Uplift Modeling

for marketing campaign analysis

Kyra Quan

Project introduction

Business objectives:

- ❑ To aid a fundraising campaign for a not-for-profit organization with a million members, by using predictive modeling to predict how many and which individuals to target.
- ❑ Maximize operating surplus

  Max (Operating Surplus) = Total Amount Raised – Cost of Calls

  Cost of contact:

Number of contacted members	Cost per member
0 – 60,000	$5 / person
> 60,000	$25 / person

Data Source

1,000,000 potential donors information which includes:

- ❑ ID data
- ❑ Socioeconomic status
- ❑ Previous donation behavior

Variable Name	Description	
ID	Member number (unique ID)	ID data
LastName	Last Name	
FirstName	First Name	
Woman	Sex (1=woman, 0=man)	Socio-demographic
Age	Age (years)	
Salary	Annual salary in USD	
Education	Highest education level	
City	Type of neighborhood	
SeniorList	Seniority for being on the VIP list	History
NbActivities	Number of participations to annual meeting	
Referrals	Number of referrals	
Recency	Number of years since last gift	
Frequency	Number of donations	
Seniority	Number of years since first donation	
TotalGift	Total Donation since a member	
MinGift	Minimum Donation since a member	
MaxGift	Maximum Donation since on the VIP list	
Contact	Direct solicitation this year (Only applicable to Round 2)	
GaveLastYear	Whether or not the individual give last year	Target
AmtLastYear	Amount given last year	
GaveThisYear	Whether or not the individual give this year	
AmtThisYear	Amount given this year	

Two-stage modeling

1. Fit a model to predict the probability P that an individual will give
2. Keeping only data from members who gave, fit a model to predict the amount given M
3. Compute P*M to determine the expected donation for each member

Stage 1

Probability of giving

P

Stage 2

Conditional amount if they give

M

Expected Donation = **P * M**

Uplift modeling

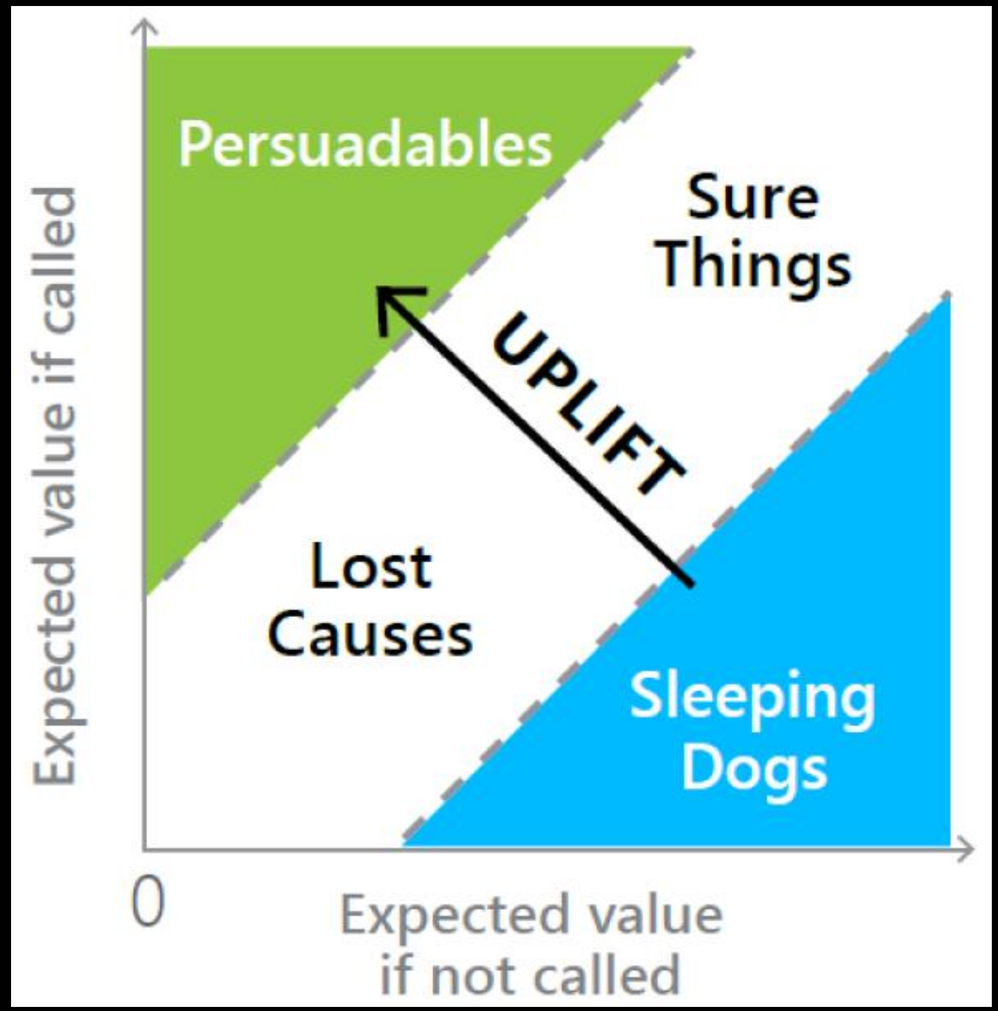
Why use uplift modeling?

❑ To make predictive model actionable.

What is uplift modeling?

1. Persuadable: are more likely to donate if called
2. Sure Things: are likely to donate whether called or not
3. Lost Causes: are unlikely to donate whether called or not
4. Sleeping Dogs: are discouraged to donate when called

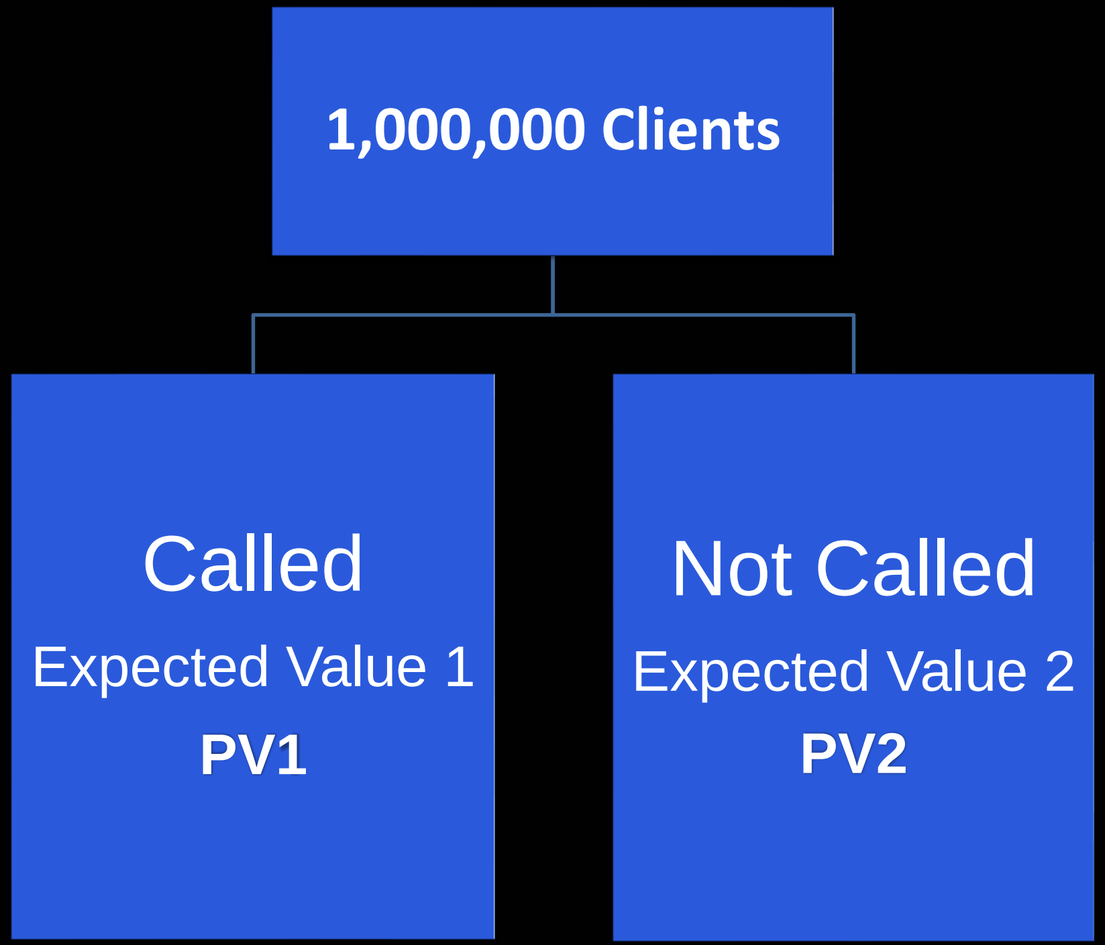
Predict the probabilities of being in each group and then calculate the uplift score.



Uplift modeling

1. Expected value if a member is contacted PV1
2. Expected value if a member is not contacted PV2
3. Compute the difference to obtain the uplift generated by contacting a member

   Uplift = PV1 – PV2



Model Building and Evaluation

Data Inspection

- ❑ 40% of the dataset contains missing values
- ❑ Missing categorical values are imputed as zero and missing continuous values are imputed as the population median

Model Characteristics

- ❑ Dataset was split into train and test set using 70-30 percentage split
- ❑ Used linear regression to predict the donation amount, and logistic regression to predict the probability of donating

Dataset	Model Type	Output Variable (Target)	Variable Type
Contacted	Logistic	GaveThisYear	Categorical
Contacted	Linear	AmtThisYear	Continuous
Not Contacted	Logistic	GaveThisYear	Categorical
Not Contacted	Linear	AmtThisYear	Continuous

Model Building and Evaluation

Variable Selection

- ❑ Backward stepwise selection method
- ❑ Reduced from 21 variables to 13 variables.

Model Performance

- ❑ Test the models on test sets, two logistic models achieved around 81% accuracy.
- ❑ Test on unseen data, 197484 clients were selected. The models yield $11M surplus.

Recommendations

For future operations

- ❑ The individuals on the contact list should be contacted in the upcoming fundraising campaign.
- ❑ A better data entry and management practice should be implemented to minimize the number of missing values.
- ❑ Socioeconomic data/variables such as job level, industry, household disposable income, debt and number of dependents would provide an extra layer of information regarding a donor’s financial situation