Kellogg School of Management
Northwestern University

Creative Gaming: Uplift Modeling

After the success of the work on propensity-to-buy modeling, Mi Haruki has asked your team to focus on uplift modeling to pick customers to target for the Zalon campaign. You plan to use the experimental control group (***Group 1***, `cg_organic_control`, i.e., the 30,000 customers who did not receive an ad), as well as ***Group 2***, i.e., the randomly chosen experimental group of 30,000 customers from the `cg_ad_treatment` data. The data for Group 2 is provided in dataset `cg_ad_random`. You will only need the data in `cg_organic_control` and `cg_ad_random` for this assignment.

Part I: Uplift Modeling Using Machine Learning (38 points)

1. Prepare the data (5 points):

a. Add a variable "ad" to `cg_ad_random` and set its value to 1 for all rows

b. Add a variable "ad" to `cg_organic_control` and set its value to 0 for all rows

c. Create a stacked dataset for the uplift analysis by combining `cg_organic_control` (Group 1) and `cg_ad_random` (Group 2). Use `cg_rct_stacked` as the name for the stacked dataset.

d. Create a training variable and ensure the proportion of "yes" and "no" values is as close as possible across the 4 sub-sets of data (i.e., training and test for ad == 0 and training and test for ad == 1). Use 1234 as the seed. Use "converted" and "ad" as the blocking variables, in that order.

   Hint: Review the uplift demo code for an example of how to set this up.

e. Check if the probability of yes/no is similar across the training and test sets for ad == 0 and ad == 1. The response rate does not have to be exactly the same in the ad == 0 and ad == 1 sections of the data. However, within ad == 0, the ratio of yes/no should be very similar for the training and test sets. Similarly, within ad == 1, the ratio of yes/no should be very similar for the training and test set.

2. Train an uplift model using Logistic Regression. Add the predicted scores for the treatment and control models, i.e., pred_control and pred_treatment, to `cg_rct_stacked` (or add them to a separate DataFrame called pred_store) and calculate the uplift score. (3 points)

3. Calculate the Uplift (%) and Incremental Uplift (%) for the uplift model (use 20 instead of the standard 10 groups) and plot performance metrics. Interpret the plots. (3 points)

   Hint: Use functions from the pyrsm package to generate the uplift table (uplift_tab), the incremental uplift (inc_uplift_plot), and the uplift plot (uplift_plot).

4. Using the `incremental_resp` column from the performance metric table created through the uplift_tab function, calculate the incremental profit you expect to make if you targeted the best 30,000 customers of 120,000 using the uplift model. (3 points)

   Hint: For every n-tile, the `incremental_resp` tells you how many incremental purchases were made when customers up to that n-tile were targeted. To extrapolate to the best 30,000 from 120,000, notice that there are a total of 9,000 customers who got the ad in `cg_rct_stacked`

5. Calculate the Uplift (%) and Incremental Uplift (%) you would get if you used a propensity model (use 20 instead of the standard 10 groups). Compare the performance metrics between the uplift and propensity models and interpret the differences. (3 points)

   Hint: Use uplift and incremental uplift plots to compare the uplift and propensity models

6. Using the `incremental_resp` column from the uplift table for the propensity model, calculate the incremental profit you would expect to make if you targeted the best 30,000 customers of 120,000 using the propensity model. How much more do you expect to make from using an uplift rather than a propensity model? (3 points)

7. Repeat steps 2-6 using a Random Forest model. Tune the model on at least two hyper parameters (8 points)

8. Repeat steps 2-6 using an XGBoost model. Tune the model on at least 3 hyper parameters. Do not use “early stopping” (10 points)

Note: Check the documentation (and use ChatGPT) for suggestions on which hyperparameters to select.

https://xgboost.readthedocs.io/en/stable/tutorials/param_tuning.html

Part II: Targeting the optimal percent of customers (24 points)

So far we have targeted 25% of model-selected customers by picking the best 30,000 out of the set of 120,000 customers. Determine if more or less than 25% of customers should be targeted.

1. What formula would you use to select customers to target using a propensity model if your goal is to maximize profits? What percentage of customers in the ad treatment test set in `cg_rct_stacked` would you target based on the propensity model? (3 points)
2. What formula would you use to select customers to target using an uplift model if your goal is to maximize incremental profits? What percentage of customers in the ad treatment test set in `cg_rct_stacked` would you target based on the uplift model? (3 points)
3. Rounding the targeting percentage numbers you calculated in 1. and 2. to the nearest 5%, use the uplift table you created for the propensity and uplift models in Part I to calculate the incremental profits you would have obtained in the test set in `cg_rct_stacked` dataset if you had targeted the optimal percentage of customers suggested by each model. (2 points)
4. Repeat steps 1-4 based on predictions from the Random Forest model (6 points)
5. Repeat steps 1-4 based on predictions from the XGBoost model (6 points)
6. Give two reasons why one model performs better than the other on incremental profit when the optimal percentage of customers is selected for each model (4 points)

Part III: Generative AI (5 points)

Generative AI (5 points): Please describe in detail how your team used Generative AI-tools like ChatGPT to support your work on this case. Provide pdfs and/or screenshots of your "discussions" with these tools and comment on what things did and did not go well. Make sure to add discussion about your thought process and how you tried to maximize the benefits from using these tools. Also add any questions you may have about the assignment and the support you received from GenAI so we can discuss these topics in class.

Note: No matter how you used Generative AI-tools, you are expected to fully understand

all elements of the case solution submitted by your group. Any group member may be called on in class to walk us through your thought process and how different parts of your code work and how you arrived at your solution.