Data Mining

February 25, 2016

Premnath Ramanathan Sriram Santhanakrishnan Vignesh Shankar

Table of Contents

Question 1

1. Which variables will you consider for modeling (and why)?

After exploring, cleaning, transforming and reducing the variables from the given dataset, there were 25 variables that seemed to be appropriate for modeling. The final set of variables were a mix of PCA components and the normal individual variables from the dataset. The following are the list of variables and the reason why they are considered for modeling:

Variable Name	Explanation
Military_1	Military and Veteran PCA Output variable. This is a single variable that represents 16 variables
Housing_1	Owner PCA Output variable. This is a representation of 5 variables
RStatus_1	Marriage Output variable. 4 variables reduced to this one variable
Ethicity_1	Ethnicity PCA Output variable. 16 variables related to ethnicity is reduced to 1 variable
Occ & Emp_1	Occupation and employment PCA Output variable
Labour_1	% Labour PCA Output variable
MailResponse_1	Types of Mail Orders PCA Output variable
AGE	Age, in our opinion, will definitely be a factor in determining whether a person would be a potential donor or not
INCOME	Income will be a factor in classification
HV2	Home Value can be a factor
CARDPROM	Represents the lifetime number of card promotions received
NUMPROM	Lifetime number of promotions received till date
CARDPM12	Number of card promotions in the last 12 months can have an impact in the final result
NUMPRM12	The number of promotions in the last 12 months can be decisive factor in classification
NGIFTALL	The number of lifetime gifts till date can determine the future donations a person can do
CARDGIFT	This variable can have be a decisive factor in classification
TIMELAG	The number of months between first and second gift may be representation fort the frequency with which the donors donate
pepstrfl	Indicates PEP Star RFA Status
totDays	Time taken to respond to mail
CLUSTER	Unique group formed using socio-economic status, urbanity, ethnicity and a variety of other demographic
wealth_rating	A persons wealth can determine whether he could be a potential donor or not
TCODE	Donor title code can have an effect on the target classification
dataSource	Source of data can play a significant role in classification
HigherChildSpending	Flag if the number of children is more than the average children per family in the country

2. Which attributes will you omit from the analyses and why.

After exploring the data, the following variables were omitted and there were various reason that led to the omission. The tabular column below outlines the set of variables that were omitted and the corresponding reason for omission.

Attribute	Reason
AC1-2	The PCA (population and age)for donors neighborhood covers the age hence AC1-AC2 would be redundant)
AGE901-907	The PCA (population and age)for donors neighborhood covers the age hence AC1-AC2 would be redundant)
ANC1-15	The ancestry data of the neighborhood does not help in predicting or deciding if the person is oriented with the specific methods followed by the society.
CLUSTER2	Cluster2 is redundant as variable CLUSTER already represents the same data
CONTROLN	Unique Identification number Is nothing but ID hence it does not help in any decision making
DMA/ADI/MSA	Code for metropolitan and designated market and does not help to make decision
DOB	Age already satisfies the details
DW1-DW9	The value were too scattered and they did not seem to make meaning patterns
EIC1-EIC16	Income field gives an insight of employment s
ETHC1	Age and PCA of ethnic ETH1-16 would give us a better picture
GEOCODE/GEOCODE2	Redundant information
HC1-16/HU1-5/HUPA1-7/HUR1-2	The data does not influence the decision tree as the data is explained in HHN1-6
HHAS1-4	Data covered under employment
HHP1	HHP2 gives better detail
HIT	Low dispersion , mostly values are 0
HV1,3,4/HVP1-6/MHUC1-2	HV2 Average home value gives the value for all the variables together
LASTGIFT	Avoiding redundancy as AVGGIFT is used
LFC1-10	Income field gives an insight of employment
MAXRAMNT/MINRAMNT	Avoiding redundancy as RAMNTALL is used
NUMCHLD	Data cover under CHILD
OCC1-13/OEDC1-7	Income field gives an insight of employment
OSOURCE	Does not influence decisions
PEC1-2	Data not relevant
RHP1-4/RP1-3	No impact on decision tree
TPE1-16/VOC1-3	Transportation & vehicles does not relate of any factors in decision making

Few variables were omitted by performing correlation matrix. The variables showed less correlation hence we used few of these variables as individual factor for the model. This was taken based on the decision tree.

Below is the correlation matrix

	VETERANS	BIBLE	CATLG	HOMEE	PETS	CDPLAY	STEREO	PCOWNERS	PHOTO	CRAFTS	FISHER	GARDENIN	BOATS	WALKER	KIDSTUFF	CARDS	PLATES	TARGET_B
VETERANS	1.00																	
BIBLE	0.21	1.00																
CATLG	0.19	0.17	1.00															
HOMEE	0.05	0.06	0.15	1.00														
PETS	0.28	0.22	0.38	0.16	1.00													
CDPLAY	0.25	0.19	0.39	0.13	0.47	1.00												
STEREO	0.25	0.28	0.35	0.10	0.40	0.53	1.00											
PCOWNERS	0.21	0.19	0.32	0.16	0.42	0.51	0.38	1.00										
PHOTO	0.18	0.15	0.20	0.06	0.22	0.28	0.35	0.23	1.00									
CRAFTS	0.20	0.23	0.26	0.07	0.35	0.27	0.34	0.22	0.22	1.00								
FISHER	0.23	0.17	0.21	0.06	0.33	0.23	0.26	0.20	0.19	0.26	1.00							
GARDENIN	0.29	0.29	0.33	0.10	0.46	0.37	0.40	0.31	0.27	0.40	0.33	1.00						
BOATS	0.11	0.07	0.10	0.03	0.19	0.16	0.15	0.16	0.16	0.14	0.30	0.14	1.00					
WALKER	0.26	0.30	0.25	0.09	0.32	0.33	0.41	0.26	0.25	0.29	0.23	0.43	0.13	1.00				
KIDSTUFF	0.07	0.11	0.04	0.04	0.05	0.04	0.06	0.05	0.04	0.07	0.04	0.04	0.03	0.04	1.00			
CARDS	0.11	0.13	0.06	0.02	0.03	0.05	0.05	0.05	0.06	0.06	0.02	0.05	0.00	0.06	0.26	1.00		
PLATES	0.11	0.06	0.05	0.03	0.05	0.04	0.02	0.03	0.04	0.04	0.02	0.03	0.02	0.04	0.15	0.22	1.00	
TARGET_B	0.01	0.02	0.01	0.00	0.01	0.01	0.01	-0.01	0.02	0.01	0.00	0.02	0.00	0.02	0.00	0.00	0.01	1.00

3. How do you clean the data, handle missing values?

The following ways were used to clean and handle the missing values. All the missing values were handled using the "Replace Missing Values" attribute.

Numeric Variables: All the missing values within numeric variables except for the 'Age' attribute were replaced with a default value of '-1'. This was done in order to have a constant pattern throughout the dataset. For age, the missing values were replaced by the mean since age followed a normal distribution pattern.

Categorical Variables: The following categorical variables were replaced by taking the mode

- Cluster
- INCOME
- DOMAIN
- WEALTH_RATING

4. What new attributes/values do you derive?

There were a lot of variables that were transformed because they were not readily available to be used in the models. For example the variables had values like 'Y' for Yes or 'X' for Yes and the blank values were mentioned to be treated as a 'No'. In these sort of situations, the data was required to be transformed. The normal scheme that was used was to map '1' to 'Y' and '0' to 'No'. In addition to the 'if' function, 'cut' function was also used to get a substring of the existing data. Refer to Exhibit 1 to find about how data was transformed using the 'Generate Attribute' operator.

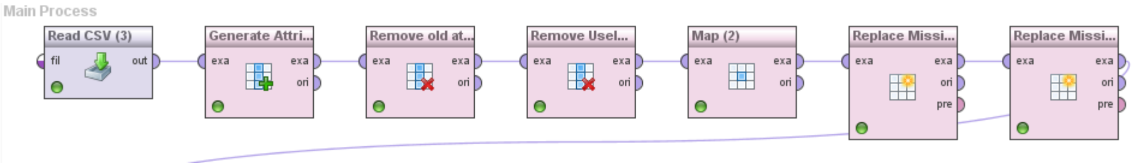
5. How do you approach data reduction and what methods for data reduction do you try?

The following were the different ways followed to perform data reduction:

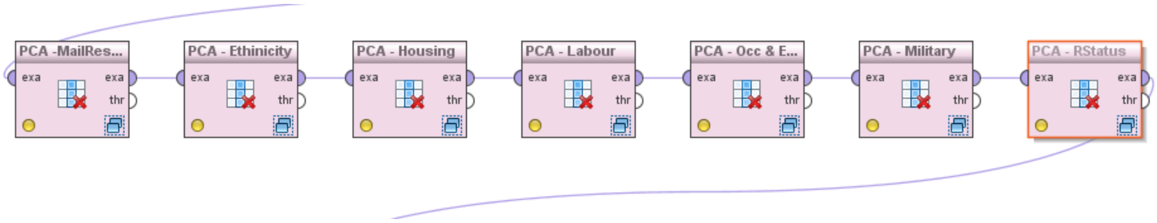
Intuition – There were certain variables which were removed based on intuition. They were simply disregarded as we thought it was not going to impact the final decision.

Distribution: One other factor that was looked upon for data reduction was the distribution. For instance, if one class has very meagre percentage compared to the other class, the variable was taken out.

Missing Values: If a variable had a lot of missing values present, then it was disregarded.



Principal Component Analysis: After reducing the number of variables through the above mentioned steps, Principal Component Analysis (PCA) was performed. Variables which were similar in nature were taken together and PCA was performed on these variables. The output from each of the PCA model was reduced to 1 irrespective of the variance, since the performance was not impacted because of the cumulative variance i.e. the performance remained the same when the cumulative variance threshold was reduced from 0.9 to 0.6 for all the variables.



6. Data exploration: Import the data, and examine the different variables – distribution of values, mean and std deviation, range of values. What do you observe?

The following is the descriptive statistics for the target variable 'TARGET_B':

There is a 65-35 split among the values in the target variable. 65% being ‘1’ and 35 % being ‘0’.

It is interesting to note this because the initial data set had only 5% in class 1 and it has been increased to 35%.

After importing the data and looking at the different variables, the one significant thing to note was the number of missing variables. The number was in the range of more than 5000. Some of these variables, as discussed earlier, were because of the way it was structured (Ex: ‘X’ or ‘Y’ for ‘Yes’ and blank values for a ‘No’). But some of the variables other than this also had a lot of missing values. The only possible solution to this was to eliminate these variables.

For variables that were continuous, there were two distributions observed from the given variables. They were either normally distributed or right skewed.

For many categorical variables, the proportions of 0’s were far greater than the number of 1’s. In these types of scenarios, the variables were not taken into consideration for analysis and it was removed.

There were also date variables that were present as part of the given dataset. All the data variables from the given dataset were not taken into consideration for data analysis and were taken out from the final set of independent variables.

7. 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?

There were a total of 7 PCAs that were performed on the dataset. Refer to EXHIBIT 2 for the different groups.

Question 2

Modeling Partitioning - Partition the dataset into 60% training and 40% validation (set the seed to 12345). [A specified seed ensures that we obtain the same random partitioning every time we run it. With no specified seed, the system clock is typically used to set the seed, and a different partitioning can result in different runs].

Consider 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

Our approach to find the best model was based on the selection of attributes and the performance results from training and validation by changing the parameters in each of the classification techniques. First we partitioned the dataset into 60% training and 40% validation. We also set the random seed value to 12345 as given in the problem to ensure that same results are obtained every time we run the process.

As the problem is dependent mainly on predicting the potential donors rather than the non-donors we set a threshold of 0.7 for both training and validation data. Which gave us an optimistic value for class precision for true 1's (correct prediction of donors) while reducing the cost incurred for falsely predicted non-donors.

The basic structure of the split validation process is available in Appendices section of this document (Please refer to Exhibit 3).

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. 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?

The different classification techniques and the parameters used for those techniques are as follows.

a) Decision Tree

We compared the results between **W-J48** operator and the **Decision Tree** operator by trying out different parameter values for both the operators on the final set of (29) attributes.

The performance comparisons of both the above mentioned operators are available in Appendices section of this document (refer Exhibit 4).

We tried using different parameters for both the operators and the significance of those parameters are as given below

Decision Tree Parameters	Best Model	Reasons/Significance
Criterion	Information_gain	Out of the four criterion, Information_gain gave us the most optimized results
Max depth	4	Decreasing the depth to 4 gave us better results. A depth less than 4 and more than 15 produced drastic results.
Pruning confidence	0.5	Default value of 0.5 for confidence gave us the best model
Pre-pruning		
Minimal gain	0.01	A minimal gain value from 0.1 to 0.001 gave us good results. But a value of 0.01 gave us the best model
Min leaf size	2	Leaf size of 1 gave us optimized output
Min size for split	2	A value ranging from 1 to 12 gave us varying results out of which a value of 2 gave us the best model.

No. of prepruning alts	3	This factor did not significantly change the accuracy of the end result. Hence, default value was used

b) <u>Logistic Regression</u>

We used W-Logistic operator on our final set of variables and trained the model to provide the desired output. We derived our best model by using the following parameter values.

W-Logistic Parameters	Best Model	Reasons/Significance
D - Turn on Debugging output	Unchecked	Debugging is turned off by default. We used the default value because there was no significant difference otherwise
R - Ridge in the log-likelihood	1	Changing the ridge value did not significantly change the performance. A value of 1 gave us the best model.
M - Maximum number of iterations	-1	Default value of -1 (until convergence) was used as other values did not change the output significantly

The performance results from this model are available in the Appendices section of this document (Refer Exhibit 5).

c) Naïve Bayes classification

We tried both Naïve Bayes and Naïve Bayes (Kernel) operators on our final set of attributes. Naïve Bayes operator gave us optimal results compared to Naïve Bayes (Kernel). Naïve Bayes produced the best model which could be because the method works better on categorical attributes. The above distinction could be due to Naïve Bayes being data dependent (Generative classifier).

The performance results from both the operators are available in the appendices section of this document (refer Exhibit 6).

The various parameters used and their significance are listed below.

Naïve Bayes Parameter	Best Model	Reasons/Significance
Laplace Correction	Checked	Laplace correction was turned on for smoothing.
Estimation mode-kernel density estimation	greedy	Default mode greedy gave us the best model. Full estimation mode produces fixed bandwidth
bandwidth selection	none	Default value of Heuristic bandwidth selection gave us a decent model. Fixed bandwidth gave us an over fit model
Minimum bandwidth	1	Value of 1.0 was used to get the best model. Other values gave us decent models
Number of kernels	17	A value of 17 for number of kernels gave us a good model. Value above 20 and below 15 gave us over fit model.
application grid size	none	Change in grid size gave us poor results.

Run each method on a chosen subset of the variables - how do you select this subset? Provide a comparative evaluation of performance of your best models from each technique.

The different subset of variables were chosen carefully by first analyzing the dataset and identifying variables that are significant and then packaging them into small subsets. A subset of the variables were chosen and run on each method to check their performance. We decided to take this subset of variables based on their significance from decision tree. The list of variables that were chosen as base subset and model applied on them are mentioned below.

Model	Training				Validation			
	Class recall		Class Precision		Class recall		Class Precision	
	True 0	True 1	True 0	True 1	True 0	True 1	True 0	True 1
Decision tree	44.83%	80.34%	81.35%	43.23%	39.55%	64.56%	66.51%	37.51%
W-J48	44.81%	79.95%	81.04%	43.10%	40.84%	72.13%	72.29%	40.66%
W-Logistics	25.36%	87.38%	79.35%	37.97%	23.27%	85.89%	74.59%	38.61%
Naïve Bayes	46.36%	72.18%	76.12%	41.31%	41.51%	67.13%	69.21%	39.20%

Does variable selection/PCA make a difference for the different models?

We tried excluding combination of PCA's and PCA's individually and checked if it affects the performance significantly. The significance of PCA's across the three techniques are given below.

Decision Tree:

We tried excluding different combinations of subset of PCAs and evaluated the model by comparing the results from our base model which includes all the PCAs. We found out that by excluding Ethnicity, RStatus, Housing and Occ & Emp PCAs gave us the best model based on performance of recall. Please refer Exhibit 7 that contains the performance of the best model using Decision Tree.

W-J48:

All variables reflecting donor interests did not lead to change in the performance but when few variables such as veterans, Pcowners improved the result. Exclusion of PCA variables such as RStatus which indicates the marital status of the neighborhood gave improved class recall but the precision % considerable went down.

W-Logistic:

Excluding both RStatus PCA and Military PCA gave us a decent model when compared to excluding other subset of PCAs but did not improve the performance obtained from our base model.

Naïve Bayes:

Excluding subset of PCA's or individual PCAs did not improve the performance when compared to the base model. Highest performance was obtained by excluding Ethnicity and Military PCAs but it was not better than the model obtained from our base set of attributes.

Question 3

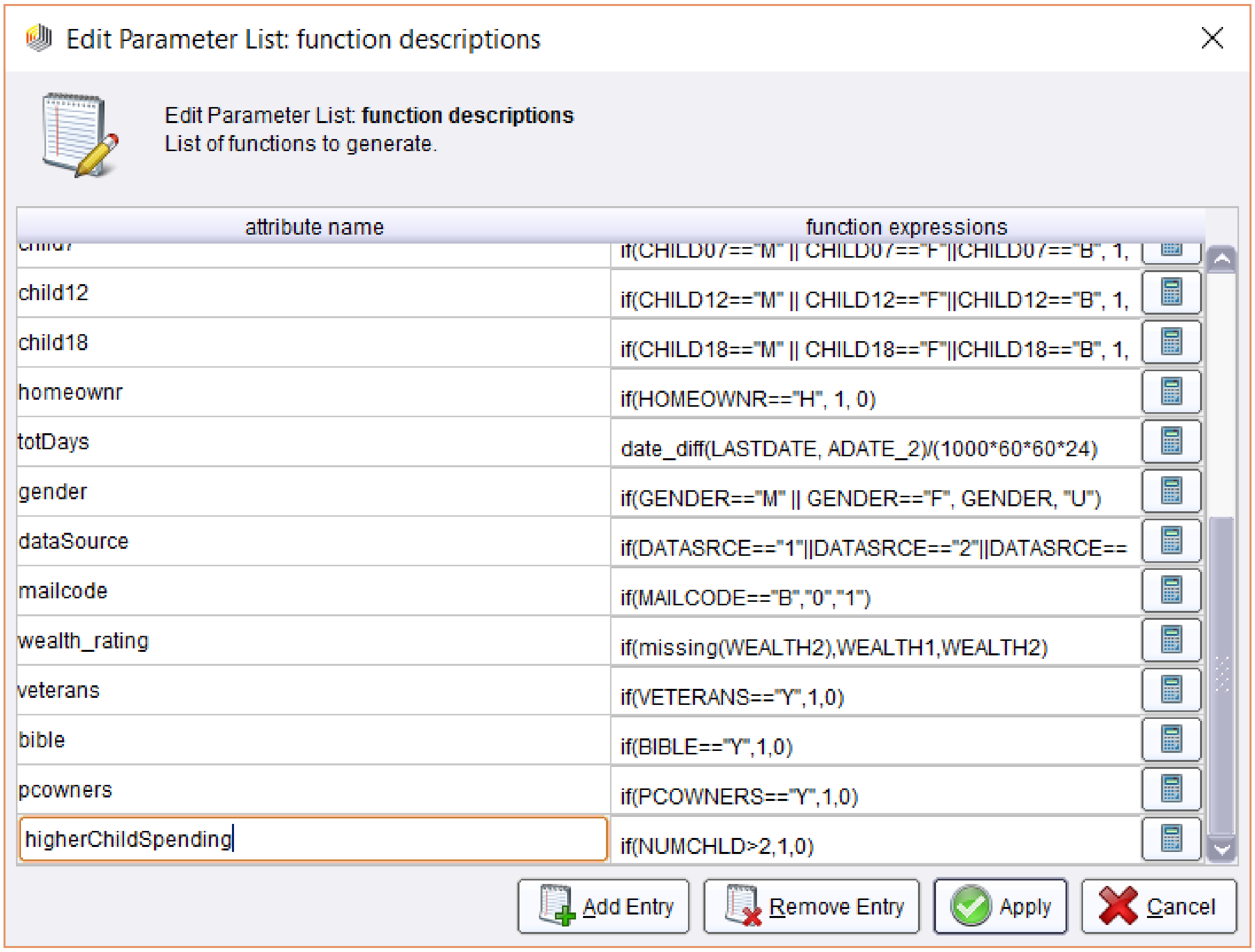
Classification under asymmetric response and cost: What is the reasoning behind using weighted sampling to produce a training set with equal numbers of donors and non-donors? Why not use a simple random sample from the original dataset? (Hint: given the actual response rate of 5.1%, how do you think 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

The weighted sampling which has 65% belonging to class 0 and 35% belonging to class 1 because 5.1% of data belonging to class 1 in the initial dataset is very meagre and taking that into consideration to build a model will not be useful in identifying the parameters that would be helpful in predicting class 1 correctly. Hence by increasing the percentage of the number of data belonging to class 1 to 35%, the accuracy with which class 1 will be predicted will be high.
The classification accuracy, in our opinion, is not a good performance metric for the purpose of maximizing profit. The best model, as we have discussed earlier as part of the second question, can be achieved by arriving at a threshold value for the confidence that can be a better performance indicator. Hence, calculating the misclassification costs by obtaining the optimal threshold would be a good metric for measuring performance. Though class recall for class 0 is very less compared to class 1, this trade-off results in the maximum profit reducing the misclassification costs.

The $ value obtained for finding out the threshold is as follows:
For Class 1: (5.1/35)*12.32 = 1.7952 (We are considering 12.32 because $0.68 would be charged for postal and hence the loss would be 12.32 and not 13)
For Class 0: (94.9/65)*0.68 = 0.9928

Appendix

Exhibit 1



Edit Parameter List: function descriptions

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

attribute name	function expressions
child7	if(CHILD07=="M" || CHILD07=="F"||CHILD07=="B", 1,
child12	if(CHILD12=="M" || CHILD12=="F"||CHILD12=="B", 1,
child18	if(CHILD18=="M" || CHILD18=="F"||CHILD18=="B", 1,
homeownr	if(HOMEOWNR=="H", 1, 0)
totDays	date_diff(LASTDATE, ADATE_2)/(1000*60*60*24)
gender	if(GENDER=="M" || GENDER=="F", GENDER, "U")
dataSource	if(DATASRCE=="1"||DATASRCE=="2"||DATASRCE==
mailcode	if(MAILCODE=="B","0","1")
wealth_rating	if(missing(WEALTH2),WEALTH1,WEALTH2)
veterans	if(VETERANS=="Y",1,0)
bible	if(BIBLE=="Y",1,0)
pcowners	if(PCOWNERS=="Y",1,0)
higherChildSpending	if(NUMCHLD>2,1,0)

Exhibit 2

PCA-1	PCA-2	PCA-3	PCA-4	PCA-5	PCA-6	PCA-7
Mail Reponses	**Ethnicity**	**Housing**	**Labour**	**Occupation & Employment**	**Military**	**Rstatus**
MAGFAML	ETH1	HU1	LFC1	EIC1	AFC1	MARR1
MAGFEM	ETH10	HU2	LFC10	EIC10	AFC2	MARR2
MAGMALE	ETH11	HU3	LFC2	EIC11	AFC3	MARR3
MBBOOKS	ETH12	HU4	LFC3	EIC12	AFC4	MARR4
MBCOLECT	ETH13	HU5	LFC4	EIC13	AFC5	
MBCRAFT	ETH14		LFC5	EIC14	AFC6	
MBGARDEN	ETH15		LFC6	EIC15	VC1	
PUBCULIN	ETH16		LFC7	EIC16	VC2	
PUBDOITY	ETH2		LFC8	EIC2	VC3	
PUBGARDN	ETH3		LFC9	EIC3	VC4	

PUBHLTH	ETH4			EIC4		
PUBNEWFN	ETH5			EIC5		
PUBOPP	ETH6			EIC6		
PUBPHOTO	ETH7			EIC7		
	ETH8			EIC8		
	ETH9			EIC9		
				OCC1		
				OCC10		
				OCC11		
				OCC12		
				OCC13		
				OCC2		
				OCC3		
				OCC4		
				OCC5		
				OCC6		
				OCC7		
				OCC8		
				OCC9		
				OEDC1		
				OEDC2		
				OEDC3		
				OEDC4		
				OEDC5		
				OEDC6		
				OEDC7		

Exhibit 3

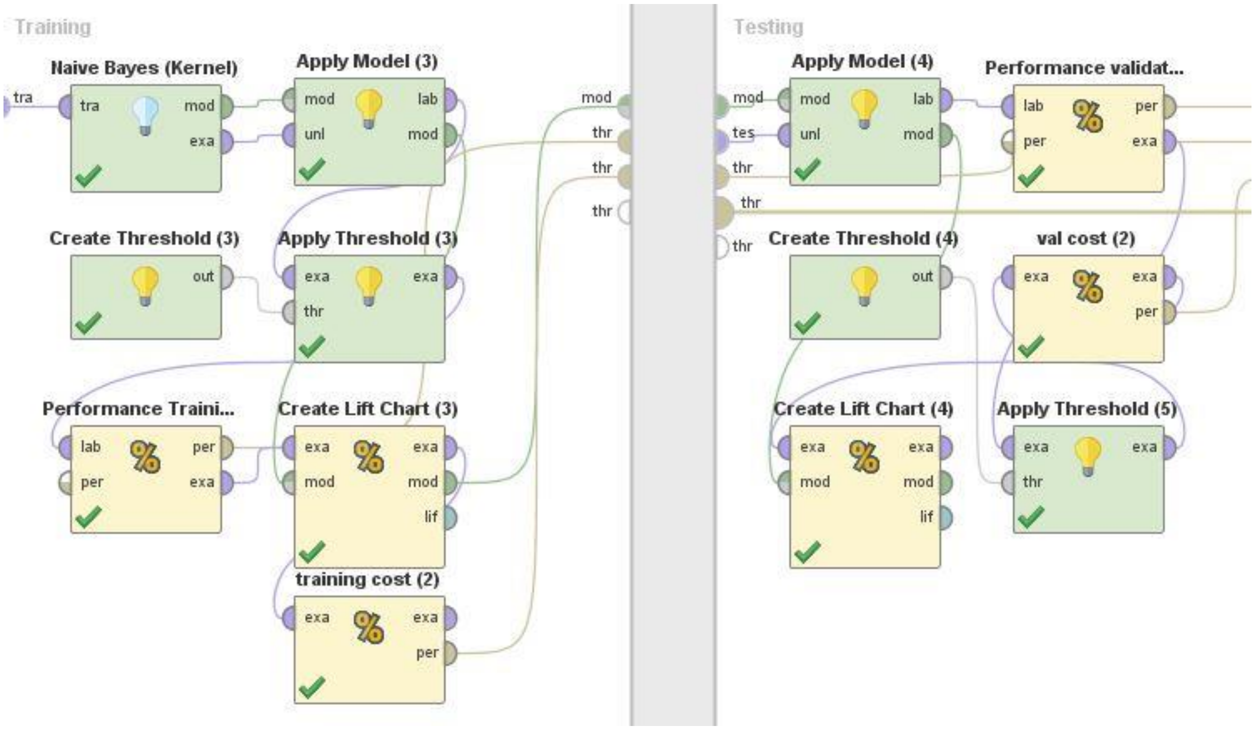


Exhibit 4

W-J48 Training	true 0	true 1	class precision
pred. 0	1765	413	81.04%
pred. 1	2174	1647	43.10%
class recall	44.81%	79.95%	

W-J48 Validation	true 0	true 1	class precision
pred. 0	1046	401	72.29%
pred. 1	1515	1038	40.66%
class recall	40.84%	72.13%	

Decision Tree Training	true 0	true 1	class precision
pred. 0	1766	405	81.35%
pred. 1	2173	1655	43.23%
class recall	44.83%	80.34%	

Decision Tree Validation	true 0	true 1	class precision
pred. 0	1013	510	66.51%
pred. 1	1548	929	37.51%
class recall	39.55%	64.56%	

Exhibit 5

W-Logistic Training	true 0	true 1	class precision
pred. 0	999	260	79.35%
pred. 1	2940	1800	37.97%
class recall	25.36%	87.38%	

W-Logistic Validation	true 0	true 1	class precision
pred. 0	596	203	74.59%
pred. 1	1965	1236	38.61%
class recall	23.27%	85.89%	

Exhibit 6

Naïve Bayes Training	true 0	true 1	class precision
pred. 0	1491	485	75.46%
pred. 1	2448	1575	39.15%
class recall	37.85%	76.46%	

Naïve Bayes Validation	true 0	true 1	class precision
pred. 0	914	349	72.37%
pred. 1	1647	1090	39.82%
class recall	35.69%	75.75%	

Naïve Bayes (Kernel) Training	true 0	true 1	class precision
pred. 0	1826	573	76.12%
pred. 1	2113	1487	41.31%
class recall	46.36%	72.18%	

Naïve Bayes (Kernel) Validation	true 0	true 1	class precision
pred. 0	1063	473	69.21%
pred. 1	1498	966	39.20%
class recall	41.51%	67.13%	

Exhibit 7

Decision Tree Validation	true 0	true 1	class precision
pred. 0	1033	509	66.99%
pred. 1	1528	930	37.84%
class recall	40.34%	64.63%	

Decision Tree Training	true 0	true 1	class precision
pred. 0	1850	414	81.71%
pred. 1	2089	1646	44.07%
class recall	46.97%	79.90%	