DATA MINING & METHODOLOGY PROJECT REPORT

Mail Response Campaign

Mohamed Jawad [A0119964R]

Suneet Prakash [A0121859N]

Jahan Balasubramanian [A0119997E]

Gokul Krishnaa [A0120066W]

Contents

1 Introduction

A Mail-Order company has a product they would like to promote. They consider a campaign offering this product for sale, directed at a given customer base. Normally, about 1 % of the customer base will be " responders", customers who will purchase the product if it is offered to them. A mailing to a million randomly-chosen customers will therefore generate about 10,000 sales. A more efficient marketing is performed by using data mining techniques.

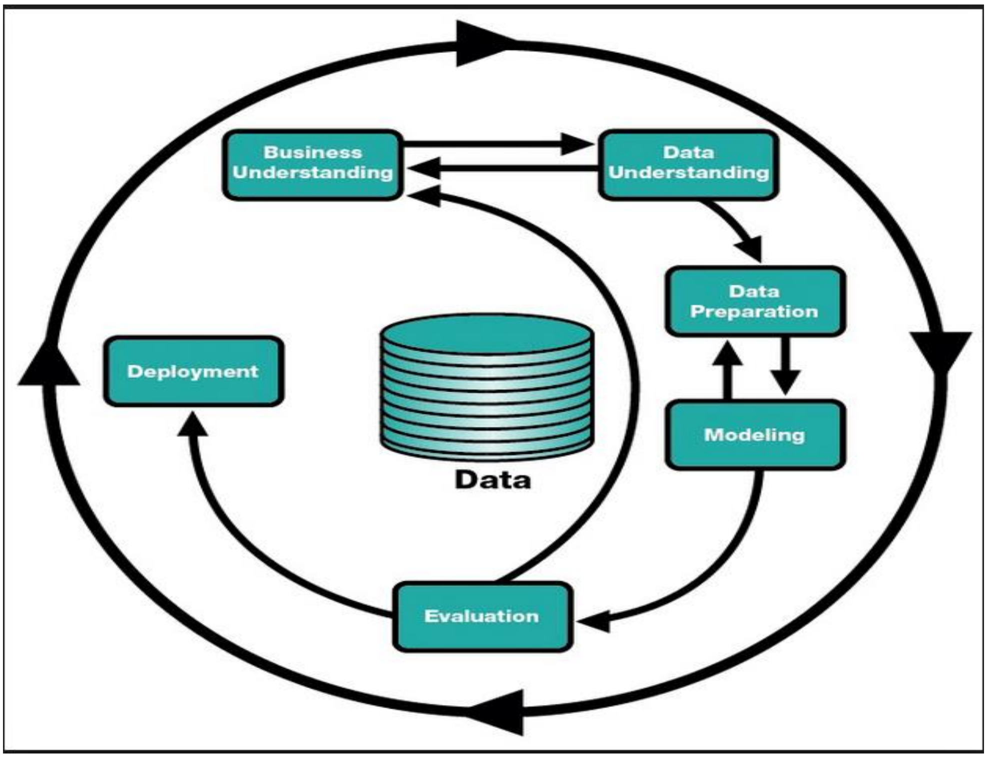
2 Objective Description:

The objective of the project is to use Data mining techniques to identify potential customers who would respond to the campaign. Since the customer base is large, an increase in the response rate from the earlier 1 % to 1.5% can reduce the cost of mailing by one third because then 1000 sales can be achieved by 666,667 mailings itself.

A secondary objective is to unearth useful patterns that might help the business to perform better.

3 Methodology applied:

We have used Cross-Industry standard Process (CRISP) methodology for data mining in order to achive the project objective.

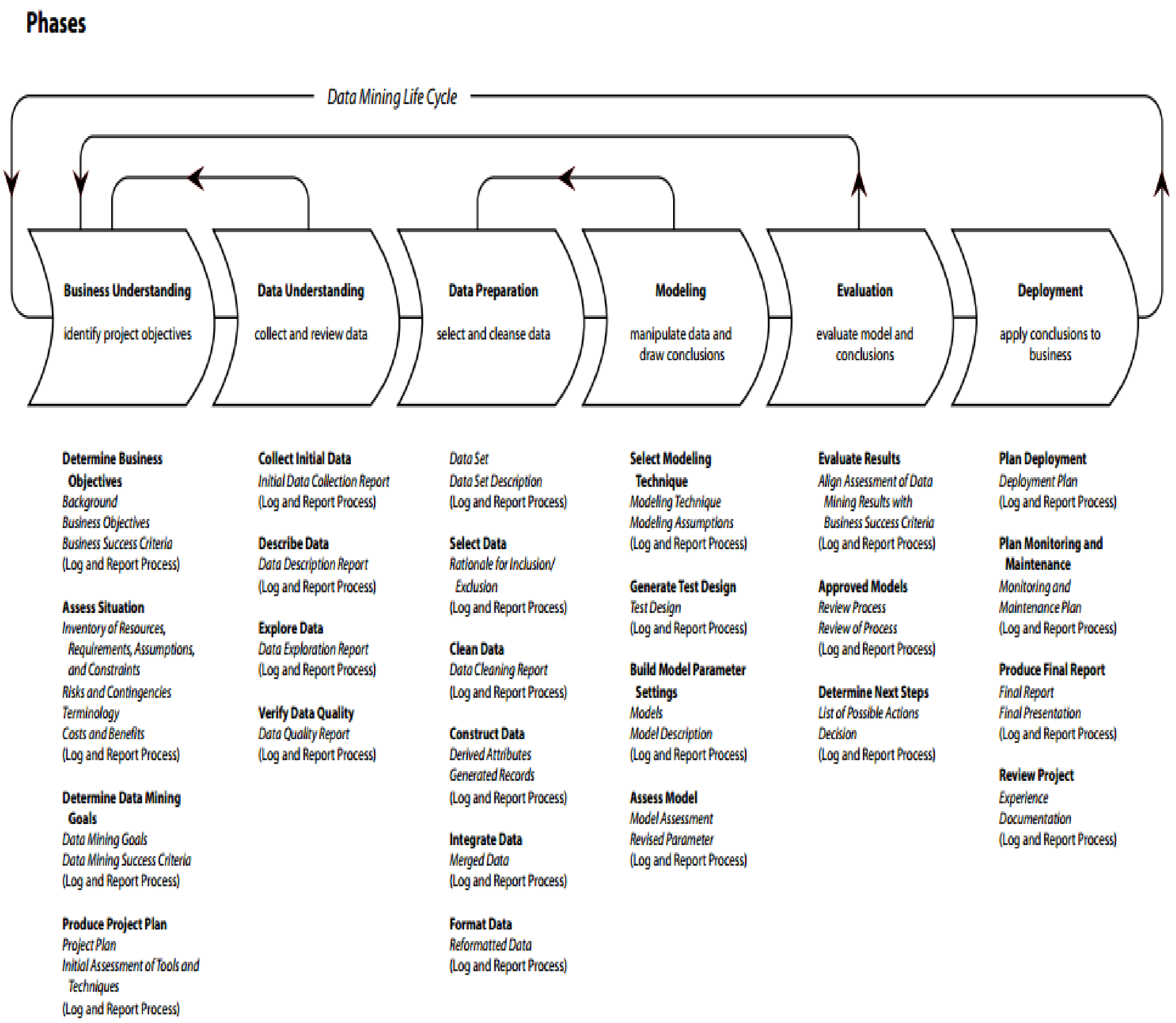


As we can see in the above diagram the CRISP methodology uses the following 6 steps:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

As shown in the above diagram the phases in CRISP methodology are iterative and are repeated multiple times depending on the evaluation result.

Please find below the detialed explanation of each phase of CRISP methodology.



4 Business Understanding

.1 Business Objective

The goal is to improve the response rate of customers in the mailing campaign conducted by the Mail-Order Company. Thus, reducing the cost of mailings to the customers by identifying potential customers who are likely to respond.

.2 Assessing the Situation

The data is available in the form of a file named project.csv. The file contains the list of the 1% responders (1079) together with 1079 randomly chosen non-responders. A total of 2158 cases are available. The software used for building the model is R and Rattle. The data is adequate with representation of all possibilities i.e customers who have responded as well as customers who have not responded to the mailing campaign. The quality of the data will be analyzed using the tools.

.3 Determine Data Mining Goals

The data mining goal is to improve the response rate of the customer by applying data mining techniques to identify the customers most likely to respond and target those customers. The goal is to increase the response rate to 1.5% from 1%, by doing so we can target 1000 customers by 66,666 mailings only. Therefore reducing the mail cost by one-third.

.4 Project Plan

Please find below the data mining project plan details.

Business requirement: 15%

Data Assessment: 25%

Data Preparation: 35%

Modelling: 15%

Model Evaluation: 10%

5 Data Understanding

We have collected the initial data from the csv file provided. The dataset is complete as it covers all the possible scenarios. We have got a set of 200 variables. The variable set were divided into multiple sections and each section were examined and analysed for their statistical properties.

The basic statistics were applied on the variables to find anomalies, outliners and noise. The relationship and correlations between various attributes were examined and the variables with high correlation values were deemed redundant and hence discarded.

The various visualization techniques such as Bar chart, Histogram and Box plot were applied to the initial dataset to formulate initial hypothesis .Based on the initial hypothesis, certain datasets were selected for further examination.

The data was further examined for its data quality. The attributes with noise and high level of skewness were dealt with the help of various transformation techniques like log and rescaling.

.1 Initial Data Collection and Data description:

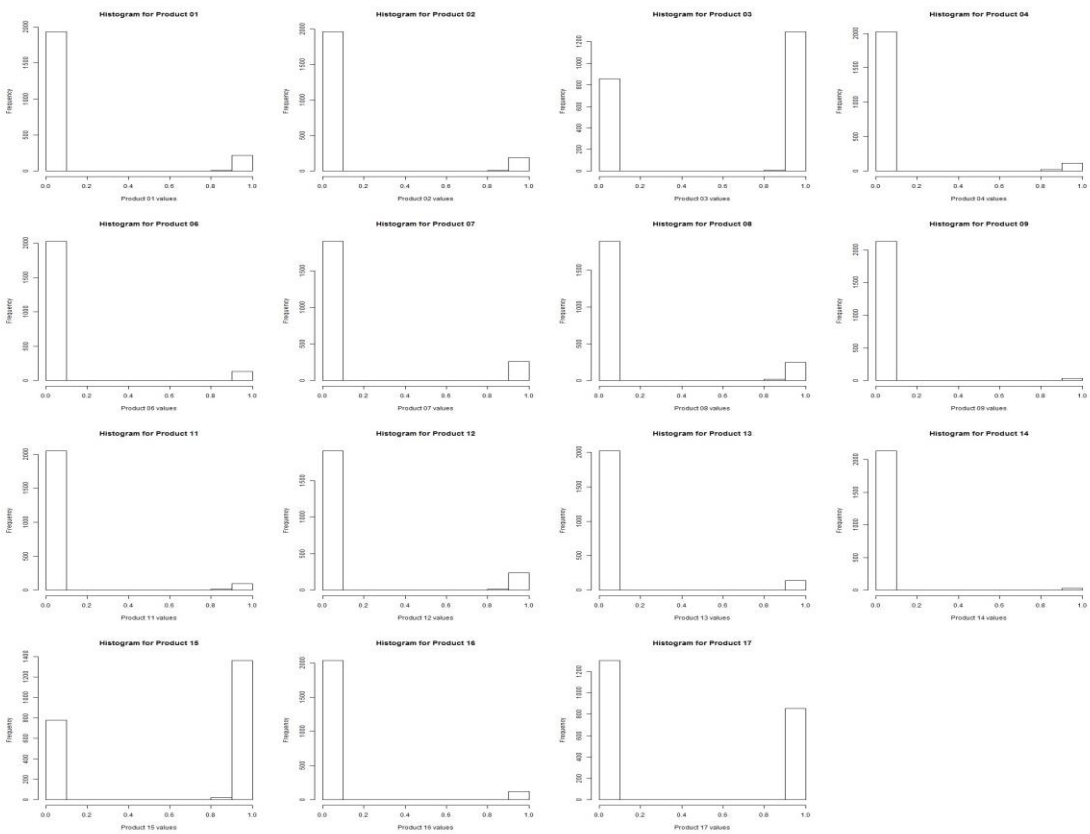
The data was collected from the csv file provided. The first 24 variables from V1 – V24 are the primary variables and are directly related to customer's profile. V35-V136 form the census variable, and v145-v199 are demographic "taxfiler" variables.

.2 Data Exploration

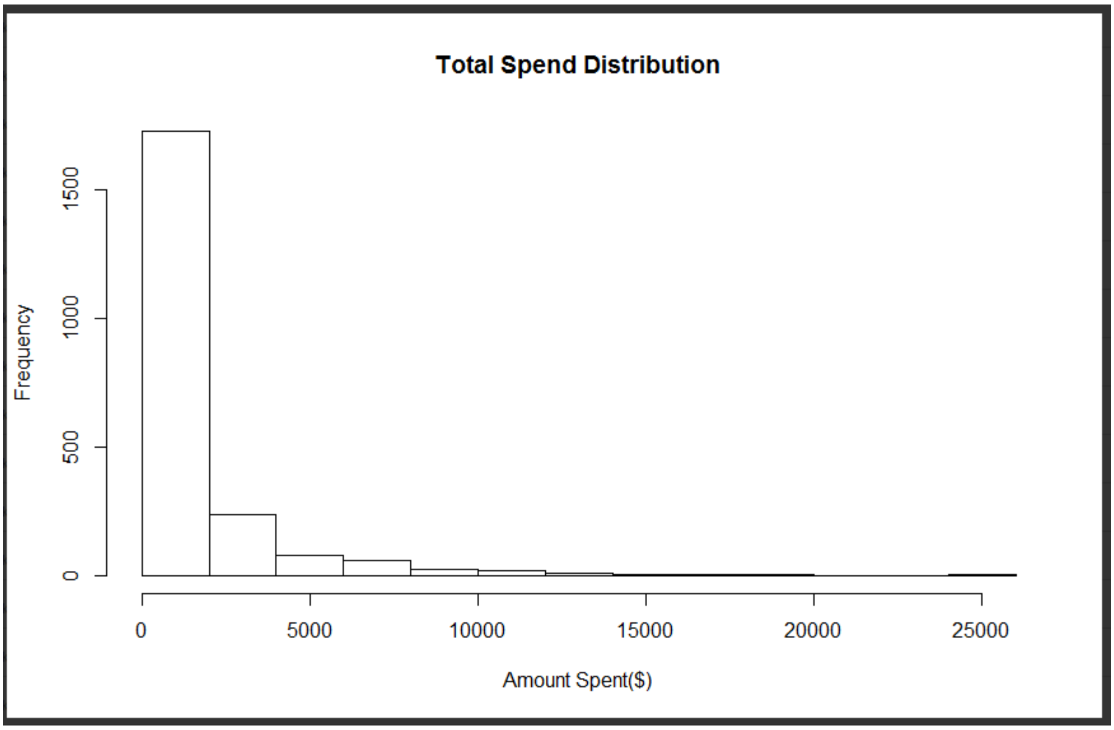
Primary Variables:

The variables v1 – v24 are directly related to the customer's profile, showing his spending on various products, transactions made by the customer and the recency with which he purchased the different products. Hence these are considered very important and informative. The following are the comments on these fields:

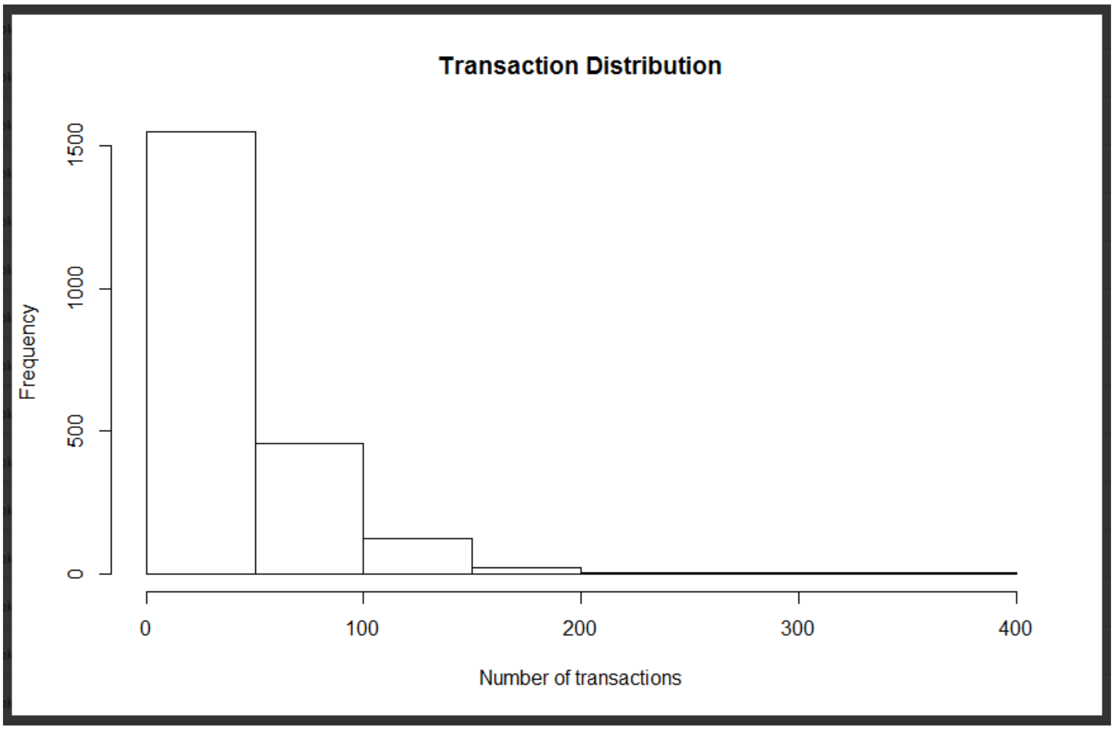
- Product recency for product 5 and 10 is missing in the given data set.
- The recency variables are positively skewed except product 3 recency and product 15 recency, which are negatively skewed.
- All the product recency variables are highly dichotomous (i.e. the data set is either grouped close to 0 or close to 1). Hence we are making it as flag variable and assigning 0 and 1 values.



- Further visualisation was performed on “Total spend” and “total transaction”.



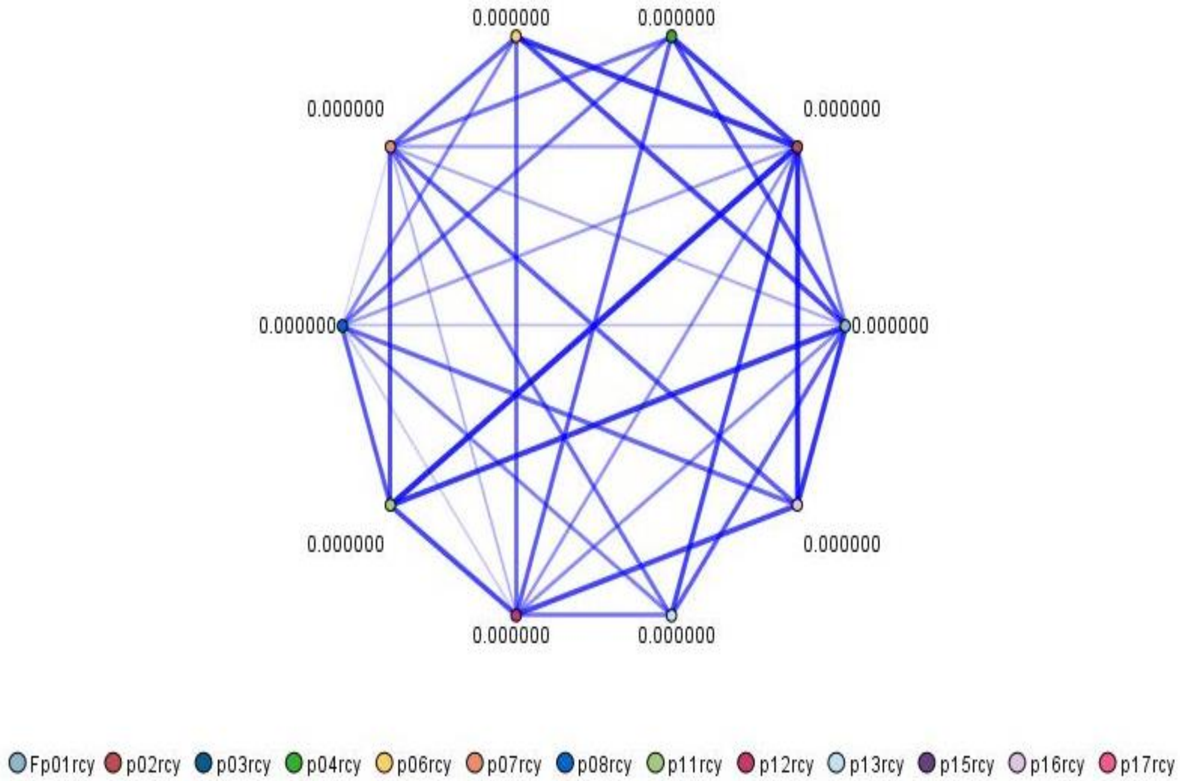
We have plotted histogram of total spend and found that the number of high spenders are very few as compared to the number of low spenders.



We have plotted above histogram of number of transaction .There are very few customers who are having high transactions. Thus from the above two histogram we can conclude that very few customers are willing to make high number of purchases or spending high sum of money.

.3 Link Analysis

We have performed the link analysis to determine the strength of relation between various products. Nodes represent the distinct products and the thickness of the lines indicates how many times they appear together.



The following tabular data illustrates the scores(confidence and support) for assocation for various products and objective. There are some interesting trends to observe

- Close to 40% of the time product 15 and 17 go together.
- Almost always when product 17 and 3 are purchased, we can expect to see product 15 also purchased.

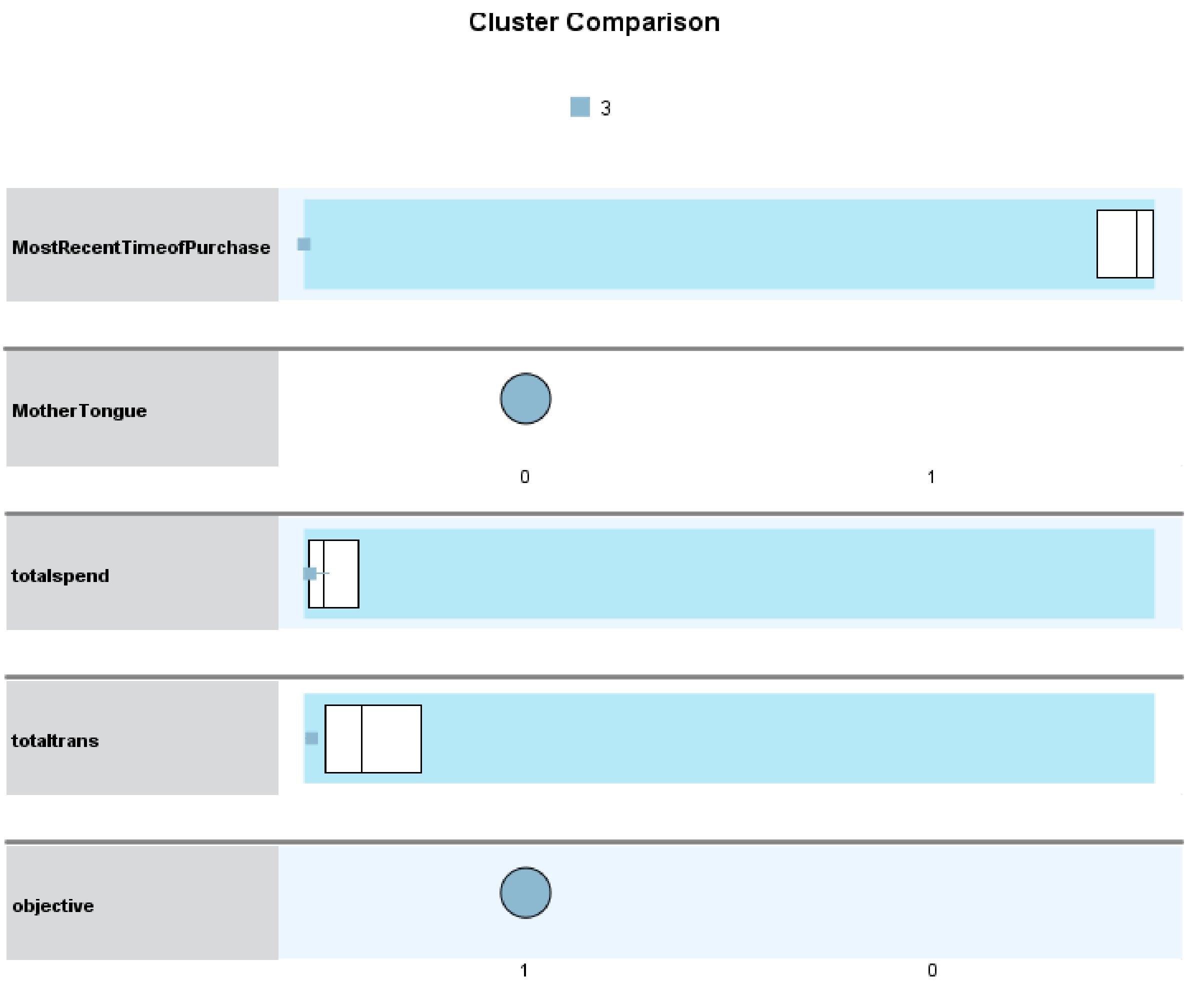
	Consequent	Antecedent	Support %	Confidence %
1	p15rcy	p17rcy and objective	24.467	99.811
2	p15rcy	p17rcy and p03rcy	23.262	99.801
3	p15rcy	p17rcy	39.666	99.766
4	p15rcy	p17rcy and objective and p03rcy	16.08	99.712
5	p03rcy	p08rcy and p15rcy	10.38	93.75
6	p03rcy	p08rcy	12.419	92.91
7	p15rcy	p12rcy	11.631	86.853
8	p15rcy	p08rcy and p03rcy	11.538	84.337
9	p15rcy	p08rcy	12.419	83.582

.4 Clustering

We employed clustering techniques to better understand the hidden and intangible patterns in the data. IBM SPSS Statistics tool was used in the clustering process. We performed a 2-step clustering and the following trends were observed.

Cluster Analysis I

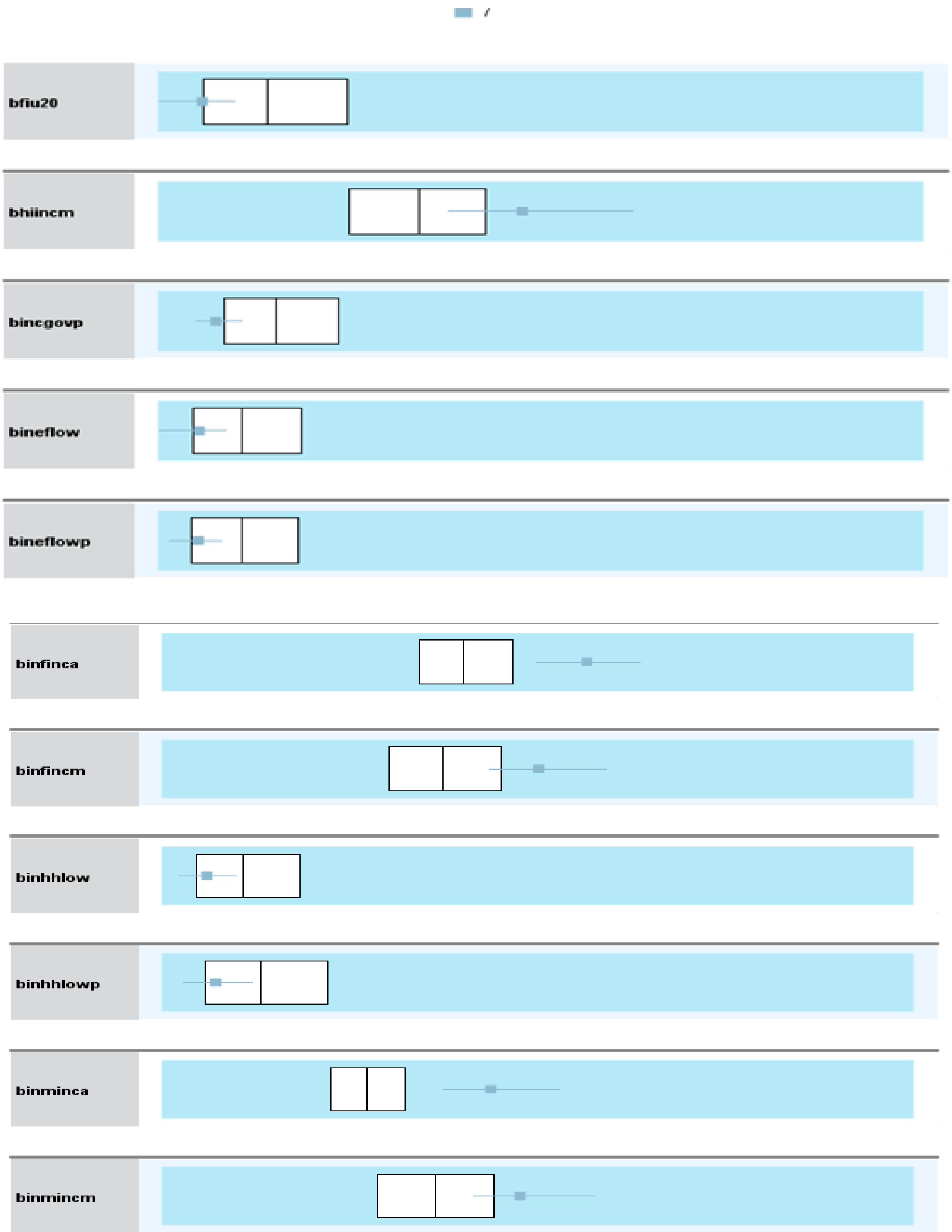
In the first analysis we investigated trends in the Recency, Frequency and Money spent, the so called RFM analysis. Most of what we observed were counter intuitive and could have potential applications in targeting customers. The variables used in this clustering approach are Most Recent time of purchase, Total Spend, and Total Transactions. What we observed was that when the recency, frequency and money spent are low, we would expect that the customer is more likely to respond to our campaign, and when the parameters are high they are unlikely to respond.



Cluster Analysis I-Fig

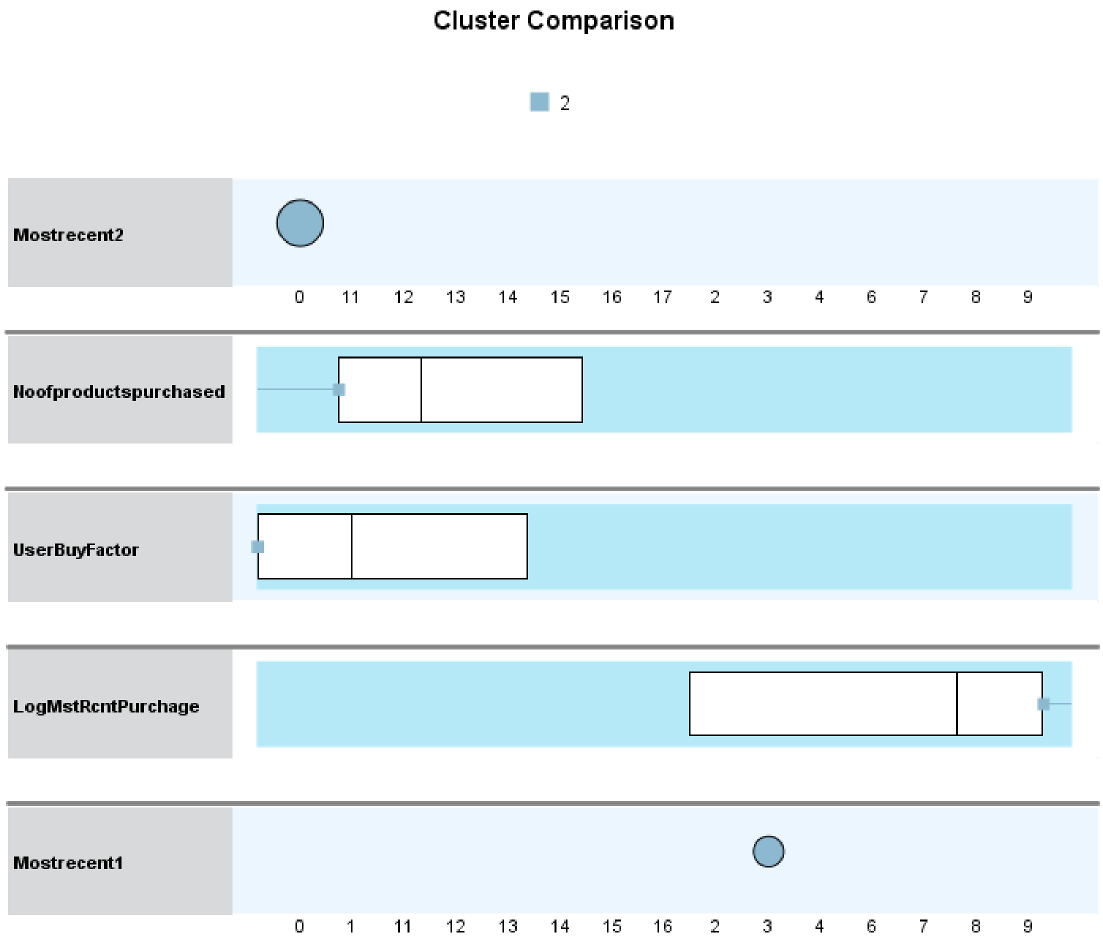
Cluster Analysis II

The second analysis tried to understand the socio economic background of the locality of potential responders. We find that people from areas that have more families under 20000 family income, more low income families, low average income of males and females in the age bracket of 15+ are more responsive to mail campaigns. The opposite of this is also true. That is to say people from areas where the above variables are high are not very likely to respond to our campaign.



Cluster Analysis III

In the next analysis, we try to figure out relationship between the number of products customer has purchased, the time of last two purchases and their potential to respond to campaigns. When number of products purchased is less, and he has only purchased one product, he is likely to respond.



6 Data Preparation

.1 Data Transformation

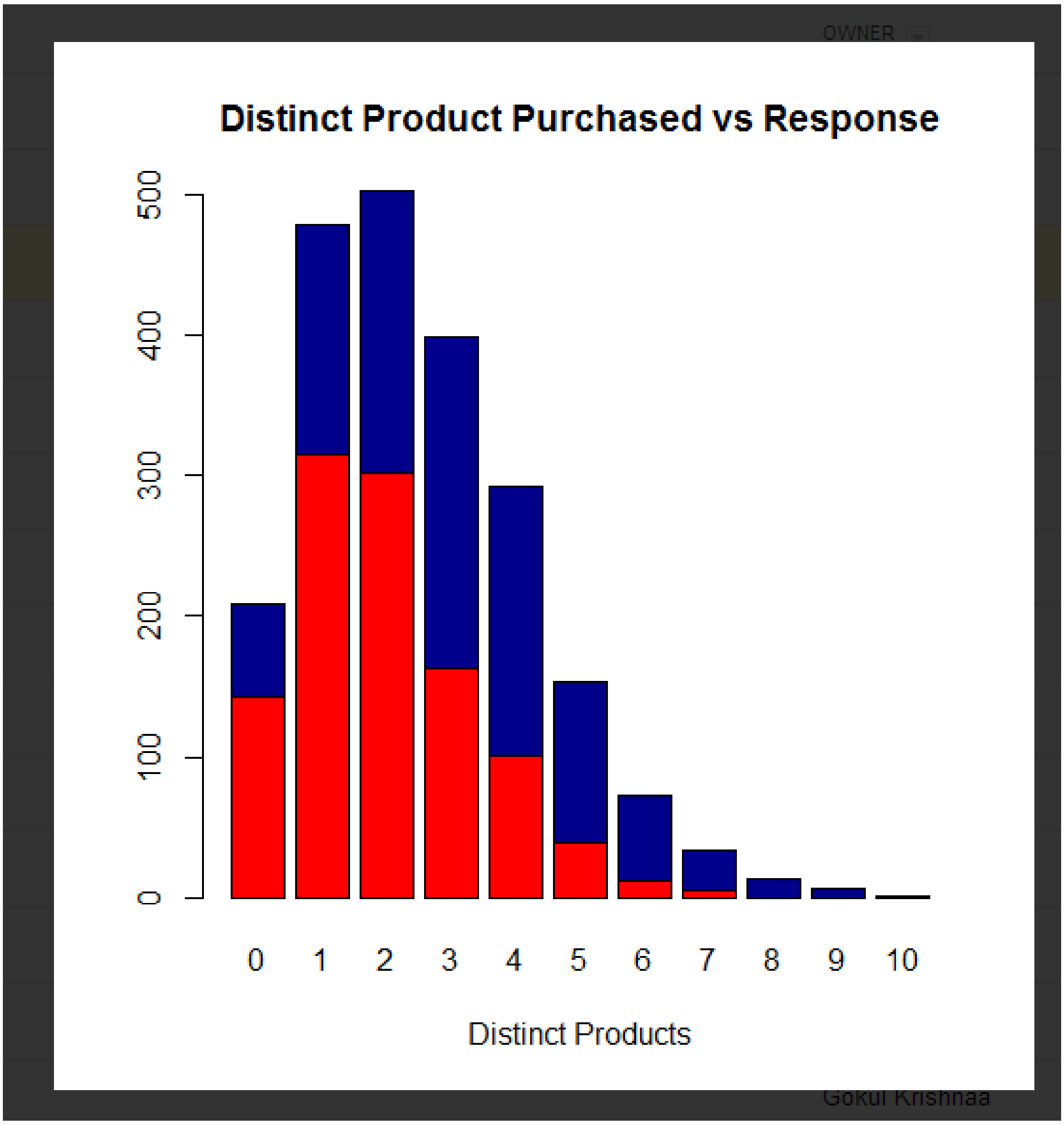
Derived Variables

Please find below the list of derived variable and there transformation rule.

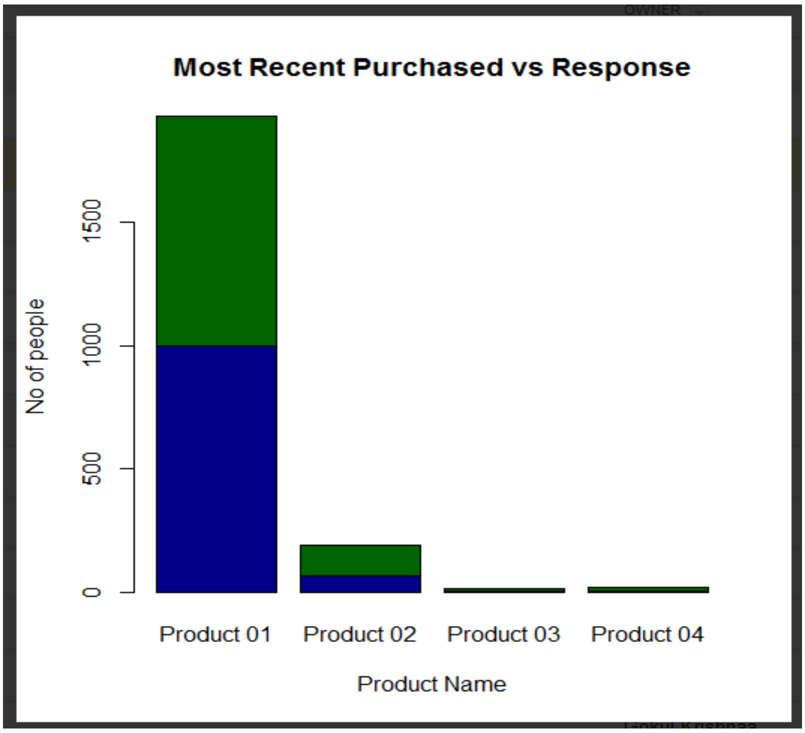
Variable after transformation	Formula for transformation
Number of distinct products purchased	Sum of number of product recency withy non zero value for a given customer
Most recent purchase	The lowest non zero value of all the product recency
Average Spend	Total spending/Total Transactions (Indicates the potential profit the customer may generate to the company)

The **number of distinct products purchased** histogram shows that most of the customer buy around 1 to 2 distinct product and very few customer buy multiple distinct products. The data distribution is highly skewed towards right.

The **number of distinct products purchased vs response** histogram shows that as the number of distinct product goes higher the probability of a customer to response to mail campaign also increases. We can therefore use this factor in order to predict customers likely to respond.



The below histogram between most recent purchased product and response does not provide any valuable information regarding the likelyhood of a customer to respond.



Transforming Numerical variables to Categorical variables:

We have used some of the numerical variables and converted them to categorical variables. This transformation will help us in reducing the variables while maintain the meaning of the overall data leading to much smaller and richer set attributes.

Numerical Variables	Categorical variable generated
tf46 Total earn T4 < $15,000 tf47 Total earn T4 $15,000 to $24,999 tf48 Total earn T4 $25,000 to $44,999 tf49 Total earn T4 $45,00 and over	Total earn
amtenglish and amtfrench	Mother tounge
betmulti and betsingle	Ethnic originSngMul
bmps and bfps	PstSecondryQualificationMF
bfsllabf and bfslplabf	LnPrntLbrFrcMemONonMbr
bimprovres and bimuk	Immigrants
binf30plus and binf7to15	IncFem15Plus
bknenglish and bknfren	KnwOffLng (knowledge of official language)

"Total taxfilers with interest & other investment income from $1 to $499","Total taxfilers with interest & other investment income from $500 to $4999","Total taxfilers with interest & other investment income from $5000 to $9999"&"Total taxfilers with interest & other investment income from $10000 and up"	**IntrInvstInc**
bndtbusser and bndtgovser	**SerIndBussGovn**
bdwmaint ,bdwmajor and bdwminor	**DwellRepMaint**
betbritish ,betenglish and betfrench	**EthOrgn(Ethnic Origin)**
bfi20to35,bfi50plus and bfiinca	**FamInc**
bhi20to35,bhi50plus and bhiu20	**HHoldInc**
bhlenglish,bhlfrench and bhlnonoff	**HomLngEngFrhNonOff**
binm15to30,binm30plus and binm7to15	**Mal15PlsInc**
"Total who claimed tuition credit","Total who claimed charitable donations" and "Total who claimed move credit"	**ClaimTutDonMove**
tf28 The number of taxfilers with income from $1 to $14,999 tf29 The number of taxfilers with income from $15,000 to $24,999 tf30 The number of taxfilers with income from $25,000 to $34,999 tf31 The number of taxfilers with income from $35,000 to $44,999 tf32 The number of taxfilers with income from $45,000 to $54,999 tf33 The number of taxfilers with income from $55,000 to $74,999 tf34 The number of taxfilers with income from $75,000 to $99,999 tf35 The number of taxfilers with income from $100,000 to $149,999 tf36 The number of taxfilers with income from $150,000 and over	**TaxfilerInc**
bsl9to13nc ,bslg9 ,bslnunivc,bslunivdeg ,bslun ivnc ,bslunivnd	**EduQual**
brlanglic,brlcathol,brlprotest,brlrcathol and brlunited	**Religion**

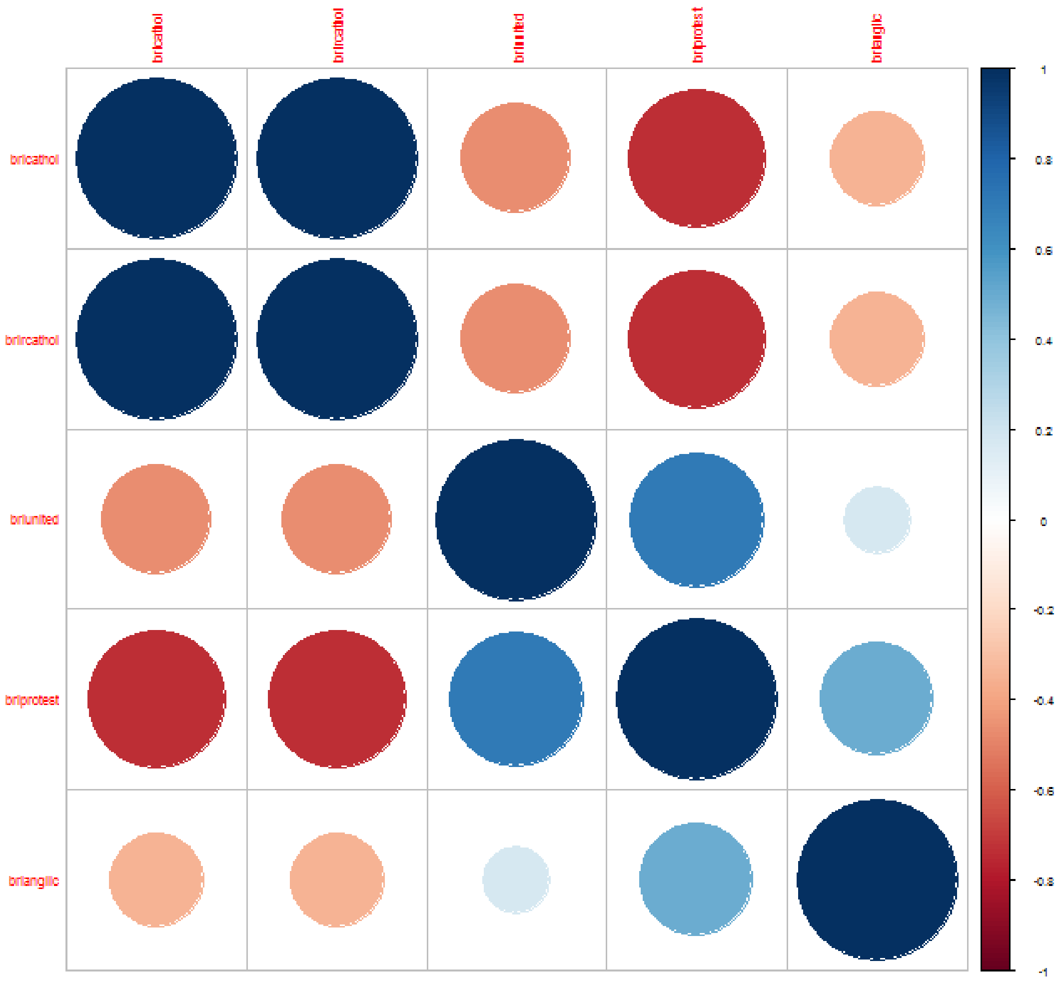
bocffabric,bocfmanage,bocfteach,bocmmanage and bocmscieng	**LineOfWrk**
bmoy1intep,bmoy5intep,bmoy5intrn,bmoy5mov and bmoy5non	**Mobility**

Variable Selection by Correlation

We are trying to capture the correlation between various input variables. Generally a high correlation means the variables are representing the same underlying meaning and therefore can be clubbed or reduced by some process.

Below are some observations about variables that are correlated.

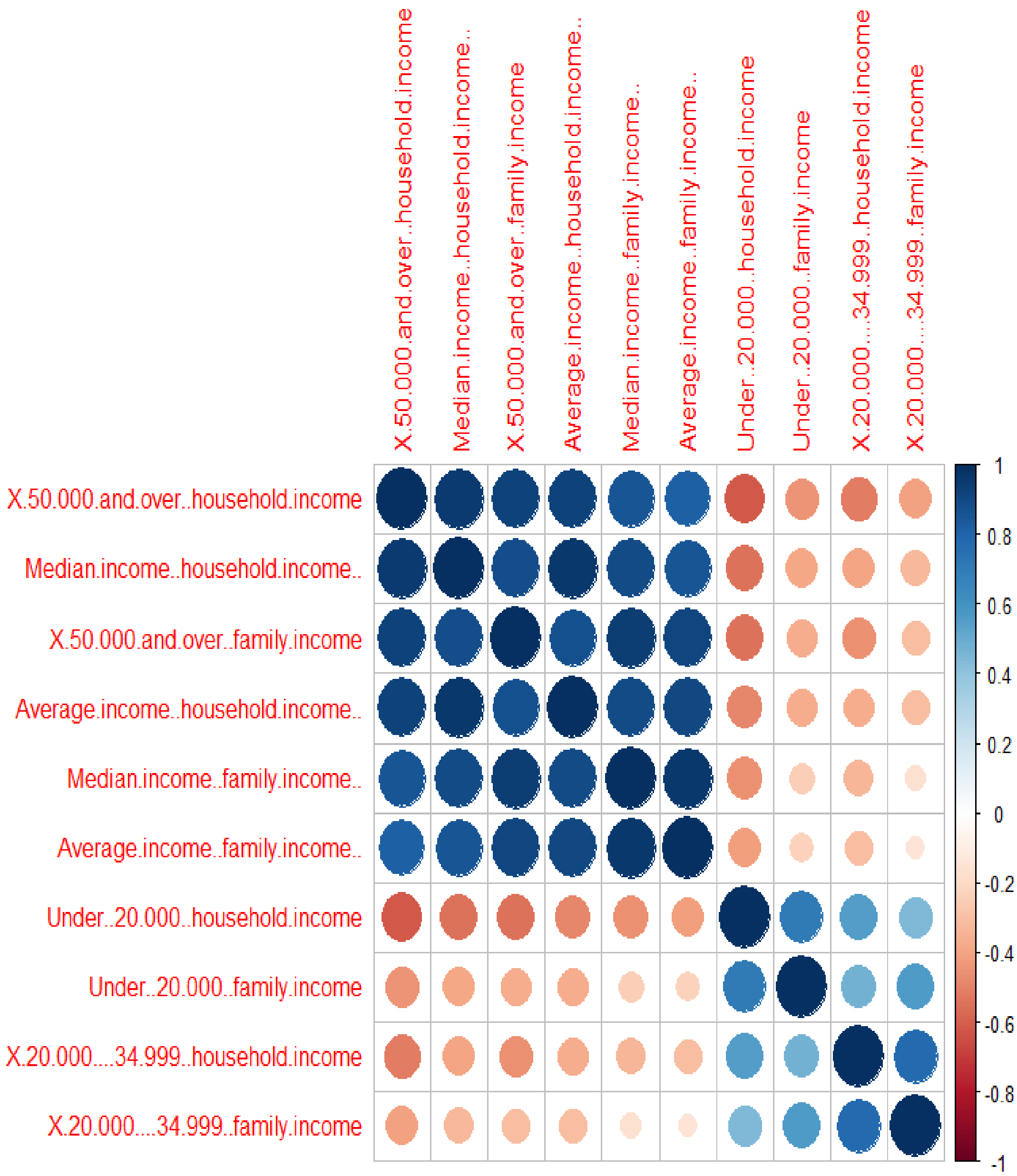
Religion Variables



We can see that Roman Catholic and religion catholic are highly correlated. Although a high correlation does not imply causality, in this case the relationship is clearly causal. Therefore we can remove the variable for Religion Catholic and retain the variable for Roman Catholic only.

Religion protestant is also highly correlated to United Church and Anglican. We can see that there is also a causality in the relationship. Hence by following a similar process we can keep variables for religion protestant and remove variable for United Church and Anglican.

Income Variables

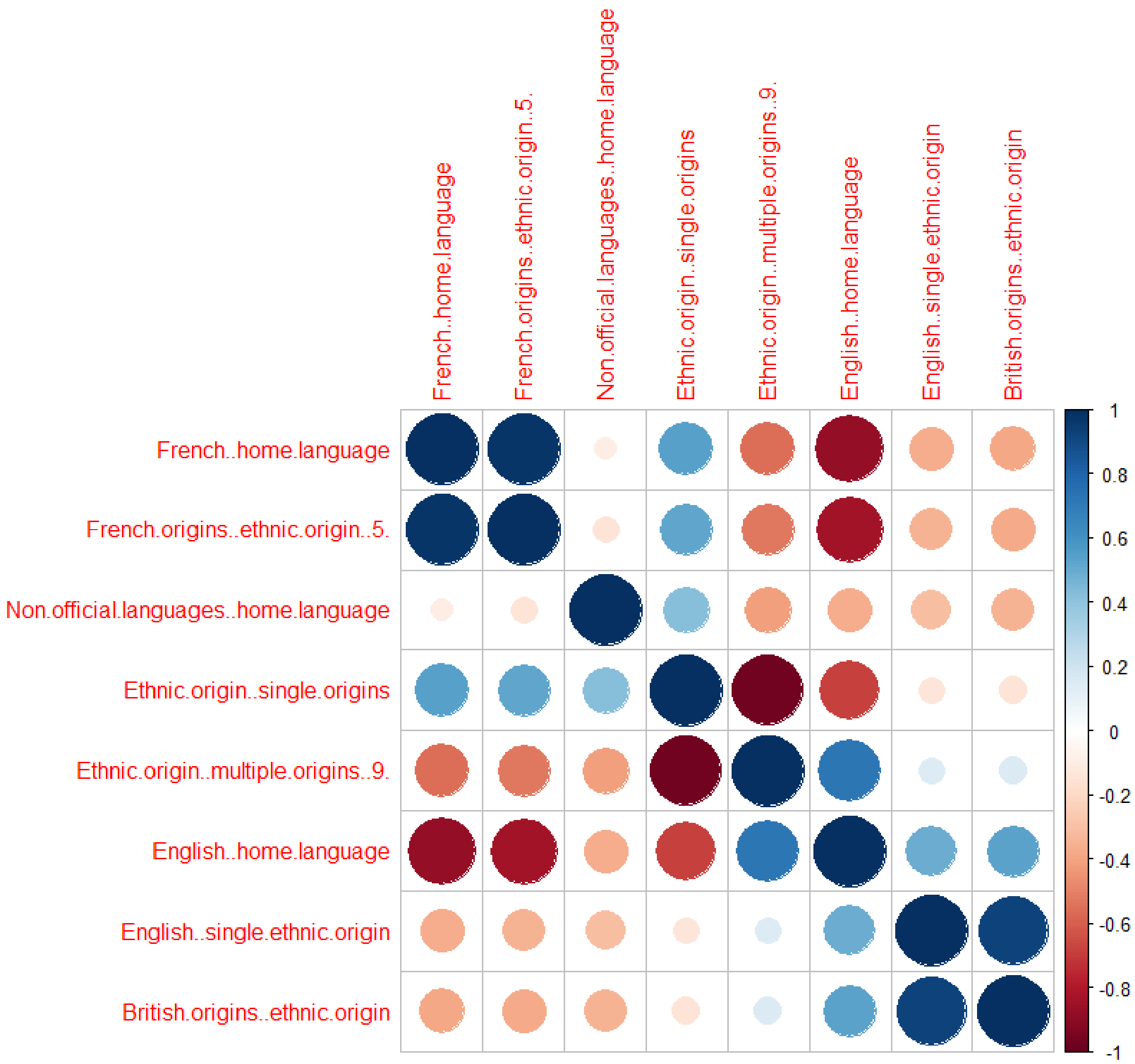


We notice from the above correlation diagram, that variable indicating "$20,000 - $34,999, family income" are strongly correlated with variables indicating "$20,000 - $34,999, household income". Also it seems logical to assume causality underlying their high correlation. Hence we will keep variables indicating "$20,000 - $34,999, household income" and remove the variable indicating "$20,000 - $34,999, family income".

Also another observation that is significant is the relation between variables for "Under $20,000, family income" and variables indicating "Under $20,000, household income". Hence we will keep "Under $20,000, household income" and remove "Under $20,000, family income". The reasoning for causality is the same as for the previous variables.

The most significant observation is the strong correlation between "$50,000 and over, household income", "$50,000 and over, family income", "Average income, household income $","Average income, family income $","Median income, family income $" and "Median income, household income $" among themselves. Hence we will keep "Average income, household income $" and remove the rest of them.

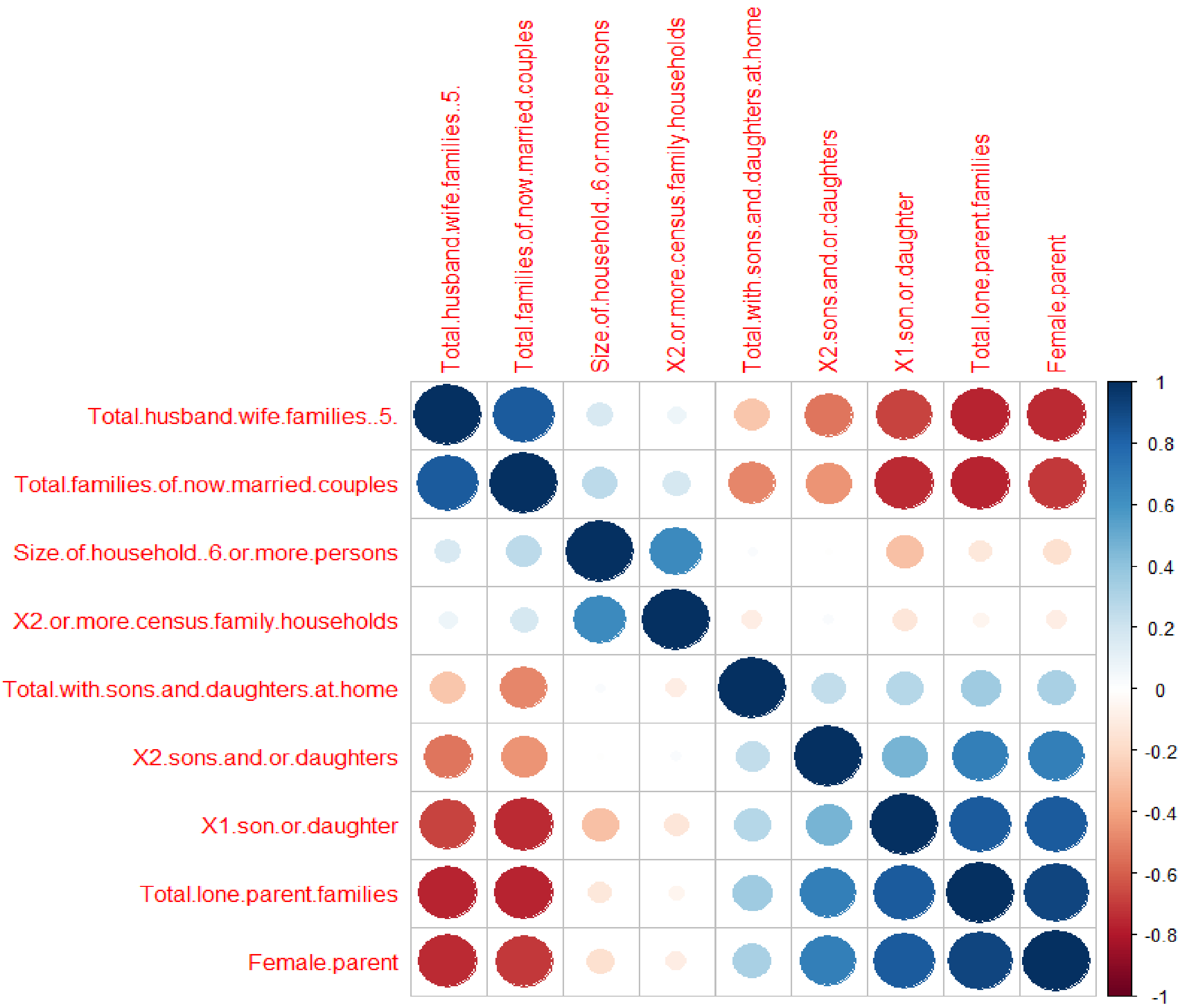
Ethnicity and Language Variables



The correlation matrix between language and ethnicity is very intuitive and mostly expected. We find that people having ethnic origin as France also have the mother tongue as French; similarly we also see the same trend with people who are ethnically British and have mother tongue as English. The causality here is implied and natural. Therefore we can keep "French origins, ethnic origin" and remove "amtfrench".

In the same spirit we carry the same process and remove " English, single ethnic origin " and retain “British origins, ethnic origin”. Another highly causal correlation is between "amtmultlin" and "amtnengnon". Hence we will keep "amtmultlin" and remove "amtnengnon".

Correlation between family types:



"Female parent" and "Total lone-parent families" are highly correlated. From this finding we can assume that most of the lone parent family consists of Single Mother. Hence we will keep "Total lone-parent families" and remove "Female parent".

"Total families of now-married couples" and "Total husband-wife families" are highly correlated. This finding is expected as husband and wife families are generally married couples. Therefore we will keep "Total families of now-married couples" and remove "Total husband-wife families".

"1 son or daughter" and "Total lone-parent families" are highly correlated. This is an interesting trend, as being a lone parent it becomes increasingly difficult to shoulder the burden of an additional child. Hence we will keep "Total lone-parent families" and remove "1 son or daughter".

Through this process of dimensionality reduction, we have taken care to see that the variability of the data set is not compromised. By applying correlation we have made sure that redundant attributes that are highly correlated are eliminated. Thereby reducing the inter correlation within the data set.

7 Modelling

The task of distinguishing respondents from non-respondents is by nature a classification problem. So the following techniques could be employed in pursuit of this task

1. Decision Tree (as implemented in rpart package using CART algorithm)
2. SVM (kernel function is Radial Basis function)
3. Random Forest
4. Ada Boosting
5. Logistic Regression
6. Neural Nets (10 hidden layers)
7. Ensemble(voting)

.1 Data Partitioning and Evaluation Strategy

We adopted the standard partitioning thumb rule and used 70% of data set for training purposes, 15% for validation and last 15% for testing.

Expected accuracy for the models is in the range of 70% to 75%.

To get a more realistic estimate of accuracy of the models, the evaluation was done on test data.

We have taken a very systematic approach to evaluation by considering various aspects of the models to quantify the accuracy of models.

1. Error Matrix
2. ROC curve
3. Lift Chart

.2 Modelling Process – Approach I

Input to the model:

Our dimensionally reduced data set consists of primary variables and principal components selected from the secondary variables. There are 15 primary components that account for maximum variability of the secondary variables. The principal components we obtained by performing PCA on the secondary variables that include census variables and taxfiler variables. We decided to settle for 15 principal components as the scree plot for the data set appears to flat out after 10 variables. We have also included some variables derived from the original data set. These derived variables are highly correlated with the response/objective.

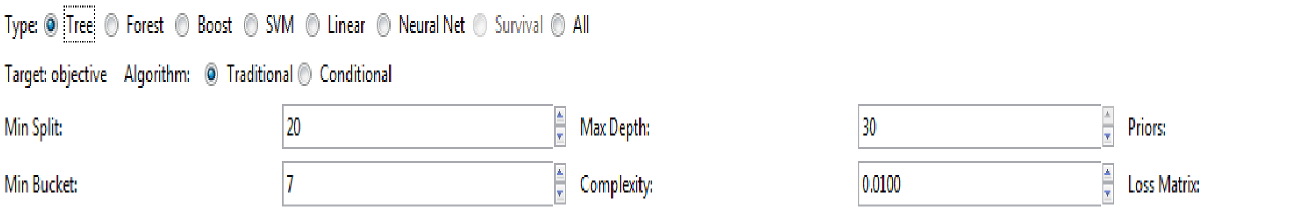
Please find below the snapshot of the dataset used for modelling

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	objective	Numeric	○	●	○	○	○	○	Unique: 2
2	p01rcy	Categoric	●	○	○	○	○	○	Unique: 2
3	p02rcy	Categoric	●	○	○	○	○	○	Unique: 2
4	p03rcy	Categoric	●	○	○	○	○	○	Unique: 2
5	p04rcy	Categoric	●	○	○	○	○	○	Unique: 2
6	totalspend	Numeric	●	○	○	○	○	○	Unique: 1341
7	totaltrans	Numeric	●	○	○	○	○	○	Unique: 171
8	p05spend	Numeric	●	○	○	○	○	○	Unique: 65
9	p05trans	Numeric	●	○	○	○	○	○	Unique: 61
10	p06rcy	Categoric	●	○	○	○	○	○	Unique: 2
11	p07rcy	Categoric	●	○	○	○	○	○	Unique: 2
12	p08rcy	Categoric	●	○	○	○	○	○	Unique: 2
13	p11rcy	Categoric	●	○	○	○	○	○	Unique: 2
14	p12rcy	Categoric	●	○	○	○	○	○	Unique: 2
15	p13rcy	Categoric	●	○	○	○	○	○	Unique: 2
16	p15rcy	Categoric	●	○	○	○	○	○	Unique: 2
17	p16spend	Numeric	●	○	○	○	○	○	Unique: 123
18	p16rcy	Numeric	●	○	○	○	○	○	Unique: 9
19	p16tenure	Numeric	●	○	○	○	○	○	Unique: 67
20	p16trans	Numeric	●	○	○	○	○	○	Unique: 48
21	p17rcy	Categoric	●	○	○	○	○	○	Unique: 2
22	Spend	Categoric	●	○	○	○	○	○	Unique: 11
23	Noofproductspurchased	Numeric	●	○	○	○	○	○	Unique: 11
24	UserBuyFactor	Numeric	●	○	○	○	○	○	Unique: 457
25	Gender	Categoric	●	○	○	○	○	○	Unique: 3
26	unempml25	Categoric	●	○	○	○	○	○	Unique: 3
27	LogMstRcntPurchage	Numeric	●	○	○	○	○	○	Unique: 173
28	lgTransSpend	Numeric	●	○	○	○	○	○	Unique: 1928
29	LonPar	Numeric	●	○	○	○	○	○	Unique: 1442
30	Mostrecent1	Categoric	●	○	○	○	○	○	Unique: 15
31	Mostrecent2	Categoric	●	○	○	○	○	○	Unique: 15
32	Mostrecent3	Categoric	●	○	○	○	○	○	Unique: 15
33	Comp.1	Numeric	●	○	○	○	○	○	Unique: 1832
34	Comp.2	Numeric	●	○	○	○	○	○	Unique: 1832
35	Comp.3	Numeric	●	○	○	○	○	○	Unique: 1832
36	Comp.4	Numeric	●	○	○	○	○	○	Unique: 1832
37	Comp.5	Numeric	●	○	○	○	○	○	Unique: 1832
38	Comp.6	Numeric	●	○	○	○	○	○	Unique: 1832
39	Comp.7	Numeric	●	○	○	○	○	○	Unique: 1832
40	Comp.8	Numeric	●	○	○	○	○	○	Unique: 1832
41	Comp.9	Numeric	●	○	○	○	○	○	Unique: 1832
42	Comp.10	Numeric	●	○	○	○	○	○	Unique: 1832
43	Comp.11	Numeric	●	○	○	○	○	○	Unique: 1832
44	Comp.12	Numeric	●	○	○	○	○	○	Unique: 1832
45	Comp.13	Numeric	●	○	○	○	○	○	Unique: 1832
46	Comp.14	Numeric	●	○	○	○	○	○	Unique: 1832
47	Comp.15	Numeric	●	○	○	○	○	○	Unique: 1832

Modelling Techniques

Decision Tree

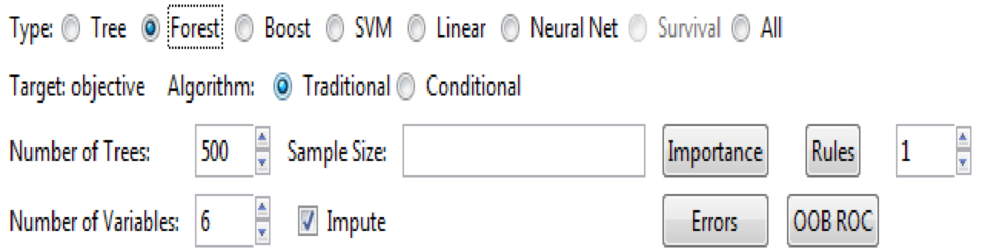
Parameter Initialization and the resulting tree nodes.




```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 1509

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 1509 751 1 (0.4976806 0.5023194)
   2) Gender>=2.5 490  92 0 (0.8122449 0.1877551) *
   3) Gender< 2.5 1019 353 1 (0.3464181 0.6535819)
     6) UserBuyFactor< 2.006503 489 229 1 (0.4683027 0.5316973)
      12) LonPar>=0.2094415 53  13 0 (0.7547170 0.2452830) *
      13) LonPar< 0.2094415 436 189 1 (0.4334862 0.5665138)
        26) Comp.3< -9843.428 50  16 0 (0.6800000 0.3200000) *
        27) Comp.3>=-9843.428 386 155 1 (0.4015544 0.5984456)
          54) Comp.12< -50.22551 131  60 0 (0.5419847 0.4580153) *
          55) Comp.12>=-50.22551 255  84 1 (0.3294118 0.6705882) *
     7) UserBuyFactor>=2.006503 530 124 1 (0.2339623 0.7660377) *

Classification tree:
rpart(formula = objective ~ ., data = crs$dataset[crs$train,
    c(crs$input, crs$target)], method = "class", parms = list(split = "information"),
    control = rpart.control(usesurrogate = 0, maxsurrogate = 0))

Variables actually used in tree construction:
[1] Comp.12       Comp.3        Gender        LonPar        UserBuyFactor

Root node error: 751/1509 = 0.49768

n= 1509

        CP nsplit rel error  xerror     xstd
1 0.407457      0   1.00000 1.06258 0.025820
2 0.017976      1   0.59254 0.59254 0.023587
3 0.014647      4   0.53262 0.59254 0.023587
4 0.010000      5   0.51798 0.58056 0.023445

Time taken: 0.34 secs
```

Random Forest

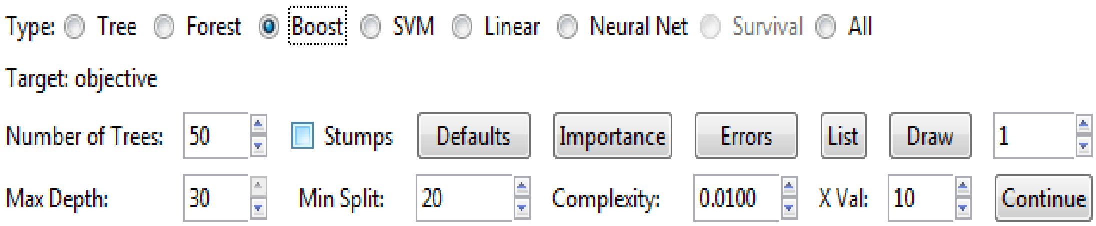
Parameter initialization and result summary




```
Summary of the Random Forest Model
==================================

Number of observations used to build the model: 1509
Missing value imputation is active.

Call:
 randomForest(formula = as.factor(objective) ~ .,
              data = crs$dataset[crs$sample, c(crs$input, crs$target)],
              ntree = 500, mtry = 6, importance = TRUE, replace = FALSE, na.action = na.roughfix)

               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 29.95%
Confusion matrix:
    0   1 class.error
0 506 245   0.3262317
1 207 551   0.2730871

Analysis of the Area Under the Curve (AUC)
==========================================

Call:
roc.default(response = crs$rf$y, predictor = as.numeric(crs$rf$predicted))

Data: as.numeric(crs$rf$predicted) in 751 controls (crs$rf$y 0) < 758 cases (crs$rf$y 1).
Area under the curve: 0.7003

95% CI: 0.6772-0.7234 (DeLong)
```

Ada Boost

Parameter Initialization and result




```
Summary of the Ada Boost model:

Call:
ada(objective ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)], control = rpart.control(maxdepth = 30, cp = 0.01,
    minsplit = 20, xval = 10), iter = 50)

Loss: exponential Method: discrete   Iteration: 50

Final Confusion Matrix for Data:
          Final Prediction
True value   0   1
         0 630 121
         1  82 676

Train Error: 0.135

Out-Of-Bag Error:  0.176  iteration= 50

Additional Estimates of number of iterations:

train.err1 train.kap1
        50         50

Variables actually used in tree construction:
 [1] "Comp.1"          "Comp.10"         "Comp.11"
 [4] "Comp.12"         "Comp.13"         "Comp.14"
 [7] "Comp.15"         "Comp.2"          "Comp.3"
[10] "Comp.4"          "Comp.5"          "Comp.6"
[13] "Comp.7"          "Comp.8"          "Comp.9"
[16] "Gender"          "lgTransSpend"    "LogMstRcntPurchage"
[19] "LonPar"          "Mostrecent1"     "Mostrecent2"
[22] "Mostrecent3"     "p01rcy"          "p02rcy"
[25] "p03rcy"          "p04rcy"          "p05spend"
[28] "p05trans"        "p06rcy"          "p07rcy"
[31] "p12rcy"          "p15rcy"          "p16spend"
[34] "p16tenure"       "Spend"           "totalspend"
[37] "totaltrans"      "unempml25"       "UserBuyFactor"
```

SVM

Parameter setting and result

```
Type: ( ) Tree ( ) Forest ( ) Boost (•) SVM ( ) Linear ( ) Neural Net ( ) Survival ( ) All

Target: objective

Kernel: Radial Basis (rbfdot)   Options:
```

```
Summary of the SVM model (built using ksvm):

Support Vector Machine object of class "ksvm"

SV type: C-svc  (classification)
 parameter : cost C = 1

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0173456635982391

Number of Support Vectors : 1151

Objective Function Value : -866.6719
Training error : 0.215374
Probability model included.

Time taken: 0.85 secs
```

Logistic Regression

Parameter setting and result

```
Type: ( ) Tree ( ) Forest ( ) Boost ( ) SVM (•) Linear ( ) Neural Net ( ) Survival ( ) All
( ) Numeric ( ) Generalized ( ) Poisson (•) Logistic ( ) Probit ( ) Multinomial
[Plot]
Summary of the Logistic Regression model (built using glm):
```

```
Call:
glm(formula = objective ~ ., family = binomial(link = "logit"),
    data = crs$dataset[crs$train, c(crs$input, crs$target)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4384  -0.8712   0.1427   0.9105   2.2875
```

Neural Network

Parameter setting and sample of result.

```
Type: ( ) Tree ( ) Forest ( ) Boost ( ) SVM ( ) Linear (•) Neural Net ( ) Survival ( ) All
Target: objective                                                                Model Builder: nnet (0/1)
Hidden Layer Nodes: 10

Summary of the Neural Net model (built using nnet):

A 47-10-1 network with 538 weights.
Inputs: p01rcy, p02rcy, p03rcy, p04rcy, totalspend, totaltrans, p05spend, p05trans, p06rcy, p07rcy, p08rcy, p11rcy, p12rcy, p13rcy, p15rcy, p16spend, p16rcy, p16tenure, p16trans, p17rcy, Spend, Noofproductspurchased, UserBuyFactor,
Gender, unempml250, unempml251, LogMstRcntPurchage, lgTransSpend, LonPar, Mostrecent1, Mostrecent2, Mostrecent3, Comp.1, Comp.2, Comp.3, Comp.4, Comp.5, Comp.6, Comp.7, Comp.8, Comp.9, Comp.10, Comp.11, Comp.12, Comp.13, Comp.14,
Comp.15.
Output: as.factor(objective).
Sum of Squares Residuals: 756.0012.

Neural Network build options: skip-layer connections; entropy fitting.

In the following table:
   b  represents the bias associated with a node
   h1 represents hidden layer node 1
   i1 represents input node 1 (i.e., input variable 1)
   o  represents the output node

Weights for node h1:
  b->h1  i1->h1  i2->h1  i3->h1  i4->h1  i5->h1  i6->h1  i7->h1  i8->h1  i9->h1
  -0.66    0.23    0.29   -0.31   -0.68   -0.36    0.27    0.23   -0.31   -0.18
i10->h1 i11->h1 i12->h1 i13->h1 i14->h1 i15->h1 i16->h1 i17->h1 i18->h1 i19->h1
   0.31   -0.02    0.29   -0.50    0.39    0.25   -0.16   -0.55   -0.52    0.25
i20->h1 i21->h1 i22->h1 i23->h1 i24->h1 i25->h1 i26->h1 i27->h1 i28->h1 i29->h1
  -0.65   -0.15   -0.03   -0.20    0.30   -0.16   -0.04    0.49    0.56    0.44
i30->h1 i31->h1 i32->h1 i33->h1 i34->h1 i35->h1 i36->h1 i37->h1 i38->h1 i39->h1
   0.41    0.51    0.38    0.22    0.47   -0.41    0.15   -0.22    0.46   -0.08
i40->h1 i41->h1 i42->h1 i43->h1 i44->h1 i45->h1 i46->h1 i47->h1
  -0.41    0.33   -0.54    0.56    0.59    0.64    0.13   -0.68
```

(Please zoom in to see the details-around 180%)

Ensemble Method

We have taken a voting approach to Ensemble method. As Data analyst we are interested in identifying and picking the best results provided by the model. We can see from our earlier stats that Random Forest is best performer of all the methods but the error can still be reduced with voting method.

In our scheme of voting method we have taken 6 models and we decide whether to send a mail promotions to the customer based on the vote of each models.

To begin with we decided to mail the user when 5 out of 6 models predicts a yes, and tried other variations, and concluded that when 3 out of 6 models predict a 'yes' the result was at its best. This turned out to be the robust model with least error rate and best False negative rate, (Which proves to be the most loss contributing factor for the Business), where you predict that the user would actually not respond to the campaign but he would have responded when a mail promotion was offered to him.

objective	rpart	ada	rf	ksvm	glm	nnet	Ensemble
0	1	1	1	1	1	1	1
0	1	1	0	0	1	0	1
0	0	0	0	0	0	1	0
1	1	1	1	1	1	1	1
0	0	0	0	0	0	1	0
0	0	0	1	0	0	1	0
1	0	0	0	0	0	0	0
0	1	1	1	1	1	0	1
1	0	0	0	0	0	0	0
0	0	1	1	1	0	1	1
0	1	1	1	1	1	0	1
1	0	1	1	0	0	1	1
1	1	1	1	1	1	1	1
1	1	1	1	0	0	1	1
0	1	0	0	1	1	1	1
1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1
0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0
1	1	1	1	1	1	0	1
0	0	0	0	0	0	1	0
1	1	1	1	1	1	1	1
0	1	1	1	0	1	0	1
0	1	0	0	0	0	1	0

.3 Modelling Process – Approach II

Here we consider a different way to construct the model where, the following steps have been performed.

1. Take the complete dataset
2. Process the data, where the variables are transformed, cleaned (as mentioned in the data preparation phase).
3. Secondary variables are categorised, and the total number of variables have been reduced to 130.
4. Variables have been reduced by performing the correlation and models have been constructed.

Data Selection

Correlation Analysis:

Variable	Pearson Correlation Coefficient	Variable	Pearson Correlation Coefficient
objective	1	bfiu20	-.143**
acffempar	-.107**	PstSecodryQualificatioMF	-0.027
acfhuswife	.087**	bfpshealth	.127**
acftotmar	.076**	bfpshuma	.089**
acfwchcom	-.104**	LPrtLbrFrcMemO1Mbr	-.050*
afem40to44	.080**	bfsmlabf	.075**
ahh6ppers	-.076**	HHoldIc_A	.117**
amtmultli	-.082**	bhiicm	.118**
MotherTogue	-.138**	bhiics	.081**
amtsigres	.088**	bhiu20	-.105**
afamrel	-.109**	HomLgEgFrhoOff	-.062**
bcwmpaid	.081**	Immigrants	-.075**
bdw46to60	-.090**	bicgovp	-.123**
bdw86to91	.100**	bieflow	-.150**
DwellRep0t	-.058**	bieflowp	-.149**
bdwperroom	-.118**	IcFem15Plus	.118**
EthicOrigiSgMul	.090**	bifica	.127**
EthicOrigiEgFr	-.054*	bificm	.108**
HholdIc	.114**	bifics	.100**
bfiicm	.114**	bihhlow	-.134**

Variable	Pearson Correlation Coefficient	Variable	Pearson Correlation Coefficient
Mal15PlsIc	.073**	uempml25	-.049*
bimfemia	.100**	LoPar	-.114**
bimica	.148**	productcout6	.283**
bimicm	.140**	teure	.149**
bimics	.079**	RRSPTaxFilersTot	-0.004
KwOffLg	0.026	ClaimTutDoMove	-0.005
blfmaempl	.093**	tf129	.086**
blfmauemr	-.075**	TaxfilerIc	.179**
blfmtuemp	-.084**	tf38	.207**
blftaempl	.079**	tf39	.199**
Mobility	-.134**	tf42	.089**
bmpscomm	.117**	EarValuelmh	.140**
bmpssocial	.095**	tf51	-.167**
bdtallid	.076**	tf52	-.112**
SerIdBussGov	.053*	tf55	-.125**
LieOfWrk	0.024	tf57	.101**
bpwmcsd	.092**	tf58	.115**
bpwmpop	.097**	tf68	.100**
bpwmusual	.111**	tf71	.155**
Religion	-.074**	tf72	.145**
tflow	-.205**	tf74	.089**
EduQual	.119**	tf75	.108**
lowincome	-.101**	tf76	.131**
highincome	.078**	tf77	.136**
productcout	.314**	tf89	.106**
tf93	.134**	tf92	.152**
tf95	.079**	tf90	.103**
tf96	.075**		

Correlation Analysis of Primary Variables

Variable	Pearson Correlation Coefficient
objective	1
p01rcy	.103**
p02rcy	.108**
p03rcy	.087**
p04rcy	.107**
totalspend	.175**
totaltrans	.136**
p05spend	-.080**
p05tras	-.076**
p06rcy	.088**
p07rcy	.111**
p08rcy	.141**
p11rcy	.093**
p12rcy	.136**
p13rcy	.086**
p15rcy	.208**
p16sped	.086**
p16rcy	.134**
p16tenure	.129**
p16tras	.101**
p17rcy	.190**
noofproductspurchased	.326**
UserBuyFactor	.328**
tflow	-.205**
Geder	-.356**
uempml25	-.049*
LogMstRctPurchage	-.185**
lgTrasSped	.123**
LoPar	-.114**
Mostrecet1	.068**
Mostrecet2	.201**
Mostrecet3	.212**

Key

	The highlighted variables are removed based on correlation.

After performing the correlation analysis, variables that have low correlation with the objective are removed and remaining predictors are taken as the model input.

Input to the Model

No.	Variable	Data Type	Input	Target	Risk	Ident
1	objective	Numeric	○	◉	○	○
2	p01rcy	Numeric	◉	○	○	○
3	p02rcy	Numeric	◉	○	○	○
4	p04rcy	Numeric	◉	○	○	○
5	totalspend	Categoric	◉	○	○	○
6	totaltrans	Numeric	◉	○	○	○
7	p05spend	Numeric	◉	○	○	○
8	p05trans	Numeric	◉	○	○	○
9	p07rcy	Numeric	◉	○	○	○
10	p08rcy	Numeric	◉	○	○	○
11	p12rcy	Numeric	◉	○	○	○
12	p15rcy	Numeric	◉	○	○	○
13	p16spend	Numeric	◉	○	○	○
14	p16rcy	Numeric	◉	○	○	○
15	p16tenure	Numeric	◉	○	○	○
16	p16trans	Numeric	◉	○	○	○
17	p17rcy	Numeric	◉	○	○	○
18	Spend	Numeric	◉	○	○	○
19	Noofproductspurchased	Numeric	◉	○	○	○
20	UserBuyFactor	Numeric	◉	○	○	○
21	acffempar	Numeric	◉	○	○	○
22	acfwchcom	Numeric	◉	○	○	○
23	MotherTongue	Numeric	◉	○	○	○
24	anfamrel	Numeric	◉	○	○	○
25	bdw86to91	Numeric	◉	○	○	○
26	bdwperroom	Numeric	◉	○	○	○
27	HholdInc	Numeric	◉	○	○	○
28	bfiincm	Categoric	◉	○	○	○

29	bfiu20	Numeric	◉	○	○	○
30	bfpshealth	Numeric	◉	○	○	○
31	HHoldInc_A	Numeric	◉	○	○	○
32	bhiincm	Categoric	◉	○	○	○
33	bhiu20	Numeric	◉	○	○	○
34	bincgovp	Numeric	◉	○	○	○
35	bineflow	Numeric	◉	○	○	○
36	bineflowp	Numeric	◉	○	○	○
37	IncFem15Plus	Numeric	◉	○	○	○
38	binfinca	Categoric	◉	○	○	○
39	binfincm	Categoric	◉	○	○	○
40	binfincs	Categoric	◉	○	○	○
41	binhhlow	Numeric	◉	○	○	○
42	binhhlowp	Numeric	◉	○	○	○
43	binminca	Categoric	◉	○	○	○
44	binmincm	Categoric	◉	○	○	○
45	blfmaunemr	Numeric	◉	○	○	○
46	Mo2ility	Numeric	◉	○	○	○
47	bmpscomm	Numeric	◉	○	○	○
48	bpwmusual	Numeric	◉	○	○	○
49	EduQual	Numeric	◉	○	○	○
50	lowincome	Numeric	◉	○	○	○
51	productcount	Numeric	◉	○	○	○
52	productcount6	Numeric	◉	○	○	○
53	tenure	Numeric	◉	○	○	○
54	TaxfilerInc	Numeric	◉	○	○	○
55	tf38	Numeric	◉	○	○	○
56	tf39	Numeric	◉	○	○	○
57	EarnValuelmh	Numeric	◉	○	○	○
58	tf51	Numeric	◉	○	○	○
59	tf52	Numeric	◉	○	○	○
60	tf55	Numeric	◉	○	○	○
61	IntrInvstInc	Categoric	◉	○	○	○
62	tf71	Numeric	◉	○	○	○
63	tf72	Numeric	◉	○	○	○
64	tf74	Numeric	◉	○	○	○
65	tf75	Numeric	◉	○	○	○
66	tf76	Numeric	◉	○	○	○
67	tf77	Numeric	◉	○	○	○
68	tf89	Numeric	◉	○	○	○
69	tf90	Numeric	◉	○	○	○
70	tf92	Numeric	◉	○	○	○
71	tf93	Numeric	◉	○	○	○
72	tf95	Categoric	◉	○	○	○
73	tf96	Categoric	◉	○	○	○
74	tflow	Numeric	◉	○	○	○
75	Gender	Numeric	◉	○	○	○
76	LogMstRcntPurchage	Numeric	◉	○	○	○
77	lgTransSpend	Numeric	◉	○	○	○
78	LonPar	Numeric	◉	○	○	○
79	Mostrecent2	Numeric	◉	○	○	○
80	Mostrecent3	Numeric	◉	○	○	○

Model Training

Using the predictors that were obtained, models are built and evaluated. The following models were tried

1. Decision Tree
2. SVM
3. Logistical regression
4. Neural networks

SVM is the only approach that gives accurate model for predicting the objective and here are results for various algorithms used for SVM.

SVM - Radial Basis approach

The kernel function used here is the Radial Basis function.

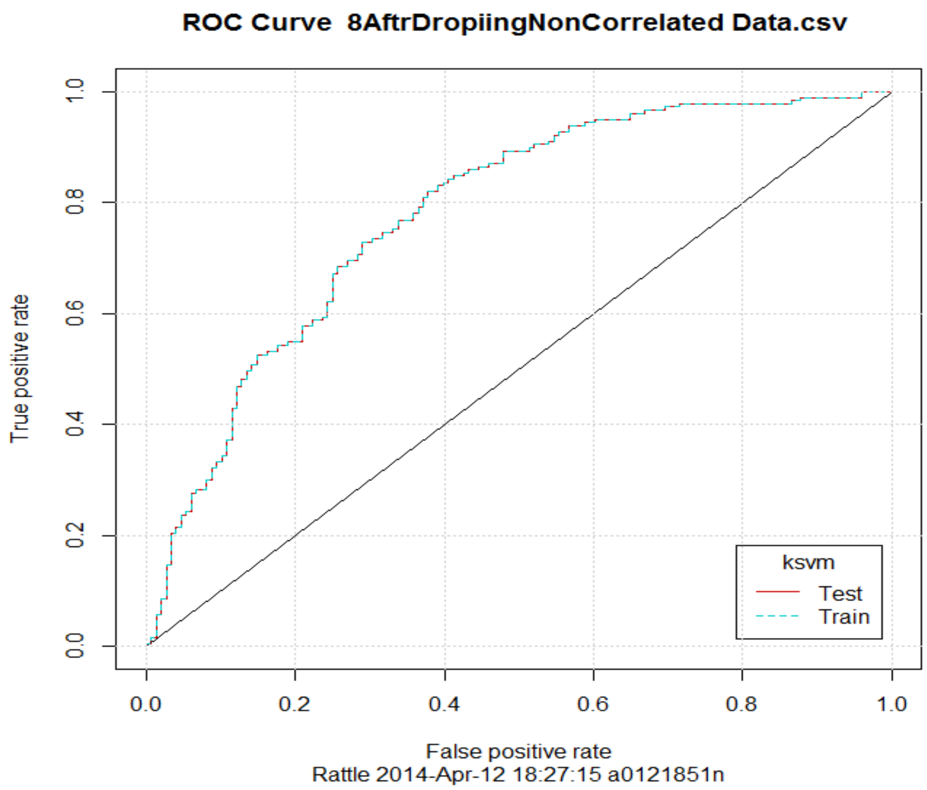
Error Matrix:

	Predicted		
Actual		0	1
	0	102	46
	1	47	130

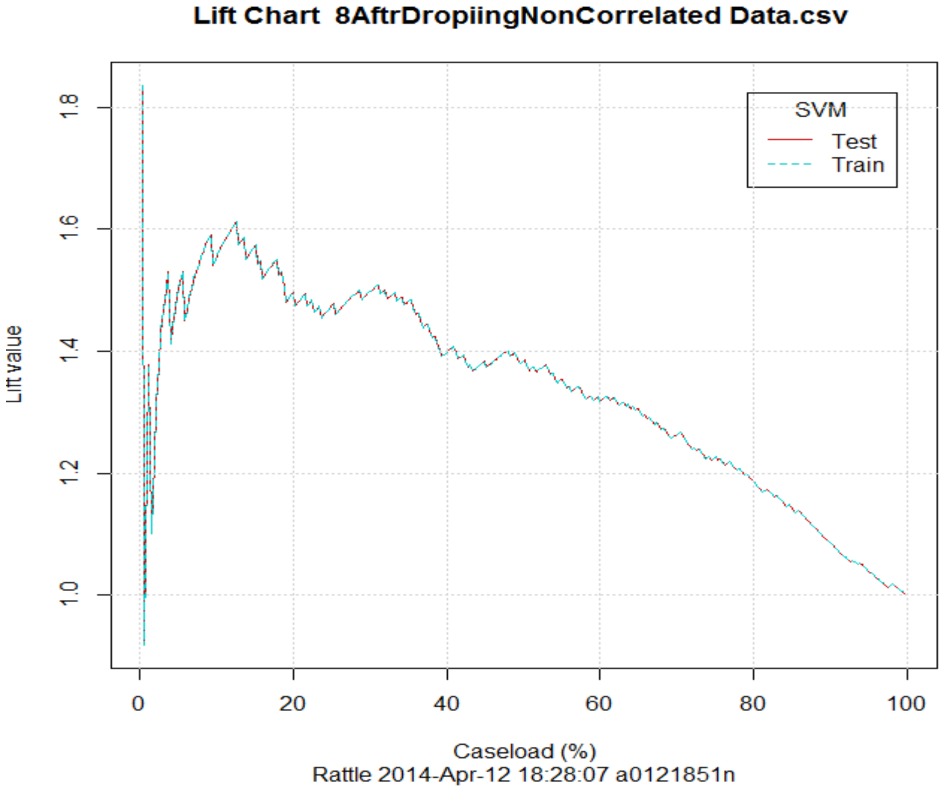
Error rate: 0.2861538

ROC:

Area under the curve: 0.7792



Lift Chart:



SVM - Laplacian method

In this approach we use Laplacian method as kernel function.

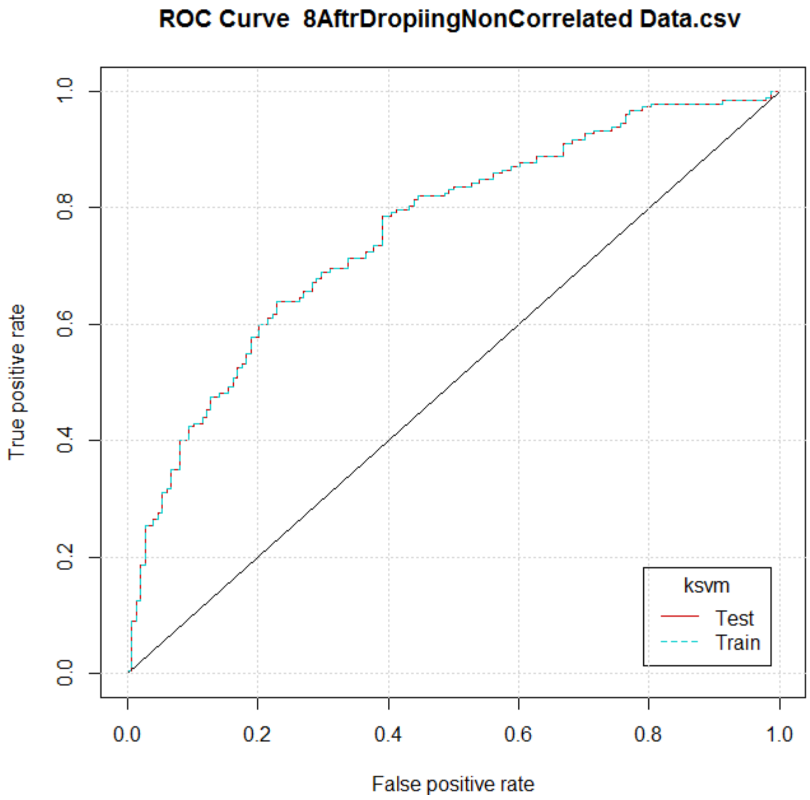
Error Matrix:

	Predicted		
Actual		0	1
	0	111	37
	1	64	113

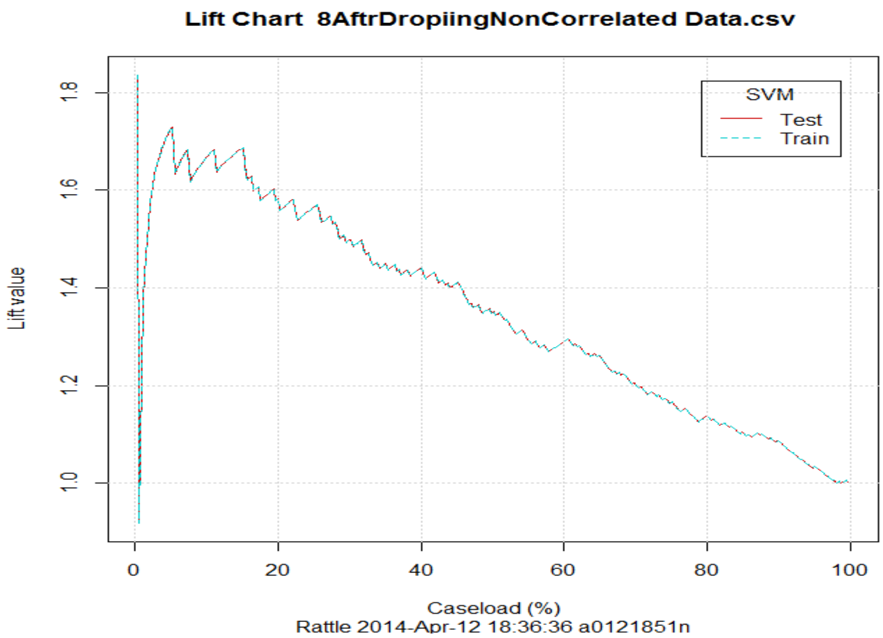
Error rate: 0.310769

ROC curve:

Area under the curve: 0.7569



Lift Chart:



SVM - Linear Dot method

The kernel method used here is Linear Dot method.

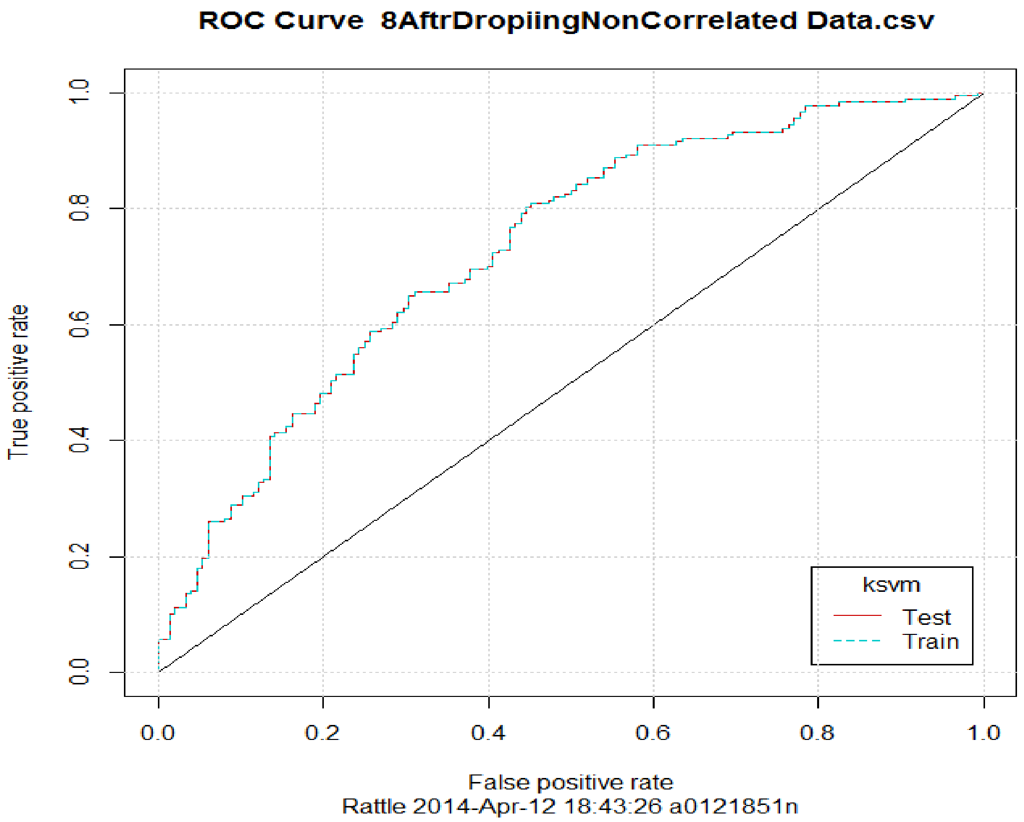
Error Matrix:

Actual	Predicted		
		0	1
	0	104	44
	1	67	110

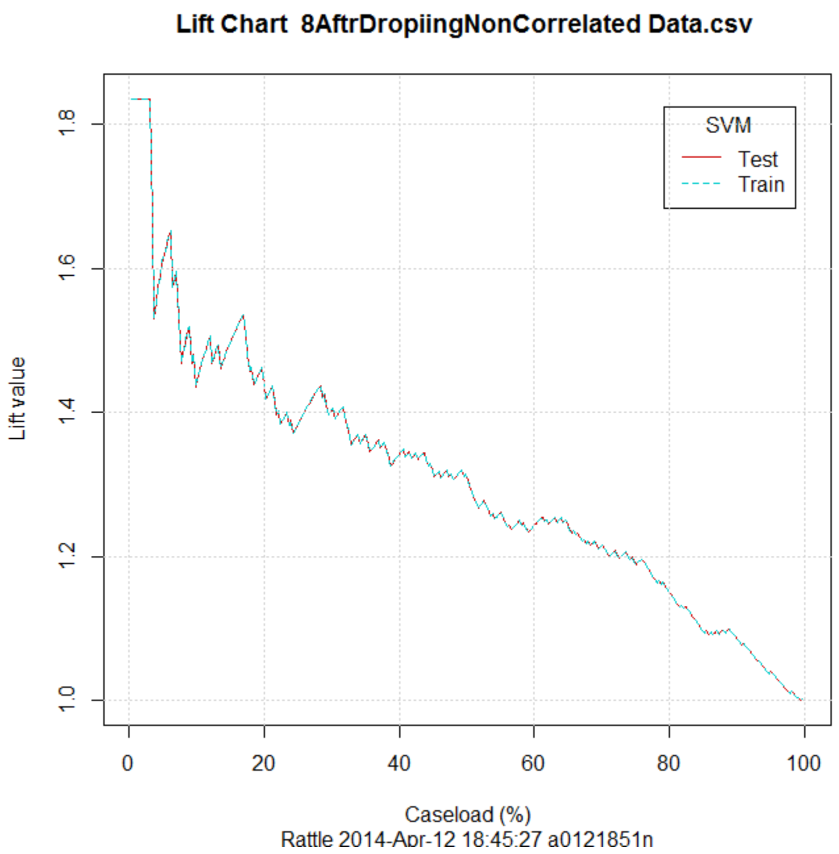
Error rate: 0.3415385

ROC curve:

Area under the curve is: 0.7288



Lift Chart:



From the above analysis, we can see that the Radial basis approach comes up with the least error rate in predicting the output, however the approach for the model is still not as good as the approach that is performed with the principal components that turns out better results with much less predictors.

8 Evaluation and Findings:

The business objective is to increase the response rate of customers. We have been able to increase the response rate. The reason for the performance is the data cleaning and transformation process which enabled us to understand relationships between the variables describing the customer's purchasing behaviour and taxfiler and census variables and thus improve the modelling.

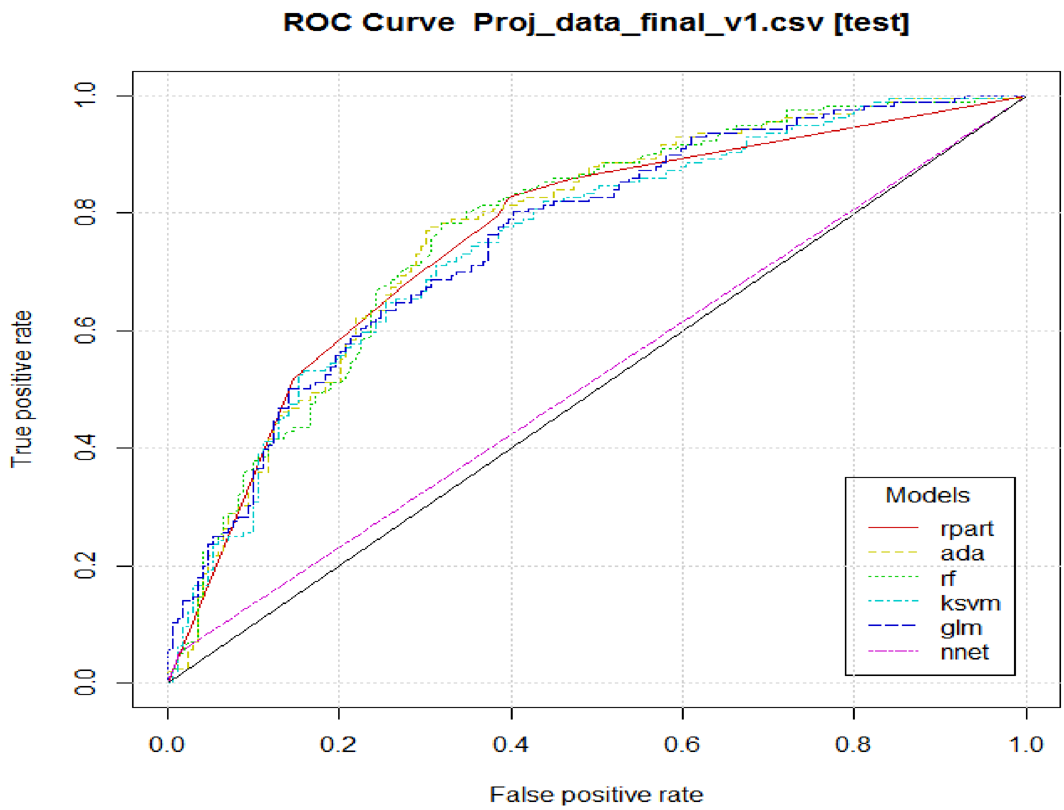
The target model accuracy is 70%-75% on the test data and we have achieved that with our accuracy of 72%.

Summary of Evaluation

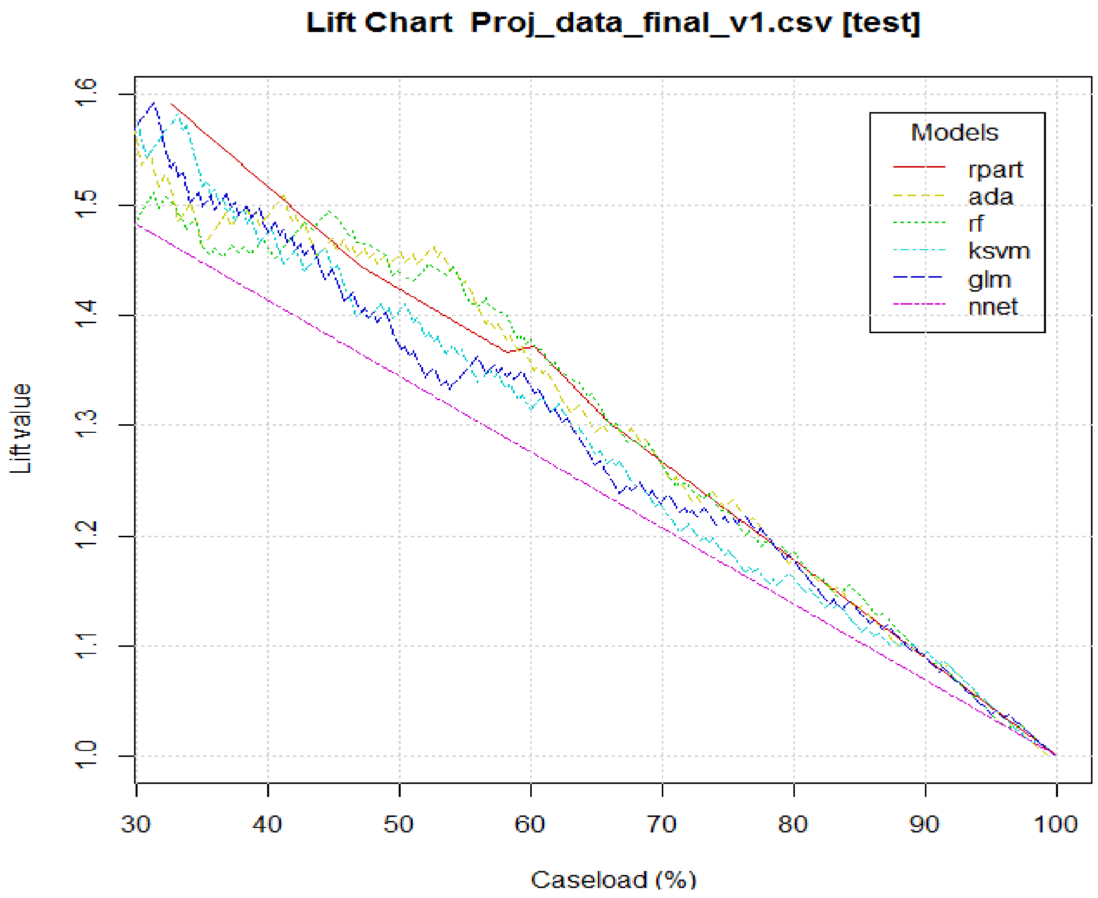
Model Name	Area Under ROC	Overall Error	False Negative %	Recall
Decision Tree	0.7589	0.2984	15	0.679487
SVM	0.7506	0.3076	14	0.717949
Random Forest	0.7701	0.2892	14	0.717949
Ada Boost	0.7687	0.2923	15	0.692308
Logistic Regression	0.7568	0.3138	17	0.653846
Neural Network	0.5197	0.4615	46	0.051282

.1 Results for ROC curve

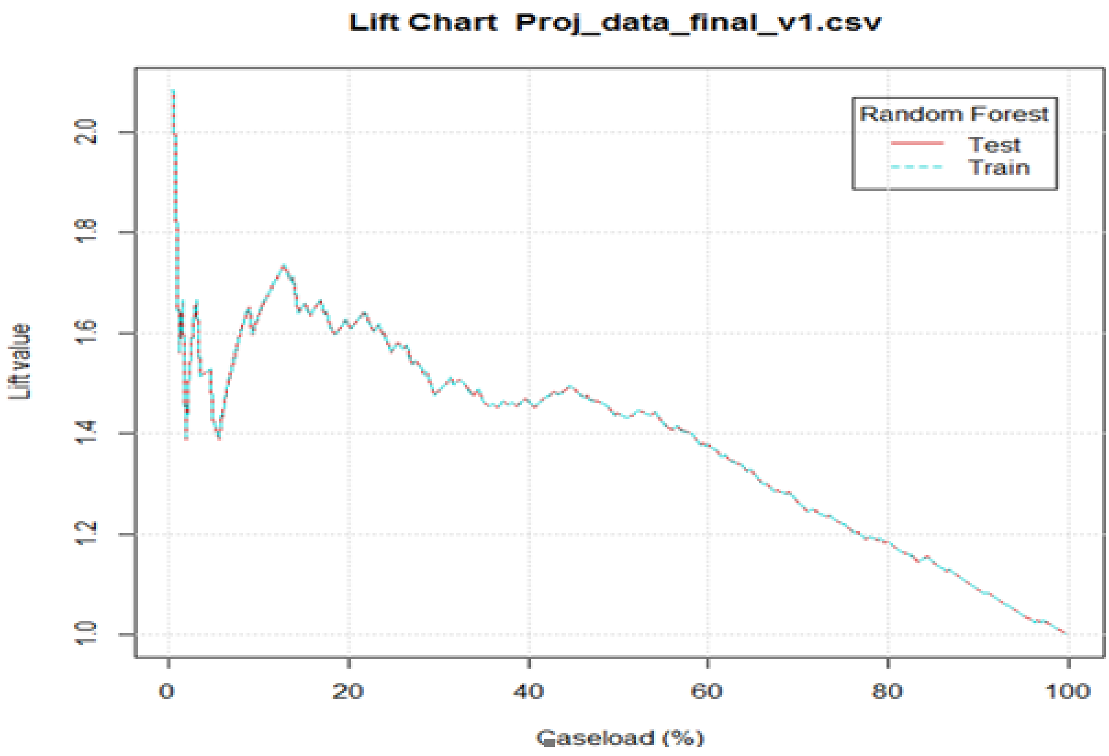
Below are the ROC plots for the different modelling techniques



.2 Results of Lift Curve



Maximum Lift Graph:



.3 Error Matrix

Error matrix for the Decision Tree model

Actual	Predicted	
	0	1
0	122	47
1	50	106

Overall error: 0. 2984615

Error matrix for the Ada Boost model

Actual	Predicted	
	0	1
0	122	47
1	48	108

Overall error: 0. 2923077

Error matrix for the Random Forest model

Actual	Predicted	
	0	1
0	119	50
1	44	112

Overall error: 0. 2892308

Error matrix for the SVM model

Actual	Predicted	
	0	1
0	113	56
1	44	112

Overall error: 0. 3076923

Error matrix for the Linear model

Actual	Predicted	
	0	1
0	121	48
1	54	102

Overall error: 0. 3138462

Error matrix for the Neural Net model

Actual	Predicted		
		0	1
	0	167	2
	1	148	8

Overall error: 0. 4615385

Error matrix for the Ensemble(voting)

Actual	Predicted		
		0	1
	0	103	58
	1	26	136

Overall Error: 0.26006192

From the ensemble method we obtain the least error rate and least false negatives which is the desired outcome based on business requirement and cost benefit analysis.

As we can see from the above observations we find that the neural network is very ill suited for our scenario and its performance compared to other techniques is not very impressive.

Also we find that the random forest has the small error rate. It is interesting to note that SVM and Random Forest minimizes this important factor in a similar manner.

.4 Cost Benefit Analysis of Data Mining

At the outset we understand from our previous mail campaign that the response rate is 1%. That is to say if we shoot out 100 mails we would expect to a response from 1 customer. The objective therefore would be to try to increase our odds of customers responding to our campaign. This would result in considerable savings in running the mailing campaign. Our objective here i.e. **success criteria** is to predict who would respond to the mail campaign by building a model.

The actual profit of the campaign is calculated as follows. For every mail sent, the company shells out some money say $1. And if the customer responds (True Positive) a profit is gained say, $50. But we need to subtract the cost of mailing the customer, so actual profit is $49 for a positive response. In the same line of reasoning, we should also consider the cost of not mailing to a customer who would actually respond (False Negative). In our case, this value turns out to be $50. It is clear that false negative severely impacts the bottom line of the campaign.

To illustrate how data mining contributes to the bottom line of the company, consider the following scenario.

Initial Stage: Running the mailing campaign without data mining models to guide us. This is actually a model of our ignorance about responders and the only knowledge we have is that the average response rate based on our previous campaigns, which is 1%. So to get 1000 respondents we need to mail 100000 customers. The cost of doing this is illustrated in the following table.

Initial Case

Population	100,000.00
No of People Mailed To	100,000.00
No of people responding	1000
Cost/mail	$1.00
Profit/Respondent	$50.00
Net - Profit	**-$50,000.00**

As we can see, this is not a very profitable campaign. To see how data mining can help us improve this, please consider the following table. This table gives the overall statistics developed from Ensemble (voting) Method

Ensemble Method

Population	100,000.00
No of People Mailed To	1,000.00
Cost/mail	1
Profit/Respondent	50
Accuracy	74.00%
Precision	70.00%
False Negative Rate	16.00%
No of people responding(estimate)	700
No of responders missed(estimate)	160
Profit from True Positive	$35,000.00
Cost of false negatives	$8,000.00
Net - Profit (Profit from responders-Cost of Missing responders- Cost of sending mail)	**$26,000.00**

So it is clear that a mail campaign driven by data mining insights can be highly profitable.