

A dark blue vertical bar is positioned on the left side of the slide. A blue arrow-shaped graphic points to the right from this bar, containing the date. In the bottom-left corner, there are several thin, curved lines in shades of blue and grey.

11/18/2015

IDS 572

Prediction of Donor and Non-Donor

In dataset, we considered the variable types as below. Row Id, Row Id. and TARGET_D role set to none.

zipconvert_2	Flag
zipconvert_3	Flag
zipconvert_4	Flag
zipconvert_5	Flag
homeowner ...	Flag
NUMCHLD	Nominal
INCOME	Nominal
gender dum...	Flag
WEALTH	Ordinal
HV	Continuous
lcmcd	Continuous
lcavg	Continuous
IC15	Continuous
NUMPROM	Continuous
RAMNTALL	Continuous
MAXRAMNT	Continuous
LASTGIFT	Continuous
totalmonths	Continuous
TIMELAG	Continuous
AVGGIFT	Continuous
TARGET_B	Flag

Figure1: variables and their types

- (a) **Statistics:** For Categorical variables, we ignored Variance, Standard Deviation and Standard Error of Mean. For flag variables, we ignored Mean, Median, Variance, Standard Deviation and Standard Error of Mean.

Variable	Mean	Min	Max	Range	Variance	SD	Median	Mode	Correlation
zipconvert_2	-	0.000	1.000	1.000	-	-	-	0.000	0.004
zipconvert_3	-	0.000	1.000	1.000	-	-	-	0.000	-0.013
zipconvert_4	-	0.000	1.000	1.000	-	-	-	0.000	-0.013
zipconvert_5	-	0.000	1.000	1.000	-	-	-	0.000	0.021
homeowner dummy	-	0.000	1.000	1.000	-	-	-	1.000	0.030
gender dummy	-	0.000	1.000	1.000	-	-	-	1.000	0.026
WEALTH	6.402	0.000	9.000	9.000	-	-	8.000	8.000	0.003
NUMCHLD	1.069	1.000	5.000	4.000	-	-	1.000	1.000	-0.046
INCOME	3.894	1.000	7.000	6.000	-	-	4.000	4.000	0.034
HV	1141.362	0.000	5945.000	5945.000	896131.383	946.642	822.000	0.000	0.023
lcmcd	388.217	0.000	1500.000	1500.000	29865.353	172.816	356.000	288.000	0.008
lcavg	432.088	0.000	1331.000	1331.000	28289.593	168.195	396.000	0.000	0.002
IC15	14.703	0.000	90.000	90.000	145.924	12.080	12.000	0.000	-0.003
NUMPROM	49.089	11.000	157.000	146.000	516.068	22.717	48.000	24.000	0.075
RAMNTALL	110.400	15.000	5674.900	5659.900	21697.270	147.300	81.000	25.000	0.020
MAXRAMNT	16.651	5.000	1000.000	995.000	493.885	22.224	15.000	10.000	-0.020
LASTGIFT	13.523	0.000	219.000	219.000	111.967	10.581	10.000	10.000	-0.080
totalmonths	31.137	17.000	37.000	20.000	17.081	4.133	31.000	31.000	-0.137
TIMELAG	6.862	0.000	77.000	77.000	30.927	5.561	5.000	5.000	0.010
AVGGIFT	10.691	2.139	122.167	120.028	55.413	7.444	9.000	15.000	-0.082

Table 1: Statistics of variables in the dataset and their correlation with Target variable

In the above table, the rows are highlighted with light gold color if the correlation between the variable and target is Strong. The rows are highlighted with light blue color if the correlation between the variable and target is Medium.

Frequency Distribution for flag variables:



Value ▲	Proportion	%	Count
0.000		78.56	2451
1.000		21.44	669

Figure2: Distribution of zipconvert_2



Value ▲	Proportion	%	Count
0.000		81.47	2542
1.000		18.53	578

Figure3: Distribution of zipconvert_3

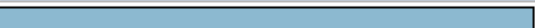

Value ▲	Proportion	%	Count
0.000		78.56	2451
1.000		21.44	669

Figure4: Distribution of zipconvert_4



Value ▲	Proportion	%	Count
0.000		61.54	1920
1.000		38.46	1200

Figure5: Distribution of zipconvert_5

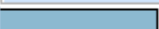

Value ▲	Proportion	%	Count
0.000		22.98	717
1.000		77.02	2403

Figure6: Distribution of homeowner dummy



Value ▲	Proportion	%	Count
0.000		39.07	1219
1.000		60.93	1901

Figure7: Distribution of gender dummy



Value ▲	Proportion	%	Count
0.000		50.0	1560
1.000		50.0	1560

Figure8: Distribution of TARGET_B

Note: Since homeowner dummy has medium correlation with the target variable, we are interested to look into the homeowner dummy proportions with TARGET_B=0 and TARGET_B=1 individually.

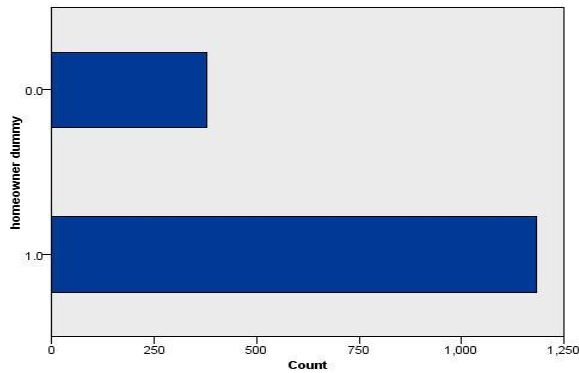


Figure9: Distribution for homeowner dummy with TARGET_B=0

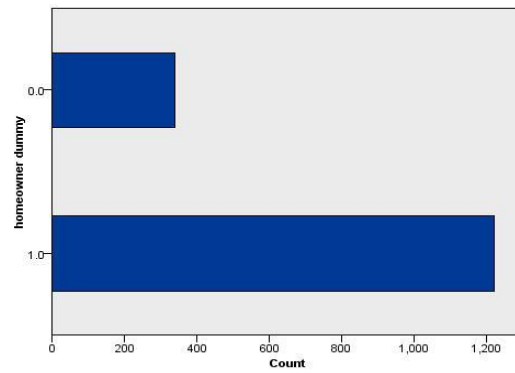


Figure10: Distribution for homeowner dummy with TARGET_B=1

We observed almost equal number of proportions for the different target variable values.

Scatter Plot for categorical and flag variables:

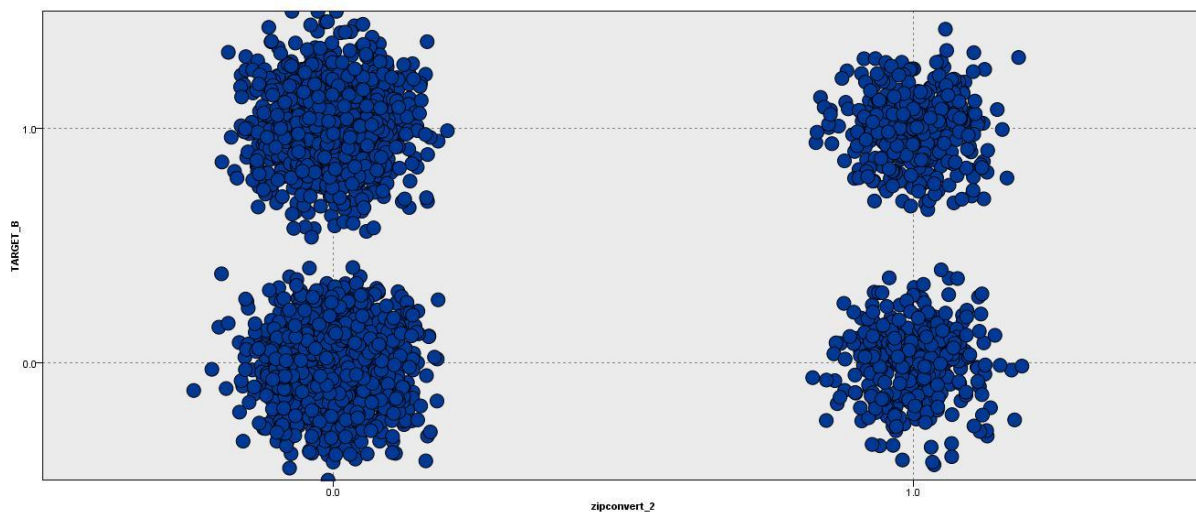


Figure11: Scatter plot for zipconvert_2 vs TARGET_B

From Figure11, we can say that the proportion of zipconvert_2=0 is high. Zipconvert_2 is almost equally divided between the target variable TARGET_B=0 and TARGET_B=1.

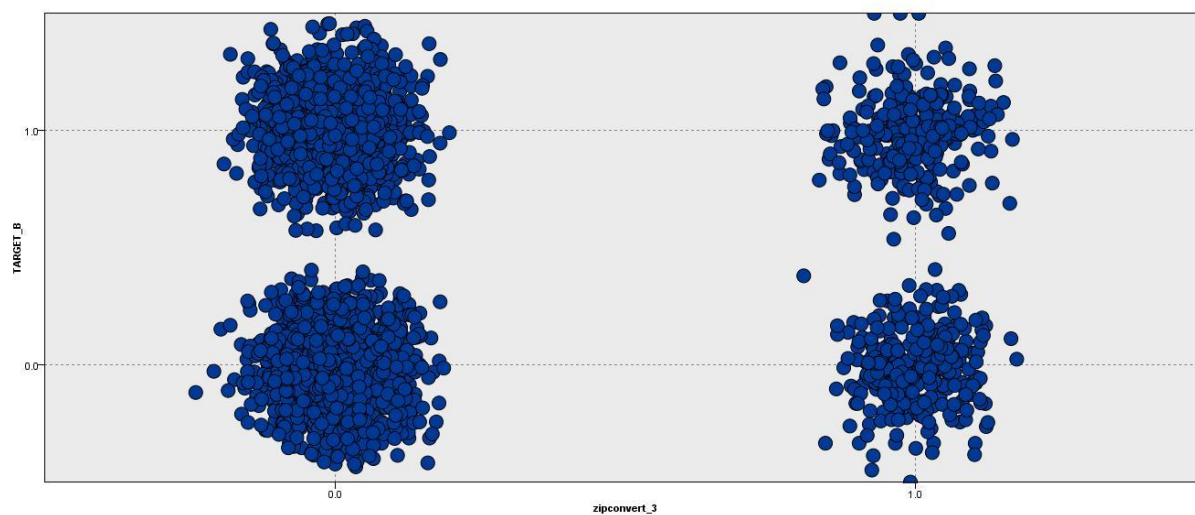


Figure12: Scatter plot for zipconvert_3 vs TARGET_B

From Figure12, we can say that the proportion of zipconvert_3=1 is very less. Zipconvert_3 is almost equally divided between the target variable TARGET_B=0 and TARGET_B=1.

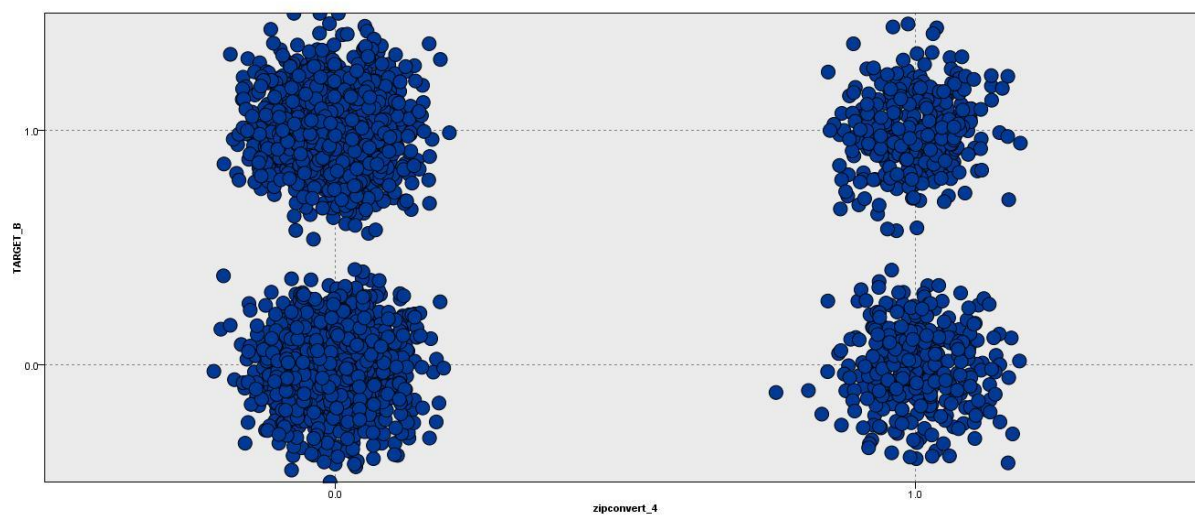


Figure13: Scatter plot for zipconvert_4 vs TARGET_B

From Figure13, we can say that the proportion of zipconvert_4=1 is very less. Zipconvert_4 is almost equally divided between the target variable TARGET_B=0 and TARGET_B=1.

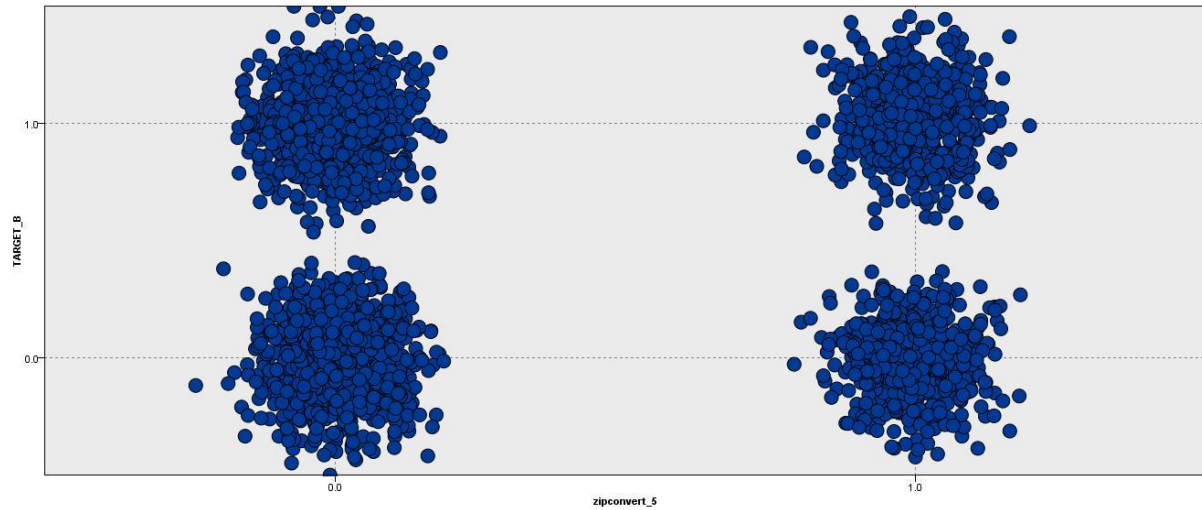


Figure14: Scatter plot for zipconvert_5 vs TARGET_B

From Figure14, we can say that the proportion of zipconvert_5=0 is high. Zipconvert_5 is almost equally divided between the target variable TARGET_B=0 and TARGET_B=1.

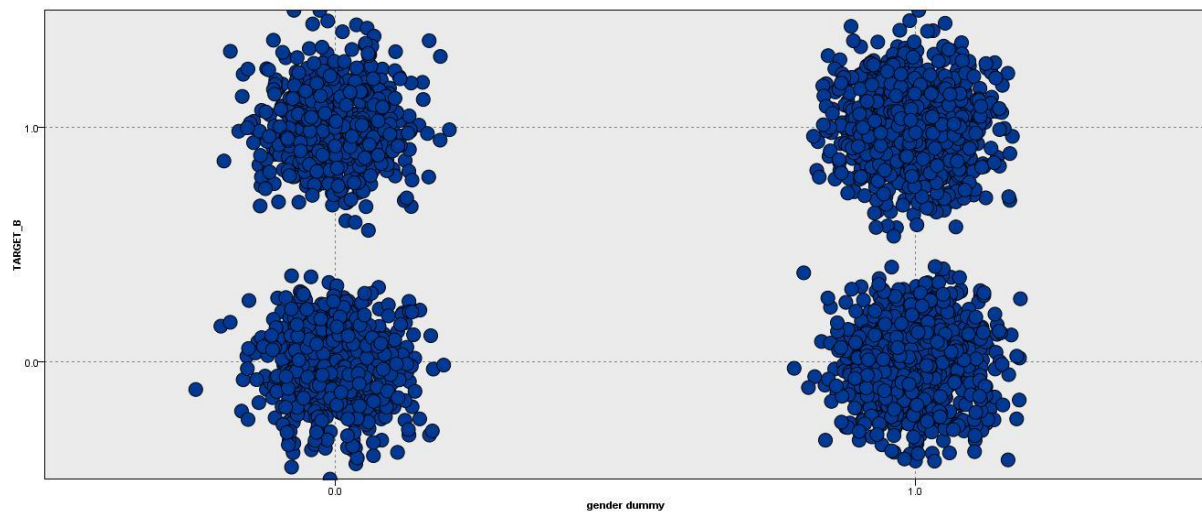


Figure15: Scatter plot for gender dummy vs TARGET_B

From Figure15, we can say that the proportion of gender dummy=1 is high. Gender dummy is almost equally divided between the target variable TARGET_B=0 and TARGET_B=1.

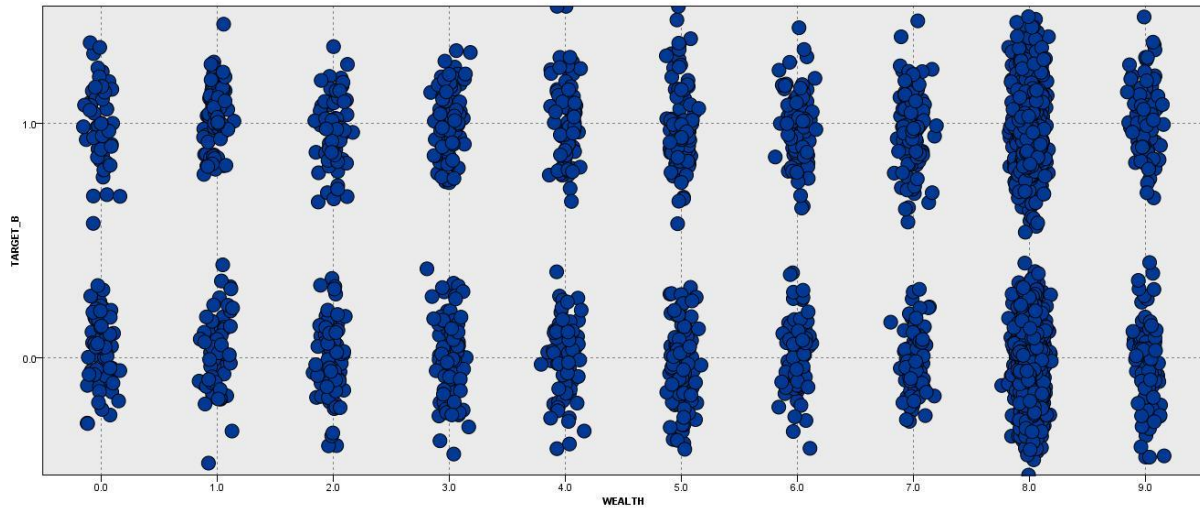


Figure16: Scatter plot for WEALTH vs TARGET_B

From Figure16, we can say that the proportion is higher for category 8. WEALTH is almost equally divided between the target variable TARGET_B=0 and TARGET_B=1.

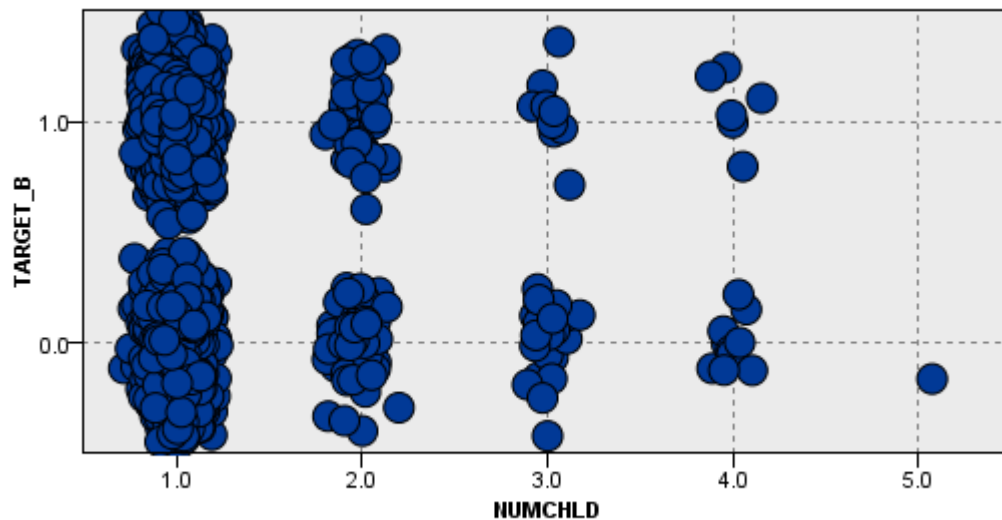


Figure17: Scatter plot for NUMCHILD vs TARGET_B

From Figure17, we observe that the NUMCHILD=1 is higher in proportion. Only very few were there with values NUMCHILD=4 and NUMCHILD=5. If NUMCHILD=5 then the value of the TARGET_B=0.

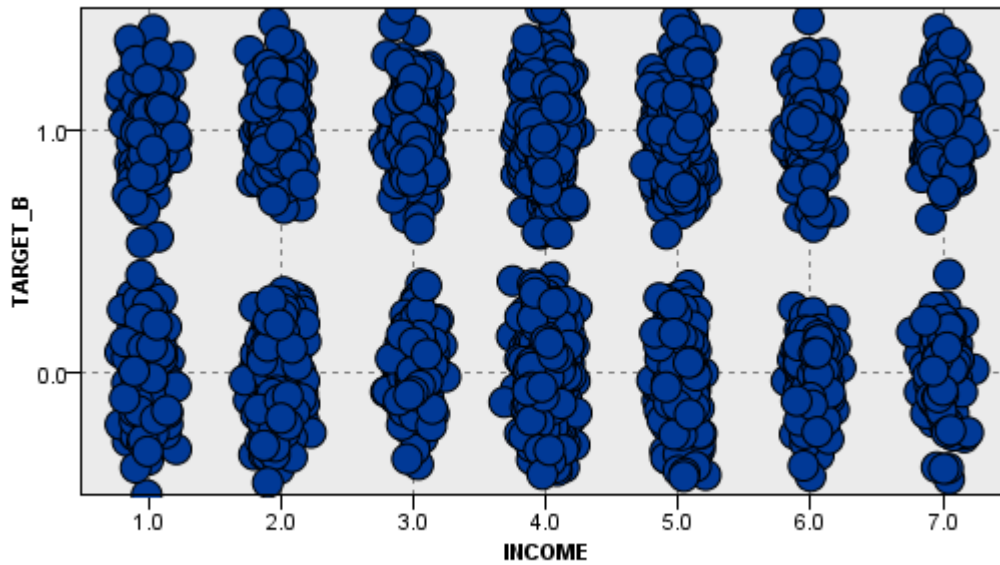


Figure18: Scatter plot for INCOME vs TARGET_B

From Figure17, we observe that the proportion is higher for INCOME=4. There is almost equal division between the TARGET_B=0 and TARGET_B=1 for all categories of income.

Histograms for continuous variables:

Note: For continuous variables, we looked into their proportions with TARGET_B=0 and TARGET_B=1 individually, as it may help us to identify the features that differentiate the target variable more.

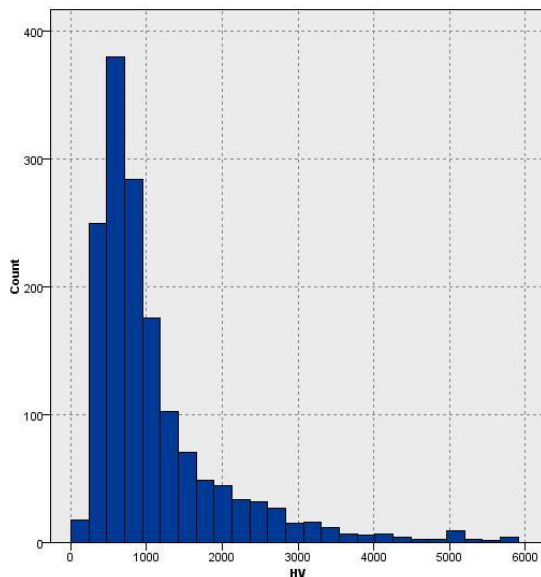


Figure19: Histogram for HV with TARGET_B=0

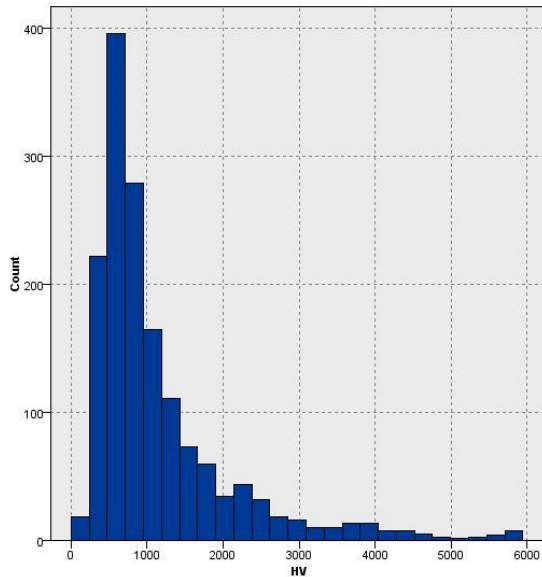


Figure20: Histogram for HV with TARGET_B=1

For HV vs TARGET_B, the proportion is higher in the range 250-1500. The proportion increases from HV range between 200 and 400, almost constant from 400-750, and then starts decreasing gradually.

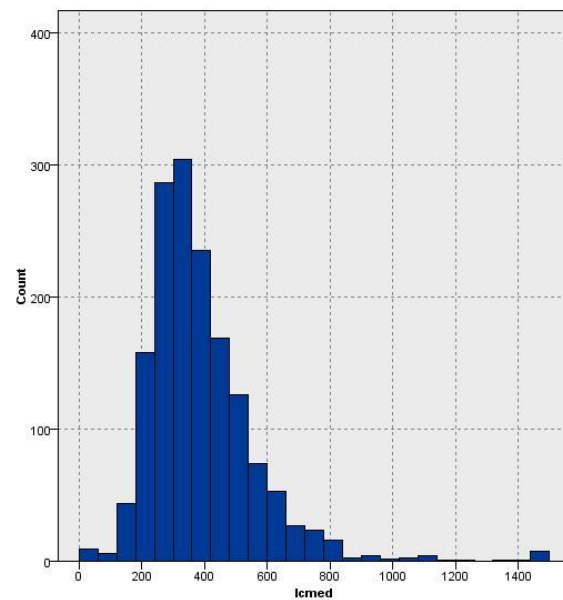
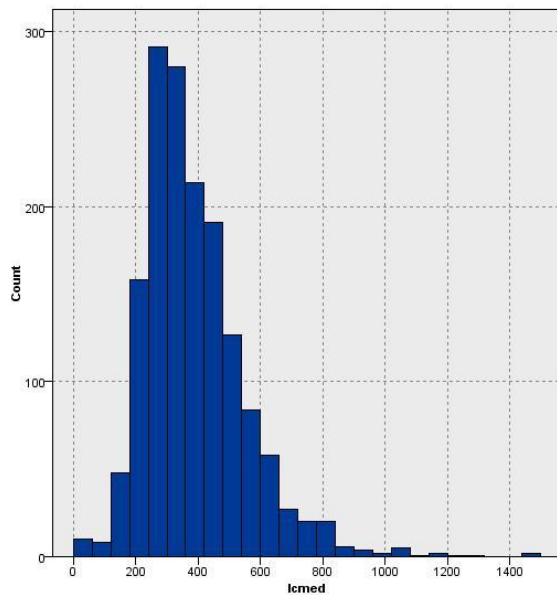


Figure21: Histogram for lamed with TARGET_B=0 Figure22: Histogram for lamed with TARGET_B=1

For lamed vs TARGET_B, the peaks points occurs at the same points of lamed. The proportion of TARGET_B increases till lamed=300 and then it starts decreasing gradually.

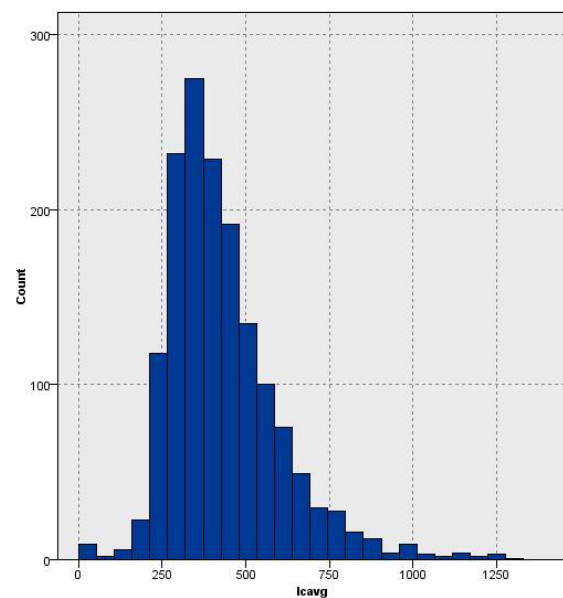
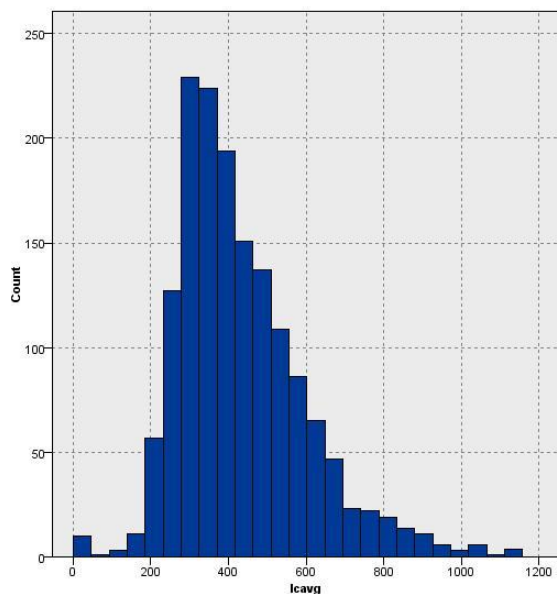


Figure23: Histogram for lcavg with TARGET_B=0

Figure24: Histogram for lcavg with TARGET_B=1

For lcavg vs TARGET_B, we didn't observe much difference between both the graphs. The proportion increases till lcavg=340 and then starts decreasing.

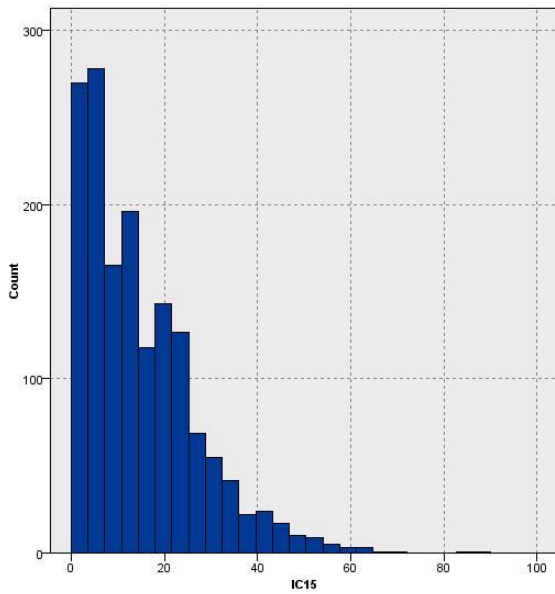


Figure25: Histogram for IC15 with TARGET_B=0

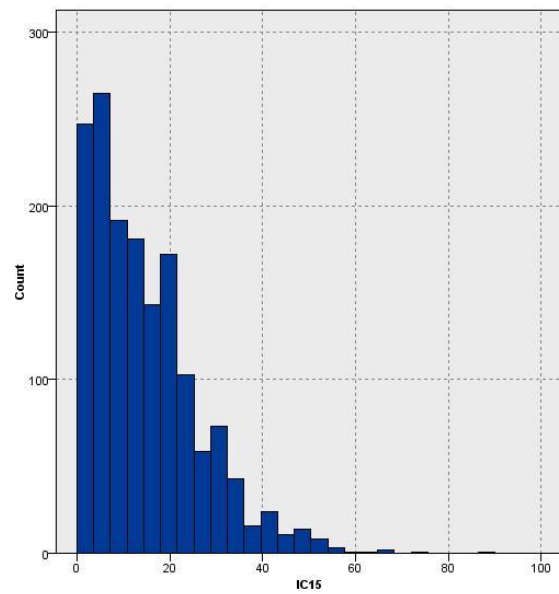


Figure26: Histogram for IC15 with TARGET_B=1

For IC15 vs TARGET_B, the peaks are higher at lower values. The proportion between IC15 and TARGET_B is inverse relationship.

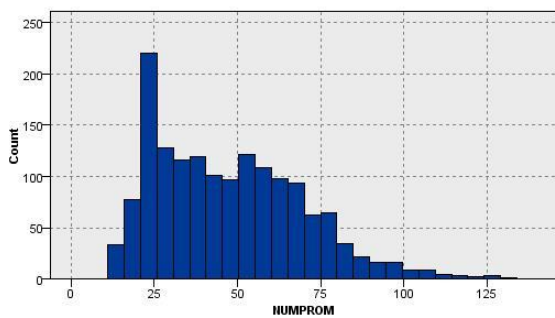


Figure27: Histogram for NUMPROM with TARGET_B=0

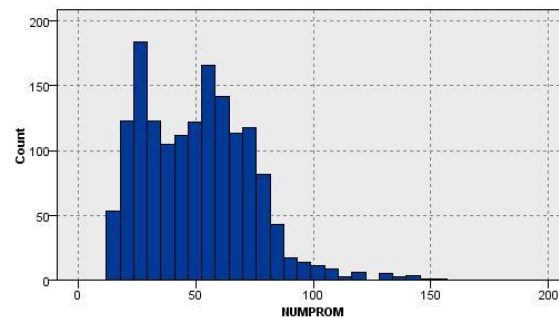


Figure28: Histogram for NUMPROM with TARGET_B=1

For NUMPROM vs TARGET_B, we can divide the graph into 3 regions (0-30, 30-70, >70). In the first region, the proportion increases with NUMPROM. In the second region, the proportion is almost constant. In the third region, proportion will decrease with the increase in NUMPROM.

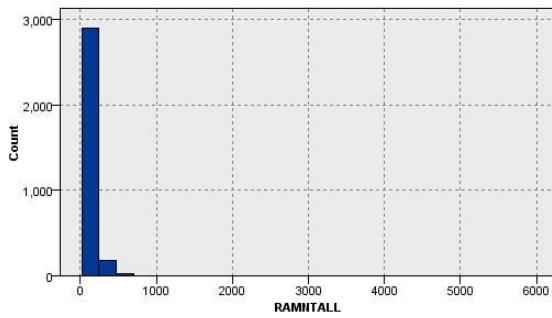


Figure29: Histogram for RAMNTALL

For RAMNTALL vs TARGET_B, the proportion is higher at a single parts only. So we haven't looked at different separate graphs.

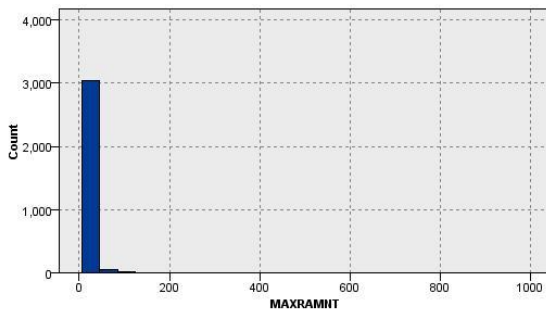


Figure30: Histogram for MAXRAMNT

For MAXRAMNT vs TARGET_B, the proportion is higher at a single parts only. So we haven't looked at different separate graphs.

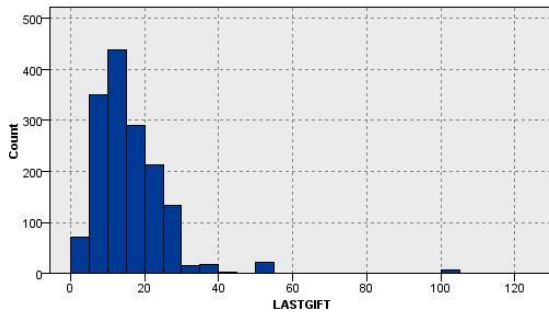


Figure31: Histogram for LASTGIFT with TARGET_B=0

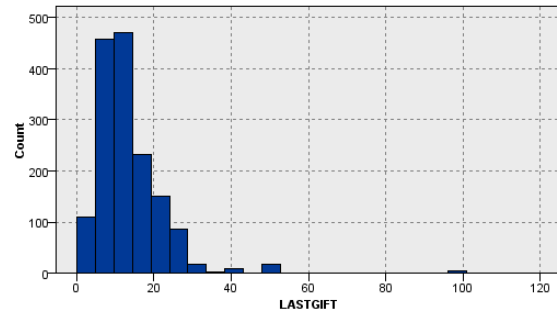


Figure32: Histogram for LASTGIFT with TARGET_B=1

For LASTGIFT vs TARGET_B, the change in the peaks for both the graphs is very less. For LASTGIFT<10, the proportion increases and then decreases gradually.

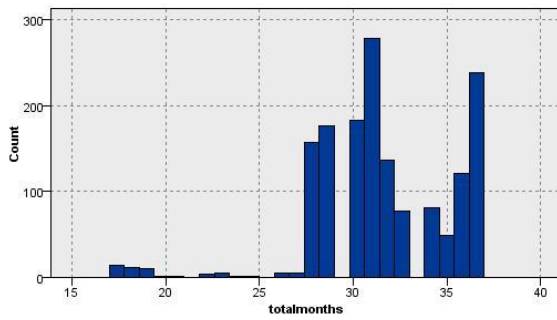


Figure33: Histogram for totalmonths with TARGET_B=0

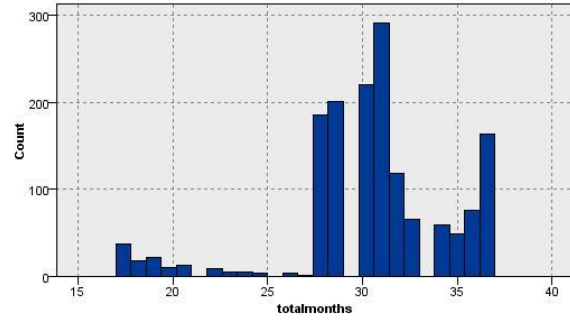


Figure34: Histogram for totalmonths with TARGET_B=1

For totalmonths vs TARGET_B, the change in the peaks for both the graphs is very less. The proportion is very less when the totalmonths<27. After that, the proportion increases till totalmonths=32 and then decreases till totalmonths=35 and then increases.

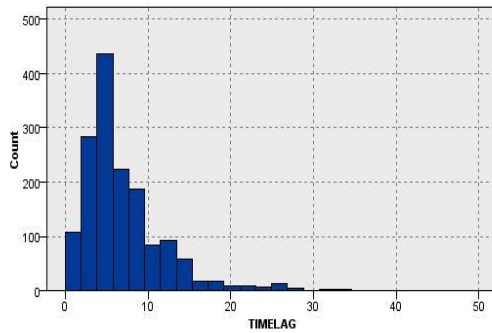


Figure35: Histogram for TIMELAG with TARGET_B=0

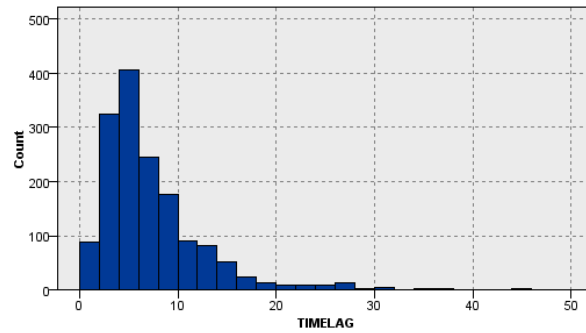


Figure36: Histogram for TIMELAG with TARGET_B=1

For totalmonths vs TARGET_B, both the graphs are almost same. The proportion increases till TIMELAG=5 and then decreases.

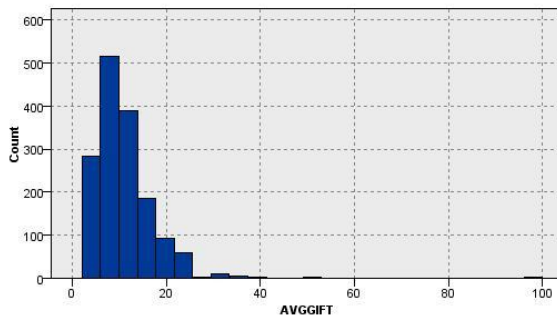


Figure37: Histogram for AVGGIFT with TARGET_B=0

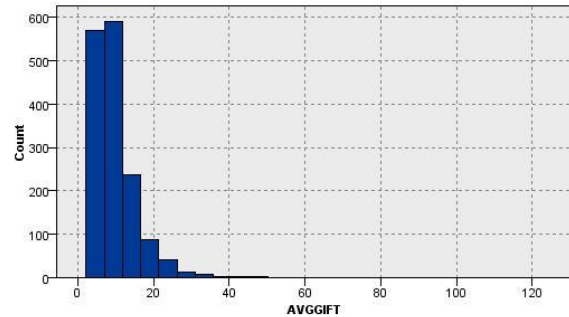


Figure38: Histogram for AVGGIFT with TARGET_B=1

For AVGGIFT vs TARGET_B, there is huge change in the peak for AVGGIT<=5 for TARGET_B=1.

'Partition' = 1_Training	0.000000	1.000000
0.000000	211	60
1.000000	154	68
'Partition' = 2_Testing	0.000000	1.000000
0.000000	147	40
1.000000	114	47

KNN:

Results for output field TARGET_B

Comparing \$KNN-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	328	66.53%	189	54.31%
Wrong	165	33.47%	159	45.69%
Total	493		348	

Coincidence Matrix for \$KNN-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	206	65
1.000000	100	122
'Partition' = 2_Testing	0.000000	1.000000
0.000000	118	69
1.000000	90	71

C&RT:

Results for output field TARGET_B

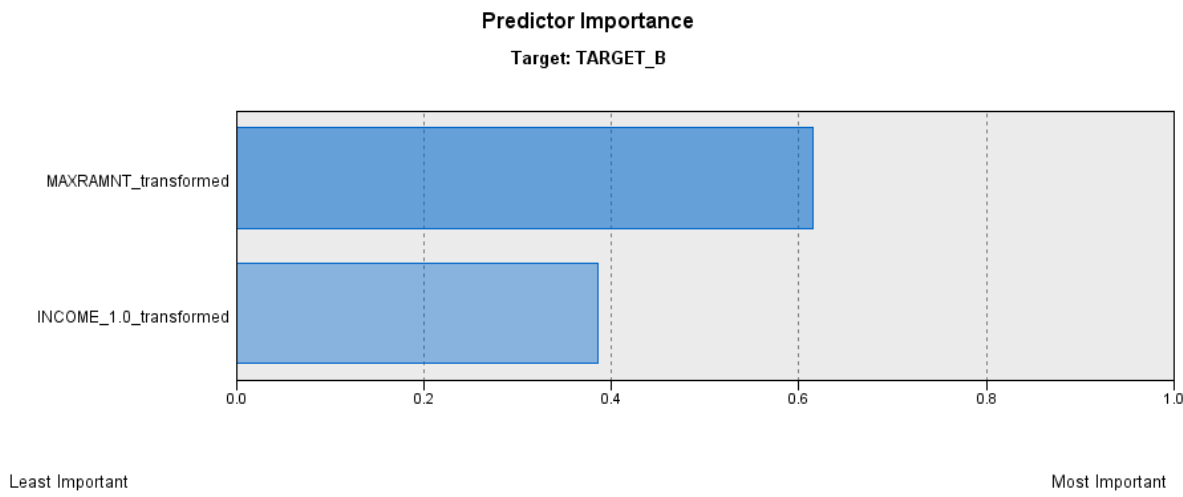
Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	271	54.97%	187	53.74%
Wrong	222	45.03%	161	46.26%
Total	493		348	

Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000
0.000000	271
1.000000	222
'Partition' = 2_Testing	0.000000
0.000000	187
1.000000	161

CHAID:



Results for output field TARGET_B

Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	297	60.24%	198	56.9%
Wrong	196	39.76%	150	43.1%
Total	493		348	

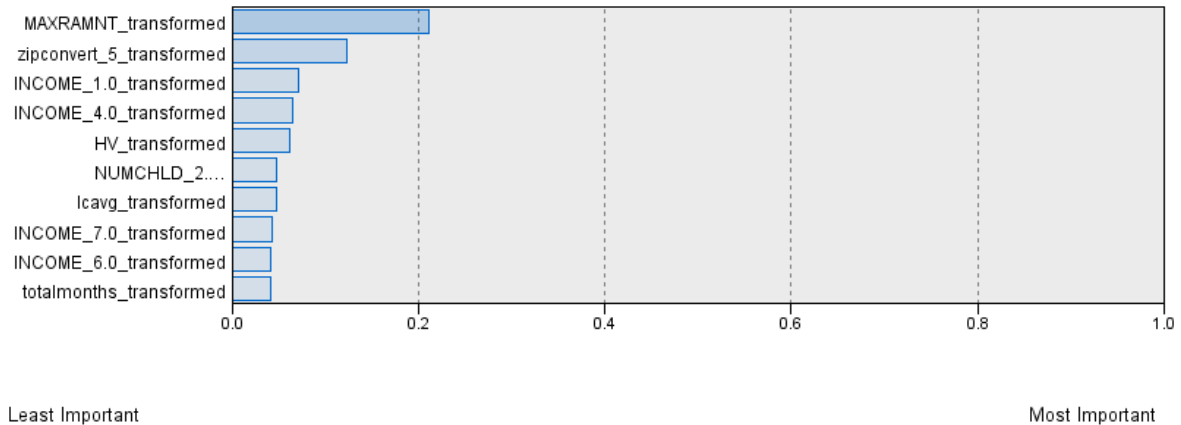
Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training		0.000000	1.000000
0.000000		215	56
1.000000		140	82
'Partition' = 2_Testing		0.000000	1.000000
0.000000		151	36
1.000000		114	47

C5.0:

Predictor Importance

Target: TARGET_B



Results for output field TARGET_B

Comparing \$C-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	416	84.38%	186	53.45%
Wrong	77	15.62%	162	46.55%
Total	493		348	

Coincidence Matrix for \$C-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	242	29
1.000000	48	174
'Partition' = 2_Testing	0.000000	1.000000
0.000000	115	72
1.000000	90	71

Bayes Net:

Results for output field TARGET_B

Comparing \$B-TARGET_B with TARGET_B

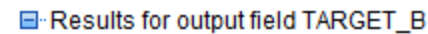
'Partition'	1_Training		2_Testing	
Correct	1	0.2%	0	0%
Wrong	492	99.8%	348	100%
Total	493		348	

Coincidence Matrix for \$B-TARGET_B (rows show actuals)

'Partition' = 1_Training	1.000000	\$null\$
0.000000	0	271
1.000000	1	221
'Partition' = 2_Testing	\$null\$	
0.000000	187	
1.000000	161	

NUMPROM>30 and NUMPROM<=70:

C&RT:



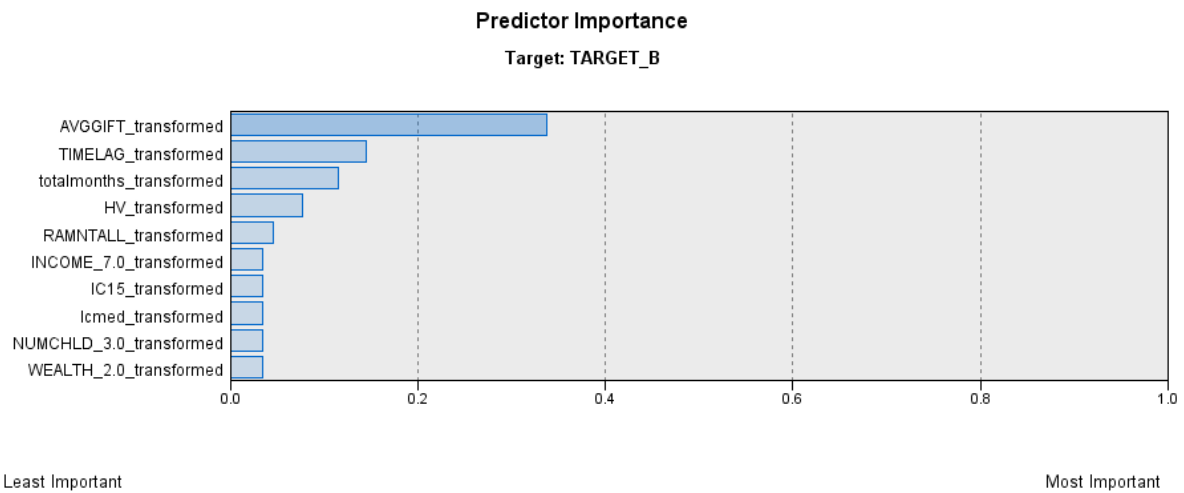
KNN:

Results for output field TARGET_B

Comparing \$KNN-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	716	69.92%	341	47.83%
Wrong	308	30.08%	372	52.17%
Total	1,024		713	

C&RT:



Results for output field TARGET_B

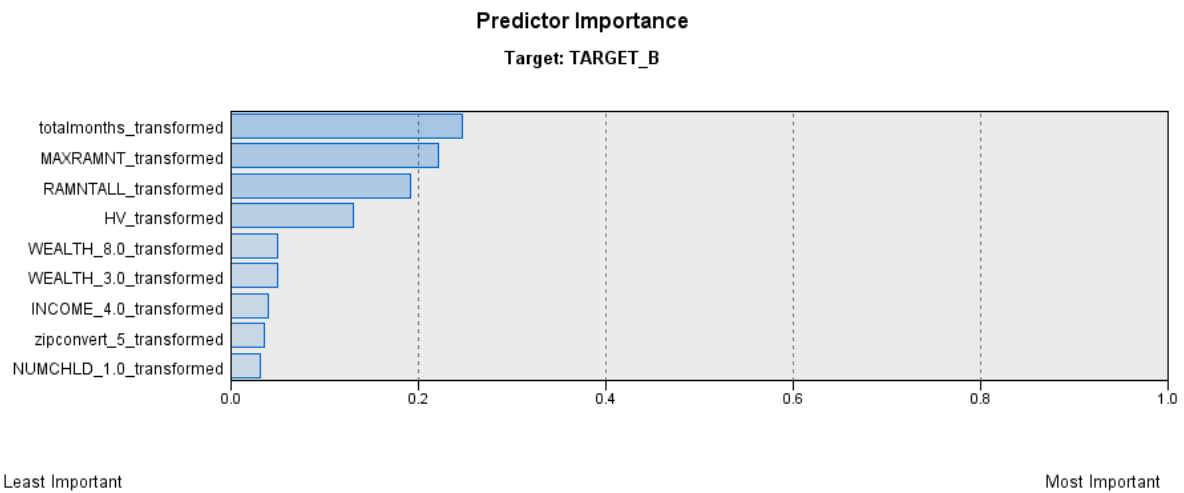
Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	610	59.57%	369	51.75%
Wrong	414	40.43%	344	48.25%
Total	1,024		713	

Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	454	57
1.000000	357	156
'Partition' = 2_Testing	0.000000	1.000000
0.000000	281	63
1.000000	281	88

CHAID:



Results for output field TARGET_B

Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	633	61.82%	382	53.58%
Wrong	391	38.18%	331	46.42%
Total	1,024		713	

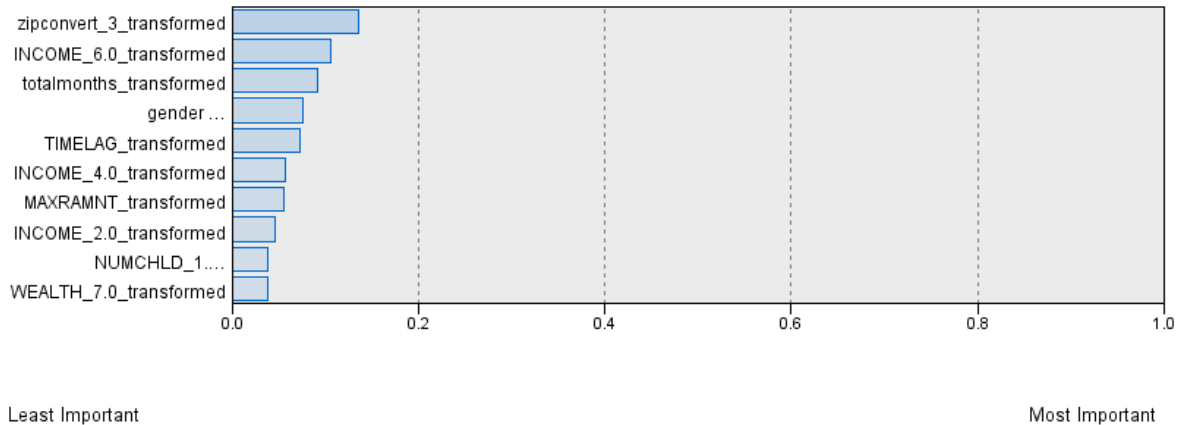
Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	298	213
1.000000	178	335
'Partition' = 2_Testing	0.000000	1.000000
0.000000	172	172
1.000000	159	210

C5.0:

Predictor Importance

Target: TARGET_B



Results for output field TARGET_B

Comparing \$C-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	888	86.72%	366	51.33%
Wrong	136	13.28%	347	48.67%
Total	1,024		713	

Coincidence Matrix for \$C-TARGET_B (rows show actuals)

'Partition' = 1_Training		0.000000	1.000000
0.000000		456	55
1.000000		81	432
'Partition' = 2_Testing		0.000000	1.000000
0.000000		185	159
1.000000		188	181

Bayes Net:

Results for output field TARGET_B

Comparing \$B-TARGET_B with TARGET_B

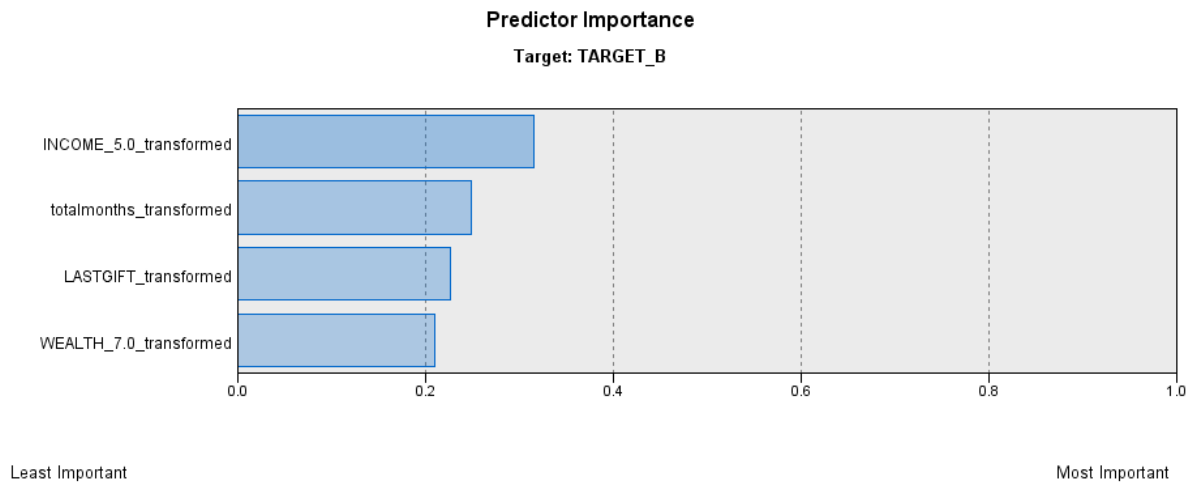
'Partition'	1_Training		2_Testing	
Correct	6	0.59%	7	0.98%
Wrong	1,018	99.41%	706	99.02%
Total	1,024		713	

Coincidence Matrix for \$B-TARGET_B (rows show actuals)

'Partition' = 1_Training		0.000000	\$null\$
0.000000		6	505
1.000000		8	505
'Partition' = 2_Testing		0.000000	\$null\$
0.000000		7	337
1.000000		2	367

NUMPROM>70:

Logistic Regression:



Results for output field TARGET_B

Comparing \$L-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	188	60.84%	117	50.21%
Wrong	121	39.16%	116	49.79%
Total	309		233	

Coincidence Matrix for \$L-TARGET_B (rows show actuals)

'Partition' = 1_Training		0.000000	1.000000
0.000000		68	72
1.000000		49	120
'Partition' = 2_Testing		0.000000	1.000000
0.000000		39	68
1.000000		48	78

KNN:

Results for output field TARGET_B

Comparing \$KNN-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	210	67.96%	113	48.5%
Wrong	99	32.04%	120	51.5%
Total	309		233	

Coincidence Matrix for \$KNN-TARGET_B (rows show actuals)

'Partition' = 1_Training		0.000000	1.000000
0.000000		60	80
1.000000		19	150
'Partition' = 2_Testing		0.000000	1.000000
0.000000		23	84
1.000000		36	90

C&RT:

Results for output field TARGET_B

Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	169	54.69%	126	54.08%
Wrong	140	45.31%	107	45.92%
Total	309		233	

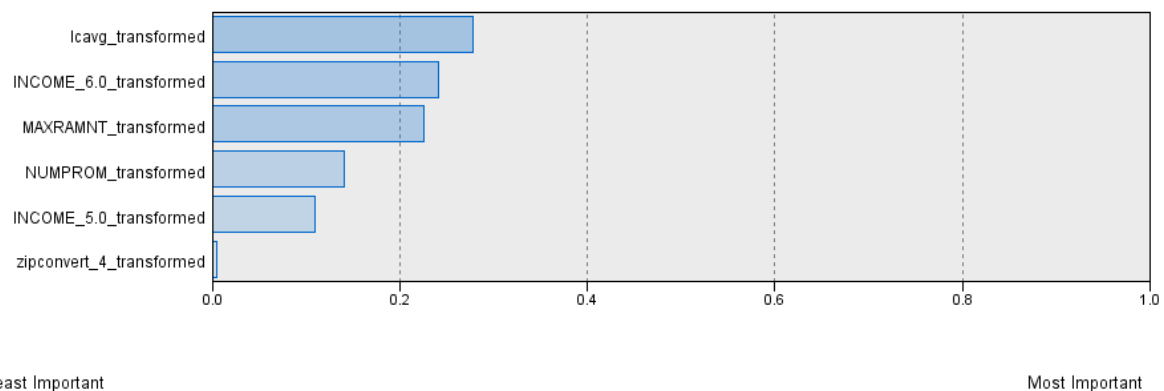
Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training		1.000000
0.000000		140
1.000000		169
'Partition' = 2_Testing		1.000000
0.000000		107
1.000000		126

CHAID:

Predictor Importance

Target: TARGET_B



Results for output field TARGET_B

Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	206	66.67%	120	51.5%
Wrong	103	33.33%	113	48.5%
Total	309		233	

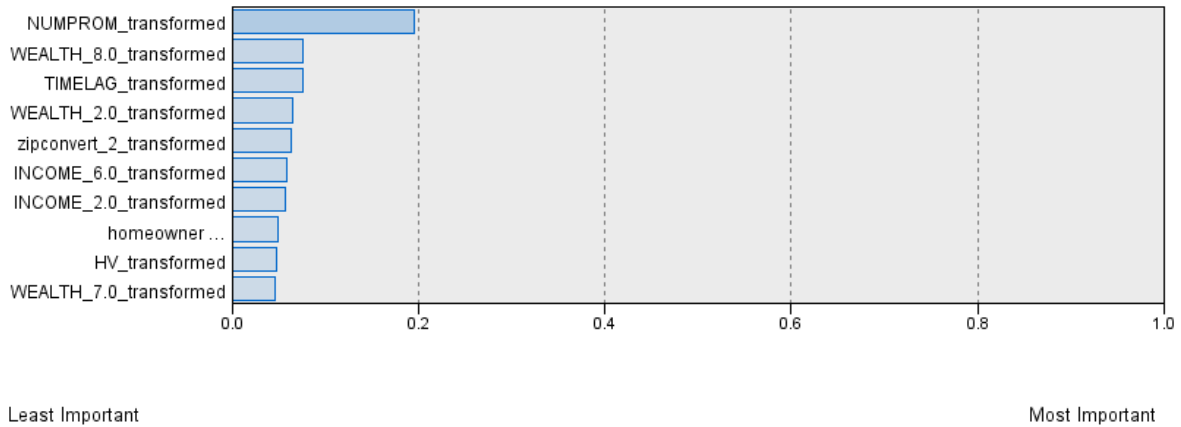
Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training		0.000000	1.000000
0.000000		64	76
1.000000		27	142
'Partition' = 2_Testing		0.000000	1.000000
0.000000		41	66
1.000000		47	79

C5.0:

Predictor Importance

Target: TARGET_B



Results for output field TARGET_B

Comparing \$C-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	271	87.7%	122	52.36%
Wrong	38	12.3%	111	47.64%
Total	309		233	

Coincidence Matrix for \$C-TARGET_B (rows show actuals)

'Partition' = 1_Training		0.000000	1.000000
0.000000		119	21
1.000000		17	152
'Partition' = 2_Testing		0.000000	1.000000
0.000000		54	53
1.000000		58	68

Bayes Net:

Results for output field TARGET_B

Comparing \$B-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	14	4.53%	3	1.29%
Wrong	295	95.47%	230	98.71%
Total	309		233	

Coincidence Matrix for \$B-TARGET_B (rows show actuals)

'Partition' = 1_Training		0.000000	1.000000	\$null\$
0.000000		7	1	132
1.000000		4	7	158
'Partition' = 2_Testing		0.000000	1.000000	\$null\$
0.000000		1	3	103
1.000000		6	2	118

Reduce the input variables:

NUMPROM<=30:

We did feature selection and the following feature selected variables are given as input to the models.

1. MAXRAMNT_transformed
2. LASTGIFT_transformed
3. Totalmonths_transformed
4. AVGGIFT_transformed
5. Homeowner dummy

Logistic Regression: The variable effecting the target is totalmonths_transformed only. Therefore, no change in output.

KNN:

Results for output field TARGET_B

Comparing \$KNN-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	346	70.18%	197	56.61%
Wrong	147	29.82%	151	43.39%
Total	493		348	

Coincidence Matrix for \$KNN-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	234	37
1.000000	110	112
'Partition' = 2_Testing	0.000000	1.000000
0.000000	148	39
1.000000	112	49

Accuracy increased with the reduced number of variables

C&RT:

Results for output field TARGET_B

Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	271	54.97%	187	53.74%
Wrong	222	45.03%	161	46.26%
Total	493		348	

Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000
0.000000	271
1.000000	222
'Partition' = 2_Testing	0.000000
0.000000	187
1.000000	161

No change with the decrease in the input variables

CHAID:

Results for output field TARGET_B

Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	291	59.03%	192	55.17%
Wrong	202	40.97%	156	44.83%
Total	493		348	

Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	208	63
1.000000	139	83
'Partition' = 2_Testing	0.000000	1.000000
0.000000	141	46
1.000000	110	51

C5.0:

Results for output field TARGET_B

Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	291	59.03%	192	55.17%
Wrong	202	40.97%	156	44.83%
Total	493		348	

Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	208	63
1.000000	139	83
'Partition' = 2_Testing	0.000000	1.000000
0.000000	141	46
1.000000	110	51

Bayes Net:

Results for output field TARGET_B

Comparing \$B-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	283	57.4%	201	57.76%
Wrong	210	42.6%	147	42.24%
Total	493		348	

Coincidence Matrix for \$B-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000	
0.000000	212	59	
1.000000	151	71	
'Partition' = 2_Testing	0.000000	1.000000	\$null\$
0.000000	149	37	1
1.000000	109	52	0

NUMPROM>30 and NUMPROM<=70:

We did feature selection and the following feature selected variables are given as input to the models.

1. MAXRAMNT_transformed
2. LASTGIFT_transformed
3. Totalmonths_transformed
4. AVGGIFT_transformed
5. NUMPROM_transformed

Logistic Regression:

Results for output field TARGET_B

Comparing \$L-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	588	57.42%	398	55.82%
Wrong	436	42.58%	315	44.18%
Total	1,024		713	

Coincidence Matrix for \$L-TARGET_B (rows show actuals)

'Partition' = 1_Training		0.000000	1.000000
0.000000		254	257
1.000000		179	334
'Partition' = 2_Testing		0.000000	1.000000
0.000000		175	169
1.000000		146	223

KNN:

Results for output field TARGET_B

Comparing \$KNN-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	721	70.41%	366	51.33%
Wrong	303	29.59%	347	48.67%
Total	1,024		713	

Coincidence Matrix for \$KNN-TARGET_B (rows show actuals)

'Partition' = 1_Training		0.000000	1.000000
0.000000		264	247
1.000000		56	457
'Partition' = 2_Testing		0.000000	1.000000
0.000000		114	230
1.000000		117	252

C&RT:

Results for output field TARGET_B

Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	615	60.06%	390	54.7%
Wrong	409	39.94%	323	45.3%
Total	1,024		713	

Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	234	277
1.000000	132	381
'Partition' = 2_Testing	0.000000	1.000000
0.000000	144	200
1.000000	123	246

CHAID:

Results for output field TARGET_B

Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	598	58.4%	397	55.68%
Wrong	426	41.6%	316	44.32%
Total	1,024		713	

Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	373	138
1.000000	288	225
'Partition' = 2_Testing	0.000000	1.000000
0.000000	245	99
1.000000	217	152

C5.0:

Results for output field TARGET_B

Comparing \$C-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	616	60.16%	408	57.22%
Wrong	408	39.84%	305	42.78%
Total	1,024		713	

Coincidence Matrix for \$C-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	333	178
1.000000	230	283
'Partition' = 2_Testing	0.000000	1.000000
0.000000	220	124
1.000000	181	188

Bayes Net:

Results for output field TARGET_B

Comparing \$B-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	585	57.13%	401	56.24%
Wrong	439	42.87%	312	43.76%
Total	1,024		713	

Coincidence Matrix for \$B-TARGET_B (rows show actuals)

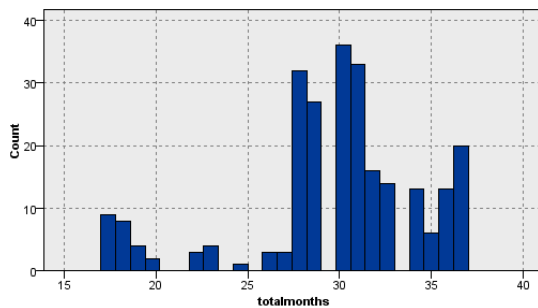
'Partition' = 1_Training	0.000000	1.000000	
0.000000	196	315	
1.000000	124	389	
'Partition' = 2_Testing	0.000000	1.000000	\$null\$
0.000000	138	203	3
1.000000	101	263	5

NUMPROM>70:

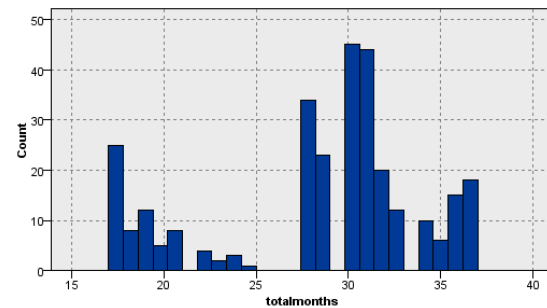
We did feature selection and the following feature selected variables are given as input to the models.

1. Totalmonths_transformed

Since we got only one variable as important, we decided to look into the histograms for totalmonths for the target variable when NUMPROM>70.



For target_b=0



For target_b=1

Input variables: NUMPROM_transformed, MAXRAMNT_transformed, LASTGIFT_transformed, totalmonths_transformed, AVGGIFT_transformed, homeowner dummy_transformed, INCOME_5.0_transformed, INCOME_6.0_transformed

totalmonths<=25:

Logistic Regression:

Results for output field TARGET_B

Comparing \$L-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	45	76.27%	27	67.5%
Wrong	14	23.73%	13	32.5%
Total	59		40	

Coincidence Matrix for \$L-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	9	9
1.000000	5	36
'Partition' = 2_Testing	0.000000	1.000000
0.000000	5	8
1.000000	5	22

KNN:

Results for output field TARGET_B

Comparing \$KNN-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	45	76.27%	24	60%
Wrong	14	23.73%	16	40%
Total	59		40	

Coincidence Matrix for \$KNN-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	5	13
1.000000	1	40
'Partition' = 2_Testing	0.000000	1.000000
0.000000	0	13
1.000000	3	24

C&RT:

Results for output field TARGET_B

Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	41	69.49%	27	67.5%
Wrong	18	30.51%	13	32.5%
Total	59		40	

Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training	1.000000
0.000000	18
1.000000	41
'Partition' = 2_Testing	1.000000
0.000000	13
1.000000	27

CHAID:

Results for output field TARGET_B

Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	44	74.58%	26	65%
Wrong	15	25.42%	14	35%
Total	59		40	

Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	6	12
1.000000	3	38
'Partition' = 2_Testing	0.000000	1.000000
0.000000	2	11
1.000000	3	24

C5.0:

Results for output field TARGET_B

Comparing \$C-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	44	74.58%	26	65%
Wrong	15	25.42%	14	35%
Total	59		40	

Coincidence Matrix for \$C-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	6	12
1.000000	3	38
'Partition' = 2_Testing	0.000000	1.000000
0.000000	2	11
1.000000	3	24

Bayes Net:

Results for output field TARGET_B

Comparing \$B-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	45	76.27%	26	65%
Wrong	14	23.73%	14	35%
Total	59		40	

Coincidence Matrix for \$B-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000	
0.000000	8	10	
1.000000	4	37	
'Partition' = 2_Testing	0.000000	1.000000	\$null\$
0.000000	5	7	1
1.000000	5	21	1

Totalmonths>25:

Logistic Regression:

Results for output field TARGET_B

Comparing \$L-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	137	52.49%	90	49.45%
Wrong	124	47.51%	92	50.55%
Total	261		182	

Coincidence Matrix for \$L-TARGET_B (rows show actuals)

'Partition' = 1_Training	1.000000
0.000000	124
1.000000	137
'Partition' = 2_Testing	1.000000
0.000000	92
1.000000	90

KNN:

Results for output field TARGET_B

Comparing \$KNN-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	200	76.63%	92	50.55%
Wrong	61	23.37%	90	49.45%
Total	261		182	

Coincidence Matrix for \$KNN-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	88	36
1.000000	25	112
'Partition' = 2_Testing	0.000000	1.000000
0.000000	43	49
1.000000	41	49

C&RT:

Results for output field TARGET_B

Comparing \$R-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	137	52.49%	90	49.45%
Wrong	124	47.51%	92	50.55%
Total	261		182	

Coincidence Matrix for \$R-TARGET_B (rows show actuals)

'Partition' = 1_Training	1.000000
0.000000	124
1.000000	137
'Partition' = 2_Testing	1.000000
0.000000	92
1.000000	90

C5.0:

Results for output field TARGET_B

Comparing \$C-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	177	67.82%	90	49.45%
Wrong	84	32.18%	92	50.55%
Total	261		182	

Coincidence Matrix for \$C-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	63	61
1.000000	23	114
'Partition' = 2_Testing	0.000000	1.000000
0.000000	35	57
1.000000	35	55

CHAID: Since CHAID didn't run for the selected variables.

Bayesian: Very less percentage of accuracy for the selected variables.

Results for output field TARGET_B

Comparing \$B-TARGET_B with TARGET_B

'Partition'	1_Training		2_Testing	
Correct	35	13.41%	15	8.24%
Wrong	226	86.59%	167	91.76%
Total	261		182	

Coincidence Matrix for \$B-TARGET_B (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000	\$null\$
0.000000	35	1	88
1.000000	31	0	106
'Partition' = 2_Testing	0.000000	1.000000	\$null\$
0.000000	15	1	76
1.000000	17	0	73

We calculated recall and precision for all the model developed above. Please refer to the excel 'hw8.xls'.

Sheet 'FullSet' refers to the models developed without reduction in the input variables.

Sheet 'ReducedSet' refers to the models developed with reduction in the input variable through the methods mentioned above.

With reduced set of input variables, the accuracy, recall, precision was better. So we proceed with the reduced set of input variables.

For NUMPROM<=30: Bayes Net has higher accuracy, recall and precision in the testing set. So we chose Bayes Net for this case.

For NUMPROM>30 and NUMPROM<=70: The accuracy is higher for models C5.0, Bayes Net and Logistic Regression in the order respectively. C5.0 has lesser recall, so we will not consider it. In the remaining two Bayes Net has higher recall, but the precision is less. Since the difference is very low, we chose Bayes Net for this case.

For NUMPROM>70 and totalmonths<=25: Recall and precision is balanced for CHAID and C5.0. So we consider CHAID.

For NUMPROM>70 and totalmonths>25: Recall and precision is balanced for KNN. So we consider KNN.

- (c) The purpose of using the stratified sampling is to produce a training data set with equal number of donors and non-donors. The reason for using this is to prevent the possibility of achieving good accuracy but with less number of true positives and profit but more opportunity cost. As per the data, the response rate is quite low and model does not have enough data to accurately recognize the donors and response is 5.1 % which is relatively low. In such scenarios where we have a balanced dataset would give good results to a certain extent. As our problem is related to marketing we have to focus on calculating the probability that customer would donate and build a model in line with our main focus to recognize true positives or the actual donors. There is a chance that we get low accuracy and higher sustainable profitability.

The parameters considered in addition to accuracy are Recall and Precision. Recall is considered as it helps to identify more no of donors. But as the recall is not only sufficient, we consider precision also as it helps us to identify the true positive with respect to true positive and false positive.

- (d) The profit is calculated as below

Original proportion is 5.1%

New proportion is 50%

Weighted profit= (((Number of actual 1, predicted 1) *13) - ((Number of actual 0, predicted 1) *0.68))*5.1/50

For weighted profit for each model, please refer to the excel 'hw8.xls' in sheets 'FullSet' and 'ReducedSet' column R.

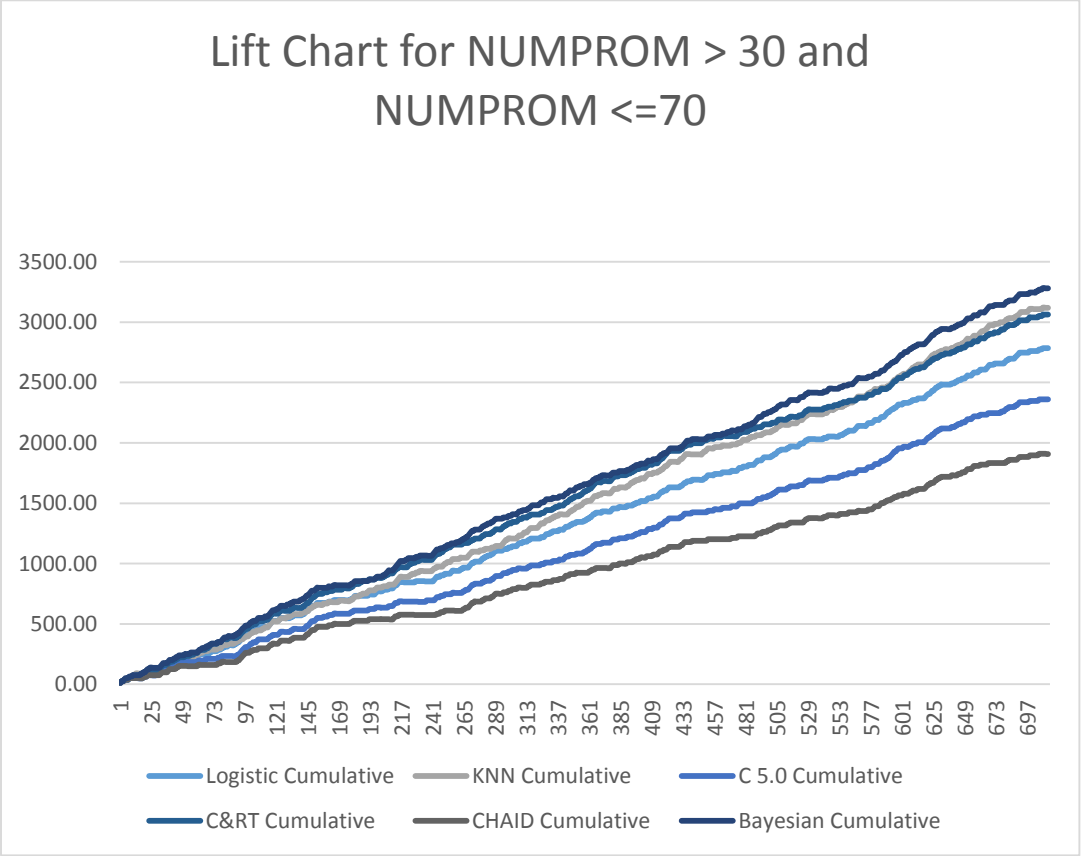
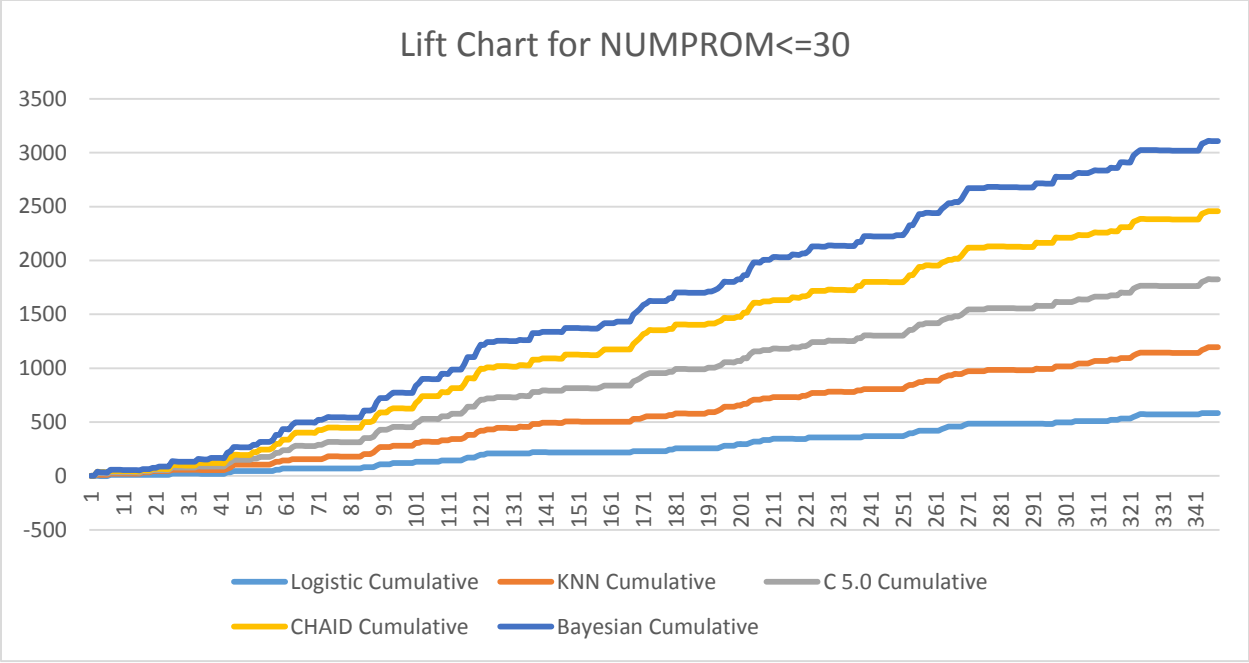
The models selected can be same with respect to profit also in the cases NUMPROM<=30, NUMPROM>30 and NUMPROM<=70, NUMPROM>70 and totalmonths<=25.

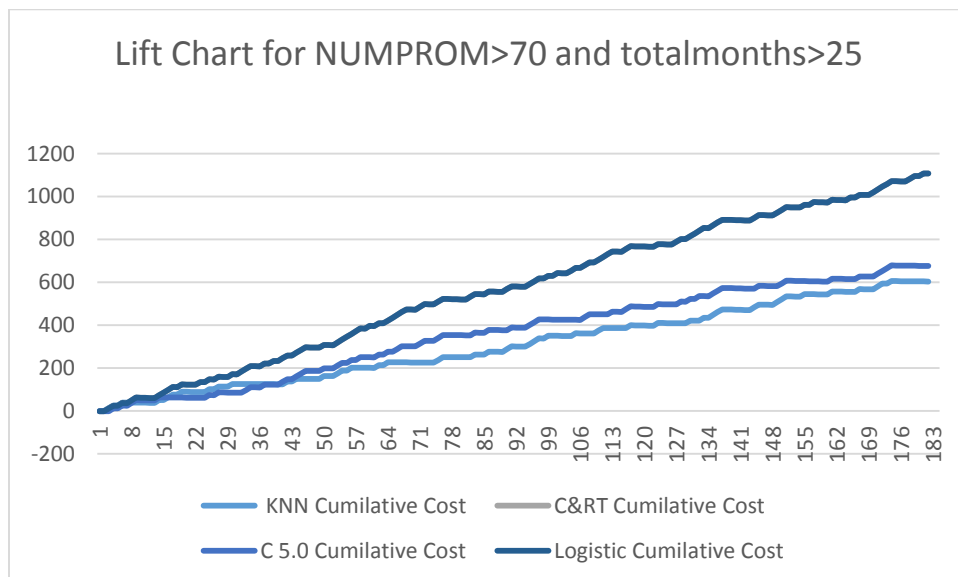
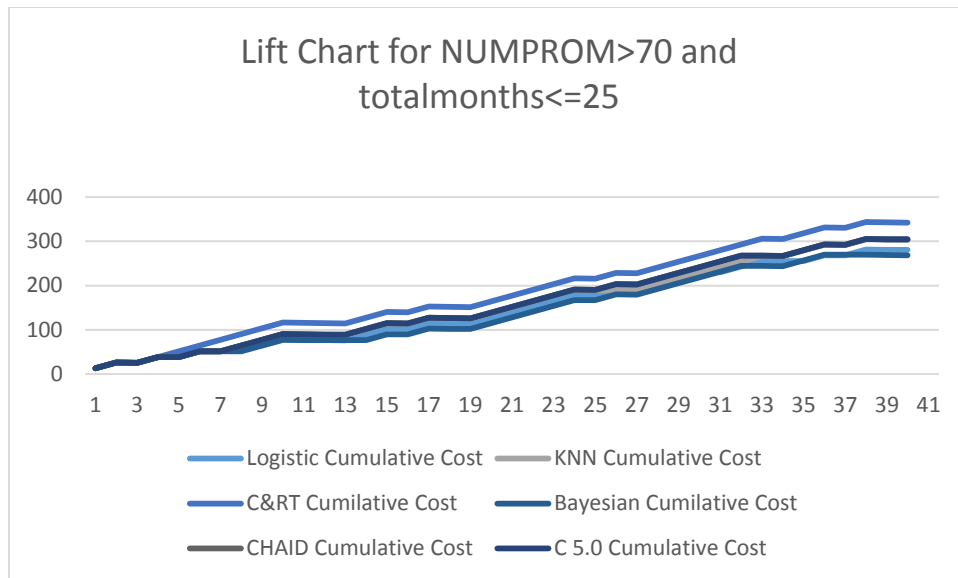
But in case NUMPROM>70 and totalmonths>25, there was huge change in the profit. If we need to choose with respect to profit also, then Logistic Regression/C&RT can be selected.

Total profit without model selected based on Profit = 66.38568 + 333.33192 + 31.06104 + 61.57536 = 492.354

Total profit with model selected based on Profit mentioned above = 66.38568 + 333.33192 + 31.06104 + 112.95888= 543.73

- (e) Charts:





(f) Best model:

For NUMPROM<=30: Bayes Net has higher accuracy, recall and precision in the testing set. The profit is also higher with value of 66.38.

For NUMPROM>30 and NUMPROM<=70: After considering profit, accuracy, recall and precision, we considered Bayes net as the best model. The profit value is 333.319.

For NUMPROM>70 and totalmonths<=25: After considering profit, accuracy, recall and precision, we considered CHAID as the best model. The profit value is 31.06. Even though this model doesn't have highest accuracy, due to recall and precision we considered this model.

For NUMPROM>70 and totalmonths>25: Since there is huge change in profit, but the accuracy difference was less, we considered total profit as the criteria. So we considered Logistic regression as the best model with profit value of 112.95.

(g) Please refer to FutureFundraising.xls for prediction results