models available these days and thus they should also be considered. The Probabilistic Model is more accurate and more flexible with a larger number of covariates, but is slightly more difficult to interpret. On the other hand, the bagged logistic regression model and the Additive Hazard model are less accurate but produce another angle of looking at things. In addition, the result from these models can cross-validate the general conclusion reached in the overall NBA lead attribution analysis.

10.2 Methodology & Approach

After extensively discussed with Francis and the BA team at the company, I conclude that the following method is the best for producing the ensembled results:

Table 1. Individual Model Weightings for the ensembled results

Model Name	Percentage
First-Touch	2%
Last-Touch	3%
Linear	5%
Position-Based	5%
Logistic Regression	5%

Time-Decay	10%
Additive Hazard	10%
Probabilistic	20%
Shapley Value	20%
Markov Chain	20%
Total	**100%**

I made the above decision largely due to the fact that after playing extensively with the models and the data the company has, the Shapley-Value, Probabilistics and Markov Chain models are among the top 3 most stable models. Their results are in line with what the company has in mind and meets their expectations. Furthermore, these three models also agree largely with the heuristic models which further strengthened their values.

10.3 Ensembled Results & Discussions

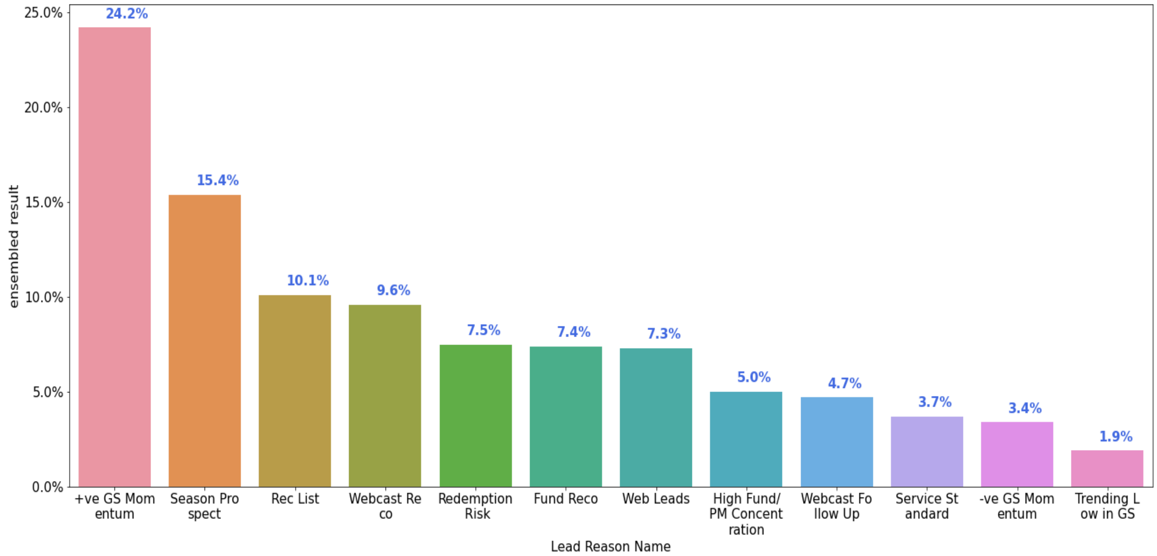


Figure 16. Ensembled Results

As shown above, the top three NBA lead reasons recommended by this ensembled model is **+ve GS Momentum, Seasons Prospect and Rec List**. This is in line with what the marketing people want at the company and we all agree to deploy this model shortly after January. By averaging the results of the 10 models, it has shown enough success in producing stable results for the NBA leads study. Compared with each single model we have before, it reduces the skewed dataset and produces a much more reliable result by integrating each model's advantages and removing their disadvantages.

XI. Conclusion

In conclusion, our ensembled model is a good starting point for the company to use as a comprehensive tool to not only study the NBA leads attributions, but also on any marketing attribution problem in general. In the future iterations of this project, we suggest the company's future coop student to focus on developing some neural networks-based (e.g LSTM-RNN-Attention Model) architectures to further improve the ensemble model's performance. In addition, there is still tons of literature online about data-driven models on attribution analysis such as improved additive hazard model by also considering a time-decay rate and customer's network in the model, which we believe will definitely add more values into our current ensembled result. All in all, we are happy with what we achieved in this project and we are ready to deploy our ensembled model onto the Company's server shortly after this report is finished.

XII. Reference

[1] L. F. Bright and T. Daugherty. Does customization impact advertising effectiveness? an exploratory study of consumer perceptions of advertising in customized online environments. Journal of Marketing Communications, 18(1):19–37, 2012.

[2] R. J. Lavidge and G. A. Steiner. A model for predictive measurements of advertising effectiveness. Journal of marketing, 25(6), 1961.

[3] L. C. Ueltschy and R. F. Krampf. The influence of acculturation on advertising effectiveness to the hispanic market. Journal of Applied Business Research (JABR), 13(2):87–102, 2011.

[4] G. Zenetti, T. H. Bijmolt, P. S. Leeflang, and D. Klapper. Search engine advertising effectiveness in a multimedia campaign. International Journal of Electronic Commerce, 18(3):7–38, 2014.

[5] B. Dalessandro, C. Perlich, O. Stitelman, and F. Provost. Causally motivated attribution for online advertising. In Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy, page 7. ACM, 2012.

[6] X. Shao and L. Li. Data-driven multi-touch attribution models. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 258–264. ACM, 2011.

[7] Yuan, J. (2020, July 21). *The Shapley Value Approach to multi-touch attribution marketing model*. Medium. Retrieved January 14, 2022, from https://medium.com/analytics-vidhya/the-shapley-value-approach-to-multi-touch-attribution-marketing-model-e345b35f3359

[8] Zhang, Ya & Wei, Yi & Jianbiao, Ren. (2015). Multi-touch Attribution in Online Advertising with Survival Theory. Proceedings - IEEE International Conference on Data Mining, ICDM. 2015. 687-696. 10.1109/ICDM.2014.130.

XIII. Appendix

13.1 Code Reference

The code for this project is being zipped into Fidelity_Charlie_ESC490.zip, inside the zip file, there's the following three file used for this project:

1. MLA_NBA_Lead_Attribution_With_Normalization_Charlie.ipynb: Code for all 10 models and results.
2. Pathe Aggregation - Charlie.ipynb: Code for reformation customer's purchasing path.
3. Final result.csv: Final result for each of the 10 models.