

IDS 572 Assignment 2



Group Details:

Name

NagaShrikanth Ammanabrolu

Suresh Sappa

Sagar Kanchi

UIN

676837954

667192596

669850639

Table of Contents

Question No. 1).....	2
Dataset Exploration	2
Some Significant variables & Observations.....	2
Attributes Considered for the Analysis	3
Attributes which we won't be considering for building our model.....	4
MISSING VALUES	5
ELIMINATION OF USELESS ATTRIBUTES	6
Question No. 2).....	7
2.1 Naïve Bayes.....	8
2.2 W-Logistic Regression.....	9
2.3 W-J48.....	10
Comparison of the different classification models and their performance	11
Question No. 3).....	12
What is the reasoning behind using weighted sampling to produce a training set with equal numbers of donors and non-donors?	12
Why not use a simple random sample from the original dataset?	12
In this case, is classification accuracy a good performance metric for our purposes of maximizing net profit?	12
If not, how would you determine the best model? Explain your reasoning.....	13
APPENDIX	14

Question No. 1)

Solution.

Dataset Exploration

The given data set is explored in three steps.

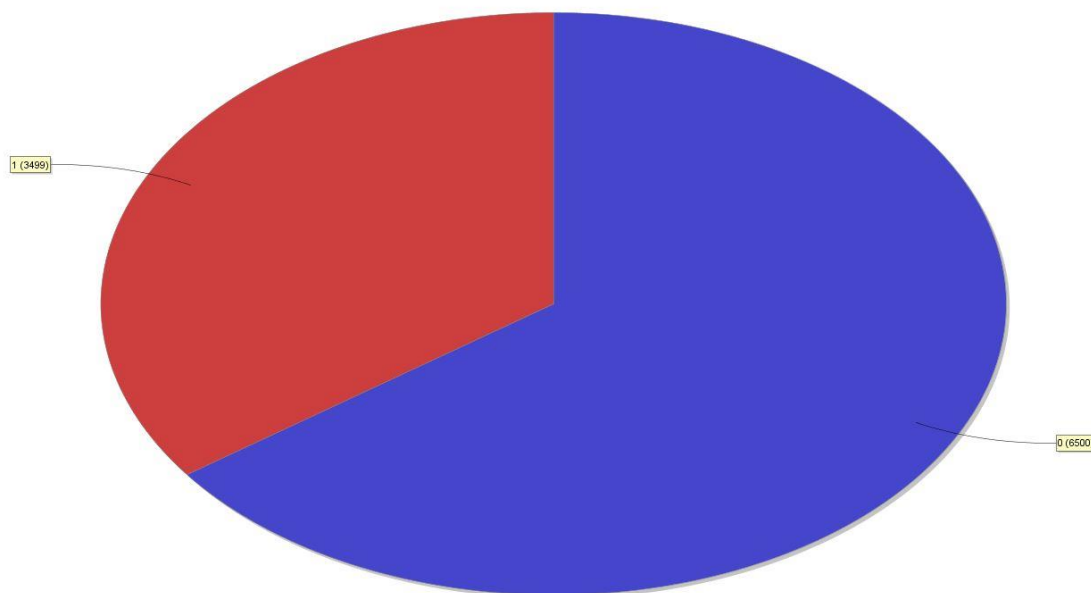
In the first step, the distribution of individual variables is checked to know the distribution of values of each variable and check for missing values and outliers. From this we figured the variables that needed transformation and those that should be included in or excluded from modeling (the variables considered and omitted are discussed in detail in the later part).

In the second step the relationship between target (dependent variables) and predictors (independent variables) is checked, which is used for feature selection.

In the third step the relationship among predictors themselves is checked, to remove the redundant variables.

Some Significant variables & Observations

TARGET_B - indicates the proportion of (donors Vs Non-donors). Plot below shows it is 35:65 i.e. approximately 35% donors and 65% non -donors.



INCOME - Income can be used as significant attribute to know the donating capacity of the person. By observing the distribution of values, we can come to know the ability of a person to be a potential donor or not.

EMPLOYMENT, WEALTH 1&2, HIGH PROPERTY VALUE – High pay receiving donors, people with good socio-economic status and with property value \geq \$200,000 (or \$150,000) can be categorized as potential targets to send our solicitations.

There are in all 480 attributes, and many attributes have missing values, many attributes are useless i.e. we don't need them in our final model. So we perform data cleaning and harmonizing step before applying data mining model on underlying data.

Attributes Considered for the Analysis

Neighborhood

Attribute Considered	Reason for Consideration
Percentage of Population from Urban, Town	This attribute provides the information on the standard of living of the Donor, which can be an important aspect for donation.
Age of the Donor	From the given data set this attribute shows the group of age of people who responded positively in earlier times.
Income of the Donor	These attributes directly reflects the donating capacity of the people from which potential donor can be classified.
Employment, Property value	

Donor Information

Attribute Considered	Reason for Consideration
DOMAIN	These attributes provides more information on the status of the Donor, which would help in targeting the prospective Donor.
HOMEOWNER	
HIT	Indicates total number of known times the donor has responded to a mail order offer other than PVA
FEDGOV	These variables imparts information on whether the donor was the veteran or any government/federal officer. Also, the data provided has the significant value.
LOCALGOV	
MALEMILI	
STATEGOV	
WWIIVETS	

TCODE	Title code of the donor
WEALTH	Wealth rating uses median family income and population statistics from each area to relative wealth within each state.
WEALTH2	

RFA

Attribute Considered	Reason for Consideration
RFA_2R	Recency/Frequency/Donation amount category shows the donors who responded to solicitations in a positive way. This can be used as a useful subset to categorize the responders from non-responders.
RFA_2F	
RFA_2A	
MDMAUD_R	
MDMAUD_F	
MDMAUD_A	

Mail order offers

Attribute Considered	Reason for Consideration
MAGFAML	As our main target variable is RESPONSE, these variables considered in this subset shows the donors who regularly responds to the mails. The donors that fall into this group can be targeted and receive a positive response.
MAGFEM	
MBCOLECT	
PUBCULIN	
PUBHLTH	
PUBNEWFN	
PUBDOITY	

Attributes which we won't be considering for building our model

Attribute Omitted	Reason for Omission
ODATE	Origin date: Date of donors first gift does not showcase any significant information on whether the donor responds to the mail or not.
OSOURCE	
MAILCODE	This tells us about whether the address is bad or good, which is not as significant as other attributes.
PVASTATE	Indicates whether the donor lives in a state served by the organization's EPVA chapter
NOEXCH	Around 99.7% of records has a value of "0" in field so it is also removed.

RECINHSE	Around 92.14 % of data under this attribute does not contain IN HOUSE record.
RECP3	Around 97.28% under this attribute does not have a P3 record
MDMAUD	It contains information on token of appreciation to donors for donating a fund of \$100+ at least once at any time in their giving history, which cannot be based upon for ensuring a response to the fundraising event as it might contain data which is obsolete.
CLUSTER	It does not show anything about prior response or donation made by the donor other than classification based on various characteristics.
AGEFLAG	Represent the exactness of age of the donor, which seems to have no relevance to our target variable.
GENDER	A donor can be a male or a female and the donation for the fundraising is not gender-specific.
SOLP3 & SOLIH	More than 99.5% data in the data set under this attribute shows a positive condition to mail
MAJOR	99.7% of records show not a major donor, which shows a pure class
GEOCODE	Geocode indicates the level geography at which a record matches the census data.
DONOR'S INTERESTS	The hobbies of the donor doesn't play significant role in determining the potential donor.
GIVING HISTORY FILE (RDATE_3 TO RMANT_24)	These variables provides the information on history of gift received from the donor in the past.
PROMOTION HISTORY FILE	
GIVING HISTORY	

MISSING VALUES

The dataset has large amount of variable (about **479** attributes) in which many of the attributes have missing values and many of them are unproductive i.e. they don't play any vital role in our final model. Thus we execute the process of data cleaning before we apply any decision techniques to our model.

We perform 'Mapping' for the attributes with missing values. This operator can be used to replace nominal values as well as numerical values and it also replaces the empty values with a significant one. For example: In a category called Hobbies, each variable takes values 'Y/N', hence mapping is done by replacing (empty values) '?' with a significant value - 'N'. We also make use of Generate attributes, this operator constructs new user defined attributes using mathematical expressions. Given below are some of the examples in which insignificant values are replaced by significant ones.

Attribute/ Name	Category	Original Value	Missing value replaced by-	Transformation Technique
Hobbies- Such as Bible, Boats, Cards, Catlg, CD Play, Collect1 etc.		Yes or No (Y/N)	N	Mapping (? - N)
recpgvg, recp3, rechinse, etc		X	0	Generate Attribute Rule: if(value = X,1,0)
Homeowner		H	0	Generate Attribute Rule: if(value = H,1,0)
Numerical values like age		1-99	-1	Replace Missing Values
Domain		C, R, S, T, U	R	Replace Missing Values AND Generate Attribute

ELIMINATION OF USELESS ATTRIBUTES

In the procedure above, we have just filled the missing values by significant values, however the number of total attributes remain the same – 479. As we made a note that all of these attributes are not needed for final analysis of our model. Therefore we eliminate some of these useless attributes. For detailed set of eliminated attributes, refer Appendix

Principal Component Analysis

After eliminating useless attributes, we are left with large amount of attributes. In practice it is not possible to consider the impact of so many explanatory attributes on one response variable i.e. 'Target_B'. Thus, we carry out 'Principal Component Analysis' for variable reduction.

Subset chosen for PCA: We applied Principal Component Analysis on 4 categories (mail order offers, donor's information, RFA and neighborhood) from entire data to get significant principal components.

Reason for doing PCA on these subsets: Donor's neighborhood has approximately 100 attributes (after performing removal steps), but all of these attributes does not create an impact on donor's contribution or donation. Hence, we performed PCA on 'neighborhood' category and reduced to 20, i.e. obtained 20 principal components out of 100. Similarly, we performed PCA on donor's information, mail order offers, RFA to get cumulative effect of all attributes into significant transformed and reduced variables.

Transformation before applying PCA: There are some categorical attributes in subsets like donor information, RFA that take nominal values for which we cannot apply PCA directly. Thus, we transformed the nominal values to numerical variables, normalize them then apply PCA.

Range transformation of PCA

Categories	Total attributes	Normalize criteria	No. of attributes after PCA
RFA	11	(Range Transformation) Threshold:	3
Mail order offers	14	(Range Transformation) Threshold:	3
Donor's information	41	(Range Transformation)	18
Donor's Neighborhood	287	(Range Transformation) No. donor's contribution or donation. Therefore, we perform PCA on entire "neighborhood" category	42
Total	353		66

(Please refer Appendix for supporting information)

Question No. 2)

Solution.

The next step after the data exploration and cleaning is to deal with the different types of classification models and techniques for data analysis. In this case, we will be splitting the data into a ratio of 60:40 for the training and the validation dataset respectively. The random seed is set to 12345. A few subset of variables are selected and the different types of classification models are applied to these data subsets.

The subsets for analysis are chosen by selecting/deselecting the different PCA blocks. The main subset in our case is taking into consideration the PCA of all the 4 subsets. The following are the test cases created using a combination of the PCA values of the 4 subsets:

- a. **All PCA values considered:** Here, the PCA values of all of the 4 subsets are considered for building the model
- b. **PCA values of 'DONOR Information' subset removed:** Here, only the Donor Information subset PCA is removed for building the model
- c. **PCA values of 'Mail order offers' subset removed:** Here, only the Mail order offers subset PCA is removed for building the model
- d. **PCA values of 'Neighborhood' subset removed:** Here, only the Neighborhood subset PCA is removed for building the model
- e. **PCA values of 'RFA' subset removed:** Here, only the RFA subset PCA is removed for building the model
- f. **PCA values of all subsets removed:** Here, all the four subsets' PCA values are removed for building the model

Below are the different types of classification techniques that have been implemented on the given dataset, for the above mentioned 6 combinations.

- 1. **Naïve Bayes**
- 2. **Logistic Regression**
- 3. **Decision Tree (W-J48)**

2.1 Naïve Bayes

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

The PCA analysis is performed on different combinations of subsets as described above. The results of the performance metrics on the different combinations of subsets is summarized below.

Scenario	Accuracy (%) for the Training Dataset	Accuracy (%) for the Test Dataset	Testing Recall Accuracy (%)
All PCA Values considered	60.99%	59.23%	50.87%

PCA values of 'DONOR Information' subset removed	61.41%	58.98%	49.2%
PCA values of 'Mail order offers' subset removed	60.81%	59.15%	51.29%
PCA values of 'Neighborhood' subset removed	54.29%	51.15%	64.35%
PCA values of 'RFA' subset removed	61.31%	58.75%	51.42%
All PCA Values Removed	54.21%	51.05%	63.86%

Summary: The above table summarizes the different combinations of subsets that we have considered for PCA Analysis and their respective performance metrics. These combination of subsets yield us the minimum number of variables, which helps in reducing the complexity of the model while not compromising on the performance of the model. As shown in the highlighted scenario above, we can see that removing the PCA values of the 'Neighborhood' subset yields a better True recall accuracy (True 1's). Although this model yields us a poor accuracy on the Testing dataset, we are more concerned with the accuracy of the True recalls. Since the true recalls indicate the actual donors who will maximize our profits, we are more interested in categorizing them as adequately as possible.

Hence, the highlighted scenario yields us a better model as compared to the other scenarios using the Naïve Bayes model.

(Please refer Appendix for supporting information)

2.2 W-Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE) or 0 (FALSE). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest and a set of independent variables.

The PCA analysis is performed on different combinations of subsets as described above. The results of the performance metrics on the different combinations of subsets is summarized below.

Scenario	Accuracy (%) for the Training Dataset	Accuracy (%) for the Test Dataset	Testing Recall Accuracy (%)
All PCA Values considered	64.10%	67.68%	21.26%
PCA values of 'DONOR Information' subset removed	67.68%	63.50%	17.37%
PCA values of 'Mail order offers' subset removed	67.66%	63.98%	18.76%
PCA values of 'Neighborhood' subset removed	67.68%	64.00%	18.21%
PCA values of 'RFA' subset removed	67.74%	64.00%	18.42%
All PCA Values Removed	62.35%	62.32%	20.34%

Summary: The above table summarizes the different combinations of subsets that we have considered for PCA Analysis and their respective performance metrics. As shown in the highlighted scenario above, we can see that considering all the PCA values yields a better True recall accuracy (True 1's). In this situation, considering all the PCA values for the analysis not only yields us the best testing recall accuracy, it also yields us a better overall test accuracy as compared to the other combination of subsets for PCA analysis.

Hence, the highlighted scenario yields us a better model as compared to the other scenarios using the W-Logistic Regression model.

[\(Please refer Appendix for supporting information\)](#)

2.3 W-J48

The W-J47 model which is a part of the WEKA extension creates a decision tree using the C4.5 algorithm. We also observed that pruning the tree doesn't have significant impact on the model we are testing here.

The PCA analysis is performed on different combinations of subsets as described above. The results of the performance metrics on the different combinations of subsets is summarized below.

Scenario	Accuracy (%) for the Training Dataset	Accuracy (%) for the Test Dataset	Testing Recall Accuracy (%)
All PCA Values considered	82.43%	57.85%	37.11%

PCA values of 'DONOR Information' subset removed	78.13%	59.13%	27.87%
PCA values of 'Mail order offers' subset removed	82.26%	58.98%	36.00%
PCA values of 'Neighborhood' subset removed	84.61%	56.67%	32.18%
PCA values of 'RFA' subset removed	81.96%	58.40%	35.79%
All PCA Values Removed	83.76%	56.90%	32.73%

Summary: The above table summarizes the different combinations of subsets that we have considered for PCA Analysis and their respective performance metrics. As shown in the highlighted scenario above, we can see that considering all the PCA values yields a better True recall accuracy (True 1's). Although this model yields us a poor accuracy on the Testing dataset, we are more concerned with the accuracy of the True recalls. Since the true recalls indicate the actual donors who will maximize our profits, we are more interested in categorizing them as adequately as possible.

Hence, the highlighted scenario yields us a better model as compared to the other scenarios using the W-J48 model.

(Please refer Appendix for supporting information)

Comparison of the different classification models and their performance

The best models from each of the three classification models are being summarized below for an overall picture of how these models perform against each other.

Classification Models	Accuracy (%) for the Training Dataset	Accuracy (%) for the Test Dataset	Testing Recall Accuracy (%)
Naïve Bayes	54.29%	51.15%	64.35%
W-Logistic Regression	64.10%	67.68%	21.26%
W-J48	82.43%	57.85%	37.11%

From the above table, we could conclude that Naïve Bayes is the classification model which yields us a better model as compared to the other models. Again, since the testing recall

accuracy are what we care for the most in the particular dataset analysis, Naïve Bayes offers the best results for the given dataset and its subsets.

Question No. 3)

Solution.

What is the reasoning behind using weighted sampling to produce a training set with equal numbers of donors and non-donors?

The reasons behind using weighted sampling to produce a training set with equal numbers of donors and non-donors is as follows:

- The main advantage of weighted sampling is that we can capture rare cases by extracting the necessary information from the given model while ensuring that the presence of bias is eliminated
- If we use simple sampling, there is a high possibility that almost all the data taken for the model would be from the non-responders as in the given data the response rate (5.1%) to non-response rate (94.9%) is 0.0537. Hence, we make use of weighted sampling to make sure that the data set is fair enough to be modelled
- Through the use of weighted sampling, the dataset can be equally divided into donors and non-donors which would result in obtaining an unbiased dataset for the model

Why not use a simple random sample from the original dataset?

A simple random sample usually does the best when the sample is not heavily positive or negative biased. In the given dataset, as the actual response rate is 5.1% and the non-response rate is 94.4%, thus the probability of using non-responder's data instead of responder's data to build the model is very high if we make use of random sampling.

Such a heavily skewed dataset arising from a simple random sample would not yield us a healthy model. To avoid this, the given dataset has been sampled with equal number of donors and non-donors. So, instead of simple random sampling, we make use of weighted sampling.

In this case, is classification accuracy a good performance metric for our purposes of maximizing net profit?

As the responders are sampled, the classification accuracy will not give us a precise model. In order to avoid this splitting, there are two ways by which we can build our model in order to maximize profit, and they are as follows:

- Confusion matrix needs to be altered to make use of original profit values.
- Weight profit in order to replicate the original data split.


If not, how would you determine the best model? Explain your reasoning.

Accuracy is taken into consideration for the best model to be chosen, but in this case accuracy might be good but the prediction might be not so good, thus for the profit to be maximized we should be targeting the probable donors.
















This can be achieved by implementing various decision techniques such as Decision trees (W-J48), Naïve Bayes, Logistic Regression, etc. These models can be then compared to the percentage of TRUE POSITIVE obtained rather than the accuracy.


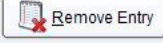

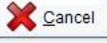
APPENDIX

Generating attributes for the missing values

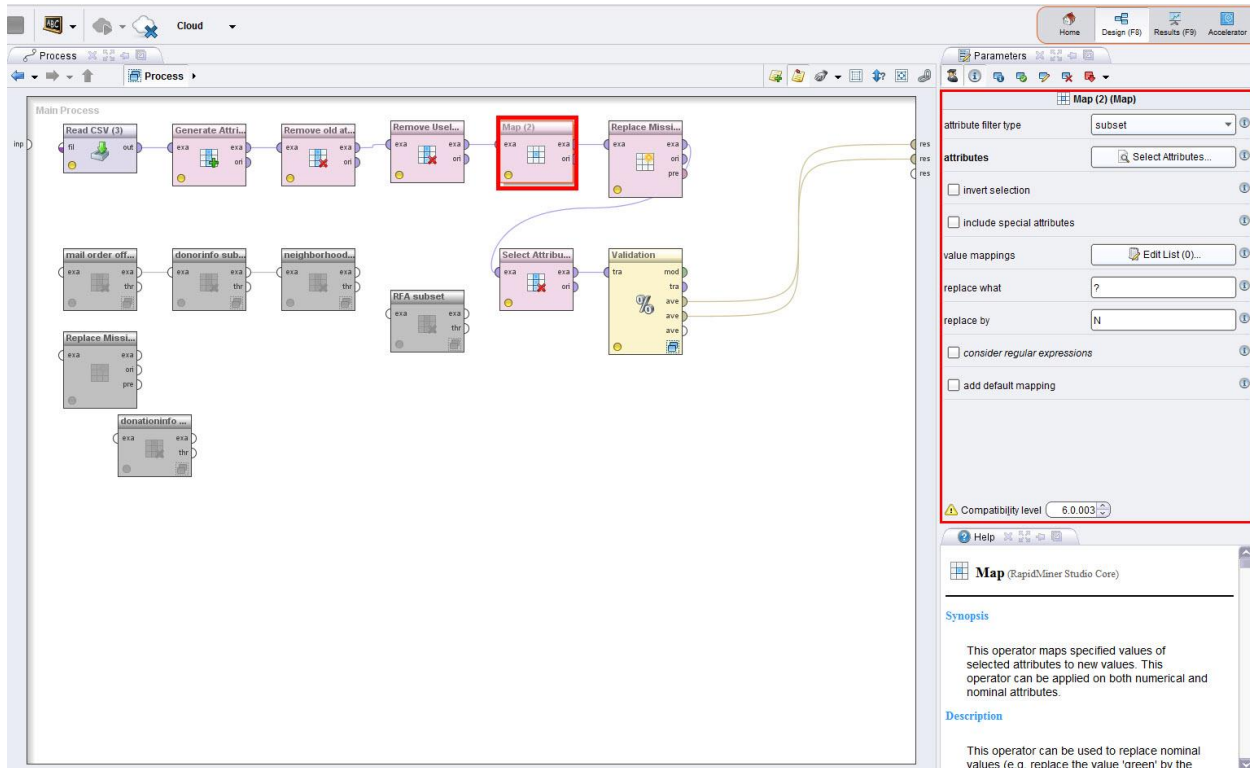


Edit Parameter List: function descriptions
List of functions to generate.

attribute name	function expressions	
EP_PvaState	if(PVASTATE=="E" PVASTATE=="P", "1", "0")	
rechinse	if(RECHNSE == "X", "1", "0")	
recp3	if(RECP3 == "X", "1", "0")	
recpgvg	if(RECPGVG == "X", "1", "0")	
recsweep	if(RECSWEEP == "X", "1", "0")	
domain	cut(DOMAIN, 0, 1)	
major	if(MAJOR=="X", "1", "0")	
pepstrfl	if(PEPSTRFL=="X", "1", "0")	
child3	if(CHILD03=="M" CHILD03=="F" CHILD03=="B", "1", "0")	
child7	if(CHILD07=="M" CHILD07=="F" CHILD07=="B", "1", "0")	
child12	if(CHILD12=="M" CHILD12=="F" CHILD12=="B", "1", "0")	
child18	if(CHILD18=="M" CHILD18=="F" CHILD18=="B", "1", "0")	
homeownr	if(HOMEOWNR=="H", "1", "0")	
totDays	date_diff(LASTDATE, ADATE_2)/(1000*60*60*24)	
gender	if(GENDER=="M" GENDER=="F", GENDER, "U")	

Mapping the missing values



Naïve Bayes model

All PCA values considered

accuracy: 60.99%

	true 0	true 1	class precision
pred. 0	2606	1007	72.13%
pred. 1	1333	1053	44.13%
class recall	66.16%	51.12%	

Training Performance matrix

accuracy: 59.23%

	true 0	true 1	class precision
pred. 0	1637	707	69.84%
pred. 1	924	732	44.20%
class recall	63.92%	50.87%	

Testing Performance matrix

PCA values of 'DONOR Information' subset removed

accuracy: 61.41%

	true 0	true 1	class precision
pred. 0	2634	1010	72.28%
pred. 1	1305	1050	44.59%
class recall	66.87%	50.97%	

Training Performance matrix

accuracy: 58.98%			
	true 0	true 1	class precision
pred. 0	1651	731	69.31%
pred. 1	910	708	43.76%
class recall	64.47%	49.20%	

Testing Performance matrix

PCA values of 'Mail order offers' subset removed

accuracy: 60.81%			
	true 0	true 1	class precision
pred. 0	2601	1013	71.97%
pred. 1	1338	1047	43.90%
class recall	66.03%	50.83%	

Training Performance matrix

accuracy: 59.15%			
	true 0	true 1	class precision
pred. 0	1628	701	69.90%
pred. 1	933	738	44.17%
class recall	63.57%	51.29%	

Testing Performance matrix

PCA values of 'Neighborhood' subset removed

accuracy: 54.29%			
	true 0	true 1	class precision
pred. 0	1865	668	73.63%
pred. 1	2074	1392	40.16%
class recall	47.35%	67.57%	

Training Performance matrix

accuracy: 51.15%			
	true 0	true 1	class precision
pred. 0	1120	513	68.59%
pred. 1	1441	926	39.12%
class recall	43.73%	64.35%	

Testing Performance matrix

PCA values of 'RFA' subset removed

accuracy: 61.31%			
	true 0	true 1	class precision
pred. 0	2588	970	72.74%
pred. 1	1351	1090	44.65%
class recall	65.70%	52.91%	

Training Performance matrix

accuracy: 58.75%

	true 0	true 1	class precision
pred. 0	1610	699	69.73%
pred. 1	951	740	43.76%
class recall	62.87%	51.42%	

Testing Performance matrix

PCA values of all subsets removed

accuracy: 54.21%

	true 0	true 1	class precision
pred. 0	1867	675	73.45%
pred. 1	2072	1385	40.06%
class recall	47.40%	67.23%	

Training Performance matrix

accuracy: 51.05%

	true 0	true 1	class precision
pred. 0	1123	520	68.35%
pred. 1	1438	919	38.99%
class recall	43.85%	63.86%	

Testing Performance matrix

W-Logistic Regression model

All PCA values considered

accuracy: 64.10%

	true 0	true 1	class precision
pred. 0	2298	1173	66.21%
pred. 1	263	266	50.28%
class recall	89.73%	18.49%	

Training Performance matrix

accuracy: 67.68%

	true 0	true 1	class precision
pred. 0	3622	1622	69.07%
pred. 1	317	438	58.01%
class recall	91.95%	21.26%	

Testing Performance matrix

PCA values of 'DONOR Information' subset removed

accuracy: 67.68%

	true 0	true 1	class precision
pred. 0	3627	1627	69.03%
pred. 1	312	433	58.12%
class recall	92.08%	21.02%	

Training Performance matrix

accuracy: 63.50%			
	true 0	true 1	class precision
pred. 0	2290	1189	65.82%
pred. 1	271	250	47.98%
class recall	89.42%	17.37%	

Testing Performance matrix

PCA values of 'Mail order offers' subset removed

accuracy: 67.66%			
	true 0	true 1	class precision
pred. 0	3623	1624	69.05%
pred. 1	316	436	57.98%
class recall	91.98%	21.17%	

Training Performance matrix

accuracy: 63.98%			
	true 0	true 1	class precision
pred. 0	2289	1169	66.19%
pred. 1	272	270	49.82%
class recall	89.38%	18.76%	

Testing Performance matrix

PCA values of 'Neighborhood' subset removed

accuracy: 67.68%			
	true 0	true 1	class precision
pred. 0	3626	1626	69.04%
pred. 1	313	434	58.10%
class recall	92.05%	21.07%	

Training Performance matrix

accuracy: 64.00%			
	true 0	true 1	class precision
pred. 0	2298	1177	66.13%
pred. 1	263	262	49.90%
class recall	89.73%	18.21%	

Testing Performance matrix

PCA values of 'RFA' subset removed

accuracy: 67.74%			
	true 0	true 1	class precision
pred. 0	3622	1618	69.12%
pred. 1	317	442	58.23%
class recall	91.95%	21.46%	

Training Performance matrix

accuracy: 64.00%			
	true 0	true 1	class precision
pred. 0	2295	1174	66.16%
pred. 1	265	265	49.91%
class recall	89.51%	18.42%	

Testing Performance matrix

PCA values of all subsets removed

accuracy: 62.35%			
	true 0	true 1	class precision
pred. 0	3221	1876	63.19%
pred. 1	363	487	57.29%
class recall	89.87%	20.61%	

Training Performance matrix

	true 0	true 1	class precision
pred. 0	2156	1234	63.60%
pred. 1	260	315	54.78%
class recall	89.24%	20.34%	

Testing Performance matrix

W-J48 model

All PCA values considered

accuracy: 82.43%			
	true 0	true 1	class precision
pred. 0	3486	601	85.29%
pred. 1	453	1459	76.31%
class recall	88.50%	70.83%	

Training Performance matrix

accuracy: 57.85%			
	true 0	true 1	class precision
pred. 0	1780	905	66.29%
pred. 1	781	534	40.61%
class recall	69.50%	37.11%	

Testing Performance matrix

PCA values of 'DONOR Information' subset removed

accuracy: 78.13%			
	true 0	true 1	class precision
pred. 0	3521	894	79.75%
pred. 1	418	1166	73.61%
class recall	89.39%	56.00%	

Training Performance matrix

accuracy: 59.13%			
	true 0	true 1	class precision
pred. 0	1964	1038	65.42%
pred. 1	597	401	40.18%
class recall	76.69%	27.87%	

Testing Performance matrix

PCA values of 'Mail order offers' subset removed

accuracy: 82.26%			
	true 0	true 1	class precision
pred. 0	3499	624	84.87%
pred. 1	440	1436	76.55%
class recall	88.83%	69.71%	

Training Performance matrix

accuracy: 58.98%			
	true 0	true 1	class precision
pred. 0	1841	921	66.65%
pred. 1	720	518	41.84%
class recall	71.89%	36.00%	

Testing Performance matrix

PCA values of 'Neighborhood' subset removed

accuracy: 84.61%			
	true 0	true 1	class precision
pred. 0	3579	563	86.41%
pred. 1	360	1497	80.61%
class recall	90.86%	72.67%	

Training Performance matrix

accuracy: 56.67%			
	true 0	true 1	class precision
pred. 0	1804	976	64.89%
pred. 1	757	463	37.95%
class recall	70.44%	32.18%	

Testing Performance matrix

PCA values of 'RFA' subset removed

accuracy: 81.96%			
	true 0	true 1	class precision
pred. 0	3525	668	84.07%
pred. 1	414	1392	77.08%
class recall	89.49%	67.57%	

Training Performance matrix

accuracy: 58.40%			
	true 0	true 1	class precision
pred. 0	1821	924	66.34%
pred. 1	740	515	41.04%
class recall	71.11%	35.79%	

Testing Performance matrix

PCA values of all subsets removed

accuracy: 83.76%			
	true 0	true 1	class precision
pred. 0	3546	581	85.92%
pred. 1	393	1479	79.01%
class recall	90.02%	71.80%	

Training Performance matrix

accuracy: 56.90%			
	true 0	true 1	class precision
pred. 0	1805	968	65.09%
pred. 1	756	471	38.39%
class recall	70.48%	32.73%	

Testing Performance matrix