Target Marketing – Fundraising

-Sheethal Heremagalur Sridhar

Background of the Case:

Dataset is provided by Paralyzed Veterans of America (PVA).

PVA is a not-for-profit organization that provides programs and services for US veterans with spinal cord injuries or disease.

Participants in the '98 CUP will demonstrate the performance of their tool by analyzing the results of one of PVA's recent fund-raising appeals.

This mailing was sent to a total of 3.5 million PVA donors who were on the PVA database as of June 1997. Everyone included in this mailing had made at least one prior donation to PVA.

All of the donors who received this mailing were one group that is of particular interest to PVA is **"Lapsed" donors**": Individuals who made their last donation to PVA 13 to 24 months ago. They represent an important group to PVA, since the longer someone goes without donating, the less likely they will be to give again. Therefore, the recapture of these former donors is a critical aspect of PVA's fundraising efforts.

However, PVA has found that there is often an inverse correlation between likelihood to respond and the dollar amount of the gift, so a straight response model (a classification or discrimination task) will most likely net only very low dollar donors.

High dollar donors will fall into the lower deciles, which would most likely be suppressed from future mailings. The lost revenue of these suppressed donors would then offset any gains due to the increased response rate of the low dollar donors.

Therefore, to improve the cost-effectiveness of future direct marketing efforts, PVA wishes to develop a model that will help them maximize the net revenue (a regression or estimation task) generated from future renewal mailings to Lapsed donors.

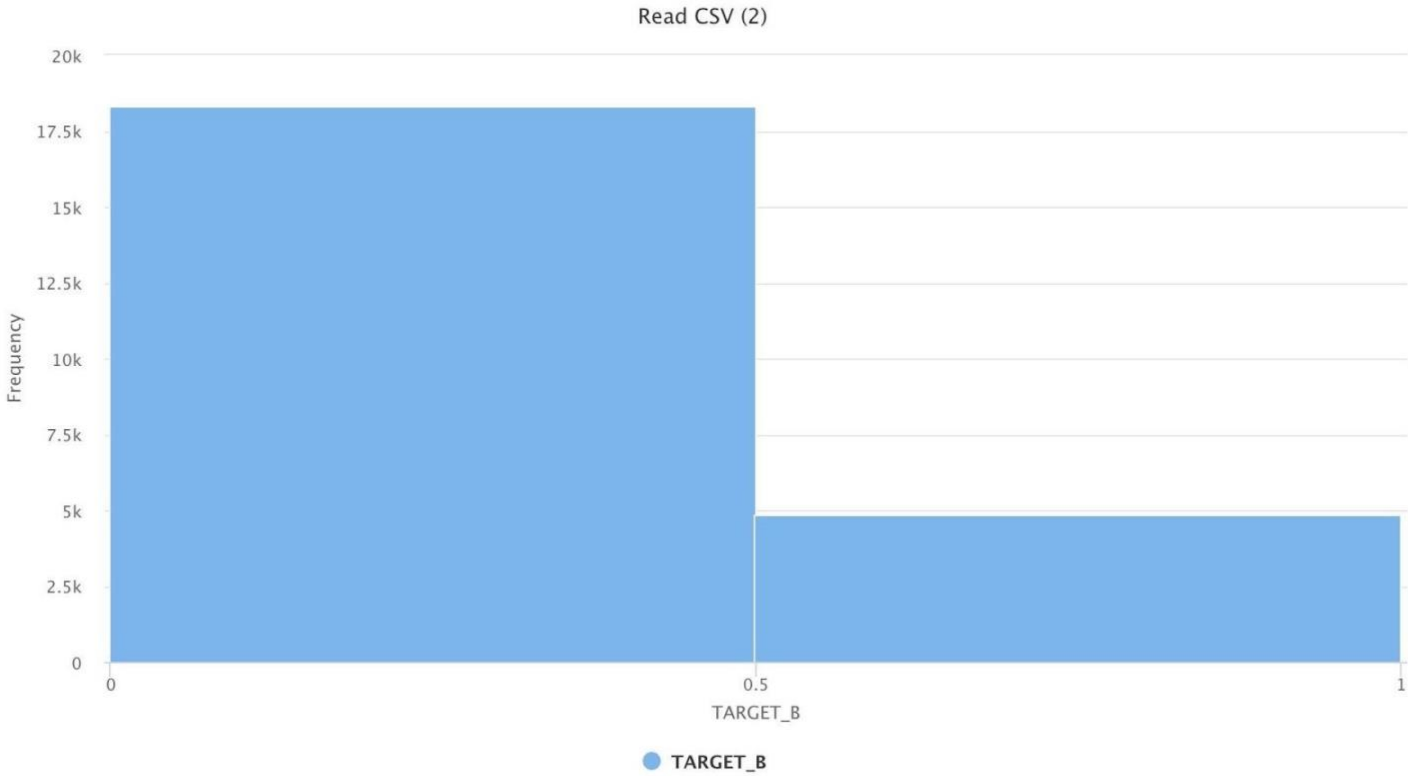
Dataset Analysis:

The dataset contains 23158 examples and 481 attributes in which many of the attributes are not useful for our modeling process. The dataset is characterized by 23% of donors and 77% nondonors represented by the following plot of the Target-B variable. We did Bivariate plots of some attributes with our response variable so that we can later make a good predictive model.

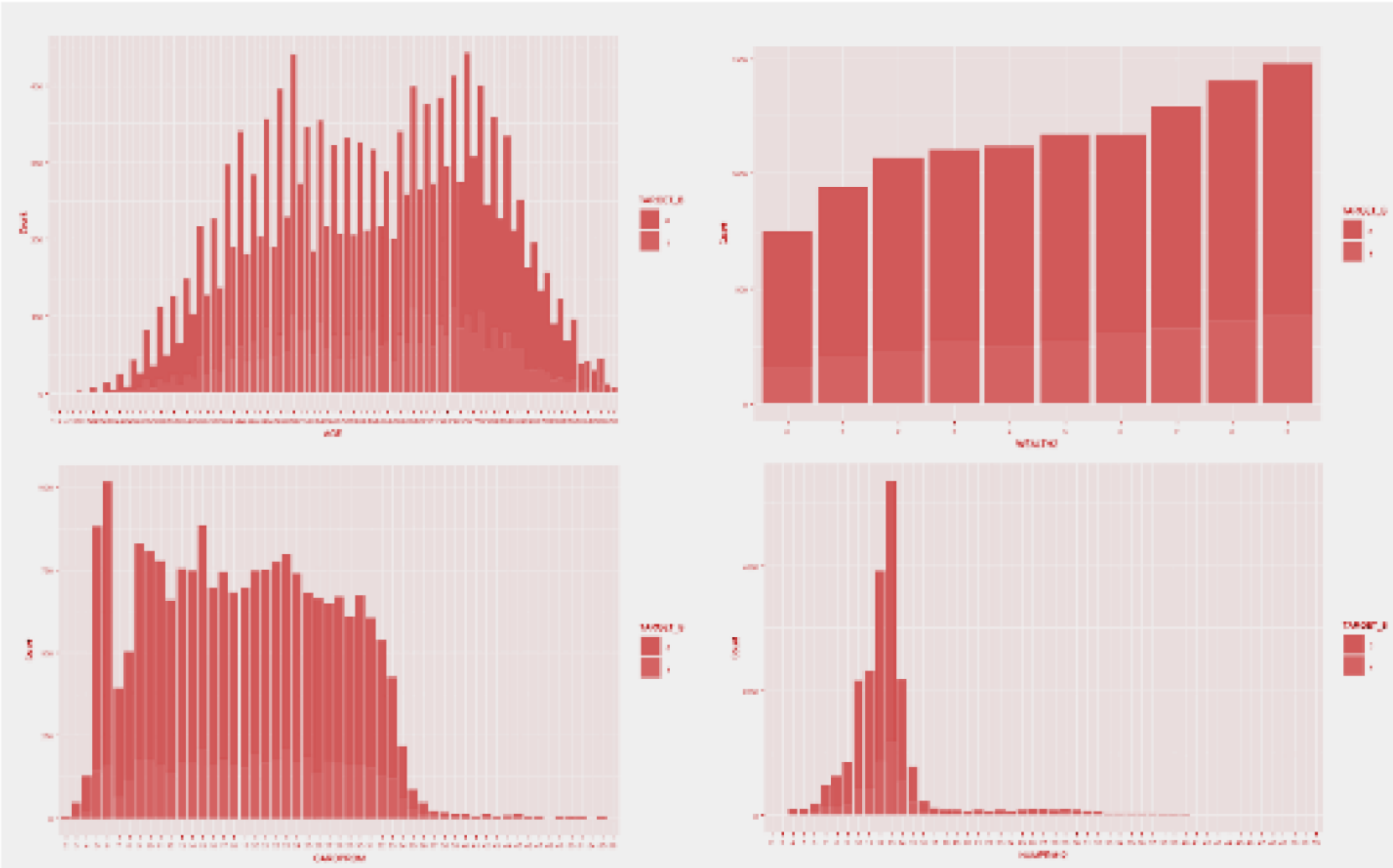
After Exploratory analysis, we found that our dependent variable is TARGET_B, We worked on the dataset to clean it and make it suitable for our later analysis (to do predictive modeling)

Dependent Variable:

We have taken TARGET_B as our dependent variable i.e.
whether the person would donate or not
0 indicates that the person will not donate and the "1" indicates the person will donate.
TARGET_B:
Count "0" = 18315
Count "1" = 4843



Below are the variable plots of Age, Card promotions received, Wealth, and a total number of promotions received in last year. The maximum number of donations are from the age group between 25 to 55, seen from the AGE plot, and the highest number of card promotions are on an average 6, which can seem from the CARDPROM plot. The plot also shows that after a number of promotions the number of donations does not change much. Affluent people tend to donate much more than people who are not wealthy, which is seen from the WEALTH plot as maximum donation count is from the 9th bin in the plot which corresponds to the wealthiest category. And finally from the NUMPRM12 plot, we can see that if the number of promotions goes beyond 13 then the donation count stars decreasing, hence we should avoid sending more than 13 promotions in a year.



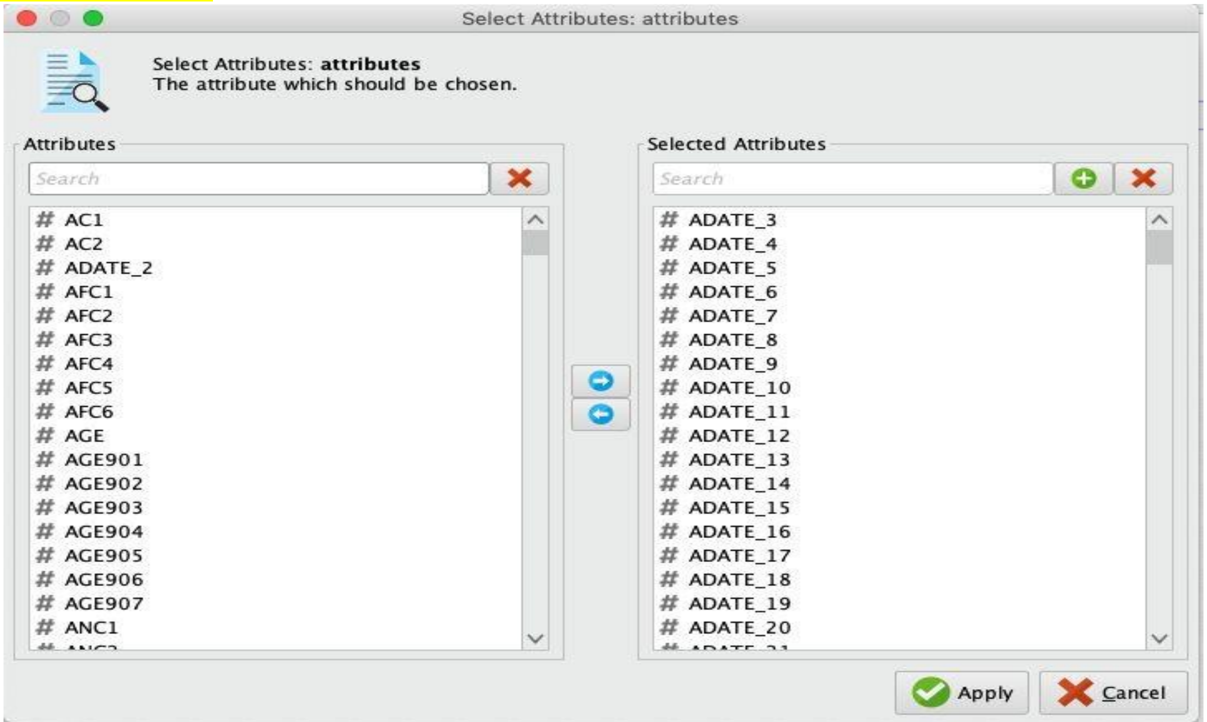
We then clean the data and reduce variables by performing PCA and random forests.

We took the dataset and performed the following steps:

1) Elimination of Unwanted Attributes
2) Generation of New Attributes
3) Mapping of Unknown or "?" values to a new value
4) Handling the Missing values
5) Reducing Attributes using Principal Component Analysis
6) Reducing Attributes using Decision Trees
7) Reducing Attributes using Random Forest

Data cleaning:

1) **Removal/deselection of attributes based on instinct:**
   We have removed many attributes that we do not require for the analysis. These attributes do not affect our dependent variable "TARGET_B" i.e. a person will donate or not After removal, only 325 attributes are left.



2) **To Generate new attributes:**
   We have generated new attributes that modify the values of some existing attributes.
   For eg. RECP3 has 'X' is some rows and is empty elsewhere.
   The new attribute set a value of 1 where there is an 'X', and 0 elsewhere.
   23 more attributes are added, After this step, we now have 348 Attributes.

3) **To Map the missing or '?' values to 'N':**
   Certain attributes have a 'Y' value indicating the presence and are empty (missing, or ? value) elsewhere.
   So, we map the '?' to 'N'.

Variables with Missing values left:
PVASTATE, RECINHSE, RECPGVG, RECSWEEP, RECP3,

PLATES	Binominal	23019	Least Y (139)	Most Y (139)	Values Y (139)
LIFESRC	Polynominal	13014	Least 1.0 (2384)	Most 2.0 (4957)	Values 2.0 (4957), 3
PEPSTRFL	Binominal	11668	Least X (11490)	Most X (11490)	Values X (11490)

Numchild- missing replace with 0
All nominal variables- urbanicity, domain ses and cluster2- Unknown category newly created Cluster 2 - 33 so remove
Remove **LIFESRC**

REAL: Cluster 532 missing values- replace with average or remove them **nominal**

Urbanicity- 532 missing - create a new missing category-U
Domainses- “
Domain remove as we generated urbancity and domainses

polynominal
Gender- 700 missing- replace u+j and missing=>U- unknown merge it
DONE on R- CONVERTED GENDER INTO FACTOR.
1= MALE | 2= FEMALE | 3= UNKNOWN

			Min	Max	Average
MBCRAFT	Real	12754	0	5	0.165
MBGARDEN	Real	12754	0	4	0.061
MBBOOKS	Real	12754	0	9	1.114
MBCOLECT	Real	12768	0	5	0.062
MAGFAML	Real	12754	0	9	0.459
MAGFEM	Real	12754	0	4	0.129
MAGMALE	Real	12754	0	3	0.067
PUBGARDN	Real	12754	0	5	0.136
PUBCULIN	Real	12754	0	4	0.141
PUBHLTH	Real	12754	0	9	0.725
PUBDOITY	Real	12754	0	8	0.230

			Min	Max	Average
PUBDOITY	Real	12754	0	8	0.230
PUBNEWFN	Real	12754	0	9	0.374
PUBPHOTO	Real	12754	0	2	0.006
PUBOPP	Real	12754	0	9	0.230
DATASRCE	Real	5173	1	3	2.492

Merge all - missing values= 12754 in each variable

MBCRAFT	Buy Craft Hobby
MBGARDEN	Buy Gardening
MBBOOKS	Buy Books
MBCOLECT	Buy Collectables
MAGFAML	Buy General Family Mags
MAGFEM	Buy Female Mags
MAGMALE	Buy Sports Mags
PUBGARDN	Gardening Pubs
PUBCULIN	Culinary Pubs
PUBHLTH	Health Pubs
PUBDOITY	Do It Yourself Pubs
PUBNEWFN	News / Finance Pubs
PUBPHOTO	Photography Pubs
PUBOPP	

			Min	Max	Average
AGE	Real	5668	1	98	61.838

Age: Replace missing with **mean = 61.838 missing values: 5668**

			Min	Max	Average
AGE	Real	0	1	98	61.838

After mapping and removing missing values the mean increased to 61.878.

			Min	Max	Average
AGE	Real	0	1	98	61.878

Same process done for other missing values for attributes such as **Income**

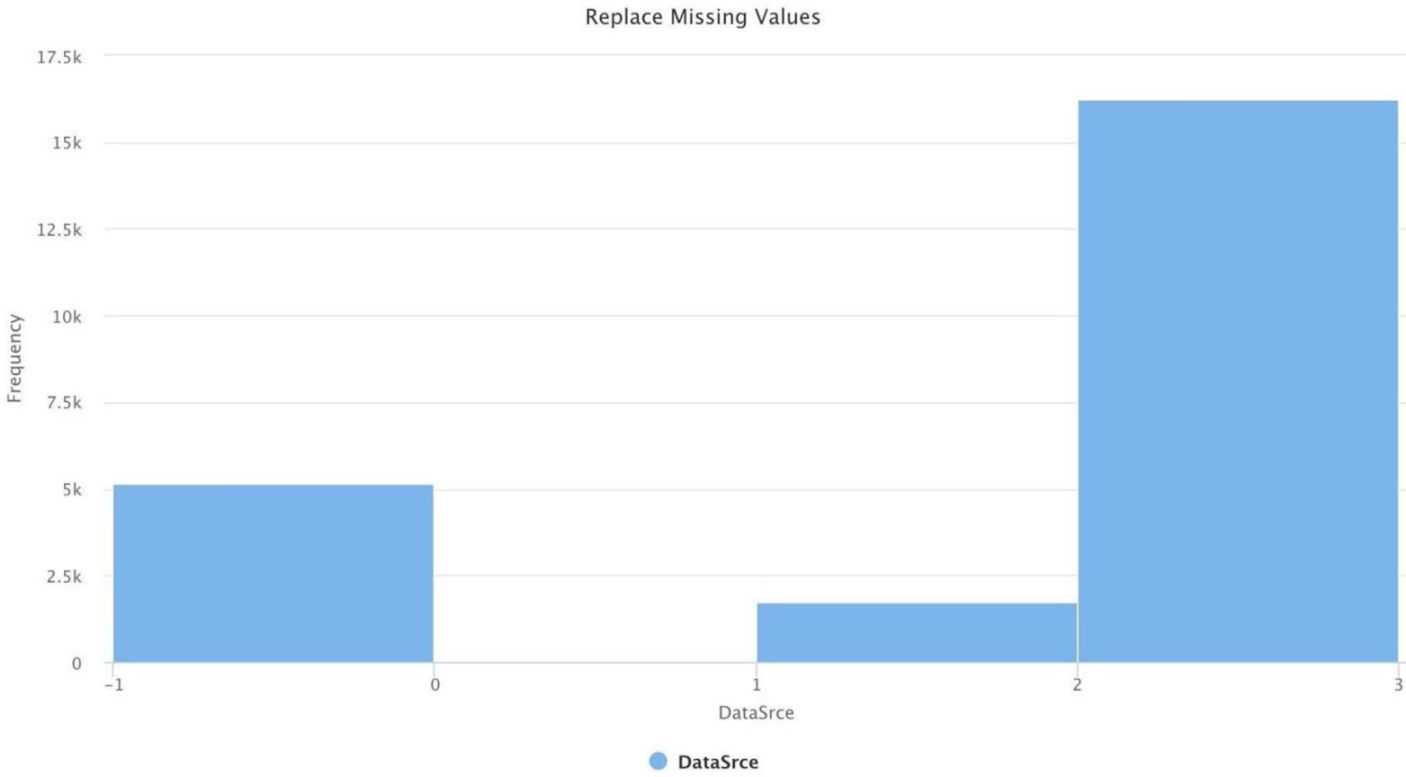
Income: Missing values: 5174, Replaced with Average= 3.922

			Min	Max	Average
INCOME	Real	0	1	7	3.922

Wealth: Merge Wealth 1 and Wealth 2 into one variable by taking the average from values <=0 else replace with maximum wealth, which will treat the missing values in both

			Min	Max	Average
wealth	Real	0	-1	9	0.906

DataSrce: Create a new category for 5173 missing values
CATEGORY 4- UNKNOWN



			Min	Max	Average
MSA	Real	33	0	9360	3560.746
ADI	Real	33	0	645	185.290
DMA	Real	33	0	881	666.095

Remove: MSA, DMA, ADI

SOLP3	Polynominal	23121	Least 01 (3)	Most 00 (17)	Values 00 (17), 12 (
SOLIH	Polynominal	21656	Least 6.0 (1)	Most 12.0 (1391)	Values 12.0 (1391),
MAJOR	Nominal	23094	Least X (64)	Most X (64)	Values X (64)
WEALTH2	Real	10507	Min 0	Max 9	Average 4.994
COLLECT1	Binominal	21811	Least Y (1347)	Most Y (1347)	Values Y (1347)
VETERANS	Binominal	20582	Least Y (2576)	Most Y (2576)	Values Y (2576)
BIBLE	Binominal	20970	Least Y (2188)	Most Y (2188)	Values Y (2188)
CATLG	Binominal	21167	Least Y (1991)	Most Y (1991)	Values Y (1991)
HOMEE	Binominal	22942	Least Y (216)	Most Y (216)	Values Y (216)
PETS	Binominal	19640	Least Y (3518)	Most Y (3518)	Values Y (3518)
CDPLAY	Binominal	20056	Least Y (3102)	Most Y (3102)	Values Y (3102)

Replace with N for missing values:
THESE ARE BINOMIAL VARIABLES

STEREO	Binominal	19992	Least Y (3166)	Most Y (3166)	Values Y (3166)
PCOWNERS	Binominal	20605	Least Y (2553)	Most Y (2553)	Values Y (2553)
PHOTO	Binominal	21956	Least Y (1202)	Most Y (1202)	Values Y (1202)
CRAFTS	Binominal	21087	Least Y (2071)	Most Y (2071)	Values Y (2071)
FISHER	Binominal	21385	Least Y (1773)	Most Y (1773)	Values Y (1773)
GARDENIN	Binominal	19792	Least Y (3366)	Most Y (3366)	Values Y (3366)
BOATS	Binominal	22649	Least Y (509)	Most Y (509)	Values Y (509)
WALKER	Binominal	20553	Least Y (2605)	Most Y (2605)	Values Y (2605)
KIDSTUFF	Binominal	22781	Least Y (377)	Most Y (377)	Values Y (377)
CARDS	Binominal	22896	Least Y (262)	Most Y (262)	Values Y (262)

COLLECT1	COLLECTABLE (Y/N)
VETERANS	VETERANS (Y/N)
BIBLE	BIBLE READING (Y/N)
CATLG	SHOP BY CATALOG (Y/N)
HOMEE	WORK FROM HOME (Y/N)
PETS	HOUSEHOLD PETS (Y/N)
CDPLAY	CD PLAYER OWNERS (Y/N)
STEREO	STEREO/RECORDS/TAPES/CD (Y/N)
PCOWNERS	HOME PC OWNERS/USERS
PHOTO	PHOTOGRAPHY (Y/N)
CRAFTS	CRAFTS (Y/N)
FISHER	FISHING (Y/N)
GARDENIN	GARDENING (Y/N)
BOATS	POWER BOATING (Y/N)
WALKER	WALK FOR HEALTH (Y/N)
KIDSTUFF	BUYS CHILDREN'S PRODUCTS (Y/N)
CARDS	STATIONARY/CARDS BUYER (Y/N)
PLATES	PLATE COLLECTOR (Y/N)

Remove ageflag, as we already have Age variable
Timelag: Replace 2214 missing values with Average= 8

(Created Variable)
AvgGapBwGifts - 2 missing- remove them-
Also, check that all the values are less than 0 which does not make any sense

After Mapping and Replacing missing Values we were left with 330 Attributes.

Name	Type	Missing	Statistics			Filter (330 / 330 attributes):
Id CONTROLN	Real	0	Min 1	Max 191779	Average 96663.476	
Label TARGET_B	Polynominal	0	Least 1 (4843)	Most 0 (18315)	Values 0 (18315), 1 (4843)	
VETERANS	Polynominal	0	Least "1" (2576)	Most "0" (20582)	Values "0" (20582), "1" (2576)	
BIBLE	Polynominal	0	Least "1" (2188)	Most "0" (20970)	Values "0" (20970), "1" (2188)	
CATLG	Polynominal	0	Least "1" (1991)	Most "0" (21167)	Values "0" (21167), "1" (1991)	
HOMEE	Polynominal	0	Least "1" (216)	Most "0" (22942)	Values "0" (22942), "1" (216)	
PETS	Polynominal	0	Least "1" (3518)	Most "0" (19640)	Values "0" (19640), "1" (3518)	
CDPLAY	Polynominal	0	Least "1" (3102)	Most "0" (20056)	Values "0" (20056), "1" (3102)	

Step 5 : Reducing Attributes using Principal Component Analysis

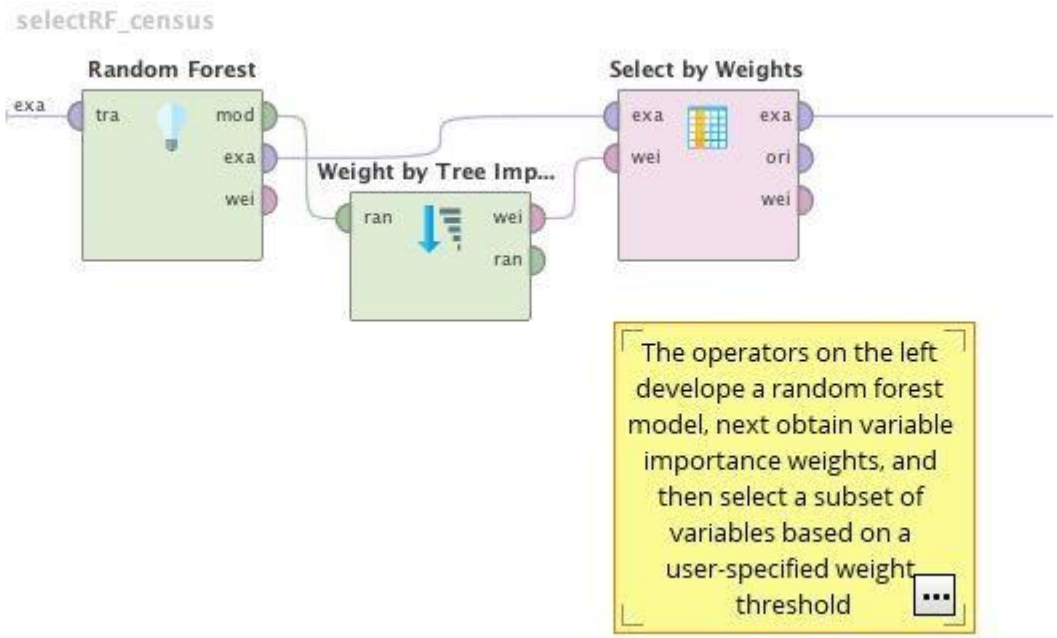
We did the PCA for dimensionality reduction (Attribute reduction).
After the previous transformations, we were left with 330 attributes. So for further processing, we need to reduce the dimensionality or the number of attributes. For that, we categorized the attributes into:

1. Donor’s hobbies and interest
2. Donor’s ability
3. Donor’s neighborhood

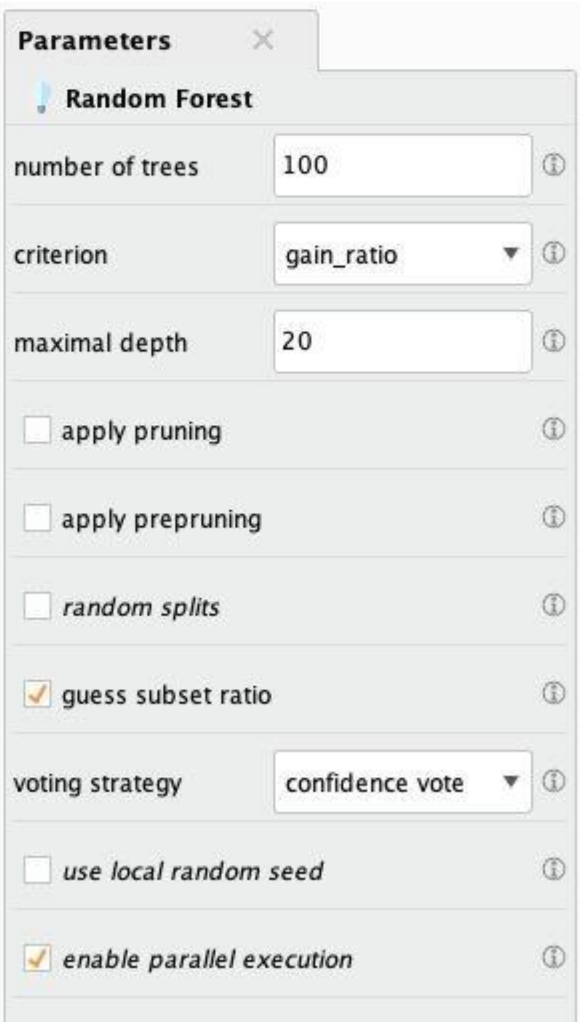
List of PCs	Attributes
PCA1: Donor’s hobbies and interest	STEREO, WALKER, PLATES, KIDS STUFF, PHOTO, CRAFTS, FISHER, HOMEE, BIBLE, BOATS, MBBOOKS, MBGARDEN, MBCRAFT, MBCOLECT, MAGMAIL, MAGFAML, PETS, PCOWNERS, PUBNEWFN, PUBHLTH, PUBDOITY, PUBGARDN, PUBQLIN, CDPLAY, etc.
PCA2: Donor’s ability	NGIFTALL, MAXRDATE, MAXRAMNT, NEXTDATE, TIMELAG, MINRDATE, LASTGIFT, CARDGIFT, RAMNTALL, MINRAMNT, etc.
PCA3: Donor’s neighborhood	RP1-RP4, HU1-HU5, TPE1-TPE9, HHN1-HHN6, SEC1-SEC5, MARR1-MARR4, IC6-IC23, ETH1-ETH16, DW1-DW9, etc.

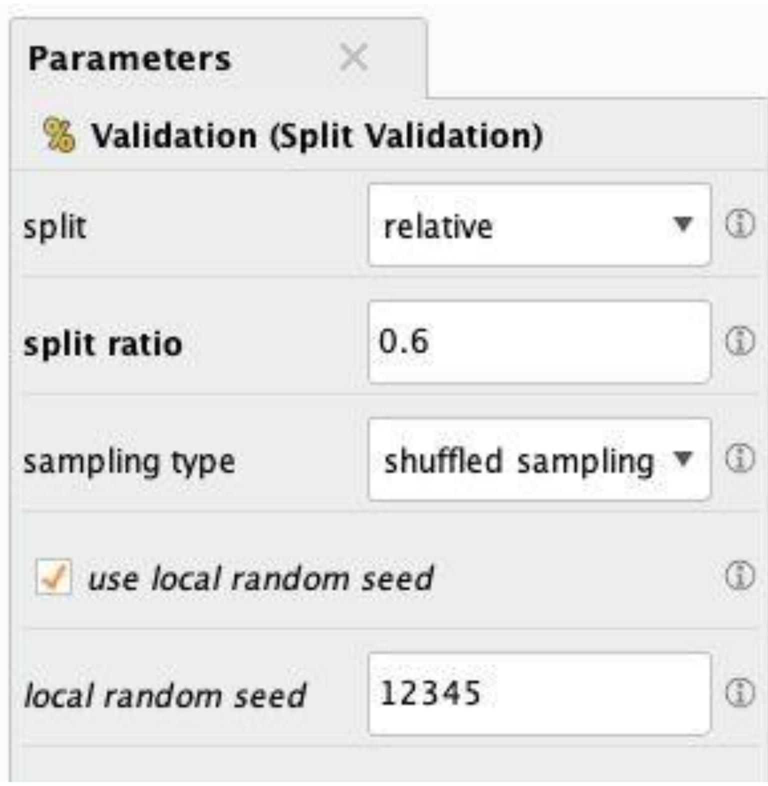
PC 1 consists of 32 variables which indicate a donor’s hobbies and interests. After PCA on these 32 variables, we reduce them to 3 principal components. Similarly, for PC 2 we had 10 variables on a donor’s history which gave us 3 principal components. Finally, for PC 3 we had 160 variables which were reduced to 10 principal components. After conducting the PC analysis we were able to reduce our dataset’s dimension to 144 variables including the 16 principal components.

Step 6: Reducing Attributes using Decision Tree and Random Forest



We did a random forest on following parameters and weighted tree by its importance. The number of trees - 100 and maximal depth - 20. We considered weight - 0.19, based on the weight selection our dataset reduced to 67 variables which are good for building the prediction model.





DECISION TREE:

TRAINING DATA		**ACCURACY: 77.16%**	
	ACTUAL 0	ACTUAL 1	CLASS PREDICTION
PREDICTED 0	9741	2884	.77
PREDICTED 1	0	3	1.0
CLASS RECALL	1.0	0.001	

TEST DATA		**ACCURACY: 76.78%**	
	ACTUAL 0	ACTUAL 1	CLASS PREDICTION
PREDICTED 0	6462	1955	.76
PREDICTED 1	0	1	1.0
CLASS RECALL	1.0	.0004	

LOGISTIC REGRESSION:

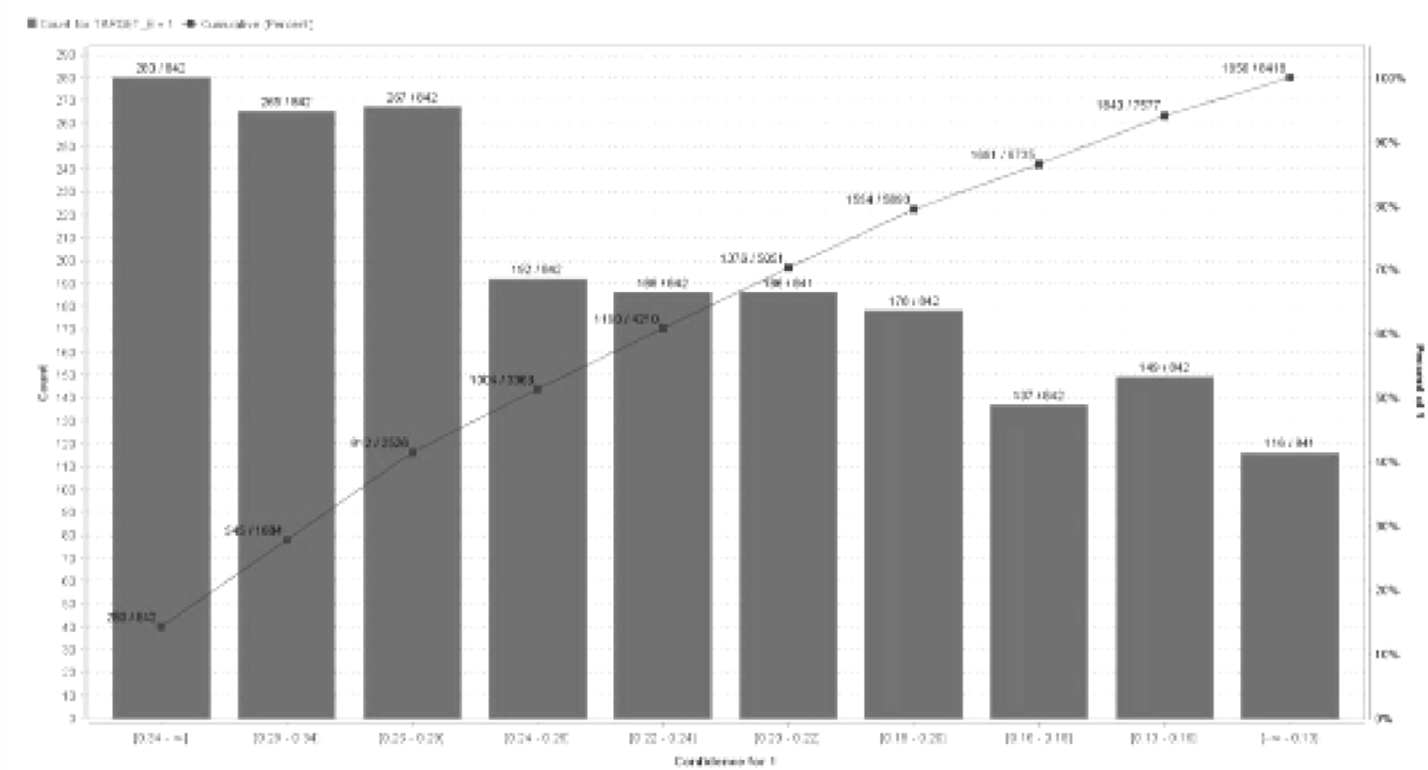
LASSO(TRAINING)		Accuracy: 52.77%	
	ACTUAL 0	ACTUAL 1	PRECISION
PREDICTED 0	4661	884	.84
PREDICTED 1	5080	2003	.28
RECALL	.47	.69	

LASSO(TEST)		Accuracy: 51.67%	
	ACTUAL 0	ACTUAL 1	PRECISION
PREDICTED 0	3047	654	.82
PREDICTED 1	3415	1302	.27
RECALL	.47	.66	

RIDGE(TRAINING)		Accuracy: 52.79%	
	ACTUAL 0	ACTUAL 1	PRECISION
PREDICTED 0	4663	885	.84
PREDICTED 1	5078	2002	.28
RECALL	.47	.69	

RIDGE(TEST)		Accuracy: 51.69%	
	ACTUAL 0	ACTUAL 1	PRECISION
PREDICTED 0	3050	654	.81
PREDICTED 1	3412	1302	.28
RECALL	.47	.67	

LOGISTIC REGRESSION (LASSO) LIFT CHART:



NAIVE BAYES

TRAINING		Accuracy: 68.25%	
	ACTUAL 0	ACTUAL 1	PRECISION
PREDICTED 0	7683	1951	.80
PREDICTED 1	2058	938	.30
RECALL	.78	.33	

TEST		Accuracy: 67.13%	
	ACTUAL 0	ACTUAL 1	PRECISION
PREDICTED 0	5054	1359	.79
PREDICTED 1	1406	598	.30
RECALL	.79	.32	

Gradient Boosting

Number of trees = 30
Maximum Depth = 6
Minimum Row = 20

TRAINING		Accuracy: 83.19%	
	ACTUAL 0	ACTUAL 1	PRECISION
PREDICTED 0	8677	1059	.89
PREDICTED 1	1066	1829	.64
RECALL	.89	.63	

TEST		Accuracy:67.71%	
	ACTUAL 0	ACTUAL 1	PRECISION

PREDICTED 0	5113	1368	.79
PREDICTED 1	1350	588	.30
RECALL	.79	.31	

Number of trees = 25
Maximum Depth = 4
Minimum Row = 10

TRAINING		Accuracy: 69.48%	
	ACTUAL 0	ACTUAL 1	PRECISION
PREDICTED 0	6986	1098	.87
PREDICTED 1	2755	1789	.39
RECALL	.72	.62	

TEST		Accuracy:62.72%	
	ACTUAL 0	ACTUAL 1	PRECISION
PREDICTED 0	4352	1027	.81
PREDICTED 1	2112	931	.31
RECALL	.79	.31	

Final model:

TRAINING		Accuracy: 79.20%	
	ACTUAL 0	ACTUAL 1	PRECISION
PREDICTED 0	11000	2890	.79
PREDICTED 1	0	5	1.0
RECALL	1.0	.0017	

TEST		Accuracy:78.98%	
	ACTUAL 0	ACTUAL 1	PRECISION
PREDICTED 0	7311	1943	.79
PREDICTED 1	4	5	.55
RECALL	.99	.0026	

Modeling Method	**Accuracy (Training)**	**Accuracy (Test)**	**Without PCA Accuracy (Training)**	**Without PCA Accuracy (Test)**
Decision Trees	77.16	76.70	77.16	76.70
Boosted Gradient Trees	69.4	62.7	72.55	62.15
Logistic Regression(Lasso)	52.77	51.67	51.46	49.96
Logistic Regression(Ridge)	52.79	51.69	51.46	49.96

Naïve Bayes	68.25	67.13	73.56	73.19
Random Forest	79.20	78.98	78.16	77.75

From the table above, taking accuracy as the performance measure, we came to the conclusion that Random forest is the best model for predicting donors.

We can see from the confusion matrix that the accuracies with and without PCA do not vary much

As from our knowledge from the start of the report, we know that the number of Donors (Target_B=1) is lower than the number of Non-Donors (TARGET_B=0).

The dataset will be biased towards Non-Donors because of the response rate of 5.1%. By using weighted sampling, we assign weight to the Donor cases and lowering it for the Non-Donor cases to reduce the bias towards non-donor cases**. As from the question statement, the losses for not identifying donors is $13 which is high compared to the cost of solicitation wrt non-potential donors, which is $0.68. Because of this, we calculate the Recall value as it is most affected. The final model is selected based on the maximum value of Recall.**