Final Project

Fundraising Case Study

MATH 60603A - Statistical Learning

Alexa Canuel (11322540)
Elliot Wyman (11317748)
Pauline Françoise Marie Malaguti (11321831)

December 19, 2022

Table of Contents

Introduction 2

Data 2

Data sets 2

Member list 2

Donation History 4

Pilot Call Group 4

Newsletter Reads 4

Social Network Usage 5

Personality Questionnaire 5

Data Cleaning and Preprocessing 6

Methods 7

Results and Conclusion 8

Appendix 9

Introduction

This project aims to assist a charitable foundation, whose mission and thus historical data encompass almost a decade of donor characteristics and activities. The discrete goal of this project is to enable the foundation to maximize donations received from their donors. Their objective is to give back as much as they can to their community during their 2021 annual campaign. To maximize donations, the foundation has agreed that making phone calls to certain persuadable donors would help them increase their donations in this upcoming campaign. As a charity, they want these calls to be as personal as possible, accomplishing that requires trained staff and the leveraging of donor data, therefore making it a non-negligible cost. The first 60,000 calls will cost them 5$ each, with any additional calls beyond the first 60,000 costing 25$. After having done a pilot study in 2020 where 100,000 donors were called at random, the results showed that depending on a person's personality traits, calling them could either encourage or discourage them from donating. This project aims to use prediction-based classification and regression algorithms to identify these two disparate groups of individuals and prepare a proposal of individuals that should be called. This paper is broken down into the following sections. First, an exploration of the donator data, including processing steps taken to prepare the data for input into machine learning models. A discussion of the methods and algorithms used to identify the user groups as well as to predict the donation amounts per individual will follow. The findings of the implementation of the above approach will be outlined in the results section.

Data

Data sets

The foundation provided the following 6 datasets for this project. Note that not all donors participate in all activities, and therefore not all donors are captured in every data set.

Members List - basic information about the members of the foundation as of 2019
Donation History - a list of all donations made up until 2020
Pilot Call Group - a set of previous donors, randomly selected to be called in 2020
Newsletter Reads – information about whether the members read the Newsletter each month in 2019
Social Network Usage - information about interactions with the foundation's social media page in 2019
Personality Questionnaire - a personality test based on 15 questions

Member list

A list of members of the foundation was collected as well as basic information about them. Members are defined as people who have donated to the foundation at least once since its inception. This list contains 1 million members who identify as male or female, ages ranging from between 16 to 90 years old, annual salaries varying ranging from 0 to 250,000 dollars, with varying levels of education and place they reside. Also provided was the year of their first donation, in explanation, the year they became a member of the foundation. Each member is identified with a unique ID, as well as their first name, last name, and email address. This list contains information from 2009 until 2019.

Figure 1 shows that there is an even distribution of male and female members in the dataset. *Figure 2* indicates that the distribution of the age of members is evenly distributed with an average age of 46 years old. *Figure 3* demonstrates that the distribution of salaries from members is skewed to the left with over 250,000 members having an annual salary of about $20,000 dollars, the mean salary being $65,516 dollars. *Figure 4* demonstrates that the majority of member have a university/college degree and *Figure 5* indicates that the places the members reside is close to an even distribution, with slightly more members living in the city and a minority living Downtown.

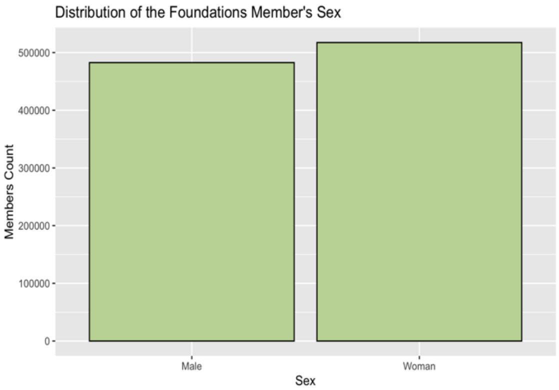

Figure 1

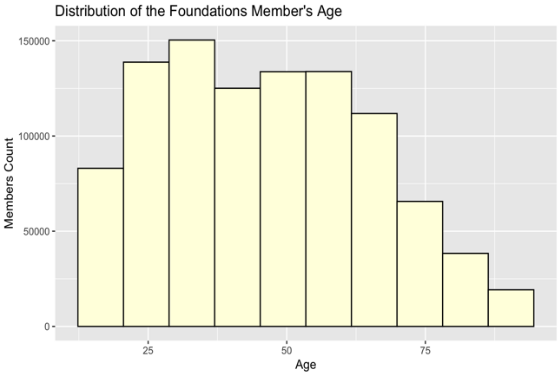

Figure 2

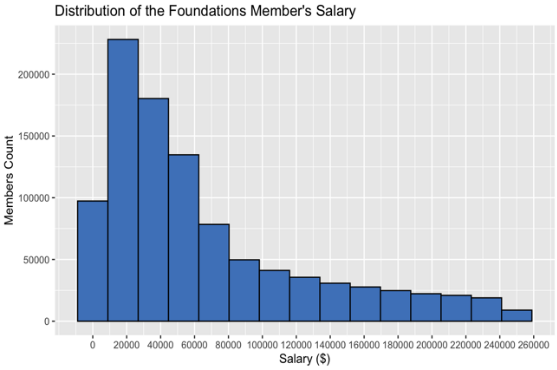

Figure 3

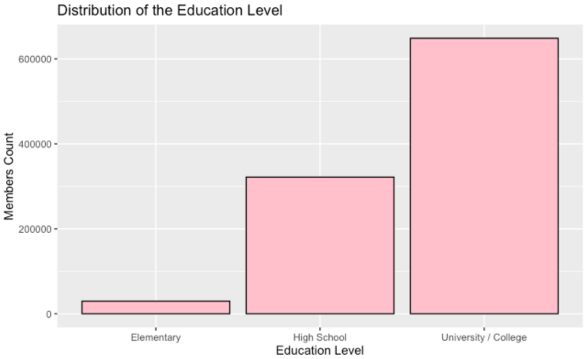

Figure 4

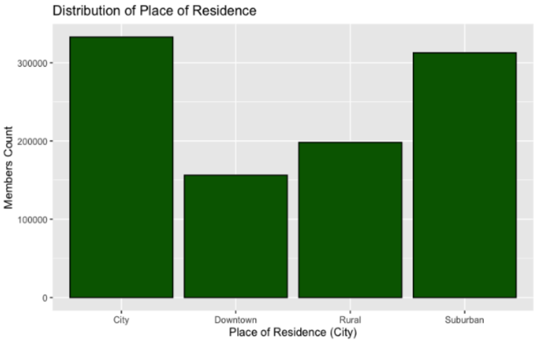

Figure 5

Figure 6 demonstrates the distribution of when the members joined the foundation. From 2009 to 2018 around 8,000 members joined per year, however in 2019 there was a spike of over 150,000 new members.

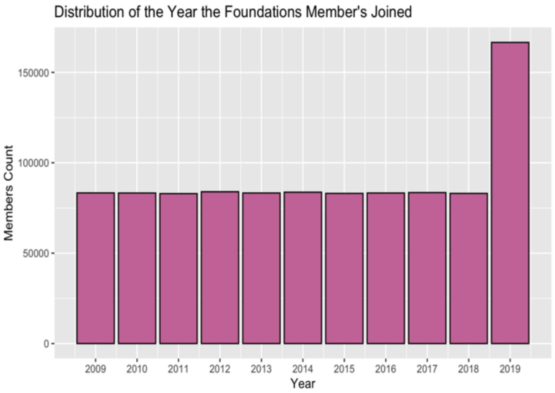

Figure 6

Donation History

A donations history list is provided from 2009 up until 2019. This list includes the dollar value of donations made by each user for every year they donated. For example, one line of the record could represent a “Jane Doe”, who donated 20$ in 2015. Another line could represent the same “Jane Doe” who donated 75$ in 2017. It is interesting to note that the minimum donation made in one year from a user is 10$ and the maximum is 10,000$ while the mean donation is 53.17$ with a median of 25$. These values indicates that the distribution is skewed to the left and that most of the values are between 0$ and 50$ donated, as seen in *Figure 7* below which is only showing the distribution from 0$ to 500$.

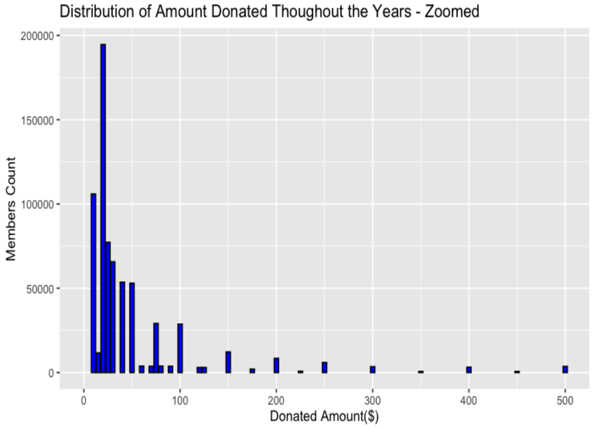

Figure 7

Pilot Call Group

With the goal of maximizing donations in 2021, the foundation ran a pilot study in 2020 by randomly selecting 100,000 members to call. This study provides information on a group of members that were called to encourage donation and indicates their reaction. As mentioned previously, someone could react positively and donate while another could decide not to donate after receiving the call.

Newsletter Reads

In 2019, the foundation sent out a newsletter each month to all 1,000,000 members. The goal with sending monthly newsletters is to update their members on what the foundation is up to, which, of course will keep member retention and ultimately encourage them to donate in the future. The count of how many members opened the email per month was recorded. Providing an indication of a member’s potential interest in the foundation. From *Figure 8*, see that the two months where the lowest number of members opened the

Newsletter were July and August, while the month with the highest opens is May, with over 230,000 members opening the email.

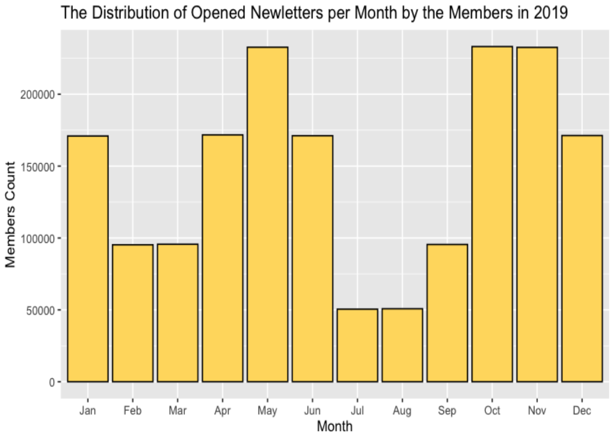

Figure 8

Social Network Usage

The foundation collected information from 400,000 social media users who have interacted with their Facebook page and posts in 2019. This dataset contains information on the number of Likes user left on the foundation's posts and the number of times they Shared these posts. In addition, if the user is a Supporter of the Facebook Page, this indicated that the user in question actively wants to see the foundations posts. This can be very valuable information on users as it gives indication that a user is invested in the foundations mission and therefore will be more likely to donate.

In terms of interactions, around three quarters of the social media users have never shared anything from the foundation. The highest number of shares from a member in 2019 is 121. In addition, all 400,000 users have liked at least one post while the maximum number of likes by one person is 260. 387,956 users out of 400,000 are not supporters of the fundraisers Facebook page and therefore making only 3% of user's Supporters of the page. It is important to note that this data set considers any Facebook user that has interacted with their page. Considerations must be made on the best way to handle this data and to match the Facebook users to the foundation members.

Personality Questionnaire

In addition to the foundation's newsletter initiative, they sent a personality trait survey to their members in the hope of getting a better understanding of the people who donate. The members had to answer 15 questions indicating on a scale of 1= "not at all" to 10= "absolutely" how much they think they possess that specific attribute. Unfortunately, the survey was filled by only 800 members. *Table 1* indicates all the personality traits the members had to rate themselves on and their mean frequency value.

Table 1: Attributes tested by Personality Trait survey and their mean value

Talkative	6.2111	Outgoing	6.059	Forgiving	5.901	Thorough	5.424	Efficient	5.468
Energetic	5.309	Helpful/unselfish	5.444	Considerate	5.598	Relaxed	6.173	Emotionally Stable	5.711
Sophisticated in Arts	5.737	Easily Distracted	5.027	Moody	5.085	Cooperative	5.425	Curious	5.196

From the table above, 3 variables that have the highest mean and therefore the attributes that were rated highest among the 800 participants. Being Talkative, Outgoing and Relaxed. The personality traits that members least associate to are Moody and Easily Distracted. Since this is a very small sample size, it would be a stretch to say that all members would rate themselves the same way.

Data Cleaning and Preprocessing

Before being able to use the data mentioned above, it is necessary to ensure that the data is appropriate for use with the models chosen and will provide reliable results.

Firstly, it is crucial to note that there were no missing values found across the data sets. In terms of pre-processing the data, it is important in the context of the problem to combine these different characteristic and behaviors of a donor, as it will help better predict if they are going to donate in the upcoming campaign. Moreover, as mentioned above, careful consideration was needed when handling social media users. Unlike the other datasets where only members were involved and where the user ID was provided to make the merge simple, this dataset contains 400k users that had various name formats. *Table 2* demonstrates the five different formats. Looking only at first and last names, there are many duplicates in both Donor Info data set and the Social Media Usage data set. The only way to guarantee a correct mapping was to use unique names only when combining the two tables, and thus many social media profiles were not used.

Table 2: Name formats in Social Network Usage Data set

<firstInitial> <lastInitial>	e.g. C K
<firstInitial> <lastName>	e.g. C Klein
<first name> <lastInitial>	e.g. Calvin K
<first name> <lastName>	e.g. Calvin Klein
"The" <lastname>s	e.g. The Kleins

In terms of the Newsletter data set, to summarize the information, every member was assigned a “read rate” that represents the number of times they opened the email in 1 year.

Furthermore, it is beneficial to perform principal component analysis to reduce the dimension of the large number of variables that represent donor traits. With 5 principal components, each a linear combination of the initial variables, nearly 100% of the variability of the data is explained, as illustrated in *Figure 9*. Therefore, reducing the dimension to 5 variables provides a very good approximation of the information contained in all 15 variables.

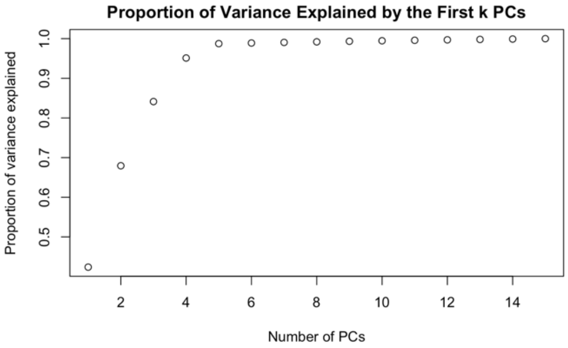

Figure 9

With these data preparations, the input matrix is composed of 1,000,000 members and 28 explanatory variables. These explanatory variables represent donor characteristics and behaviors.

Methods

The methodology used is called uplift modeling which consists of implementing two phases. The first phase estimates the probability of donating any dollar amount, provided the donor is met with an intervention, in this case a call. This prediction is calculated by comparing the difference between logistic regression models trained on each of the treatment group (those who were called during the pilot study) and the control group (those who were not called during the pilot study). The second phase consists of building a linear regression model to predict donation amount. Finally, donors suitable for intervention (call) can then be identified based on the combination of classification and regression outputs.

In determining the classification of donors, based on intervention, some terminology needs to be introduced. The concepts of sleeping dogs are people that the foundation should not call because they have a negative reaction to the campaign, sure things and lost causes are people for whom the probability of donation is indifferent to the campaign, so it is important not to lose call resources on these categories. Therefore, the foundation is interested in the persuadable category since they are individuals who will only donate following a call. The following table provides a visual representation of these classes.

Table 3: Donor Categories

	Donation	No Donation
Treatment (Call)	Persuadable	Sleeping dogs
Control (No call)	Sure things	Lost causes

Multi-stage modeling was employed to determine the donor classes. It consists of calculating a value τ_{au}, based on the following formula.

$$\tau_{au} \quad = \quad p_T - p_c$$

p_c: *Probability of a member donating from the control group*
p_T: *Probability of a member donating from the treatment group*

Logistic regression models are used to estimate p_T and p_c. The pilot group contacted in 2020 are used as the training set individuals for the model that predicts p_T, consisting of 100,000 individuals, where the target is their donation amount in 2020 (1 corresponds to a donation > \$0; 0 corresponds to no donation). Similarly, a second logistic regression model, used to estimate p_c was fitted to the 900,000 individuals not targeted in the pilot study, these 900,000 Individuals comprised the training set for the model that predicts p_c.

Both models were given 28 covariates to predict the likelihood of donating. The following variables were used: age, education, gender, salary, place of residence, newsletter read rate, the 5 principal component variables that represent the personality traits of a person, donations made in the last ten years, 3 social network variables and the year of joining the foundation.

The values p_T and p_c are then obtained by performing a prediction on all 1 million members, using each model. The classes, persuadable, sure things and lost causes, and sleeping dogs can now be identified by the value τ_{au}.

- $\tau_{au} > 0$: Persuadable
- $\tau_{au} = 0$: Sure things and lost causes
- $\tau_{au} < 0$: Sleeping dogs

In the second phase of this methodology, a linear regression approach was used to predict the donation amount which is continuous. The covariates used in this model are the same as those used in the logistic regression model used to calculate donation probability.

Once the donation amounts were predicted for the persuadable, the donations were divided into two groups. Those between $5 and $24 and those over $25. Of the individuals in the first group, only the first 60,000 were kept. This decision was based on the fact that the first 60,000 calls cost the association $5, and the aim was to make a profit from these calls. All the people in the second group were recommended to be called because their donation is $25 or more and will not generate a loss for the foundation. As the aim was to maximise net donations, the potential donations of sure things were also estimated to predict the total amount of donations.

Results and Conclusion

With the help of the uplift model, 423,485 members were predicted to be persuadable and are therefore recommended to be called. As the number of persuadable is greater than 60,000, there should be a discrimination against individuals called after the first 60,000. In the explanation of the project, recommendations were made only to call further individuals if their prediction donation was more than $25, the cost of the call. It turns out that all the leftover persuadable were expected to donate more than $25 dollars, and the discrimination consideration did not remove any individuals.

As the goal of the project was to maximize net donations, the expected donations of the sure things (certain donors) were also computed. Summing the donations of the sure things and the persuadable, the expected net donation for the charity total is $24,727,040. The cost of making the calls is calculated as:

60,000 at $5, plus an additional 363,485 (423,485-60,000) at $25

Therefore, the net profit for the charity is estimated to be **$15,339,915**.

While the estimates found through the experiments conducted in this project are promising, the correctness of the models and their results will remain unknown until the deployment of the strategy. One method for improving the accuracy of the models could be to use different parameters. During the modelling of the project, it was decided to use all the covariates available in the model, but a subset of these could boost model performance. Another potential improvement to the model is selecting a higher threshold for a member being a persuadable, the current methodology uses a threshold of 0.0, to prevent misclassifying donors, and preventing an intervention being presented to a sleeping dog, this threshold could be increased to 0.3 for example.

Finally, with the help of the pilot study, donor characteristics and behaviors, a call list called Result_collection.csv was created identifying all the members the foundation should call to increase their net profit.

Appendix

Data set name	**Filename**
Donor Info	Memberlist.csv
Donation History	DonationHistory.csv
Pilot Call Group	ListContacted2020.csv
Newsletter Reads	NewsletterRead.csv
Social Network Usage	SocialNetworkUsage.csv
Personality Questionnaire	sampleBig5_Questionnaire.csv
Call list	Result_collection.csv