R Notebook

Code

Project Overview

Welcome to the Draper and Associates Marketing Team! We have recently been contracted by Arch & Harbor, a large department store chain throughout the United States. A&H recently conducted their first email marketing test, and would like us to analyze the results, and develop a data-based strategy for targeted marketing emails going forward.

We have been provided with the following information about this test campaign:

- The campaign was conducted on 64,000 customers that purchased from A&H at least once in the last year
- Customers were randomly assigned to 1 of 3 different groups
    - 1/3 received an email promo featuring Men's merchandise
    - 1/3 received an email promo featuring Women's merchandise
    - 1/3 received no email promotions
- The Men's/Women's promos were not tied to the gender of the customer
- The emails were all sent on the same day, and sent once
- Each email costs $0.10 to send
- Impact results were tracked over the two week window following the email campaign send date. Three different impact measures were tracked:
    - Website Visits: Whether the customer visited the A&H website in the two week observation window
    - Purchases: Whether the customer purchased any merchandise in the two week observation window
    - Customer Spend: The actual dollars spent by the customer in the two week observation window

A&H has provided us with a dataset that contains the treatment group, performance data, and demographic data for 50,000 of the 64,000 customers in the test (they have held back a test set for their own internal performance assessment).

The following dictionary defines that data:

- Time_Since_Purchase: The number of months since customer's last A&H purchase
- Historical_Spend: The total dollar amount customer has spent historically in the last year at A&H
- Historical_Spend_Group: A bucketing of Historical_Spend values
- Mens: Indicates customer purchased mens merchandise within last year
- Womens: Indicates customer purchased womens merchandise within last year
- Region: Classifies customer residency region as Urban, Suburban, or Rural
- Was_New_Customer: Indicates if customer made their first purchase in last 12 months
- Prior_Purchase_Channel: Indicates the channel the customer made their purchases from (Phone, Web, or Multichannel)
- Estimated_Monthly_Income: An estimate of customer's monthly income, using a model built by an A&H intern
- EmailViabilityScore: A score related to the validity of the customer's email address
- Email_Type: Treatment group customer was in
- AnyTreatment: An aggregated treatment indicator, that combines the mens and womens promos into a single treatment
- WebVisit: A binary indicator for whether customer visited website in 2 week performance period
- Purchased: A binary indicator for whether customer made a purchase in 2 week performance period
- Spend: The total dollars spent by the customer in 2 week performance period

Tasks To Be Completed

Data Exploration

A&H would like you to conduct an exploratory analysis of the data, and generate a set of insights gained. In particular, they would like answers to the following questions:

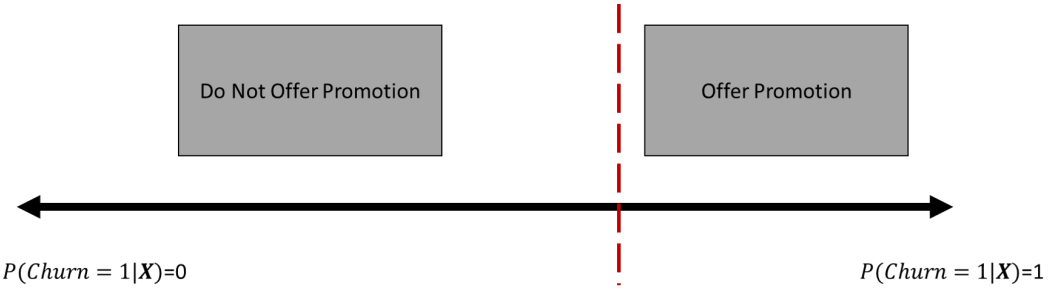
1. Overall, how effective is sending email promotions to customers? Is it net profitable?
2. Which email campaign (Mens or Womens), overall, has the best incremental performance lift?
3. Was either the Mens or Womens campaign more effective at driving previous phone purchasers to the website?
4. Is there a "best" time since last purchase for sending an email promo?
5. Does a customer's previous spend amount influence their response to email promos? Is there a group that isn't profitable to send promos to?

What other interesting or odd trends do you observe in the data?

Modeling

A&H would like us to use the provided data, and the insights we discovered, to develop a model and implementation strategy for optimizing the profitability of future email promo campaigns.

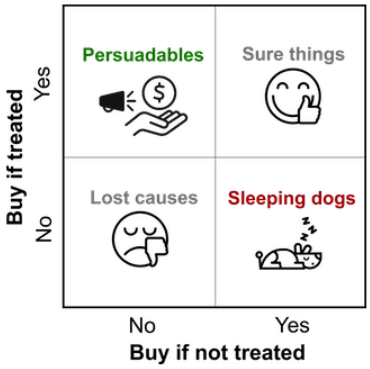
Traditionally, if a business wanted to determine who they should take some action on or treat in some way, they would predict the probability of some outcome (such as churn or conversion) or the expected outcome value (such as revenue or loss). As an example, a cable television provider may build a model that predicts the probability that a customer will cancel their service in the next few months, and may offer promotions or marketing to customers with the highest probabilities of churn.



While some of these customers may become less likely to churn because of the promotions, it is also possible that the promotion could make a customer more likely to churn (maybe they forgot they were paying for the service, and you just reminded them)! For reasons such as these, it is better to predict the *expected change* in outcome value or probability that will result from taking the action or applying the treatment, and this is precisely what the field of uplift modeling does.

Uplift modeling works in situations such as ours, where there is randomized treatment data. The randomization of the treatment allows us to marginalize over all other demographic or potential confounding attributes in such a way that we can assert causality between application of the treatment, and the change in outcome.

In the uplift modeling framework respondents/customers fall into 4 groups based on the causal impact the treatment has on them, and these groups are shown below.



1. Persuadable	• Who only respond to the marketing action because they were targeted
2. The Sure Things	• Who would have responded whether they were targeted or not
3. The Lost Causes	• Who will not respond irrespective of whether or not they are targeted
4. Do Not Disturbs	• Who are less likely to respond because they were targeted

We see that the “sleeping dogs” are customers who would have responded favorably, had you not treated them, such as the cable purchaser above, who forgot he was still paying for your service; these are customers we want to explicitly avoid treating. Less dangerous, but still to be avoided are the “Lost Causes” and “Sure Things”. These are the customers who will not be moved by the treatment, either for good or for bad, and treating them will incur cost without any incremental gain. The Sure Things can be even more unfavorable if the treatment involves a coupon of some kind, as you will have given an unnecessary discount to someone who would have purchased anyway, effectively cannibalizing your own sales.

Clearly, it is the persuadables that we want to identify and target, those customers who require the treatment in order to have a favorable response. In uplift modeling, these are customers who are predicted to have an uplift>0, that is, they are incrementally more likely, or more profitable, if treated, versus if not.

There are many types of uplift modeling algorithms, and the following packages in R or Python should contain everything you need, along with overviews on how to build them.

- Uplift Random Forest and CCIF in the Uplift R Package
- https://www.aboutwayfair.com/tech-innovation/pylift-a-fast-python-package-for-uplift-modeling (https://www.aboutwayfair.com/tech-innovation/pylift-a-fast-python-package-for-uplift-modeling)
- https://github.com/Minyus/causallift/ (https://github.com/Minyus/causallift/)

You will want to make sure to:

- Select the right features to include in your model
    - Watch out for variables which are predictive of the treatment indicator -> These violate the randomization assumption, and bias the model
- Choose the right target variable (Which outcomes do we care about impacting? Site visits, Purchases? Etc?)
- Choose the right algorithm for the target variable you choose
    - Some algorithms only work with binary targets, and not continuous ones
- Optimize the algorithm in some way
- Determine how well your model predicts uplift (Qini or Area Under the Cumulative Gains Curve is fine)
- Propose a strategy that uses the model to determine who should be sent future email promos
    - Be sure to include profitability estimates/impacts of your strategy and compare against no promos, and randomized promos
- Time permitting, propose a design for a second iteration of the randomized test, which helps to fill in gaps uncovered during exploratory analysis

Your submission should include the following:

- A notebook (R or Jupyter) that shows your analysis steps and results, and any pre-processing you do to your data
- A completely reproducible model object

- An overview of your model’s performance
- A strategy proposal for using the model, with profitability impact

Suggested Project Breakdown

You will be given two hours to complete this exercise, and we suggest adhering, roughly, to the guidelines below:

- Data Exploration and Cleaning (20-30 Mins)
- Algorithm Research (10 Mins)
- Model Development, Optimization, and Testing (45-55 Mins)
- Strategy Development and Impact Estimation (10-15 Mins)
- Documentation and Wrap-Up (5-10 Mins)

You make work in teams of up to 3, and we suggest delegation of various responsibilities, such as algorithm research.