Target Marketing - Fundraising

IDS 572: DATA MINING ASSIGNMENT-2

ASHUTHOSH GOWDA
LAAVANYA GANESH
AKSHAY MERCHANT

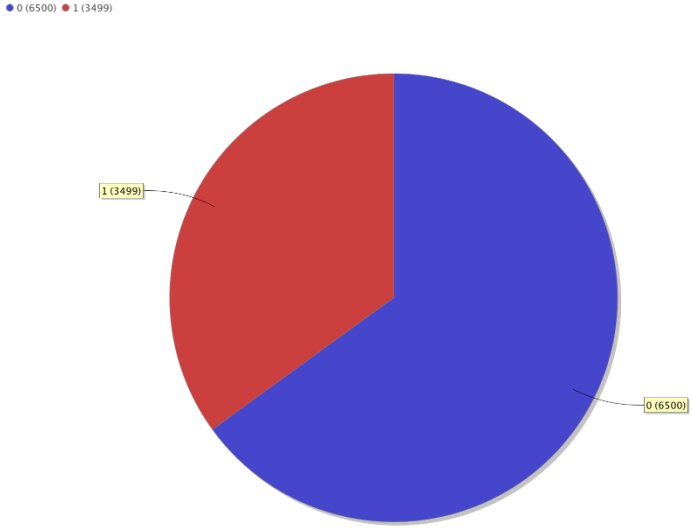
Q.1:

The dataset has many variables – some (most?) of them may not be useful for our purpose. Your first task is to clean and explore the data, determine missing values and how you might handle these, which variables you think need not be considered, which should be transformed, etc. This is a major task – and can take time, much more than the modeling step that comes next. You will find below a list of subset of variables that someone found useful. 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 collegeEducation, 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 why) 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?

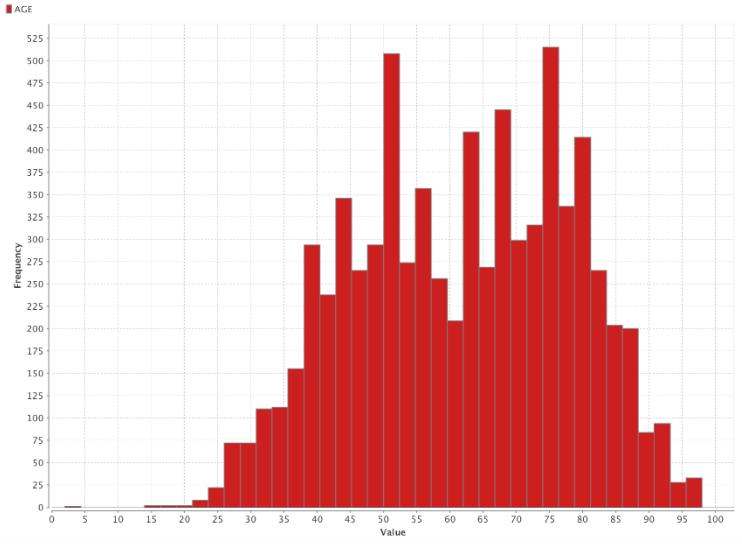
Solution:

On loading the PVA dataset with the Read excel operator with TARGET_B as label, we explored the below observations:

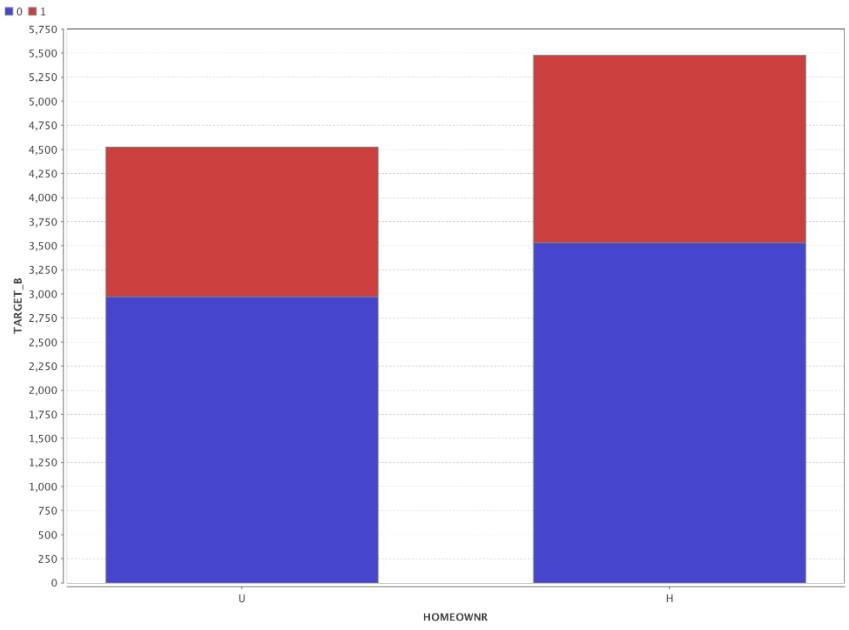
- Distribution of Donors(1) to Non-Donors(0) is given by TARGET_B attribute and the distribution of 0 to 1 is 6500: 3499.



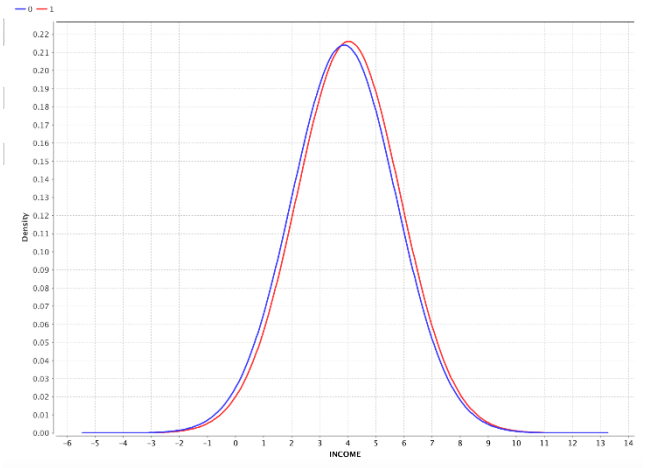
- We also explored the distributions of certain variables AGE, GENDER, HOMEOWNER and WEALTH2 against the TARGET_B label and found the below observations:
    1. **AGE**-AGE has 2477 missing values out of the 9999 values that are present which is approximately 25% missing values. From the distribution chart, most of the donors start donating from the age of 25 and gradually increase. There are hardly any donors below the age of 25.



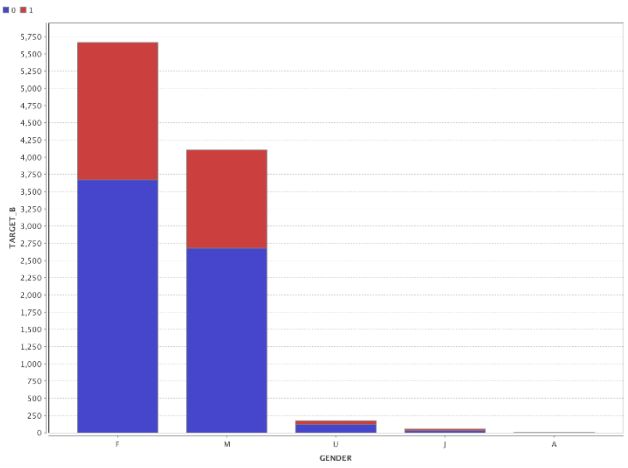
    2. **HOMEOWNR –** HOMEOWNR has 2401 missing values out of the 9999 values that are present which is approximately 25% missing values**.** The majority of cases did own homes. Proportion of Donors and Non-Donors is similar in both Home-owner and Unknown categories of approximately 34%:66%



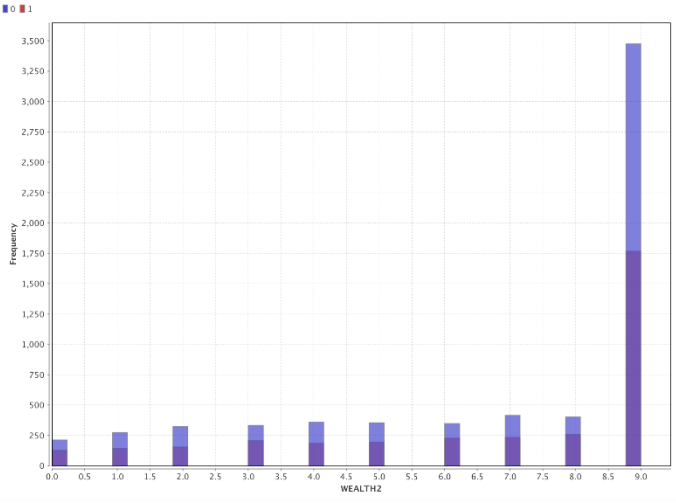
3. **INCOME** -INCOME has 2269 missing values out of the 9999 values that are present which is approximately 23% missing values. Distribution of Donors and Non-Donors is normal over Income.



4. **GENDER** – GENDER has 299 missing values out of the 9999 values that are present which is very minimal of approximately 3% missing values. Out of 5 classes F, M, U, J and A more than 97% data is shared by Male and Female classes. The proportion of Donors to Non-Donors is almost the same in Females and Males of approximately 65%:35%



5. **WEALTH2**- WEALTH2 has 4516 missing values out of the 9999 values that are present which is approximately 50% missing values. Distribution of TARGET_B is skewed towards the highest rating.



DATA SELECTION:

We decided to focus our efforts on the set of attributes that would provide the most meaningful information and would have relevance on predicting donation probability. After going through the data dictionary we decided to classify each attribute on the basis of our domain knowledge and intuition as **47 RELEVANT** (important), **137 IRRELEVANT** (not beneficial for prediction of outcome) and **297 ATTRIBUTES FOR PCA.** We have attached an Excel Workbook, where we have provided the list of 481 attributes, their classification and the reason for classifying them into the above 3 categories in the DATA SELECTION Worksheet.

VARIABLE TRANSFORMATIONS:

ATTRIBUTE	NEW ATTRIBUTE	TRANSFORMATIONS
TCODE	NEW TCODE	We analyzed the given dataset values from the data dictionary and thought that different donor titles can be combined to form one numeric categorical value. 0=Unknown; 1=Mr. ; 2= Mrs.; 3= Mr. and Mrs. ; 4= Miss; 5= Military
NOEXCH, RECINHSE, RECP3, RECPGVG, RECSWEEP,PEPSTRFL	NOEXCH, RECINHSE, RECP3, RECPGVG, RECSWEEP,PEPSTRFL	transformed the value X to 1 to convert it into numeric categorical variable
DOMAIN	DOM1,DOM2	splitting the first and second character of DOMAIN S=1; C=2; T=3; U=4; R=5
CLUSTER, CLUSTER2	CLUSTER, CLUSTER2	We have combined Row labels into numeric categories by using Excel Pivot Table.
GEOCODE,GEOCODE2	GEOCODE,GEOCODE2	We have combined Row labels into numeric categories by using Excel Pivot Table.

ATTRIBUTE	NEW ATTRIBUTE	TRANSFORMATIONS
GENDER	GENDER	Transform J and A to U. Gender of the person holding joint account is not known. U=0; M=1; F=2
WEALTH1, WEALTH2	WEALTH2	Wealth2 rating uses median family income and population statistics from each area to index relative wealth within each state. The segments are denoted 0-9, with 9 being the highest income group and zero being the lowest. Each rating has a different meaning within each state. Wealth1 simply contained the description: wealth rating. **Since Wealth1 has a similar description to Wealth2 we know that there is a lot of information overlap, which allows us to justify dropping the less informative of the two which is Wealth1.** **For missing values, we copy the WEALTH1 values if present in WEALTH1 or replace it with the Median of the State the value corresponds to.**
MBCRAFT, MBGARDEN, MBBOOKS, MBCOLLECT, MAGFAML, MAGFEM, MAGMALE, PUBGARDN, PUBCULIN, PUBHLTH, PUBDOITY, PUBNEWFN, PUBPHOTO and PUBOPP	RESPONSEtoOFFERS	We have added up the values of MBCRAFT, MBGARDEN, MBBOOKS, MBCOLLECT, MAGFAML, MAGFEM, MAGMALE, PUBGARDN, PUBCULIN, PUBHLTH, PUBDOITY, PUBNEWFN, PUBPHOTO and PUBOPP to give a new single attribute

ATTRIBUTE	NEW ATTRIBUTE	TRANSFORMATIONS
SOLP3, SOLIH	SOLIH	Since the Solicit Limitation Code P3 and in House have similar meaning for their values, we know that there is a lot of information overlap. **So we retain SOLIH. For its missing values, we first copy values from SOLP3 and the remaining replace with 999 as an unlimited random value**
RFA_3, RFA_5,RFA_13	RFA3left, RFA3mid, RFA3right, RFA5left, RFA5mid, RFA5right, RFA13left, RFA13mid, RFA13right	splitting the 1st, 2nd and 3rd character of RFA_3, RFA_5 and RFA_13

TCODE TRANSFORMATIONS:

Binned value	New Category	Previous values which are clubbed
0	Unknown and miscellaneous	0,9,22,30,93,202,980,39002,45
1	Mr	1,36,40,100
2	mrs	2
3	Mr and Mrs	72;1002;4002;14002
4	Miss	3,28,28028
5	Military	4,13,14,18,24,116

GEOCODE TRANSFORMATIONS:

VARIABLE	GEOCODE							
OLD	1	4	3	14	2	12	5	Blank
NEW	1	2	3	4	5	6	7	8
VARIABLE	GEOCODE2							
OLD	A		B, (blank)		C		D	
NEW	1		2		3		4	

CLUSTER2 TRANSFORMATION:

VARIABLE	CLUSTER2				
NEW BINS	**1**	**2**	**3**	**4**	**5**
OLD	5	2	1	17	22
	8	3	6	24	25
	12	4	9	29	28
	13	7	10	33	30
	16	11	14	35	37
	20	18	15	36	41
	27	21	19	40	43
	32	23	31	44	47
	42	26	34	45	50
	62	46	38	48	51
	(blank)	49	39	52	54
		53	57	55	60
		56	59	58	
				61	

CLUSTER TRANFORMATION:

VARIABLES	CLUSTER2				
NEW BINS	**1**	**2**	**3**	**4**	**5**
OLD	3	1	8	2	6
	4	5	15	10	7
	13	11	17	25	9
	14	12	20	27	19
	22	16	23	30	21
	24	18	26	31	32
	28	33	29	37	41
	48	38	34	39	44
	52	40	35	43	45
	53	46	36	49	47
	(blank)		42	51	50

a) HANDLING MISSING VALUES

The Relevant Variables did have a few missing values. Based on intuition, we filled the missing values in the most appropriate way we considered fit.

Attribute/ Category Name	Values Present	Missing value replaced by-
NEW TCODE	1,2,3,4,5 and blanks	0 blanks and unknown =0
NOEXCH, RECINHSE, RECP3, RECPGVG, RECSWEEP, PEPSTRFL	X and Blanks	Blanks=0
GENDER	0,1,2	0 (since missing values we have replaced with U)
WEALTH2	0-9	Values were substituted with WEALTH1; Median of Wealth2 grouped by State
NUMCHILD	1,2,3,4,5, blanks	Blanks=0
SOLIH	0,1,2,3,4,5,6 and blanks	SOLIP; Blanks=999
AGE, INCOME	numeric	Median value

Principal Component Analysis:

After eliminating useless attributes and necessary transformations, we were left with 297 odd variables. On these 297 variables we performed a technique for variable reduction called 'Principal Component Analysis'.

Attribute chosen for PCA: We apply PCA on 5 subsets (PERCENT valued attributes, NUMERIC [median, average, number] valued attributes, PROMOTOTION ATTRIBUTES, VETERAN ATTRIBUTES and GOVERNMENT ATTRIBUTES) from entire data to get significant principal components.

Reason applying PCA on these subsets:

Reduce the dimensionality. For example, there are 244 PERCENT Valued Attributes, but not all impacting the outcome significantly, so we wanted to consider the influence of only those attributes which are significant and hence we set the threshold to 95%. Similarly, we perform PCA on 42 NUMERIC VALUED, 5 PROMOTION, 3 VETERAN and 3 GOVERNMENT Attributes to get cumulative effect of all variables into few transformed and reduced variables. **After performing a PCA, we have reduced the 297 variables to 40 variables namely PC_1-PC-27, PC-28-PC-34, PROM1, PROM2, VET1, VET2, GOV1 and GOV2**

DATA REDUCTION:

We now had 38 RELEVANT transformed attributes plus the 40 PCA Transformed Variables, which gave us a total of 78 attributes on which we run the Variable Importance R Code. We have used the "importance()" function in Random Forest, to get list of Variables arranged in the order of least Important to Most Important.

After getting the list of Variables in Order of least importance to most importance, we used Backward Selection method to get the best variables, that would be useful for our modelling.

FINAL ATTRIBUTE SET USED FOR MODELLING					
AGE	MALEMILI	RFA13left	pc_1	pc_12	pc_23
AVGGIFT	NEW.TCODE	RFA13mid	pc_2	pc_13	pc_24
CARDPM12	NUMPRM12	RFA3left	pc_3	pc_14	pc_25
CLUSTER	PEPSTRFL	RFA3mid	pc_4	pc_15	pc_26
DOM1	PROM1	RFA5left	pc_5	pc_16	pc_28
DOM2	PROM2	RFA5mid	pc_6	pc_17	pc_29
GEOCODE2	RECINHSE	RFA5right	pc_7	pc_18	pc_30
GOV1	RECPGVG	SOLIH	pc_8	pc_19	pc_31
GOV2	RECSWEEP	VET1	pc_9	pc_20	pc_32
HPHONE_D	RESPONSEtoOFFERS	VET2	pc_10	pc_21	pc_33
INCOME	RFA_2A	WEALTH2	pc_11	pc_22	pc_34

Q2.

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 the following classification techniques on the data:

• Decision Trees (you can use J48, or any other suitable type of decision tree)

• Logistic Regression

• Naïve-Bayes

• Random forest

• Boosted trees

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? Run each method on a chosen subset of the variables - how do you select this subset? (Be sure NOT to include "TARGET−D" in your analysis.)

Provide a comparative evaluation of performance of your best models from each technique. Does variable selection/PCA make a difference for the different models?

Answer

The subset of variables used in the models below was selected by performing Random Forest and computing the Mean Decrease by Gini. Then Backward Selection was used to eliminate the least important variables.

After the variables were selected, we split them into a train set and validation set with a split ratio of 60:40. We have also set the Random Seed to 12345.

Decision Trees

We chose the Decision Tree Classifier instead of the J48 Classifier since it gave better a better accuracy and model for the dataset. The best model we got was using Criteria: Gini Index, Max Depth: 35, Confidence: 0.5, Minimum Gain: 0.0015, Minimum Leaf Size: 6, Minimum Size of Split: 40, Number of Pre-pruning Alternatives: 3.

Below is the Confusion Matrix for the Validation Data for the selected variables.

Accuracy=60.22%	True.0	True.1	**Class Precision**
Pred.0	1928	958	66.81%
Pred.1	633	481	**43.18%**
Class Recall	75.28%	**33.43%**	

To compare, we took the whole dataset and performed Decision Tree Classification on the same parameters. The Confusion Matrix below shows the output of the validation data.

Accuracy=59.48%	True.0	True.1	**Class Precision**
Pred.0	2005	1065	65.31%

Pred.1	556	374	**40.22%**
Class Recall	78.29%	**25.99%**	

As you can see, when we replace the attributes selected by Random Forest with the whole dataset, we get a reduction in Accuracy, Recall and Precision.

Changes In Parameter Values

We used Gini Index as the split criteria as the others did not give good models. Gain Ratio and Accuracy gave a really low Class Recall while Information Gain had a lower Accuracy and Class Recall when compared to Gini Index.

For Max Depth, Confidence, Min Sample size we saw that initial changes improved the model. But after a certain point (i.e Max Depth: 35, Confidence: 0.5, Min Sample Size: 40), there was no change in the Class Recall, Precision and Accuracy of the model.

Boosted Trees

For Boosted Trees, we used Gradient Boosted Trees and set the parameters as follows, Number of Trees: 40, Maximal Depth: 3, Min Rows: 10, Min Split Improvement: 0.0, Number of Bins: 20, Learning Rate: 0.1, Sample Rate: 1, Distribution: Bernoulli.

We got the following Confusion Matrix by running the Gradient Boosted Trees Operator for the attributes we selected

Accuracy=57.55%	True.0	True.1	**Class Precision**
Pred.0	1439	576	71.41%
Pred.1	1122	863	**43.48%**
Class Recall	56.19%	**59.97%**	

By running the Gradient Boosted Trees Operator for the whole dataset, we got the following Confusion Matrix.

Accuracy=53.52%	True.0	True.1	**Class Precision**
Pred.0	1156	454	71.80%
Pred.1	1405	985	**41.21%**
Class Recall	45.14%	**68.45%**	

When we use this model on the complete dataset, even though there is a very high increase in Recall, there is a decrease in Accuracy and Precision.

Changes In Parameter Values

We found that as we initially started to increase the No of Trees, we started getting better models. But when we set No of Trees > 50, the Accuracy, Precision and Recall Started to decrease. (The difference between No of Trees: 50 and No of Trees: 40 was a 4% drop In Class Recall, which was why we preferred No of Trees: 40).

We tried different values for Max Depth, but found that, at Max Depth: 3, we got a really high Class Recall, which is why we preferred it to other Max Depth values.

Naïve Bayes

We ran Naïve Bayes Classification on the attributes we had selected and got the following Confusion Matrix

Accuracy=59.35%	True.0	True.1	**Class Precision**
Pred.0	1663	728	69.55%
Pred.1	898	711	**44.19%**
Class Recall	64.94%	**49.41%**	

Then we ran the Naïve Bayes Classifier on the complete dataset and got the Confusion Matrix shown below

Accuracy=51.85%	True.0	True.1	**Class Precision**
Pred.0	1168	533	68.67%
Pred.1	1393	906	**39.41%**
Class Recall	45.61%	**62.96%**	

When we used the attributes selected by the Random Forest Variable Importance Function, we got a pretty good Accuracy, Recall Rate, and Precision.

For the complete dataset, we observed a much higher Recall, but that came at the expense of Accuracy and Precision.

Changes In Parameter Values

We got the same output with and without Laplace Correction.

Logistic Regression

After Attribute Selection/PCA'ing the variables we got the following confusion matrix while using logistic regression. We decided on the following parameters Solver: Auto, Use Regularization: Yes, Lambda: 0.005, Lambda Search: Yes, Number of lambdas: 10, Lambda Min Ratio: 1, Early Stopping: No, Alpha: 0.1, Standardize: No.

Accuracy=49.23%	True.0	True.1	**Class Precision**
Pred.0	784	254	75.53%
Pred.1	1777	1185	**40.01%**
Class Recall	30.61%	**82.35%**	

When we used Logistic Regression on the whole dataset, we got the following confusion matrix.

Accuracy=55.12%	True.0	True.1	**Class Precision**
Pred.0	1273	507	71.52%
Pred.1	1288	932	**41.98%**
Class Recall	49.71%	**64.77%**	

When logistic regression is performed on the whole dataset, we get a significantly higher accuracy, a better precision, but a significantly lower Recall.

Changes In Parameter Values

We tried different solver's but preferred AUTO since it gave a higher Accuracy compared to the other Solvers.

We found that Logistic Regression with regularization gave better outcomes. Then we tried different lambda values and found Lambda: 0.005 to be the best one as it gave a marginally better Accuracy and Class Precision above 40%.

Random Forest

We performed Random Forest, first, on the selected attributes, then on the complete dataset and obtained the two Confusion Matrices shown below.

We performed Random Forest in R, with number of trees: 100, Interaction depth: 10

Shown below is the Confusion Matrix for Random Forest performed on the attributes selected by us.

Accuracy=65.58%	True.0	True.1	**Class Precision**
Pred.0	1812	883	67.23%
Pred.1	134	170	**55.93%**
Class Recall	68.89%	**16.14%**	

Below is the Confusion Matrix Output for Random Forest performed on the complete dataset.

Accuracy=64.84%	True.0	True.1	**Class Precision**
Pred.0	1856	964	65.80%
Pred.1	90	89	**49.72%**
Class Recall	4.62%	**8.45%**	

Both of them have pretty low Recall rates, but the attributes we selected give a better accuracy and precision than the complete dataset for Random Forest.

Changes In Parameter Values

For Random Forest we mainly changed two parameters, number of trees and interaction depth.

For number of trees we tried different Number of Trees ranging from 50 to 500, and found the most optimal Number of Trees to be 500. Below 500 trees, the output gave a lower number of Predicted 1's that were True 1's(True Positives).

We found that Iteration.depth: 10 gave the best outcome, as above and below that number, the number of False Positives and False Negatives were higher, which affected Recall and Precision.

Best Model

From the models above, we felt that the Boosted Trees Model was a pretty good model as it had a high Recall and Precision and a decent Accuracy.

Q3.

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.

Solution:

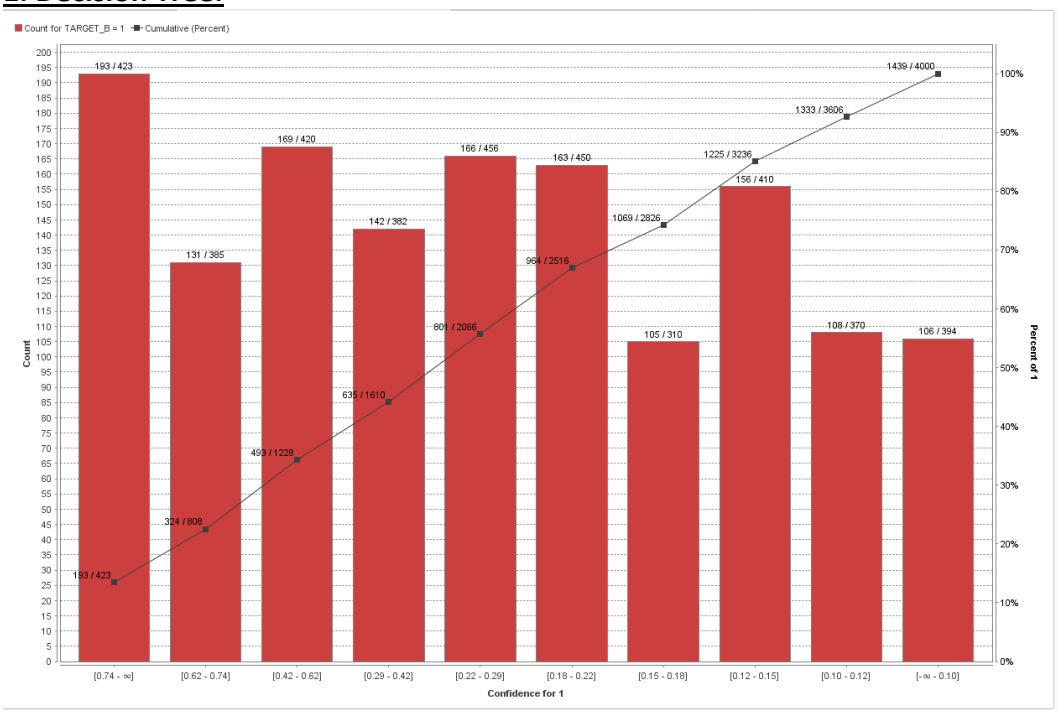
The original Dataset has an actual response rate of 5.1%. Since, this Dataset Sample under represents Donors, a model built out of it would be biased towards Non-Donors, and we would fail to get a good prediction on the Unseen data, as most of the Rows, would get predicted as Non- Donors. Hence, we Over-Sample the Donor Sample and Under-Sample the Non-Donor Sample to get a balanced number of Cases between the two. A model built out of this would give a healthy model which could Predict at an acceptable accuracy.

Classification Accuracy is perhaps not the best performance metric for the purpose of maximizing net profit. Since, this model is built to predict the Donors from the Non-donors, we can consider, Precision and Recall as a better Performance metrics to maximize net profit.

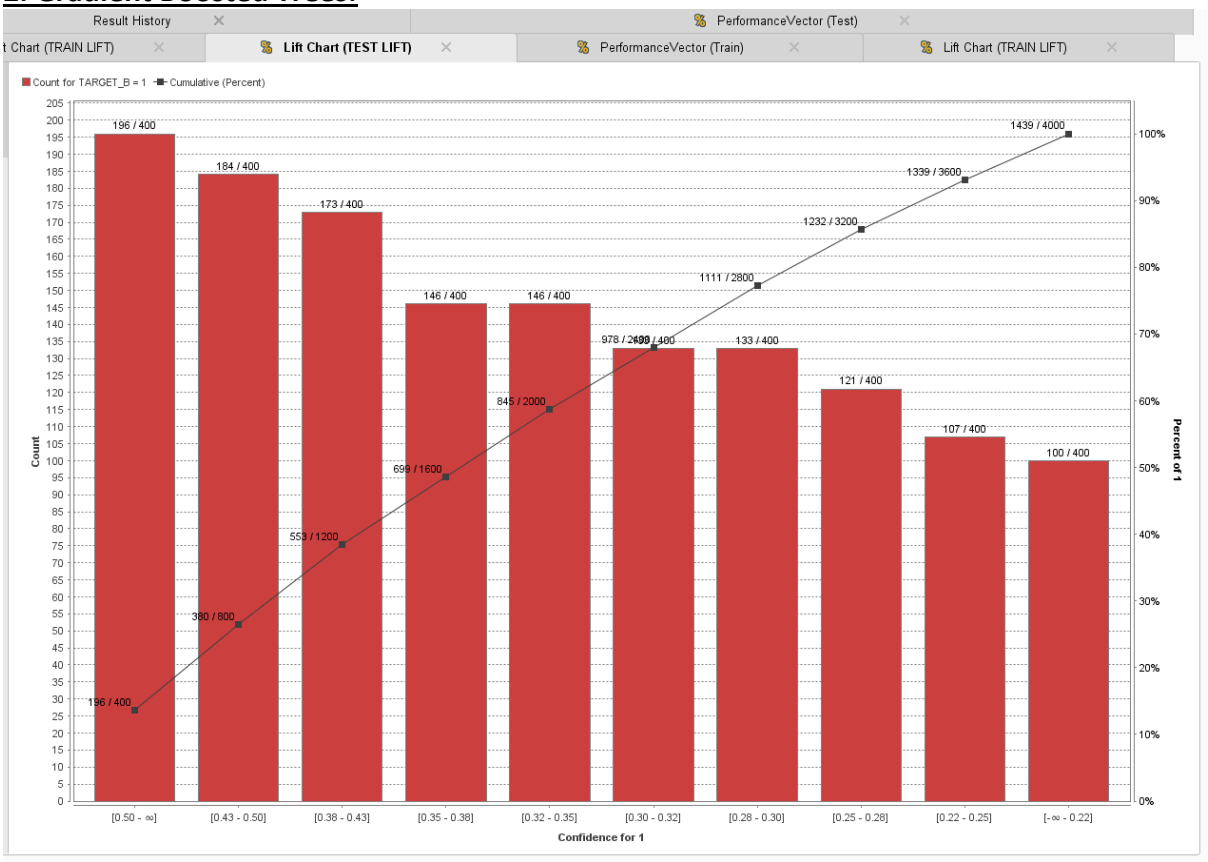
Precision would give us the Percentage of True Donors, for every Predicted Donor.
Recall would give us the Percentage of Donors Predicted out of the overall Donors.
A good model is one which has a high Precision and high Recall.

We could also use the LIFT CHARTS to evaluate a good model.
Here are a few LIFT CHARTS from the model that we have created:

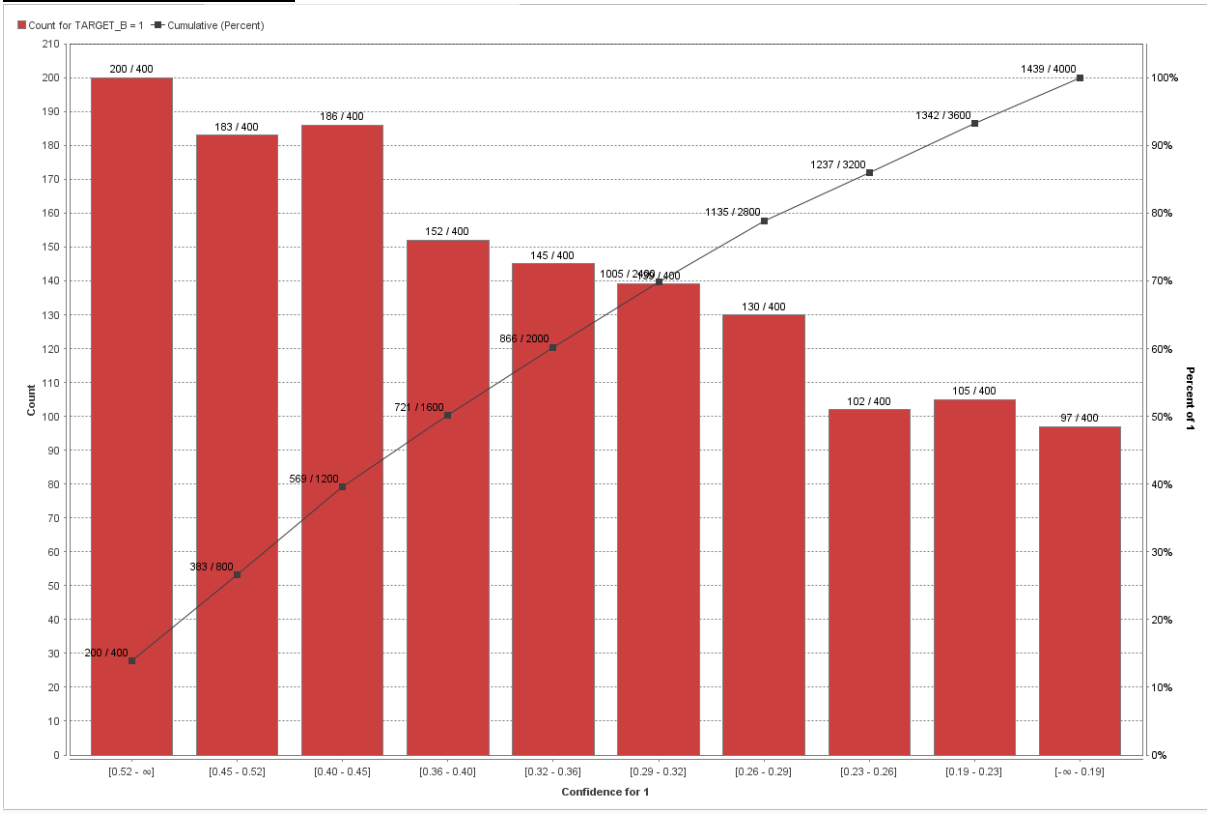
1. Decision Tree:



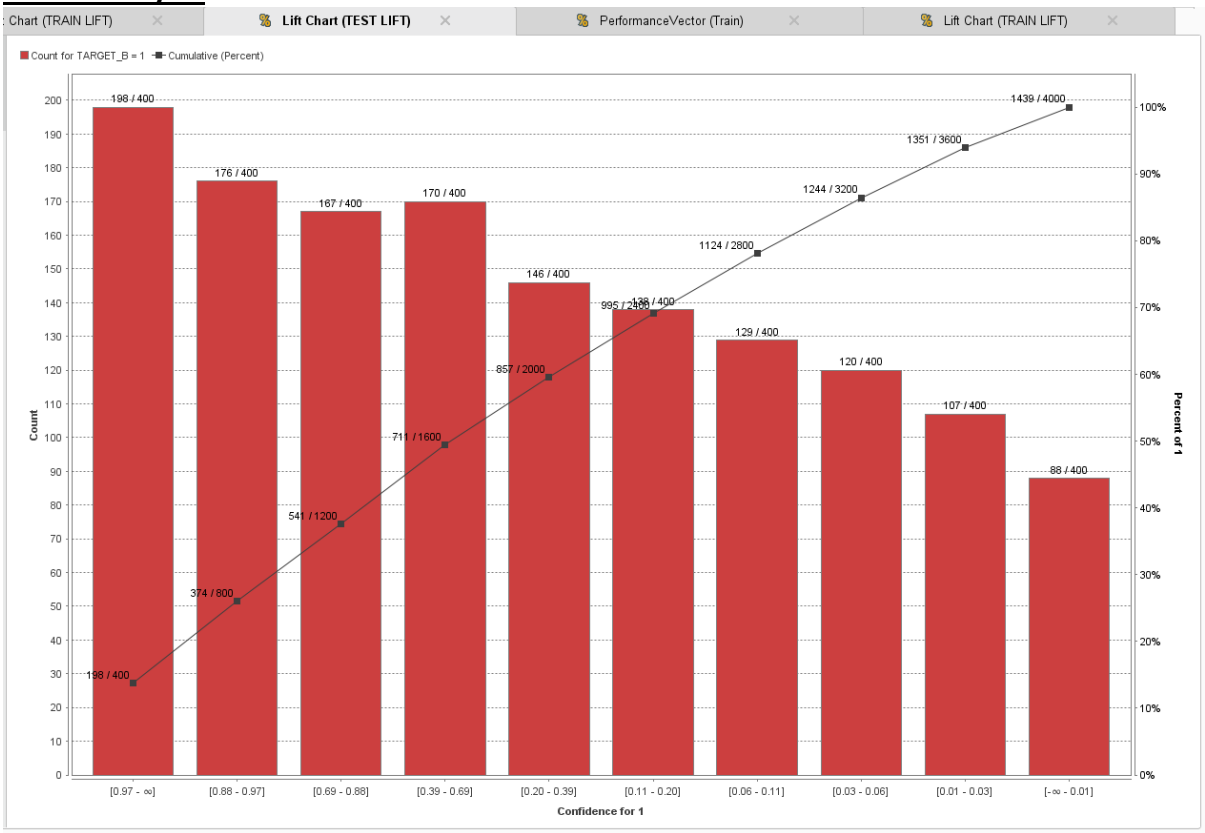
2. Gradient Boosted Trees:



3. Logistic Regression:



4. Naïve Bayes:



From this we can Conclude that Naïve Bayes, Logistic Regression, Boosted Trees do give a good Precision.

But based on Precision and Recall, we would like to conclude that Gradient Boosted Tree gives the Best Model.