PREDICTIVE MODELING II ASSIGNMENT

Aniruddh Kumar
Ishaan Goel
Eric Weijie Gu
Xiao Quan
Piranavan Easwaramoorthy
Huzaifah Sohail

NOVEMBER 18, 2019

MBAN 5210A

Prepared for Cristina Anton

To: Marketing Management
From: Special Analytics Team
Date: November 18, 2019
Subject: Who to Contact for The Upcoming Fundraising Campaign

To the valuable members of our marketing team,

Our organization has been able to stand strong for the last 12 years due to the sacrifices and collaborative efforts made by its many individuals. In relation, our team would like to continue to stand strong by making the same efforts alongside our marketing team to support the upcoming fundraising campaign. Regarding this, the following document has been created to explain the strategy we have come up with and its implementation for the fundraising campaign. To ensure clarity, we will define any technical terminology that may potentially be difficult to understand.

The Business Problem
For the upcoming campaign, the organization is looking to initiate a direct contact campaign as one of their marketing activities. The target population of this campaign is one million potential donors whom we have data for. The objective is to maximize the donation amount received while minimizing the cost associated with contacting individual members. Hence, the business problem revolves around finding a method to maximize the operating surplus associated with the direct contact campaign, or in other words, finding out which individuals and how many of them to target.

The Analytical Problem
Upon understanding the business problem, our analytics team found it optimal to use the data available about each potential donor to build predictive models that predict the amount of future donor donations based on whether they are contacted by the organization or not. As a simple definition, predictive modeling essentially uses data recorded from a number of variables (predictors) to predict a certain outcome. Because of our current donor data, we know that there are individuals who are likely to donate regardless of whether they are contacted or not, and there is also the possibility that individuals may only donate if they are contacted. Hence, the analytical challenge was to determine the value created by contacting each individual, otherwise known as the uplift. In this case, the uplift is determined by finding out how much each donor will donate if they are contacted or not contacted (expected donation value), which would be determined through the predictive models that we have built.

Data Background
The available information prior to the data modeling stage includes one million potential donor names, their socioeconomic status, and their previous donation behavior. The information includes a total of 18 variables such as name, gender, age, salary, education level, type of neighborhood, seniority, participation level, past donation pattern, and if they have been contacted by the organization. 12 of these variables are continuous, which are numeric variables that can have an infinite number of values between any two values. Five of these variables are categorical, which means they can only take on a limited number of possible values that assign each observation to a nominal category. Please refer to the Appendix section for a comprehensive list of the variables (Figure 1). Overall, 10 percent of the donors have been contacted, over 60 percent of all donors have achieved university / college level education, and more than 85 percent of the donors did not

donate last year or this year. Additionally, we have identified missing values in the following variables: seniority, frequency, recency, total gift amount, minimum gift amount and maximum gift amount. The missing values could be due to multiple reasons, such as a lack of donation history or data mismanagement.

The Overall Modeling Process

The methodology our team used to achieve the final deliverable follows a three-step process:
1) We first divided the dataset into two separate subsets based on whether the donor had been contacted or not. The dataset of contacted donors contains 100,000 records and the dataset of non-contacted donors contains 900,000 records.
2) For each dataset, we built regression models to predict the probability of the donor donating this year, as well as the amount they would donate this year. The multiplication of the two figures is equal to the expected value for each donor. This step of our method output the expected amount each donor would give for when we do and do not contact them (Figure 2).
3) We then found the difference between the two figures (expected amount if we contact versus if we do not contact). The difference in values indicated the uplift of each donor – how much additional value we would expect from them once we contacted them. Then, using this uplift for each donor and assuming the cost of contacting a donor was $25, we filtered the list to capture all donors with an uplift greater than $25. This was done to ensure that the donors who were contacted would result in a positive operating net surplus. Below are the calculations that were needed to create the optimal contact list (see Figure 2 for further clarification).

EC: Expected donation amount if contacted = Probability an individual will donate if contacted * The predicted amount they will donate if contacted
ENC: Expected donation if not contacted = Probability an individual will donate if not contacted* The predicted amount they will donate if not contacted
Uplift: Value created by contacting = EC - ENC

Model Approach

<u>Data Inspection</u>
As is needed with any dataset, the first step was to inspect the dataset to understand each variable, as well as the kind of values that were expected. In addition, the number of missing values was analyzed for the team to decide whether it needed to be cleaned up using imputation, as described in the next section.

<u>Imputation</u>
To clean and organize the datasets, we computed different imputation methods to handle missing values. Typically, records with missing values can be ignored or be filled with another value, such as mean or median. For all four models, missing values appeared to be significant. We tested replacing all missing values as zero, as well as replacing missing categorical values as zero and missing continuous values as the population median. Rather than imputation we could also have chosen to ignore the missing values, but as a team we decided to go with imputing. This was due to the concern of losing significant amounts of data. Since SAS model building tends to ignore

missing values in fitting regression/logistic models, approximately only 35 percent of the dataset would have been available for model building due to the lack of complete data records.

Model Characteristics

After splitting the original dataset into two subsets, we needed to predict both the probability of donating and the donation amount for each subset. To predict the donation amount, we used linear regression. To predict the probability of donating, we used logistic regression. This is because, for an output that is a continuous number (i.e./amount donated), a linear model should be used, and for an output that is categorical in nature (i.e/whether or not a person will donate), a logistic model should be used. To calculate the uplift amount for each donor, creating two linear and two logistic models were necessary. Below is a table providing an overview of the characteristics of the models that were built.

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

Variable Selection

During the process of model building, we chose the backward stepwise selection method as the variable selection technique. This entailed starting off with a full model that included all the variables from the dataset. From there, the variable that was the least significant had been permanently removed from the model. This process was repeated until every variable left in the model produced a p-value less than the stopping criteria, indicating the variable's significance. The full model that we started out with consisted of 21 variables, and after the backwards stepwise selection method was completed, we ended up with 13 variables. It was also vital to understand if there was any sort of collinearity within the variables. For example, we did observe collinearity between MaxGift and TotalGift, as a higher MaxGift amount would result in a higher TotalGift amount. Dropping unnecessary variables would thus help in improving our model, and similarly, adjusting for collinearity amongst the variables would help stabilize the model.

Model Testing & Output

As is the standard, we split each of the datasets into training and testing subsets. This allowed us to make a model based on a training dataset, and then to test the model on the testing dataset, to see how well the model would perform on predicting outcomes. A 70-30 percentage split was done for the training and testing data split, respectively. For the regression model, we used SAS, a statistical software suite, which offers multiple powerful methods to build a model. In our case, we opted for the backward selection model to arrive at the most "computationally correct" model.

The four models produced the probability of donation as well as the donation amount for each subset. We combined the outputs by multiplying the probability and the amount, creating the expected value for each group. With the two lists of expected values, we then compared them side-by-side. The difference between those two lists of expected values gave us the uplift for each donor. We then ranked the donors by the uplift amount to identify the top donors to target. With the uplift available, we decided to filter the list to those who had an uplift value greater than $25. In other

words, we are going to contact the list of people who, if we contact, will donate $25 or more than they would have.

Model Performance

For each model that we built, we concluded the error rates from the testing datasets. For the two logistic models, we achieved around 70 percent accuracy. For linear models we used Train ASE (Average Squared Error) and Validation ASE to conclude them being good acceptable models that did not overfit. Overfitting of a model increases bias, which restricts our ability to deploy the model to predict future test sets.

Recommendation & Key Considerations

Based on the analysis and modeling done, we were able to create a donor list whose uplift value was greater than $25 per donor. The Special Analytics Team's recommendation is that the individuals on the list should be contacted in the upcoming fundraising campaign, which will ultimately maximize the net raised funds. Furthermore, we identified areas of improvement for future campaigns. The missing values in the dataset created constraints on the accuracy of the model. 60 Percent of the observations had missing values. To make up for the missing data, several assumptions had to be made regarding its imputation. To aid in creating a more accurate model, a better data entry and management practice should be implemented to minimize the number of missing values. Additionally, other socioeconomic data/variables could be collected from donors and added to improve the model. We have determined that 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, which proved to be significant in predicting the final outcomes.

For future operations, the foundation should focus on 1) the top-line and 2) the cost. The optimal top-line performance can be realized through both monetary value and donor quantity. As time passes, the foundation will be able to capture more donation behavior data. Every year, the foundation can utilize similar models to determine a new list of valuable target donors. While this method will continue to maximize operating surplus, the foundation should also explore other initiatives that influence donors' behavior, such as member rewards programs and price anchoring. Through rewards programs, the foundation can incentivize members to donate more and can collect more user data to improve future predictive models. With price anchoring, we can maximize each transaction utilizing consumer psychology.

On the cost side, it is essential to audit and cut unnecessary expenses and to utilize economies of scale. Minimizing the cost will allow the foundation to increase margins and to target a larger number of donors. A full review of the foundation's portfolio of initiatives could help determine the profitability of each initiative. By targeting and scaling the initiatives with higher profitability, the foundation can then continue to maximize and sustain its net gain. We believe that, with the power of data, we are able to maximize net surplus in any given situation. By engaging with donors through additional programs and initiatives, we can then enhance the situation by maximizing our potential.

Appendix A: Preferred Grade Distributions

Model performance	Model write-up
40%	60%

Appendix B: Figures

Figure 1: Variable List

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	

Figure 2: Uplift

Calculate Uplift

1. EC: Expected donation if contacted = PGivingContact * PredContact
2. ENC: Expected donation if not contacted = PGivingNoContact * PredNoContact
3. Uplift: Value created by contacting = EC - ENC

	A	B	C	D	E	F	G	H	I	J
1	ID	Gave	AmtGave	PredContact	PredNoContact	PGivingContact	PGivingNoContact	EC	ENC	Uplift
2	2422073	0	0	368,9250731	368,9250731	0,734872067	0,233744654	271,1127	86,23426	184,8785
3	2394415	1	40	267,5369606	267,5369606	0,851819193	0,227434215	227,8931	60,84706	167,0461
4	2201020	1	40	263,0737339	263,0737339	0,854196836	0,232647544	224,7168	61,20346	163,5133
5	2940637	1	20	283,3737778	283,3737778	0,735778966	0,162365502	208,5005	46,01013	162,4903
6	2561134	1	10	365,1596021	365,1596021	0,825811442	0,383343509	301,553	139,9816	161,5714
7	2387866	1	20	274,6629921	274,6629921	0,713165121	0,130874294	195,8801	35,94633	159,9337