1. Business Problem

The objective of this report is to describe the work done for the foundation's upcoming fundraising campaign. Specifically, the report will focus on how the foundation can optimize its donation collections through its direct contact campaign. Only donations from current members will be collected, which amounts to 1,000,000 potential donors. The only associated cost to the campaign is the cost per call. The first 60,000 calls bear a $5 cost per phone call, while any further calls have a $25 cost. To aid the analysis, the foundation provided a dataset about its members, and it contains information such as ID data, socioeconomic status, and previous donation behavior. The goal is to use this information to maximize the operating surplus, which is calculated by the total amount raised minus cost of calls.

Marketing strategies can sometimes have the unintended consequences of annoying its targets and in turn, alienating them from the goal of the advertisement. Optimal marketing strategies want to avoid targeting members who would respond negatively to the target ad or would be unaffected by it at all. In this marketing campaign, members would fall to one of four segments:

- **Persuadables:** members who only donate because they were called.
- **Sure Things:** members who would have donated whether they were called or not.
- **Lost Causes:** members who will not donate irrespective of whether or not they are called.
- **Sleeping Dogs:** members who are less likely to donate because they were called.

Avoiding sure things, lost causes, and, especially, sleeping dogs is critical, while finding the persuadables is essential. The analysis must account for this unique challenge associated with direct marketing.

2. Analytical Problem

The business problem above requires an analytical methodology that would help maximize the operating surplus. To framework the requirements needed to find the solution, the problem was broken into two parts:

1. Based on the information from the dataset, what are each member's donation probability and donation amount given that they were contacted and not contacted?
2. Taking the insights found from question one, what is a member's incremental expected donation based on making a call?

Predictive response modeling can help tackle the first problem. Logistic regression and multiple linear regression analysis can be used to determine the probability and amount that a member will donate next year, respectively, for both scenarios. Then, uplift modeling can be used to judge the effectiveness of the campaign's activity on a person and, in turn, tackle the second problem. Uplift modeling helps determine the behavior of a target through the incremental effect (uplift) of direct contact activity. Each call should have a positive uplift, and it is calculated as follows:

$$Uplift = (Pr[donation|contact] - Pr[donation|nocontact]) \times [amount] = EC - ENC$$

The model will not only measure the effectiveness of contacting a member, but also compare it against a lack of action for the same member, allowing optimal campaign activity to be determined.

3. Overview of Final Model

The final model specified for submission was the 21st model developed (preceding models will be discussed later). This following five steps will explain the final model:

3.1. Step 1

The data were explored, cleaned and prepared for analysis. Two statistical techniques were required: missing value imputation, and the creation of dummy variables. For the first technique, several variables, such as recency, the total donated, maximum donated, and minimum donated, appear as null in the dataset for members who had not donated in previous years. For these variables, a SAS code was used to transform the null values to zero values, allowing the program to function correctly. For the second technique, a categorical variable is broken down into a dummy variable for each category except one. That one category will act like the status quo category and does not require a variable. For example, for the variable of *city*, three dummy variables were used to represents the category types “city”, “rural” and “downtown”, and that left “suburban” representing that status quo. If a member is living in a rural area, the variable for rural would have a value of one, and the other dummy variables would equal zero. If a member is living in the status quo area, suburban, the three dummy variables would have a value of zero. In total, eight variables were changed: 1) Variables of *Recency*, *Frequency*, *Seniority*, *TotalGift*, *MinGift*, and *MaxGift* had null values imputed, and 2) *City* and *Education* were changed to dummy variables. As a result, 1) null values had meaning in the model, and 2) the programming process was more efficient.

3.2. Step 2

A logistic regression model was created to identify which members should be contacted. This model analyzes variables and interactions between variables to calculate the probability of one outcome (e.g., a member choosing to donate) versus another (the member choosing not to donate). The data was split into two datasets to create this model: a training set to develop the model, and a validation set to evaluate the performance of the model. The data had a 70-30 split between training and validation. However, the response variable in question, whether or not the member donated this year, is skewed, with only 15% donating. A stratified sample was used to create an optimal training environment, ensuring that both sets contained a rate of 15% of members donating.

The model also considered interactions between all variables, and a forward selection algorithm was run to find all the possible 2-level interactions for consideration. These interactions, plus the member variables, were evaluated through two selection techniques: fast-backward elimination, and all-possible automated.

Both selections resulted in fairly reliable models based on training data, each with concordance rates of approximately 73% (a measurement of how many observed outcomes match the logic of the model). Based on validation data, both selections produced an AUC of approximately 0.7312. AUC is a measure of quality for binary classification model and can take a value between 0.5 (random classifier) and 1.0 (the most accurate classifier). However, the all-possible technique produced a model with better fit statistics and is simpler of the two due to having fewer parameters (48 vs 53), and so each member was evaluated based on their probability of donating using this model. This was done on two variations of the same dataset: first assumed that all members were contacted, and the other assumed that none of them have. This allowed the model to evaluate a member's probability of donating in two of said scenarios.

3.3. Step 3

A linear regression model was developed to predict how much a member will donate. The training dataset included only members who had donation data. Such filtering is necessary because including those who did not donate would result in training based on many values of zero, which would greatly skew the model. The overall model assumes that members who are contacted will donate, so training the linear regression model this way is appropriate. To determine predictor variables to include in the model, a stepwise model chose variables based on the fit statistic of AIC. With the model, scoring was done on the two datasets from Step 2. This allowed an expected donation amount to be calculated in the event of both contact and no contact, which helped to identify customers who were not worth contacting (e.g. "sure things").

3.4. Step 4

Uplift was then calculated by using the insights developed from the two possible scenarios. To reiterate from before, uplift is derived by finding the expected donation if a member is contacted (EC), then the expected donation if the member is not contacted (ENC), and finally finding the difference between the first from the second. This dataset was then sorted, so that members were ranked by uplift, in descending order.

3.5. Step 5

Two conditions were used to determine if a member would be selected for a contact:

1. Uplift value is at least 30*.
2. The probability of donating, if contacted, is at least 45%.

*At uplift of 30, the operational surplus is $5 per call (later calls will cost $25). The high threshold is used to offset the bias that will exist within the predictions.

The conditions resulted in 166,984 contact points. Regarding the performance of the model, accuracy is relatively high, as the final model currently ranks as the third-highest submission of all the teams analyzing this problem (as of the written of this report). While the creation of 21 models would usually result in overfitting, the final model is not expected to experience this problem. This is because the overall model is identical, in terms of its predictors and uplift cutoff, compared to the 12th model developed. However, the cutoff for the probability of donating if contacted was added as an additional requirement to ensure that fewer "sleeping dogs" or "lost causes" were contacted.

4. Conclusion and Future Consideration

The organization is 12 years old, so there should be more data on historical donations that could have been provided. Many members have donated multiple times in the past; proven by the fact that the *Frequency* variable in the provided dataset is as large as 10 for some members. This means some members have donated ten times in the last 12 years, but we do not have the sums of these donations to help us predict future donation amounts. There could be additional temporal relationships that might be uncovered with studying the donation amount over time for members who have that historical data. In general, we could encourage the organization to provide all available data on its members, including data that the organization might believe would have no relationship with donations. Predictive modeling can uncover relationships in data that do not initially sound intuitive or logical.

Other external sources of data can also be consulted to make better decisions. General economic indicators can be examined to see if they provide any predictive insights into donations. For example, members may forego donations purely on the basis that there is an economic recession. Many economic indicators exist that can be tested, but a short example could include GDP growth, stock market performance, interest rates, and unemployment rate. In addition to providing insights into the health of an economy, some of these insights may be correlated with the net worth of members. *Salary* is only one component of a member's net worth, and other elements such as financial assets and real estate are affected by economic indicators such as stock market performance and many others. Since a member's decision whether to donate and the amount donated could be influenced by changes in net worth, this is another reason to consult economic indicators.

For the future, we could encourage the organization to consider other forms of contacting specific clusters of members. Calling costs increasing to $25 after contacting 60,000 members means the organization will lose money if the uplift from calling a member will be less than $25. However, cheaper forms of communication, such as an email or an automated message, might be more appropriate for lower uplift members. Not only would new forms of contact have a possibility of increasing the probability of donation, but they would also be providing a rich source of data for the organization to study; the effects of other methods for contacting members can be studied identically to calling. Possible future actions could include cheaper forms of communication for members with a lower uplift.