



# TARGET MARKETING (PVA)

---

Developed a data mining model to improve the cost effectiveness of their direct marketing campaign. Using the facts available from the dataset we develop a classification model that can effectively capture donors so that the expected net profit is maximized.

NISHANTH SINGAMSETTY



...

## Data Exploration and Elimination of Attributes

The dataset has 480 variables data we can observe that there are few attributes which contains redundant information and might not be helpful for the prediction of a donor. Our first task is data cleansing and exploration of the manually, determining the missing values and the potential ways to handle and transform the variables to be considered.

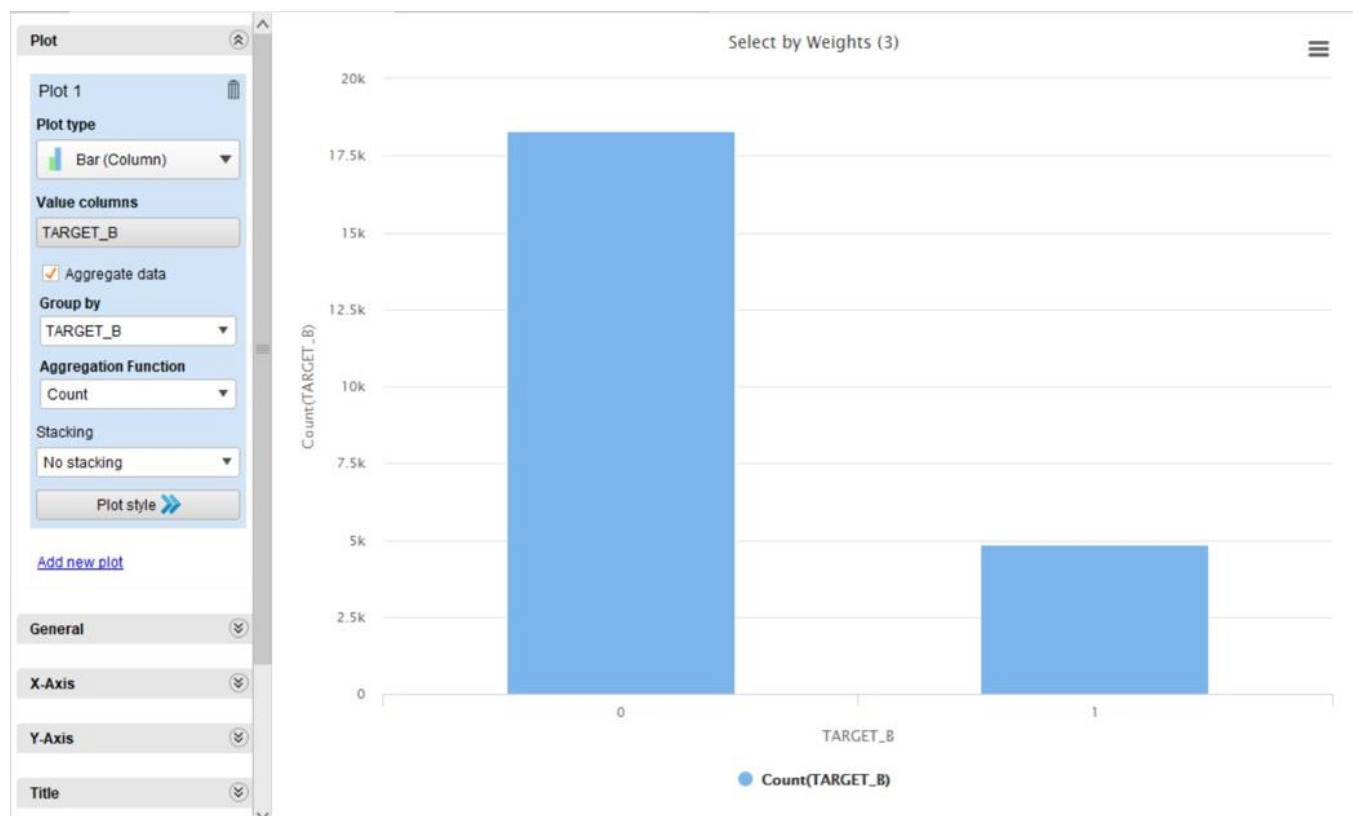
- Data exploration: We will import the data and examine the different variables manually. We observe the potential transformations that can be made, the ways to handle missing values and variables to consider while performing PCA (Principal Component Analysis) and why. The distribution, mean, std. deviation and range of values are calculated. We will also use classification techniques like decision trees, logistic regression (using Ridge and Lasso), Random Forest and Boosted Trees to see if we get any help to determine which feature variables to include in the predictive model for donors.
- Data cleansing – After exploration we observe that many variables have 'empty' values in many rows. In these some of the values may be missing, empty values may carry information (like for a variable college Education, empty values may indicate no-college-education which can be interpreted as a specific value). These missing values maybe handled by calculating the average, median or most occurring value of the feature variable (this method is only applicable for few feature variables). In the dataset we observe some of the rows having many variables with missing data, such rows are removed.

As we proceed answering different questions, we will give the outline of data cleansing steps undertaken (reason for doing) steps that we perform (and why) and the different feature variables taken into consideration after performing the different classification techniques. We will provide detail justification for using the different procedures and the metrics for the analysis.

1. The dataset has many variables – some (many?) 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 modelling step that comes next. You will find below a list of subsets of variables that were found useful in earlier analysis.

### Dataset Introduction

The dataset contains 23158 observations with 478 regular attributes and 2 special attributes. Our 2 special attributes are Id variable CONTROLN and response variable (label) TARGET\_B. We found that there were 4843 responses for 1(donor) and 18315 for 0(non-donor).



#### (a) Which attributes will you omit from the analyses and why?

We initially created random forests, weighted attributes by tree importance and used select by weights to analyze which variables are important and which are not. From our observations, we concluded that it is important to omit some attributes as listed below because they will not be useful in our predictions. However, relying solely on the random forest model for data cleaning and preparation is insufficient. Thus, we will use our intuition to understand and other methodologies to analyze variables at a deeper level.

## Methodology

- **Relevance:** We examined the variable explanation for each variable given at [https://kdd.ics.uci.edu/databases/kddcup98/epsilon\\_mirror/cup98dic.txt](https://kdd.ics.uci.edu/databases/kddcup98/epsilon_mirror/cup98dic.txt) and eliminated variables on the basis of their relevancy to the target variable. For example
  - Cluster – There are 53 distinct categories in the variable however it is not defined to what each category means.
  - ADATE\_XX - We removed the set of these date attributes as they are not significant to our analysis.
  - RDATE\_XX – We removed them as they were irrelevant. It also contains some missing values.
  - CHIL, CHILC, CHILD – The attributes retaining to children does not seem relevant to our analysis.
  - EIC, OCC, ODEC – A person’s employment and occupational attributes seem irrelevant to our analysis.
  - TCODE – The title and marital status of a person seems a bit irrelevant for our goal of prediction
  - ZIP, ZIP CODE and STATE – The demographic information does not seem relevant.
  - OSOURCE - This describes the origin source which is not really a related data to what we are trying to predict using our model.
  - RAMNT 2-23 – These attributes do not seem relevant to our analysis.
  - RFA 3-24 was removed as they do not pertain to 97NK mailing, which determines our target variable.
  - RFA\_2R – This attribute had a single value throughout the observations and hence was removed.
  - HC, DW, HU, HUPA, HUR, HV, HVP, MHUC, HHAGE, HHN, IC – These two attributes kind does not seem to affect or play an important role in building the right model.
- **Missing Values:** Variables that have high % of missing values that cannot be imputed using other techniques
  - NUMCHLD – the variable has values ranging from 1 to 6. However, it has ~18k values missing.
  - MDMAUD – this was removed as it contains missing values.
  - SOLIH, RECPGVG – These attributes have more than 90% missing values.
- **Redundancy:** There were variables that gave similar information as other variables. Therefore, we removed these variables. Example
  - The following variables indicate the number of known times the donor has responded to other types of mail order offers (ex – MBCRAFT, MBBOOKS, PUBOPP). There is a similar variable HIT – indicating the number of times the customer has responded to other promotional mails. There HIT is an aggregated summary of these variables.
  - AGE901-AGE907, AGE901-AGE907, HHAGE, DOB – After including main Age attribute, other attributes relating to age cause redundancy.

## (b) How do you clean the data, handle missing values?


Missing values do not always mean that the data is missing because of which it would be incorrect to ignore or delete the missing observations. In our dataset, we had a lot of variables that have missing values in their observations which have to be handled well by either removing them or transforming them to meaningful data based on what the variable infers.

The dataset has the following types of missing observations:


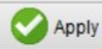
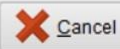
- **Numeric Attributes and Real Attributes:** Numeric Attributes and Real Attributes like INCOME, AGE, WEALTH1, WEALTH2, RAMNTALL, SEC, CLUSTER, etc. which were handled by replacing the missing value with the average value.
- **Binomial Attributes:** Binomial Attributes like PETS, CATALG, CRAFTS, BOATS, etc. were handled by replacing the missing value with an "N" as they represent no interests.
- **Polynomial Attributes:** SOLIH, LIFESRC, etc. are some of the polynomial attributes which were handled by replacing the missing value by a default value 0.0.

## (c) What new attributes/values do you derive?

To standardize our dataset, new attributes have been generated from existing attributes. Here is the list of new attributes that was created to better perform the analysis.

 Edit Parameter List: **function descriptions**  
List of functions to generate.

attribute name	function expressions
EP_PvaState	if(PVASTATE=="E"    PVASTATE=="P", "1", "0")
rechinse	if( RECHINSE == "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")
homeownr	if(HOMEOWNR=="H", "1", "0")
totDays	date_diff(LASTDATE, ADATE_2)/(1000*60*60*24)
avgcresp	if (CARDPROM > 0, CARDGIFT/CARDPROM, 0)

 Add Entry
  Remove Entry
  Apply
  Cancel

Note: For categorical variables having two/multiple levels and missing values (PVASTATE, RECHINSE, RECP3, RECSWEEP, PEPSTRFL, HOMEOWNR, MAJOR, GENDER) we transformed them into corresponding 1's and 0's. Also, we split some attributes to derive information and stored them into new attributes (urbanicity, domainSES). Furthermore, new attributes like avgcresp,



avgallresp, lastToMaxGiftRatio, maxToMinGiftRatio, avgGapBetwGifts were created by averaging their value. After generating these new attributes by the “Generate Attributes” operator in RapidMiner, we deleted the old variables by “Remove old attsSelect Attributes”.

**(d) Which variables will you consider for modelling (and why)? How do your findings relate with the variable subsets given in Tables 1 and 2. Explain how you approach data reduction, variable selection? What methods do you try and find useful? Summarize your findings.**

To understand our model, it is important to understand our approach. Firstly, we explored the data. Secondly, we cleansed the data and finally, we performed data reduction using Principal Components Analysis (PCA).

- **Data exploration:** We loaded the data using Read CSV template of the Rapid Miner. To analyze and identify the variables important we created Random Forest, weighted attributes by tree importance and used select by weights. The observations we obtained concluded that omitting some attributes (listed below) is important as they are not useful in our predictions. We use our intuition to understand and implement other methodologies to analyze variables at a deeper level as being dependent on Random Forest model for data cleansing and preparation is not enough.

**Interpretations of the some of the variables are as follows:**

**INCOME:** The income range is from 1 to 7, and the mean is 3.922. This implies that a larger set of people have income in the 3rd and 4th quantile. After analyzing the box plot, we can infer that most people have an income lies between the lower quartile and median.

**WEALTH1:** The mean of WEALTH1 is 5.345, and it ranges from 0 to 9. After analyzing the bell curve, we can infer that the data is nearly normally distributed.

**AGE:** Age ranges from 1 to 98, with the average age being 61.838. Because the average age is 61.838, anyone above this age group is less likely to contribute to the donation.

**VETERANS:** The veterans attribute is a categorical field and is distributed over 11.12% Y's and 88.88% N's.

**BIBLE:** The BIBLE attribute is a donor interest and the data has about 9.44% people who have an interest in BIBLE reading.

**HOME:** This attribute indicates whether the donor/non-donor works from home. About 0.932% of them work from home.

**What do you observe, and does this help in determining which variables to focus further attention on? What variable transformations do you make (and why)?**

### **How do your findings relate with the variable subsets given in Tables 1 and 2?**

Post our analysis, we observed that there are many matching variables with the variable subsets given in Tables 1 and 2. Variables such as income, wealth, gender etc are essential as they will dictate how accurate our model eventually turns out to be.

But, some variables in the tables such as HV, NUMCHLD etc. are omitted from analysis because they were either irrelevant, had missing values or were redundant and thus would not help us build and predict accurate models which is the goal.

#### **Variable selection:**

After we cleaned our data, we removed unwanted attributes and added new attributes. From 480 variables, we are now left with only 113 variables.

- Out of these 113 variables, 2 are special variables - id variable CONTROLN and response variable TARGET\_B.

Some of the most useful variables we find are:

- INCOME: This indicates as an important variable as it gives the information of household income, which helps us in analyzing if the house has the potential to donate.
- DOMAIN: The geographic region of the receiver can be used to estimate for a potential donation from the region. The Domain attribute is broken into 2 new attributes - UrbanCity and DomainSES that analyzes the attribute in a deeper level.
- AGE: This a significant variable that helps in determining if the receiver can donate or not.
- MAJOR: This tells us who all the major donors in the dataset are.
- PVASTATE: This variable indicates if the donor lives in a state served by the organization's EPVA chapter. It likely that a donor living in such area will donate.
- MDMAUD: To decide if a donor would donate or not, it is better to have an idea about their donation frequency, which makes it an important variable.

#### **Variable transformations:**

The binary variables like PVASTATE, RECINHSE, RECP3, RECPGVG, PEPSTRFL, HOMEOWNR, and MAJOR etc. had 'X' as values which denoted the presence and rest of the missing values represented absence. So, we created new attributes where '1' denoted the presence and '0' the absence. The variable GENDER had four values ('M', 'F', 'U', 'J'). So, if the GENDER value was 'M' or 'F', we retained those values as GENDER itself and imputed other missing values as 'U'. The variable DOMAIN consisted of 2 bytes (1st: URBANICITY and 2nd: SOCIO-ECONOMIC STATUS). So,



we extracted each of the bytes as a new variable from the original variable namely, Urbanicity and domainSES.

- **Data Cleansing:**

We cleaned the missing values by exploring the value and type of attributes, handling them as below:

- Missing values of real and numeric variables: We imputed missing values to the average values.
- Missing values of binominal variables: The attributes that have 'Y' and 'N' describe donor interest. We identify some values as missing and were not indicated by the donor. Therefore, we can conclude they were not interested in the activity. So, we use MAP operator to convert the missing values to 'N'.
- Missing values of polynomial variables: The attributes like LIFESRC, SOLIH values have actual missing values and hence we imputed these missing values by 0.0.

**Random forests and/or decision trees can help determine which variables to include in a predictive model for donors. Principal Components Analysis (PCA) can also help in data reduction. Explain what you do, findings, and whether/how you think this is useful. How do you approach data reduction? What methods for data reduction do you try?**

- We have eliminated most of the variables during our data exploration and cleaning steps. For these steps, we used a combination of random forests and intuition.
- Furthermore, we implemented a data dimension reduction technique called Principal Components Analysis (PCA) to reduce the number of variables by reducing the redundancy among those variables.

**Perform Principal Components Analysis (PCA) – Our reason to use this technique to arrive at our findings.**

We performed Principal Components Analysis on the below five sets of variables:

- Variables reflecting donor interests (COLLECT1, VETERANS, BIBLE, CATLG, HOMEE, PETS, CDPLAY, STEREO, PCOWNERS, PHOTO, CRAFTS, FISHER, GARDENIN, BOATS, WALKER, KIDSTUFF, CARDS, PLATES) – these attributes reflect donor hobbies and combining them to form a subset of attributes help to further analyze them, reducing redundancy. This analysis gives 3 Principal Components that reflect the important features needed in the model.
- Variables reflecting the number of known times the donor has responded to other types of mail order offers (MBCRAFT, MBGARDEN, MBBOOKS, MBCOLECT, MAGFAML, MAGFEM, MAGMALE, PUBGARDN, PUBCULIN, PUBHLTH, PUBDOITY, PUBNEWFN, PUBPHOTO, PUBOPP) – these attributes reflect the donor response to types of mail order and combining them gives 3 Principal Components which we consider in our model for analysis.
- Variables reflecting the occupation of donors (OCC1 - OCC13) – These features describe the occupation of donors and we reduced their dimensionality by applying PCA and got 3 Principal Components.





- Variables reflecting the income of households (IC1 – IC23) – These attributes describe the household income in the range <\$15,000 to >=\$150,000 distributed across IC6 to IC23, and IC1 – IC5, which describes the average and median of household income. When PCA is run it gives 3 Principal Components which we have considered for the model.
- Variables reflecting ethnicity (ETH1 – ETH16) – These attributes reflect the characteristics of the donor neighborhood, we combined the 3 Principal Components which give the most important features needed to build model.

## 2. Modeling

**Partitioning - Partition the dataset into 60% training and 40% validation.**

**Consider the following classification techniques on the data:**

- **Decision Trees**
- **Logistic Regression, using Ridge and Lasso.**
- **Random forest**
- **Boosted trees**

**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?**

**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. Consider confusion matrix and related measures, lift, ROC (and any others you find useful)**

**Explain the performance measures you find useful for comparing models.**

**Which model(s) will you use and why.**

Post data cleaning and exploration, **we obtained a set of 90 variables.** We used these variables for our model. Using the multiply operator in rapid miner, we made multiple copies of this data. We did a 60-40 split on data using set seed (12345) and ran different models as mentioned below:

## Decision Tree

We ran a few different variations of a Decision tree model. However, the below model with the chosen parameters has the best overall performance.

Model	Accuracy on Training	Accuracy on Test	Parameter (Maximal Depth)	Parameter (Pruning)	Parameter (Pre-pruning)	Parameter (Confidence)
Initial Model	78.87%	78%	20	TRUE	TRUE	0.2
Model DT	79.26%	78.93%	15	TRUE	FALSE	0.2
Other Parameters for Model DT	Minimal gain = 0.1 Minimal leaf size = 2 Minimal size for split = 4					

### Accuracy measures

On training data:

accuracy: 79.26%

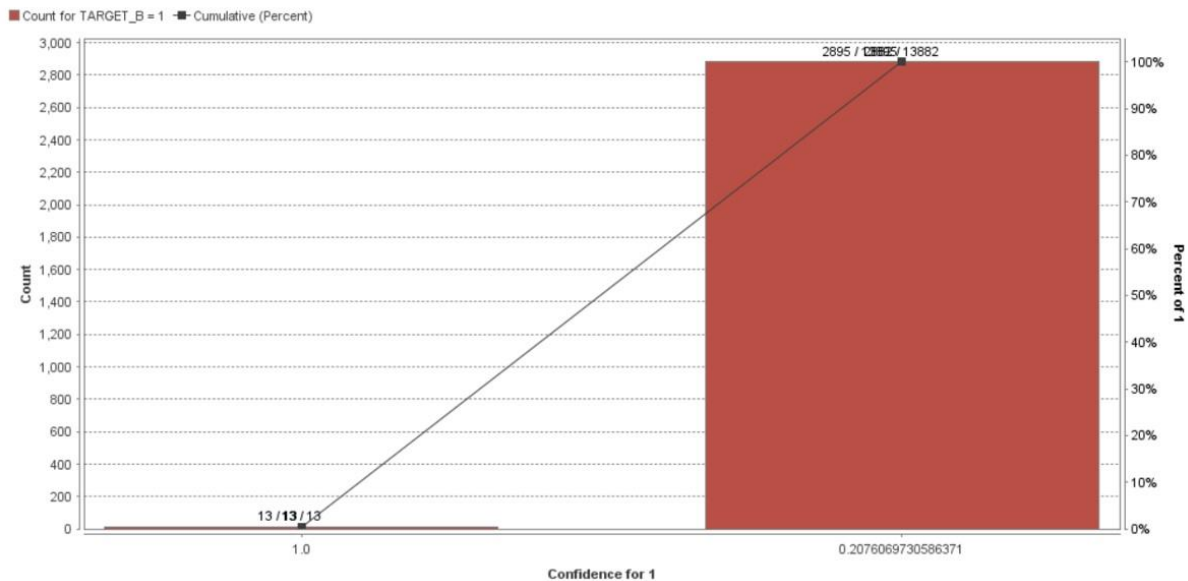
	true 0	true 1	class precision
pred. 0	11000	2882	79.24%
pred. 1	0	13	100.00%
class recall	100.00%	0.45%	

On test data:

accuracy: 78.93%

	true 0	true 1	class precision
pred. 0	7307	1944	78.99%
pred. 1	8	4	33.33%
class recall	99.89%	0.21%	

## Lift Chart Test Data:



## Few Performance Measures

AUC: 0.501 (positive class: 1), precision: 35.71% (positive class: 1), recall: 0.52% (positive class: 1), lift: 158 %, specificity: 99.89% (positive class: 1)

## Logistic Regression, using Ridge and Lasso

We built logistic regression model using values of alpha as 0 and 1

alpha = 0.0 represents ridge regression

alpha = 1 represents lasso.

Lambda is the penalty introduced to the coefficients, and we tested for lambda value as 0.005.

The accuracy for the 2 models is:

Model	Accuracy on Train	Accuracy on Test	Lambda	Alpha
Model Lasso	55.03%	53.79%	0.005	1
Model Ridge	55.58%	54.19%	0.005	0

Logistic regression using ridge gives the better performance over lasso, as the accuracy on test data is 54.19%.

## Accuracy measures of Ridge LR

### On training data:

accuracy: 55.58%

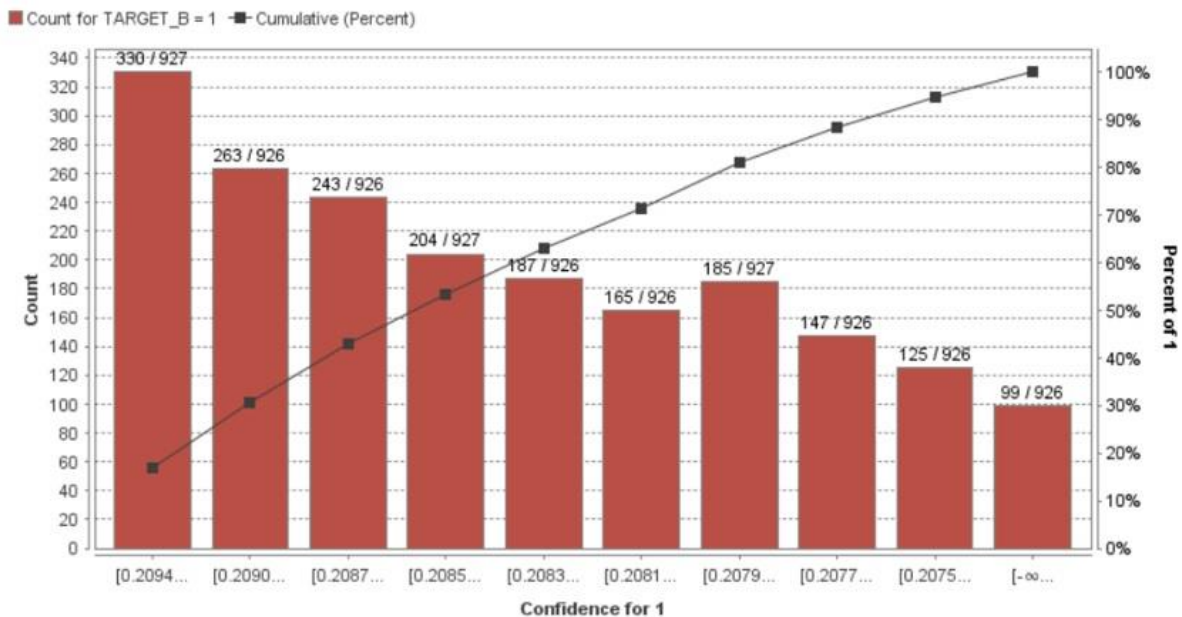
	true 0	true 1	class precision
pred. 0	5784	956	85.82%
pred. 1	5216	1939	27.10%
class recall	52.58%	66.98%	

### On test data:

accuracy: 54.19%

	true 0	true 1	class precision
pred. 0	3765	693	84.45%
pred. 1	3550	1255	26.12%
class recall	51.47%	64.43%	

### Lift Chart for Test Data:



## Performance measures

AUC: 0.618(positive class: 1), precision: 26.12% (positive class:1), sensitivity: 64.43%, specificity: 51.47% (positive class: 1).

## Random Forest

We developed two Random Forest models by changing the number of trees and maximal depth. The better accuracy was given by Model 2, which was built by number of trees as 150 and maximal depth as 15.

Model	Accuracy on Training	Accuracy on Test	Parameter (No. of Trees)	Parameter (Criterion)	Parameter (Maximal Depth)
Model 1	79.05%	78.5%	100	Gini_Index	20
Model 2	87.88%	78.98%	150	Gini_Index	15

### Accuracy measures

On training data:

accuracy: 87.88%

	true 0	true 1	class precision
pred. 0	11000	1684	86.72%
pred. 1	0	1211	100.00%
class recall	100.00%	41.83%	

On test data:

accuracy: 78.98%

	true 0	true 1	class precision
pred. 0	7314	1946	78.98%
pred. 1	1	2	66.67%
class recall	99.99%	0.10%	

### Performance Measures

AUC: 0.599 (positive class: 1), specificity:99.99% (positive class:1), precision 66.67%% (positive class: 1)



## Boosted Trees:

We implemented Boosted trees by using Gradient boosted models.

Model	Accuracy on Training	Accuracy on Test	Parameter (No. of Trees)	Parameter (Maximal Depth)	Parameter (Min rows)
Model 1	68.91%	63.69%	20	5	10
Model 2	94.94%	72.83%	40	10	15

We see that Model 2 performs better with an accuracy of 72.72% on test data. The optimal parameters which give good test accuracy are – Number of trees = 40, Maximal Depth = 10, Min rows = 15.

### Accuracy Measures

#### On training data:

accuracy: 94.89%

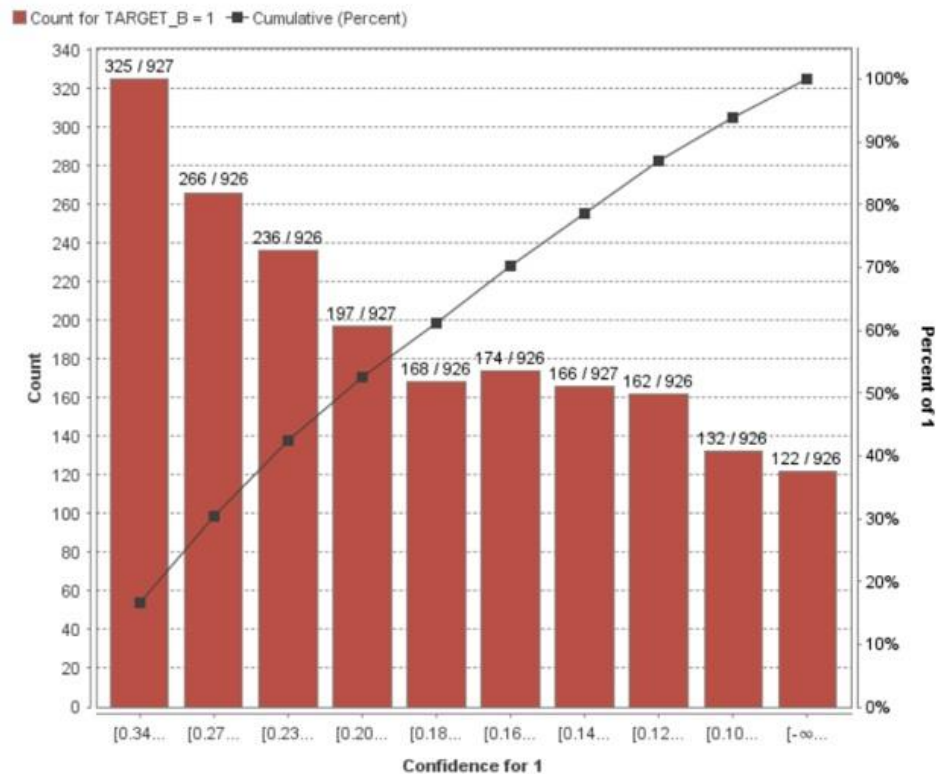
	true 0	true 1	class precision
pred. 0	10773	483	95.71%
pred. 1	227	2412	91.40%
class recall	97.94%	83.32%	

#### On test data:

accuracy: 72.83%

	true 0	true 1	class precision
pred. 0	6217	1419	81.42%
pred. 1	1098	529	32.51%
class recall	84.99%	27.16%	

### Lift Chart for Test Data:



### Performance Measures

AUC: 0.606 (positive class: 1), precision: 32.51% (positive class: 1), recall: 27.16% (positive class: 1), specificity: 84.99% (positive class: 1)

From analyzing the accuracy of all the models built above, we can see that Random Forest and Decision Trees give the best model accuracy on the test data. This is also evident from the performance measures that we have obtained.

**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.**

Sampling weights yield accurate population estimates for the main parameters of interest. If we wish to use our sample to calculate a descriptive statistic that accurately measures the true value in the population, then we need to weight. After all, this is the original purpose of sampling weights – to reverse the distortion imposed by the differential sampling probabilities.

The reason behind using weighted sampling to produce a training set with equal numbers of donors and non-donors is the fact that the original dataset is highly skewed with an actual response rate as 5.1%. There is a high probability that if a random sample is selected, it will mostly contain non-responders which will in turn prevent us from being able to predict the important variables accurately. If we used the simple random original dataset, the results obtained would be highly skewed data with poor classification accuracy.

Classification accuracy is an important performance measure. However, it is important that we also use other measures like Recall and Precision as classification accuracy is not the only criterion for calculating the maximum profit. Other techniques like Logistic regression, decision trees, random forests etc. can be used to build models to obtain better and accurate results.