IDS 572 - Data Mining for Business

Assignment 2: Target Marketing Fundraising

Team Members:

Mrunal Ghorpade (677441117)

Neha Chimata (655196210)

Tanvi Sethi (672107527)

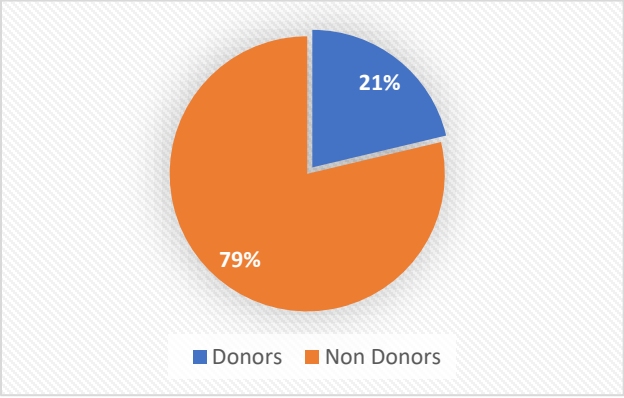
INTRODUCTION:

A national veteran's organization wishes to develop a data mining model to improve the cost effectiveness of their direct marketing campaign. 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. Using these facts, we take a sample of this dataset to develop a classification model that can effectively capture donors so that the expected net profit is maximized. Weighted sampling is used, under-representing the non-responders so that the sample has a more balanced number of donors and non-donors.

ASSIGNMENT QUESTIONS:

1) You will find below a list of subsets of variables that were found useful in earlier analysis. Which variables will you consider for modeling (and why)? Which attributes will you omit from the analyses and why. How do you clean the data, handle missing values? What new attributes/values do you derive? How do you approach data reduction? What methods for data reduction do you try? Data cleaning - certain variables have 'empty' values in many rows. Some of these may be actual missing values, while the empty values may carry information (e.g. for a variable like college Education, empty values may indicate no-college-education which can be coded as a specific value). Some variables carry separate information in different bytes..... Outline the data cleaning steps that you perform (and rationale). Data exploration: Import the data, and examine the different variables – distribution of values, mean and std deviation, range of values. What do you observe? What variable transformations do you make (and why)? Perform Principal Components Analysis (PCA) – which variables do you include for PCA (give your reason). Do decision trees help determine which variables to include in a predictive model for donors? How? (You 2 can try the Weight by Information Gain and other related operators under Feature selection operators in RapidMiner). Also try Random Forest based variable selection.

The dataset contains 16111 data points and 480 attributes. We observe that the donors to non-donors ratio is 1:4 (i.e.) data has 21% donors (TARGET–B = 1) and 79% non-donors (TARGET–B = 0) and a graphical representation of the same is given below:



Missing Values:

Yes, 371 columns had missing values after removing certain columns based on intuition and are treated with the help of "Replace Missing Values Operator"," Map Operator" and "Impute Missing Values Operator".

VARIABLES	ACTION
BIBLE,BOATS,CARDS,CATLG,CDPLAY,COLLECT1,CRAFTS,FISHER,GARDENIN,HOMEE,KIDSTUFF,PCOWNERS,PETS,PHOTO,PLATES,STEREO,VETERANS,WALKER	Used Map Operator to map "?" to "N"
MAGFAML,MAGFEM,MAGMALE,MBBOOKS,MBCOLECT,MBCRAFT,MBGARDEN,MDMAUD,MDMAUD_A,MDMAUD_F,MDMAUD_R,PUBCULIN,PUBDOITY,PUBGARDN,PUBHLTH,PUBNEWFN,PUBOPP,PUBPHOTO,SOLIH,domainSES,urbanicity	Used Impute Missing Values Operator with the help of KNN to fill the missing values.
All the Numeric Variables with missing values	Replaced missing values with special value -1.

Variable selection:

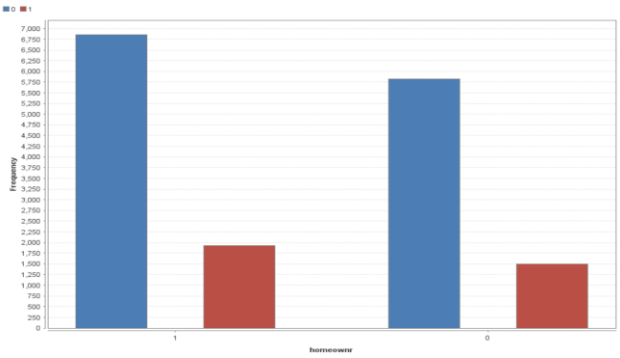
The list of potential base variables chosen based on intuition are given below:

VARIABLE	RESASON
AGE	Donation pattern can be observed in certain age-group of donors.
HOMEOWNR	This variable chosen because response to mail orders differed based on house ownership of donor.
GENDER	The variable was chosen because it was observed that there were more female donors than male.
TIMELAG	This variable helps us in understanding the frequency of donation by a donor.
DOMAIN	It was observed that number of donors varied significantly based on Socio-economic level of the neighborhood.
HIT	The number of times a donor has responded to mail orders before, is a crucial variable for predicting the response of a future mail order.
MAJOR	Major donors have been observed to have good response to mail orders. It is important to identify the size of this population.
WEALTH2	It was observed that there was a significant difference in the distribution of number of donors in each category.
INCOME	Household income of the donor will indicate the donating capacity and hence important in determining the response to the mail.
NUMCHILD	Donation pattern varied based on number of children of the owner.
NUMPROM, NUMPRM12	Response patterns were observed to vary if the number of promotions sent, especially in the last 12 months, were high.

The statistical summary of the base variables chosen are given below

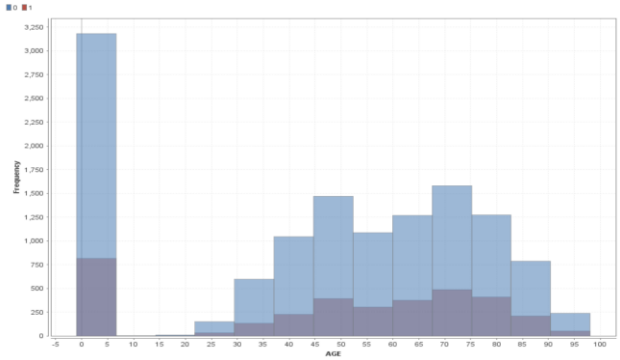
VARIABLE	MISSING	MINIMUM	MAXIMUM	MEAN	STD DEV
NUMPROM	0	5	170	47.455	22.804
wealth2	7308	0	9	5.026	2.801
NUMCHILD	14053	1	5	1.507	0.78
Income	3634	1	7	3.916	1.86
Hit	0	0	241	3.376	9.553
Age	3995	1	98	61.721	16.545
time lag	1537	0	89	8.031	6.231
numprm12	0	2	58	12.845	4.522

Observations:



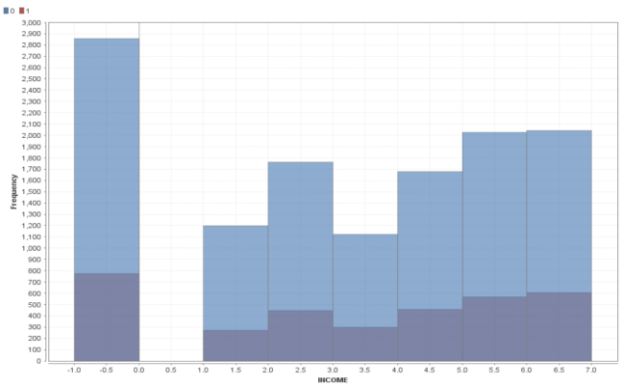
Home Owner:

It can be seen that, people who own houses donated more frequently when compared to those who donot own houses.



Age:

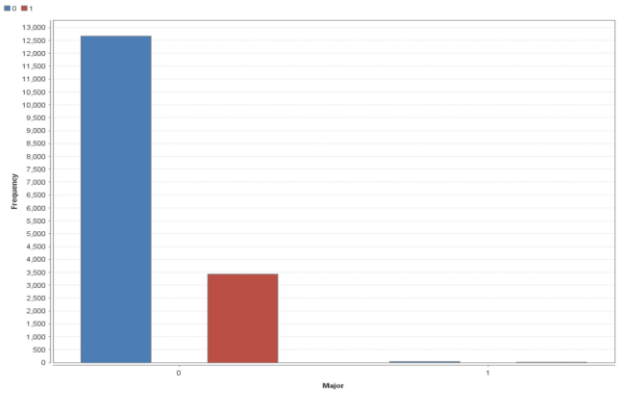
Apart from the missing values, a normal distribution can be observed for both Age of Donor and Donation Response.i.e. Donors of all age groups can be targeted for promotion.



Income:

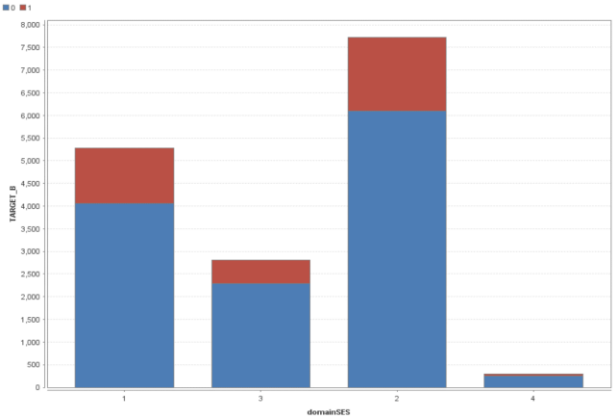
Except the missing values, An Increasing trend can be observed in the income and the donation response variables.

Donors from higher income groups are more likely to respond to a donation mail.



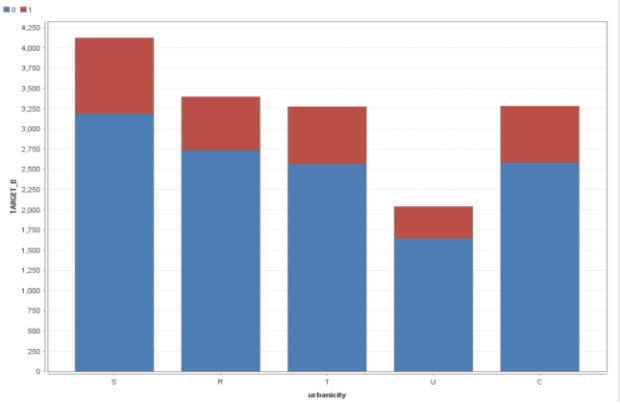
Major:

Number of major donors are 31 and 16080 are not major donors. It can be seen from the below plot that 21.76% of people who are not major donors are responders.



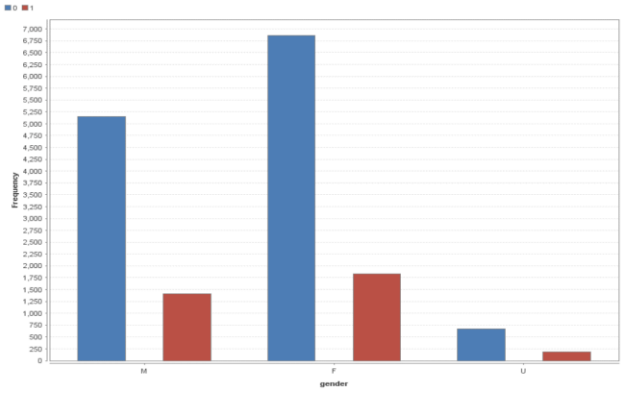
Domain:

Responses are greater from the highest and average Social economic status neighborhoods.



Urbancity:

It can be seen that, except for the urban level of the Urbanicity all others have similar percentage of response.



Gender:

(Post Transformation)

It was observed that there were more female donors than male.

Variable Transformations:

Certain columns are transformed to improve readability and interpretability as follows:

NEW VARIABLE	TRANSFORMATION EQUATION
EP_PvaState	if(PVASTATE =="E" ||PVASTATE=="P", "1", "0")
rechinse	if(RECINHSE == "X", "1","0")
recp3	if(RECP3 == "X", "1", "0")
recsweep	if(RECSWEEP == "X", "1", "0")
urbanicity	cut(DOMAIN, 0, 1)
pepstrfl	if(PEPSTRFL=="X", "1", "0")
child3	if(CHILD03=="M" || CHILD03=="F"||CHILD03=="B", "1", "0")
child18	if(CHILD18=="M" || CHILD18=="F"||CHILD18=="B", "1", "0")
homeownr	if(HOMEOWNR=="H", "1", "0")
totDays	date_diff(LASTDATE, FISTDATE)/(1000*60*60*24)
avgcresp	if (CARDPROM > 0, CARDGIFT/CARDPROM, 0)
avgallresp	if (NUMPROM > 0, NGIFTALL/NUMPROM, 0)
lastPromToGiftGap	date_diff(LASTDATE, MAXADATE)/(1000*60*60*24)
lastToMaxGiftRatio	if (MAXRAMNT>0, LASTGIFT/MAXRAMNT, 0)
maxToMinGiftRatio	if(MINRAMNT > 0, MAXRAMNT/MINRAMNT, 0)
avgGapBetwGifts	if (NGIFTALL > 0, (date_diff(LASTDATE, FISTDATE)/(1000*60*60*24))/NGIFTALL, 0)
domainSES	cut(DOMAIN,1,1)
mailcode	if(MAILCODE == "B", 0, 1)
child07	if(CHILD07=="M" || CHILD07=="F"||CHILD07=="B", "1", "0")
child12	if(CHILD12=="M" || CHILD12=="F"||CHILD12=="B", "1", "0")
recpgvg	if(RECPGVG == "X", "1", "0")
gender	if(GENDER=="M"||GENDER=="F",GENDER,"U")

Variable Omission:

Post data exploration and understanding, certain variables are omitted from the further analysis. The variables and the reason behind omitting them are as follows:

VARIABLE	REASON
ODATEDW, OSOURCE, TCODE, NOEXCH, AGEFLAG, DATASRCE, SOLP3, GEOCODE2,NEXTDATE,HPHONE_D,GEOCODE2, NGIFTALL , CARDGIFT	Attributes not relevant due to one of the following reasons: 1. Irrelevant information related to donor (e.g. Salutation) 2. Date of the second gift which is not needed as it cannot be used for future prediction 3. Use of such field in the model does not make sense w.r.t to the response variable 4. Redundant information as similar information can be obtained form other variables

DOB	Age of donor is available
CLUSTER	CLUSTER2 is the code i.e. the nominal symbolic field and is redundant as filed CLUSTER is being used in the model.
ADATE_2 -> ADATE_24	Date on which mails were send, most of which were sent more than one year ago and its too much of information which is available in other variables
RAMNT_3 -> RAMNT_24	Amount donated during previous campaigns which have been summarized in other variables making it irrelevant to the current response predictions.
RFA_2 -> RFA_24	RFA Status during previous campaigns which have been summarized in other variables making it irrelevant to the current response predictions.
RDATE_3 -> RDATE_24	Date on which previous donations were made, most of which were done more than one year ago

VARIABLE SELECTION BASED ON PCA, RANDOM FOREST AND DECISION TREES:

1) Principal Component Analysis (PCA):

PCA helps in reduction of variables by performing orthogonal transformation (by creating eigen vectors which represents new dimensions) and obtaining a set of new variables referred to as Principal Components. The new variables are linearly uncorrelated and can be less than or equal to the original number of variables. The first component of PCA has the highest variance i.e. contains maximum information the second component contains 2nd largest variance and so on. Based on the cumulative variance of PCs and maximum number of PC's to be incorporated we select how many Principal Components should be selected for our model.

Since the number of variables is high, it is important to reduce them but without compromising on the information content. The variables that are used for PCA are grouped into 3 parts according to their relationship with one another and based on the relevance to our target variable i.e. TARGET_B. On each group we have performed PCAs separately.

Details about the PCAs is as follows:

PCA 1:

In this PCA we have 18 variables (as listed below) which tell us about the donor's interest. As these variables are nominal so we converted it to numeric and then PCA component was applied. We have considered PC's till cumulative Variance = 0.7.

Input variables to PCA 1 (Donors Interest):

VARIABLES	
BIBLE	HOMEE
BOATS	KIDSTUFF
CARDS	PCOWNERS
CATLG	PETS
CDPLAY	PHOTO
COLLECT1	PLATES
CRAFTS	STEREO
FISHER	VETERANS
GARDENIN	WALKER

Output:
From 18 variables, 7 principle components were generated.

PCA 2:

In this PCA we have 14 variables (as listed below) which tell us about the times a donor has responded to other types of mail order offers. As these variables are nominal so we converted it to numeric and then PCA component was applied. We have considered PC's till cumulative Variance = 0.7V.

Input variables to PCA 2 (Donors Response):

VARIABLES	
MAGFAML	PUBDOITY
MAGFEM	PUBGARDN
MAGMALE	PUBHLTH
MBBOOKS	PUBNEWFN
MBCOLECT	PUBOPP
MBCRAFT	PUBPHOTO
MBGARDEN	PUBCULIN

Output:
From 14 variables, 7 principle components were generated.

2) Random Forest:

Random forest consists of several decision trees. Every node in these decision trees is split on a single variable amongst m randomly chosen variables, based on which attribute gives the purest node. The weight of each attribute is calculated by analysing the split points of Random Forest model. Attributes with higher weights are considered more relevant and important.

We have used the Random Forest model on the variables related to donors neighbourhood and they are as follows:

VARIABLES			
AC1	EIC1 -16	HUPA1 - A7	OCC1 - 13
AC2	ETH1 -5,8,9,13,14,16	HUR1 -2	OEDC1 - C6
ADI	ETHC1 - C5	HV1 – 4	POBC1
AFC4	HC1- 8, 10-19,21	HVP1 -6	POBC2,
AFC5	HHAGE1 -3	IC1 – 23	POP90 1-3. C1-C5
AGE901 -7	HHAS1 – 4	LFC1 -10	RP1 - 4
ANC1 -10	HHD1 – 12	MARR1 -4	SEC1 - 5
DMA	HHN1 -6	MC1 – 3	TPE1 - 4, 8-13
DW1 -8	HHP1 -2	MHUC1	VC1 - 4
EC1 -8	HU1 – 5	MSA	VOC1 -3

The following criterion was used while implementing the Random Forest model:

Number of Trees: 100

Criterion: Gain ratio

Maximum Depth: 20

After the variables are weighted we selected only those who have weights greater than 0.15, after which amongst 235 input variables 113 variables were selected.

Even after applying the random forest model the number of variables is high. Hence, we applied an PCA after random forest model. As most of the variables are in percentage, range transformation (0.0 to 1.0) is applied to ensure the principal components are only scaled to smaller value. We have considered PC's till cumulative Variance = 0.8.

Output:

From 235 variables, selected 113 variables with the use of Random forest model and amongst these 113 variables, 8 principle components were generated.

3) Decision Tree:

Different types of decision trees (Gini, Gain Ratio, Information Gain) were applied on the entire dataset (after treating the missing values) with an objective of identifying few important attributes which could be used in our classification model. However, the results were unsubstantial, due to which identifying important attributes using decision trees was not possible.

2) Modeling

Partitioning - Partition the dataset into 60% training and 40% validation (set the seed to 12345).[A specified seed ensures that we obtain the same random partitioning every time we run it. With no specified seed, the system clock is typically used to set the seed, and a different partitioning can result in different runs]. Consider different classification techniques on the data. (Decision Trees (you can use J48, or any other suitable type of decision tree)/ Logistic Regression, using Ridge and Lasso/ Naïve-Bayes/Random forest/ Boosted trees) Run each method on a chosen subset of the variables - how do you determine which variables to include in the data for modeling with each of the methods above? Consider whether the different methods above. Incorporate mechanisms for variable selection? Be sure to test different parameter values for each method, as you see suitable. What parameter values do you try for the different techniques, and what do you find to work best? (Be sure NOT to include "TARGET–D" in your analysis. (why?)) Provide a comparative evaluation of performance of your best models from each technique. Does variable selection/PCA make a difference for the different models?

We partitioned the PVA dataset into 60:40 ratio and set the random seed to 12345, to ensure same random partitioning every time the dataset is run.

We train our model using different classification technique and different subset of variables with each technique to get the best model to predict the donors. Since the misclassification cost of the donors who responded is $13 and misclassification cost of who did not respond is $0.68, we focus on getting higher rates of true cases of donors i.e. the recall rate of class 1. The various techniques and the outputs on different subsets used are outlined below:

1. Decision Trees (J48 operator)
2. Logistic Regression Using Ridge and Lasso
3. Naïve Bayes Classifier
4. Random Forest
5. Boosted Trees

The subset of variables which gives the best recall rate with each classification technique would be selected to be compared among other models.

1. Decision Trees (J48 Operator)

J48 operator was used with the following parameters:

PARAMETERS	VALUES
U	Unchecked
C	0.25
M	5
R	Unchecked

N	
B	Checked
S	Unchecked
L	Checked
A	UnChecked

The results of the decision tree are as follows:

VARIABLES	ACCURACY		RECALL RATE (CLASS 1)	
	TRAINING	TEST	TRAINING	TEST
All Variables	91.95	69.09	74.08	21.06
Base Variables+PCA1 + PCA2+ Random Forest (Donor's Neighborhood)	87.61	72.1	52.94	18.85
Base Variables+PCA1+PCA2	93.21	69.01	78.71	21.73
Base Variables+ PCA2 + Random Forest (Donor's Neighborhood)	86.89	73.9	47.88	17.81

With the above results, we observed that the accuracy is highest for the model trained with Base Variables, PCA1 and PCA2, however the recall rate is low for all the models.

2. Logistic Regression

LASSO

The parameters chosen for the best model and its output with various subsets of variables are:

PARAMETERS	VALUES
Solver	Auto
Use Regularization	Yes
Lambda	0.01
Alpha	1
Standardize	Yes
Add Intercept	Yes
Remove Collinear Columns	Yes

The results from the logistic regression model with LASSO are as follows:

VARIABLES	ACCURACY		RECALL RATE (CLASS 1)	
	TRAINING	TEST	TRAINING	TEST
All Variables	52.56	52.11	70.04	69.4
Base Variables+PCA1 + PCA2+ Random Forest (Donor's Neighborhood)	51.47	50.65	71.68	69.55
Base Variables+PCA1+PCA2	52.56	52.11	70.04	69.4
Base Variables+ PCA2 + Random Forest (Donor's Neighborhood)	51.47	50.65	71.68	69.55

With the above results, it can be observed that model with base variables and all principal components (PCA1, PCA2 and Random Forest) give the best recall rate for the donors.

Ridge Regression

Using the same parameters as LASSO and alpha = 0, we get the following results:

VARIABLES	ACCURACY		RECALL RATE (CLASS 1)	
	TRAINING	TEST	TRAINING	TEST
All Variables	67.06	61.95	59.2	47.01
Base Variables+PCA1 + PCA2+ Random Forest (Donor's Neighborhood)	54.83	52.59	71.15	66.96
Base Variables+PCA1+PCA2	68.21	63.36	56.74	44.35
Base Variables+ PCA2 + Random Forest (Donor's Neighborhood)	57.11	67.63	54.81	62.68

On comparing the Ridge regression with LASSO, we can say that, the LASSO gives us a slightly better model in terms of accuracy and recall rate when the base variables and all the principal components (PCA1, PCA2 and Random Forest) are taken into consideration.

3. Naïve Bayes Classification:

With laplace correction checked, we obtained the following results using this classification which works on the Bayes Theorem.

VARIABLES	ACCURACY		RECALL RATE (CLASS 1)	
	TRAINING	TEST	TRAINING	TEST
All Variables	57.62	56.31	55.63	52.77
Base Variables+PCA1 + PCA2+ Random Forest (Donor's Neighborhood)	72.01	71.28	30.59	29.71
Base Variables+PCA1+PCA2	57.82	56.98	55.73	52.85
Base Variables+ PCA2 + Random Forest (Donor's Neighborhood)	71.97	71.34	30.64	29.56

The best model obtained was with data consisting of the base variables and Principal components of donor's response to other mails(PCA2) and donor's interests(PCA1).

4. Random Forest

The parameters considered while implementing Random forest as the modelling operator are as follows:

PARAMETERS	VALUES
Number of Trees	200
Criterion	Gain_Ratio
Maximal_Depth	50
Apply Pruning	Yes
Confidence	0.25
Apply Prepruning	Yes
Minimal Gain	0.1

Based on the results below, it is observed that, the recall rate for all the possible parameters and variables came out to be 0.

Hence, we will not prefer random forest for modelling.

VARIABLES	ACCURACY		RECALL RATE (CLASS 1)	
	TRAINING	TEST	TRAINING	TEST
All Variables	78.52	79	0	0
Base Variables+PCA1 + PCA2+ Random Forest (Donor's Neighborhood)	78.52	79	0	0
Base Variables+PCA1+PCA2	78.52	79	0	0
Base Variables+ PCA2 + Random Forest (Donor's Neighborhood)	78.52	79	0	0

5. Boosted Trees:

We used Adaboost operator with Decision stump since Decision stump are very efficient when used with Boosted trees. The parameters used and results are as below:

ADABOOST	
Iterations	30
DECISION STUMP	
Criterion	Information Gain
Minimum Leaf Size	5

The results obtained are as follows:

VARIABLES	Accuracy		Recall Rate (Class 1)	
	Training	Test	Training	Test
All Variables	78.52	79	0	0
Base Variables+PCA1 + PCA2+ Random Forest (Donor's Neighborhood)	78.52	79	0	0
Base Variables+PCA1+PCA2	78.52	79	0	0
Base Variables+ PCA2 + Random Forest (Donor's Neighborhood)	78.52	79	0	0

The above observation made it clear that adaboost cannot be used as a classification technique to predict the donors.

Comparison Between the Models That Gave Us Results:

MODELS	VARIABLES	Accuracy		Recall Rate (Class 1)	
		Training	Test	Training	Test
LASSO	Base Variables + PCA1+PCA2+ Random Forest (Donor's Neighborhood)	51.47	50.65	71.68	69.55
RIDGE	Base Variables + PCA1+PCA2+ Random Forest (Donor's Neighborhood)	54.83	52.59	71.15	66.96
J48	Base Variables + PCA1+PCA2	93.21	69.01	78.71	21.73

Observations and Conclusions

- Among the above models, logistic regression with LASSO gave us the best model in terms of recall rate and i.e. 69.55%. It used the base variables and all the PC components i.e. Principal components from donor's interests(PCA1), responses to other mails (PCA2) and random forest with PCA of donor's neighborhood.
- With different variable subsets, different classification techniques gave different accuracy and recall results except for random forest and ada-boost which gave 0 as the recall rate given any combination of the parameters.
- The subset of variables including base variables and all the principal components from PCA and Random forest gave the best results in two of the three selected models.

3) Classification under asymmetric response and cost: What is the reasoning behind using weighted sampling to produce a training set with equal numbers of donors and non-donors? Why not use a simple random sample from the original dataset? (Hint: given the actual response rate of 5.1%, how do you think the classification models will behave under simple sampling)? In this case, is classification accuracy a good performance metric for our purposes of maximizing net profit? If not, how would you determine the best model? Explain your reasoning.

Actual Response rate is 5.1%, so the data set is biased towards Non-Donors and Donors have been under represented. If simple random sampling technique is used to predict significant attributes, it will yield too few Donors w.r.t Non-Donors i.e. generate more number of 0's (non-donors) than 1's(donors). In such case, no reliable conclusions can be drawn from the dataset.

To avoid this problem, we use weighted sampling. Weighted sampling assigns weightage to each data in such a way that Donor Responders who are rare cases get larger weight and the Non-Donors who have more number of cases get lower weights.

Usually to choose a model classification accuracy is supposed to be a good performance measure. But in our case, we want a model that gives us maximum profit. The cost of failing to identify a potential donor ($13) is much higher than sending invites to non-donors (0.68). So here the motivation is to identify as many donors who will respond correctly as possible. This cannot be done by looking at the accuracy. So, the criterion used is Recall and we pick a model to be the best model based on the recall rate.

From the analysis in Question 2, We obtained the best model for this dataset through Logistic Regression with LASSO with the highest Recall rate of 69.55%.