**MSBA 434 - Data Mining & Visualization**

**Direct-Maling Marketing Campaign Project**

**Classification Models**

*Conducted by Louie Tran*

# **Introduction**

## **Problem Summary**

The purpose of this project is to help a veterans’ organization to develop a predictive model to improve the cost-effectiveness of their direct marketing campaigns. The organization, with its in-house database of over 13 million donors, is one of the largest direct-mail fundraisers in the United States. According to their recent mailing records, the overall response rate is 5.1%. Out of those who responded (donated), the average donation is $13.00. Each mailing, which includes a gift of personalized address labels and assortments of cards and envelopes, costs $0.68 to produce and send. Thus, this is not cost-effective because the expected profit from each mailing is $13.00 x 5.1% - $0.68 = -$0.017. To help this organization improve their campaign’s effectiveness, I will develop a classification model using data from the most recent mailing campaign that best captures the likely donors so that the expected net profit is maximized.

## **Outline Approach**

In this project, I will use two different models, K-NN and Naive Bayes on the same training data, using the same group of predictor variables. To derive a better solution, I will compare these two models’ predicted expected net profits, their accuracy, and lift charts on the same training data set. After that, the final model will be used to predict a new dataset that the model was not trained on to come up with a list of potential donors.

# **Dataset Preliminary Analysis**

## **Dataset Overview**

The direct-mailing marketing campaign sample dataset contains 3,120 observations that detail characteristics of sampled individuals with weighted sampling being used so that the sample has equal numbers of donors and non-donors. From these observations, I will randomly divide this dataset into two sets: training data with 60% observations and validation data with 40% observations from the original dataset; training data is going to be the set that my classification models are built upon. I also set the seed of both models to 1234 to have the same random training dataset.

Overall, in this dataset, there are 24 variables, 8 of which will be used in the process of building K-NN and Naive Bayes Models. The predictor variables are mixtures of both numerical and categorical variables.

## **Why is oversampling appropriate?**

Using weighted sampling is a normal practice in classification problems. In classification models, the disparity in the frequencies of the observed target variables can have an influential negative impact on the overall fitting of a model. A normal random sample can potentially weigh heavily towards any particular class that has more observations in the dataset. In this case, specifically, the non-donors are more common and if we didn’t use the weighted sampling practice, it could produce a biased result, which is something we always want to avoid in data analysis.

## **Variables Selection**

Out of 20 predictor variables, I decided to use 7 of them for this project, meaning that both K-NN and Naive Bayes will use the same set of variables. To understand which variables have a correlation with or have a significant impact on the target variable (TARGET\_B), I used logistic regression on TARGET\_B and every predictor variable individually to achieve their p-values, which were used to rank the significance of all predictor variables. The table below shows the result of the 7 variables that I picked. With a p-value of 0.05, **INCOME** and **HOMEOWNER** did not make the cut but since their p-values are borderline, I decided to incorporate them into this project.

|  |  |
| --- | --- |
| **Variable Name** | **P-Value ( <= 0.05)** |
| NUMCHLD | 0.0112 |
| homeowner.dummy | **0.0971** |
| INCOME | **0.0614** |
| LASTGIFT | 0.0000124 |
| AVGGIFT | 0.00000969 |
| NUMPROM | 0.0000331 |
| totalmonths | 0.0000000000000472 |

## **Response Variable**

In this project, I’m only focusing on the classification response variable which represents whether or not a mail-campaign recipient donated to the organization (TARGET\_B), and the other response variable which records the amount donated (TARGET\_D) will be ignored.

## **Categorical Predictor Variables**

**Number of Children (NUMCHLD**) is a categorical variable with four levels: 1, 2, 3, and 4, which represents the number of children each veteran has.

**Homeowner (homeowner.dummy)** is a binary variable with 1 representing people who own a house and 0 representing people who do not own a house.

**INCOME** indicates the veterans’ household income. It is not a natural categorical variable but was divided into eight categories.

## **Numerical Predictor Variables**

Out of the seven chosen variables to use in this project, four of them are natural numerical variables.

**LASTGIFT** records the amount of the most recent gift in dollar

**AVGGIFT** represents the average amount of gifts to date in dollar

**NUMPROM** indicates the lifetime number of promotions received to date

**TOTALMONTHS** shows the number of months from the last donation to July 1998, which was the last time the case was updated

## **Net Profit Calculation**

Because in this dataset, weighted sampling was used; therefore; to accurately calculate the expected net profit of each mail-campaign recipient, I will have to divide each row’s net profit by the oversampling weights applicable to the actual category of the row. After calculation, the expected net profit for a donor is $1.257 and for a non-donor is -$1.283, which will be re-coded as:

**(Non-Donor Net Profit) np\_0 = -1.283**  **(Donor Net Profit) np\_1 = 1.257**

# **Models Building and Validation**

## **K-NN Classification**

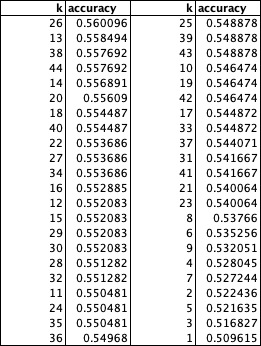
### **Variables Standardization**

Because all the variables used in this project are records of different units and measurements; variable standardization is necessary to ensure that no single variable can have a dominant effect on the target variable because of its original form and all variables will be equally considered in this model.

### **K Value**

To find the best k value to use in this method, I started by determining the largest k value I want to check. Based on my research with several credible sources, which say that the k value for a data set would be equal to the square root of the total observations, I have calculated my largest k value to be 44. In R, I ran a for loop to achieve a list of the accuracy of these 44 k values, which are represented in the table below.

Based on the table below, k with a value of 26 yields the highest accuracy, which is 56%; however, to create a more simple model and to avoid potentially over-fitting data, I picked 13, the value with the second-best accuracy, which is 55.8%, to be my k-value.



### **Net Profit**

**Net profit formula** = #Donors \* NP\_1 + #Non-Donors \* NP\_0

**Non-Donor Net Profit** : np\_0 = -1.283

**Donor Net Profit** : np\_1 = 1.257

After running the K-NN algorithm in R, below is the table showing the prediction result used on the training data set and the actual data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Prediction** | | **Training** | |
|  | Non-Donors | Donors | Non-Donors | Donors |
| Count | 943 | 929 | 915 | 957 |
| Percentage | 53.7% | 46.3% | 48.9% | 51.1% |
| Net Profit/Type | -$1,217.074 | $1,167.419 | -$1,180.936 | $1,202.604 |
| Total Net Profit | -$49.655 | | $21.668 | |
| Difference | $71.323 | | | |

## **Naive Baye’s Classification**

### **Variables Transformation**

First of all, I added a new variable called YEAR by dividing all elements of TOTALMONTHS by 12 and rounding them up by assigning those new values to column YEAR in the training dataset. In this model, because to use Naive Baye’s, every variable has to be categorial, I first transformed all numerical variables and INCOME into categorial by binning them into 3 categories: “Low”, “Medium”, “High” since the variables’ units or in dollars using “OneR” package. After that, I use the function factor() to convert all variables to factor type.

### **Accuracy**

To measure the accuracy of this method, I first ran Naive Baye’s algorithm on the training data set to get a list of predicted values. Next, I used the function confusionMatrix() to compare the result of the predicted values and that of the training dataset to obtain the accuracy of this method, which is 53.9%.

### **Net Profit**

Net profit formula = #Donors \* NP\_1 + #Non-Donors \* NP\_0

NP\_0 = -1.283

NP\_1 = 1.257

After running the Naive Baye’s algorithm in R, below is the table showing the prediction result used on the training data set and the actual data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Prediction** | | **Training** | |
|  | **Non-Donors** | **Donors** | **Non-Donors** | **Donors** |
| **Count** | 960 | 912 | 915 | 957 |
| **Percentage** | 51.3% | 48.7% | 48.9% | 51.1% |
| **Net Profit/Type** | -$1,239.014 | $1,146.056 | -$1,180.936 | $1,202.604 |
| **Total Net Profit** | -$92.958 | | $21.668 | |
| **Difference** | $114.626 | | | |

## **Models Comparison**

|  |  |  |
| --- | --- | --- |
|  | **Prediction** | |
|  | **K-NN** | **Naive Baye’s** |
| **Net Profit** | -$49.655 | -$92.958 |
| **Difference** | $71.323 | $114.626 |
| **Accuracy** | 56.5% | 53.9% |

### 

### **Lift Charts**

### 

Between these two lift curves produced based on two models: K-NN and Naive Bayes, there is a distinct difference in the depths of them. K-NN’s chart drops significantly two times lower and sharper as well as picks up faster than that of Naive Bayes.

### **Final Model**

Based on the summary table above, I can see that not only K-NN’s accuracy is higher than that of Naive Bayes, the later result of the two models’ expected profits compared with the actual one has proven that K-NN is more accurate since its result of -$49.7 is closer the actual data, which is roughly $21.7. Not only that, among the two lift charts, KNN’s chart seems to be more dominant with a deeper and sharper drop and a quicker pickup. Therefore, out of the two models, I have decided to used K-NN to predict the new dataset.

## **Prediction**

### **Data Transformation**

Similar to the process done with the training data in the analysis of the K-NN model, I will first standardize all the variables except for the response variable to ensure that every variable will be equally considered in this algorithm.

### **Results**

This table below represents the prediction for future fundraising plans after running K-NN with k value = 13 in R. After sorting the list of the probability of potential donors, I would go through approximately 60% of the list because this number is close to the donor rate produced by K-NN. I decided to add a few more percents to make minimize the chance of missing any potential donors.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Target Variable** | **Count** | **Percentage** | **Net Profit/Type** | **Total Net Profit** |
| **Donors** | 1141 | 57.1% | $1,433.826 | $325.1665 |
| **Non-donors** | 859 | 43.9% | -$1,108.66 |

# **Conclusion**

The result produced by the K-NN model (k=13) indicates that almost 60% of the mail recipients will donate, which is a good sign. However, in order to predict the net profit more accurately, the organization should develop a predictive model to predict how much a donor would donate based on the given information to understand which factors have the most impact. To maximize the profit, even more, understanding which factors are the most important is not enough, they need to re-evaluate their marketing campaign to see which elements need improvement and what else they can do to incentivize people to donate.