Prasanna Rao

BINGO BONUS:

The following bingo bonus related were attempted

- Used decision trees in R to create rules. This was implemented in SAS
- Recreated as much of the program as you can in “R”
- Used few SAS Macros, but not completely.

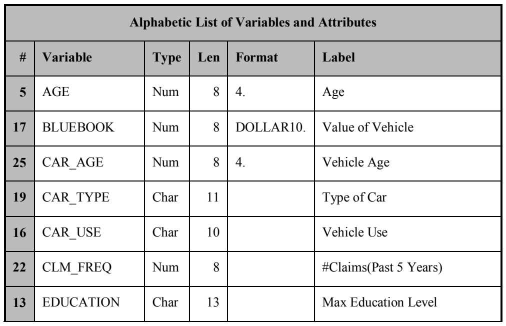
INSURANCE LOGISTIC REGRESSION PROJECT

The primary purpose of the assignment was to analyze 8161 records to predict the probability of a car crash. Secondary part of this assignment is to come up with a model to predict the cost for a person crashing their car. Logistic regression was performed using forward, backward and stepwise using series of variables deemed fit for the model based on exploratory data analysis. These model based on AIC and log likelihood was run on test data set created by using random to determine if the model was adequate to be deemed good enough to a car crash and the amount associated with each crash.

1. DATA EXPLORATION

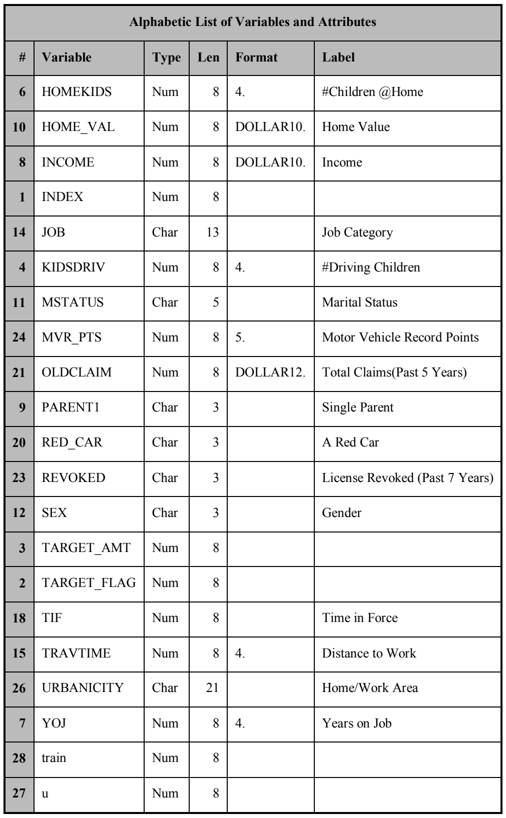
Data exploration is a critical component of data analysis. The intent of the data analysis was to understand any relationships in the underlying data, find missing values, find any outliers, influential points. The data was split into 70% training and 30 % testing data .

1) EDA: Using SAS Proc means, the following data was obtained.
   A) Attributes



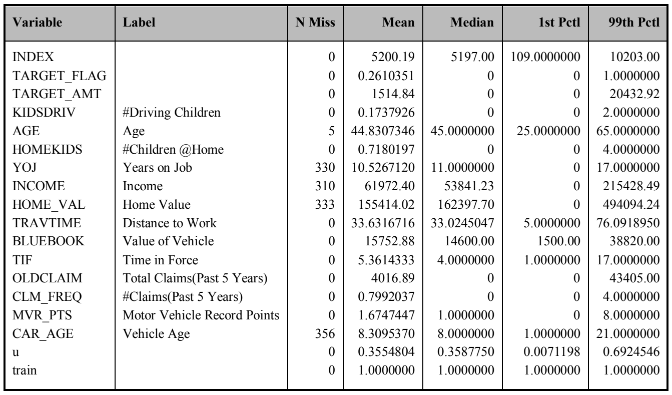
Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Label
5	AGE	Num	8	4.	Age
17	BLUEBOOK	Num	8	DOLLAR10.	Value of Vehicle
25	CAR_AGE	Num	8	4.	Vehicle Age
19	CAR_TYPE	Char	11		Type of Car
16	CAR_USE	Char	10		Vehicle Use
22	CLM_FREQ	Num	8		#Claims(Past 5 Years)
13	EDUCATION	Char	13		Max Education Level

Prasanna Rao



Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Label
6	HOMEKIDS	Num	8	4.	#Children @Home
10	HOME_VAL	Num	8	DOLLAR10.	Home Value
8	INCOME	Num	8	DOLLAR10.	Income
1	INDEX	Num	8		
14	JOB	Char	13		Job Category
4	KIDSDRIV	Num	8	4.	#Driving Children
11	MSTATUS	Char	5		Marital Status
24	MVR_PTS	Num	8	5.	Motor Vehicle Record Points
21	OLDCLAIM	Num	8	DOLLAR12.	Total Claims(Past 5 Years)
9	PARENT1	Char	3		Single Parent
20	RED_CAR	Char	3		A Red Car
23	REVOKED	Char	3		License Revoked (Past 7 Years)
12	SEX	Char	3		Gender
3	TARGET_AMT	Num	8		
2	TARGET_FLAG	Num	8		
18	TIF	Num	8		Time in Force
15	TRAVTIME	Num	8	4.	Distance to Work
26	URBANICITY	Char	21		Home/Work Area
7	YOJ	Num	8	4.	Years on Job
28	train	Num	8		
27	u	Num	8		

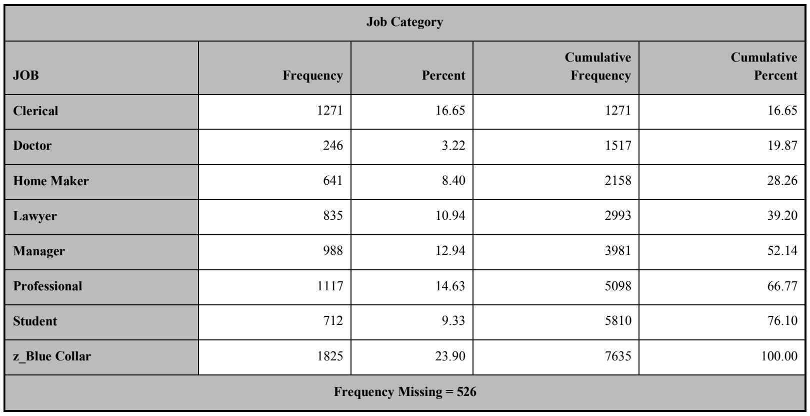
a. Mean / Standard Deviation / Median



Variable	Label	N Miss	Mean	Median	1st Pctl	99th Pctl
INDEX		0	5200.19	5197.00	109.0000000	10203.00
TARGET_FLAG		0	0.2610351	0	0	1.0000000
TARGET_AMT		0	1514.84	0	0	20432.92
KIDSDRIV	#Driving Children	0	0.1737926	0	0	2.0000000
AGE	Age	5	44.8307346	45.0000000	25.0000000	65.0000000
HOMEKIDS	#Children @Home	0	0.7180197	0	0	4.0000000
YOJ	Years on Job	330	10.5267120	11.0000000	0	17.0000000
INCOME	Income	310	61972.40	53841.23	0	215428.49
HOME_VAL	Home Value	333	155414.02	162397.70	0	494094.24
TRAVTIME	Distance to Work	0	33.6316716	33.0245047	5.0000000	76.0918950
BLUEBOOK	Value of Vehicle	0	15752.88	14600.00	1500.00	38820.00
TIF	Time in Force	0	5.3614333	4.0000000	1.0000000	17.0000000
OLDCLAIM	Total Claims(Past 5 Years)	0	4016.89	0	0	43405.00
CLM_FREQ	#Claims(Past 5 Years)	0	0.7992037	0	0	4.0000000
MVR_PTS	Motor Vehicle Record Points	0	1.6747447	1.0000000	0	8.0000000
CAR_AGE	Vehicle Age	356	8.3095370	8.0000000	1.0000000	21.0000000
u		0	0.3554804	0.3587750	0.0071198	0.6924546
train		0	1.0000000	1.0000000	1.0000000	1.0000000

Income , HOME_VAL, YOJ,AGE and CAR_AGE were found to have missing values among the numeric variables.The missing values were imputed in Data preparation stage using Decision Tress.

Categorical Variable Missing values



Job Category				
JOB	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Clerical	1271	16.65	1271	16.65
Doctor	246	3.22	1517	19.87
Home Maker	641	8.40	2158	28.26
Lawyer	835	10.94	2993	39.20
Manager	988	12.94	3981	52.14
Professional	1117	14.63	5098	66.77
Student	712	9.33	5810	76.10
z_Blue Collar	1825	23.90	7635	100.00
Frequency Missing = 526				

Among categorical variables, only JOB had missing values.

b) **Outliers**

KIDSDRIV , HOMEKIDS , YOJ , INCOME , HOME_ VAL , TRAVTIME ,BLUEBOOK , TIF , MVR_PTS , & OLDCLAIM were found to have outliers and they were imputed to their respective 99% and 1% .

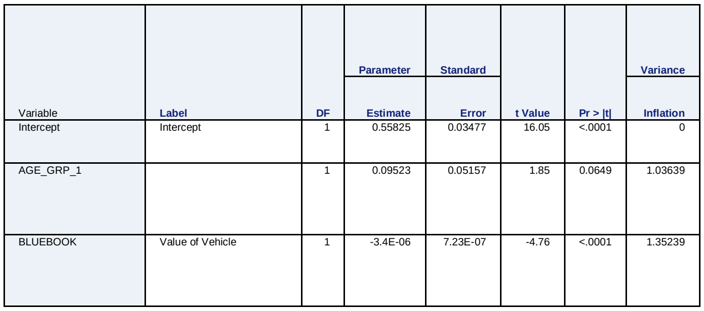
C) **Correlations:**

Moderate Correlation was found between the following variables.

a) KIDSDRIV & HOMEKIDS
b) AGE & HOMEKIDS
c) HOME_VAL & INCOME

These variables when regressed on TARGET_FLAG and then on TARGET_AMT were found to have VIF those variables were found to be negligible. Hence we can rule out correlations amongst these variables.

Correlation check



			Parameter	Standard			Variance
Variable	Label	DF	Estimate	Error	t Value	Pr > |t|	Inflation
Intercept	Intercept	1	0.55825	0.03477	16.05	<.0001	0
AGE_GRP_1		1	0.09523	0.05157	1.85	0.0649	1.03639
BLUEBOOK	Value of Vehicle	1	-3.4E-06	7.23E-07	-4.76	<.0001	1.35239

CLM_FREQ_2		1	0.12527	0.01552	8.07	<.0001	2.20168
IMPUTED_CAR_AGE		1	-0.01161	0.00208	-5.57	<.0001	2.97728
IMPUTED_HOME_VAL		1	-1.62E-07	6.23E-08	-2.6	0.0093	2.3572
IMPUTED_INCOME		1	-4.66E-07	1.87E-07	-2.5	0.0126	2.78649
IMPUTED_YOJ		1	-0.00339	0.00136	-2.48	0.0131	1.16063
KIDSDRIV	#Driving Children	1	0.06716	0.01039	6.46	<.0001	1.08805
MCAR_TYPE_1		1	-0.08317	0.01278	-6.51	<.0001	1.22374
MCAR_TYPE_3		1	-0.02373	0.01491	-1.59	0.1114	1.2228
MCAR_TYPE_4		1	0.0347	0.01775	1.96	0.0506	1.16463
MCAR_USE_2		1	-0.12488	0.01177	-10.61	<.0001	1.24215
MEDUCATION_3		1	0.05007	0.01828	2.74	0.0062	2.0911
MJOB_5		1	-0.10006	0.01513	-6.61	<.0001	1.1727
MREVOKED_N		1	-0.17244	0.01813	-9.51	<.0001	1.33772
MSTATUS_Y		1	-0.0576	0.01448	-3.98	<.0001	1.93678
MURBANICITY_1		1	0.27895	0.0141	19.79	<.0001	1.24499
MVR_PTS_2		1	0.02472	0.01271	1.94	0.0519	1.16348
MVR_PTS_3		1	0.23028	0.04033	5.71	<.0001	1.06632
OLDCLAIM	Total Claims(Past 5 Years)	1	-4.9E-06	1.1E-06	-4.44	<.0001	2.25226
MPARENT1_N		1	-0.08633	0.01803	-4.79	<.0001	1.41879
TIF	Time in Force	1	-0.00853	0.00125	-6.85	<.0001	1.00554
TRAVTIME	Distance to Work	1	0.00238	0.000332	7.19	<.0001	1.03547

D) MISSING VALUES

INCOME, HOME_VAL, YOJ, CAR_AGE, AGE and JOB were found to have missing values. These were calculated using decision TREES. The decision tree was run in R using TREE package and the logical conditions generated using decision trees was implemented in SAS.

DATA PREPARATION

a) INCOME, HOME_VAL, YOJ, CAR_AGE, AGE and JOB were found to have missing values. These were calculated using decision TREES. The decision tree was run in R using TREE package and the logical conditions generated using decision trees was implemented in SAS

INCOME Decision Tree

```
2) HOME_VAL < 256145 4884 5.481e+12 44720
 4) JOB: Clerical, Home Maker, Student 2036 7.710e+11 20970
  8) JOB: Home Maker, Student 1013 1.937e+11   8014 *
  9) JOB: Clerical 1023 2.387e+11 33810 *
 5) JOB: Doctor, Lawyer, Manager, Professional, blue Collar 2848 2.741e+12 61690
 10) EDUCATION: <High School, high School 1144 3.656e+11 47220 *
 11) EDUCATION: Bachelors, Masters, PhD 1704 1.975e+12 71410
  22) HOME_VAL < 29182.5 679 1.228e+12 89510
   44) EDUCATION: Bachelors, Masters 561 5.825e+11 82520 *
   45) EDUCATION: PhD 118 4.871e+11 122800 *
  23) HOME_VAL > 29182.5 1025 3.775e+11 59420
   46) HOME_VAL < 190553 328 7.858e+10 40560 *
   47) HOME_VAL > 190553 697 1.273e+11 68300 *
3) HOME_VAL > 256145 1161 1.523e+12 114800
 6) HOME_VAL < 364773 901 2.966e+11 99620 *
 7) HOME_VAL > 364773 260 3.013e+11 167300
 14) HOME_VAL < 452014 178 5.709e+10 150300 *
 15) HOME_VAL > 452014 82 8.067e+10 204300
```

HOME_VAL Decision Tree

```
2) MSTATUS: Yes 3597 3.802e+13 197300
   4) INCOME < 60259 2128 9.278e+12 134800
    8) JOB: Student 328 6.001e+11 19750 *
    9) JOB: Clerical, Doctor, Home Maker, Lawyer, Manager, Professional,  Blue Collar 1800
3.543e+12 155800
     18) INCOME < 28291.5 636 8.073e+11 115300 *
     19) INCOME > 28291.5 1164 1.125e+12 177900 *
   5) INCOME > 60259 1469 8.393e+12 287900
   10) INCOME < 119164 1177 1.975e+12 259000
    20) INCOME < 89820 749 6.868e+11 237800 *
    21) INCOME > 89820 428 3.557e+11 296300 *
   11) INCOME > 119164 292 1.503e+12 404000 *
 3) MSTATUS: z_No 2448 3.470e+13 80720
   6) INCOME < 74991.5 1697 1.200e+13 57580 *
   7) INCOME > 74991.5 751 1.974e+13 133000 *
```

CAR_AGE Decision Tree

```
 2) EDUCATION: <High School, z_High School 2827 33630 4.175 *
3) EDUCATION: Bachelors, Masters, PhD 3218 80300 11.210
       6) EDUCATION: Bachelors 1740 26570 8.869 *
        7) EDUCATION: Masters, PhD 1478 32940 13.970 *
```

YOJ Decision Tree

```
2) INCOME < 2.5 504      0 0.00 *
3) INCOME > 2.5 5541 42960 11.45
   6) HOMEKIDS < 1.5 4203 30240 11.07
     12) MSTATUS: Yes 2455 18380 11.55 *
     13) MSTATUS: z_No 1748 10520 10.40 *
   7) HOMEKIDS > 1.5 1338 10240 12.63 *
```

b. Transform data (by putting it into buckets)

The following continuous variables were divided into buckets

a) Age group : Young and elderly drivers are considered risky and hence they were divided into two groups

   1) AGE_GRP_1 consisting drivers from ages below 25 to ages above 65
   2) AGE_GRP_2 consisting drivers from between 25 and 65.

b) MVR_PTS: It was divided into two groups depending upon crashes and mvr points accumulated.

   1) MVR_PTS_1 consisting of points lesser than or equal to 2.
   2) MVR_PTS_2 consisting of points greater than 2.

c) CLM_FREQ : Based on claims vs crashes , CLM_FREQ was divided into two groups

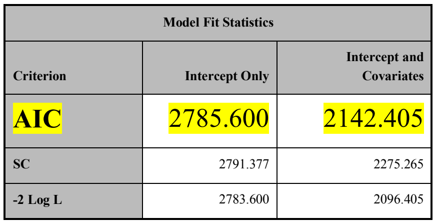
2. BUILD MODELS

Using the training data, models were built. 70% of the given data set was divided into training and testing. Using a random seed generator, training set comprised of 30 % of records whereas testing had 70 % of records. Data preparations done for testing were replicated for testing dataset.
The model generated from training were used on the testing dataset to generate the lift and gain charts.

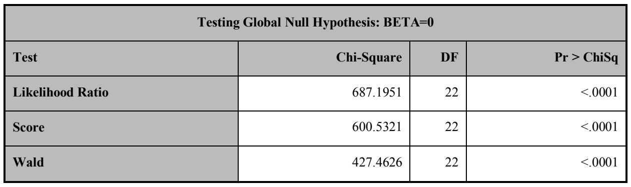
Model 1: Using backward selection all variables were used in the model. In order to use all variables to begin with, backward selection was used

```
Proc logistic data=&SCORETEST. Descending plots (only)=roc(id=prob);
 model1: model TARGET_FLAG (ref="0") =AGE_GRP_1 AGE_GRP_2 BLUEBOOK CLM_FREQ_1
    CLM_FREQ_2 HOMEKIDS IMPUTED_CAR_AGE IMPUTED_HOME_VAL IMPUTED_INCOME_1
    IMPUTED_INCOME_2 IMPUTED_INCOME_3 IMPUTED_YOJ KIDSDRIV MCAR_TYPE_1
    MCAR_TYPE_2 MCAR_TYPE_3 MCAR_TYPE_4 MCAR_TYPE_5 MCAR_TYPE_6 MCAR_USE_1
    MCAR_USE_2 MEDUCATION_1 MEDUCATION_2 MEDUCATION_3 MEDUCATION_4
    MEDUCATION_5 MGENDER_M MGENDER_N MJOB_1 MJOB_2 MJOB_3 MJOB_4 MJOB_5
    MJOB_6 MJOB_7 MJOB_8 MREVOKED_N MREVOKED_Y MSTATUS_N MSTATUS_Y
    MURBANICITY_1 MURBANICITY_2 MVR_PTS_1 MVR_PTS_2 MVR_PTS_3
    M_CAR_AGE_FLAG M_HOME_VAL_FLAG M_INCOME_FLAG M_JOB_FLAG M_YOJ_FLAG
    OLDCLAIM MPARENT1_N MPARENT1_Y TIF TRAVTIME /selection=backward
    Roceps=.1 LACKFIT;    Output out=pred1 p=phat;
```

Output AIC :

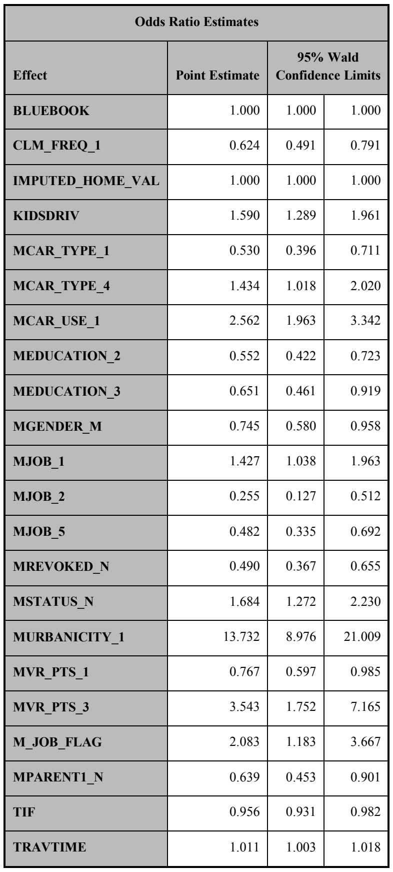


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	2785.600	2142.405
SC	2791.377	2275.265
-2 Log L	2783.600	2096.405



Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	687.1951	22	<.0001
Score	600.5321	22	<.0001
Wald	427.4626	22	<.0001

ODDS Ratio

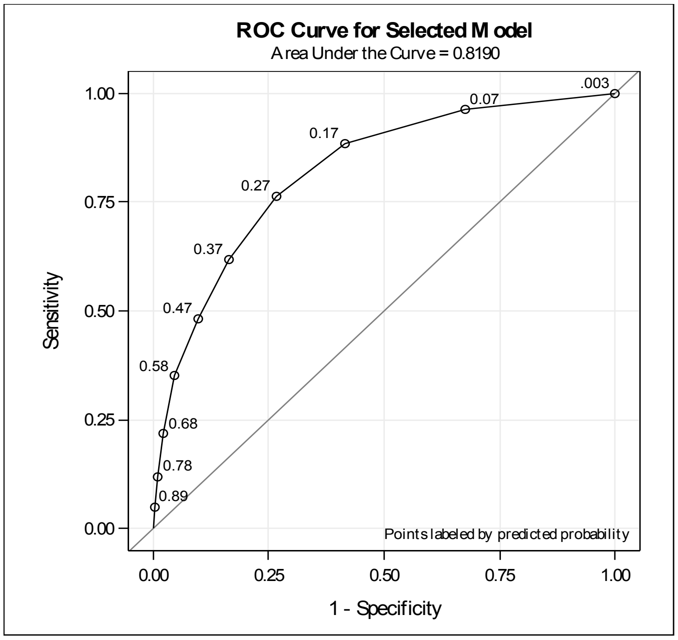


Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
BLUEBOOK	1.000	1.000	1.000
CLM_FREQ_1	0.624	0.491	0.791
IMPUTED_HOME_VAL	1.000	1.000	1.000
KIDSDRIV	1.590	1.289	1.961
MCAR_TYPE_1	0.530	0.396	0.711
MCAR_TYPE_4	1.434	1.018	2.020
MCAR_USE_1	2.562	1.963	3.342
MEDUCATION_2	0.552	0.422	0.723
MEDUCATION_3	0.651	0.461	0.919
MGENDER_M	0.745	0.580	0.958
MJOB_1	1.427	1.038	1.963
MJOB_2	0.255	0.127	0.512
MJOB_5	0.482	0.335	0.692
MREVOKED_N	0.490	0.367	0.655
MSTATUS_N	1.684	1.272	2.230
MURBANICITY_1	13.732	8.976	21.009
MVR_PTS_1	0.767	0.597	0.985
MVR_PTS_3	3.543	1.752	7.165
M_JOB_FLAG	2.083	1.183	3.667
MPARENT1_N	0.639	0.453	0.901
TIF	0.956	0.931	0.982
TRAVTIME	1.011	1.003	1.018

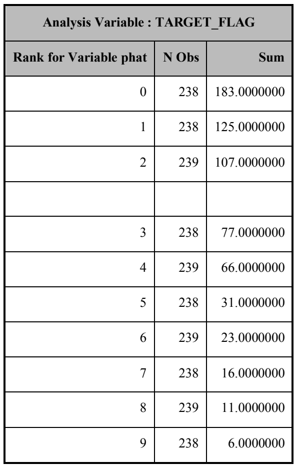
ROC Curve

The model generated above was used to generate

1) ROC curve :



2) Gains and KS statistic :
   Ranked data:



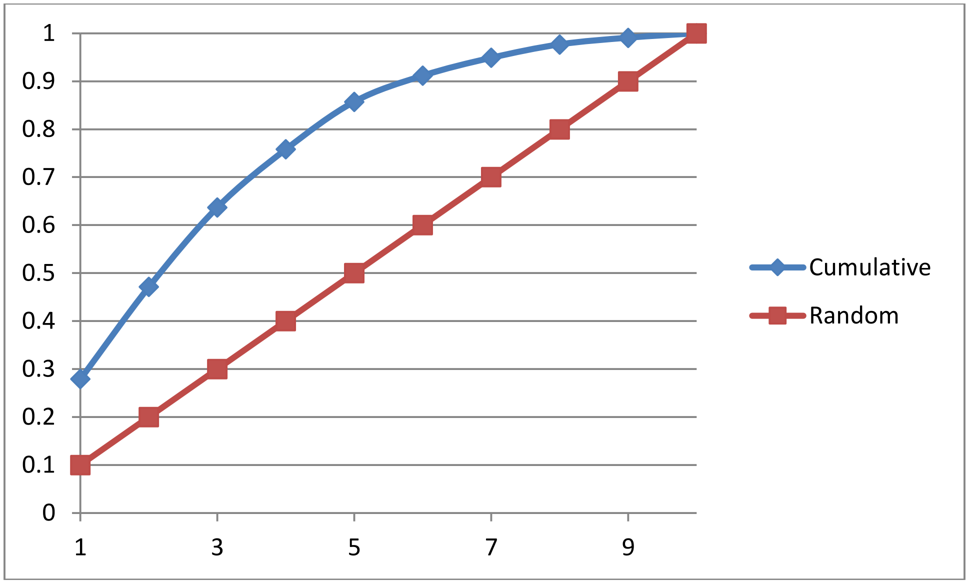
Analysis Variable : TARGET_FLAG		
Rank for Variable phat	N Obs	Sum
0	238	183.0000000
1	238	125.0000000
2	239	107.0000000
3	238	77.0000000
4	239	66.0000000
5	238	31.0000000
6	239	23.0000000
7	238	16.0000000
8	239	11.0000000
9	238	6.0000000

Gain Chart

GROUP	OBS	Responses	Total OBS	Total Responses	Theoretical Responses	LIFT	Cumulative	RANDOM	Difference

1	238	183	238	183	64.3918	2.842	0.28372	0.1	18%
2	238	125	476	308	64.3918	1.9412	0.47752	0.2	28%
3	239	107	715	415	64.3918	1.6617	0.64341	0.3	34%
4	238	77	953	492	64.3918	1.1958	0.76279	0.4	36%
5	239	66	1192	558	64.3918	1.025	0.86512	0.5	37%
6	238	31	1430	589	64.3918	0.4814	0.91318	0.6	31%
7	239	23	1669	612	64.3918	0.3572	0.94884	0.7	25%
8	238	16	1907	628	64.3918	0.2485	0.97364	0.8	17%
9	239	11	2146	639	64.3918	0.1708	0.9907	0.9	9%
10	238	6	2384	645	64.3918	0.0932	1	1	0%

KS Statistic: 37 %, lift 2.84



Model 2: Using backward selection, using the following variables as reference, logistic regression was performed on rest of the variables:

MVR_PTS_1: Since it has minimum number of points
AGE_GRP_1: Risky drivers
CLM_FREQ_1: Fewer number of crashes.
IMPUTED_INCOME_3: Maximum income group which is expected to have fewer number of crashes
MEDUCATION_4: PhD’s who have fewer number of crashes compared to rest.

```
Proc logistic data=&SCORETEST.  Descending plots (only)=roc(id=prob);

  model2: model TARGET_FLAG (ref="0")=AGE_GRP_2 BLUEBOOK CLM_FREQ_2 HOMEKIDS
    IMPUTED_CAR_AGE IMPUTED_HOME_VAL IMPUTED_INCOME_1 IMPUTED_INCOME_2
    IMPUTED_YOJ KIDSDRIV MCAR_TYPE_1 MCAR_TYPE_3 MCAR_TYPE_4 MCAR_TYPE_5
    MCAR_TYPE_6 MCAR_USE_2 MEDUCATION_1 MEDUCATION_2 MEDUCATION_3
    MEDUCATION_5 MGENDER_M MGENDER_N MJOB_1 MJOB_3 MJOB_4 MJOB_5 MJOB_6
    MJOB_7 MJOB_8 MREVOKED_N MREVOKED_Y MSTATUS_N MSTATUS_Y MURBANICITY_1
    MVR_PTS_2 MVR_PTS_3 M_CAR_AGE_FLAG M_HOME_VAL_FLAG M_INCOME_FLAG
    M_JOB_FLAG M_YOJ_FLAG OLDCLAIM MPARENT1_N MPARENT1_Y TIF TRAVTIME
    /selection=backward roceps=.1 LACKFIT;
   Output out=pred1 p=phat;
Run;
```

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	2785.600	2146.561
SC	2791.377	2290.974
-2 Log L	2783.600	2096.561

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	687.0393	24	<.0001
Score	604.5464	24	<.0001
Wald	431.0667	24	<.0001

Odds ratio for the significant variables:

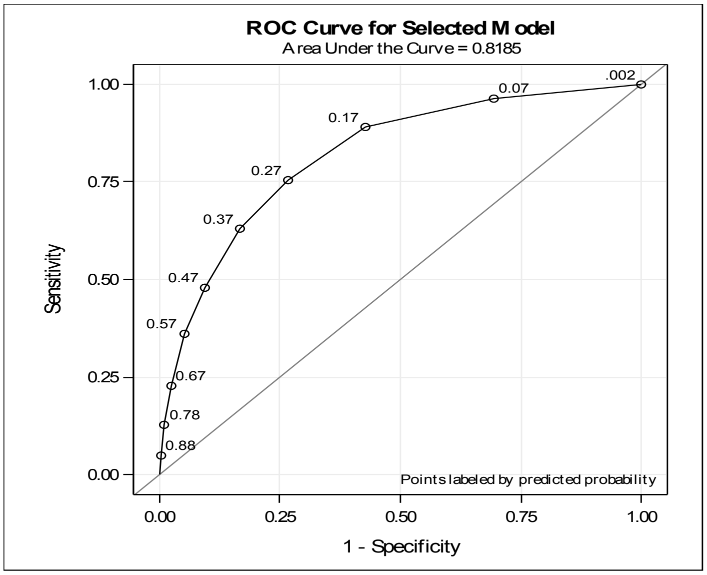
Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
AGE_GRP_2	0.153	0.063	0.374
CLM_FREQ_2	1.850	1.488	2.300
IMPUTED_HOME_VAL	1.000	1.000	1.000
KIDSDRIV	1.661	1.346	2.049
MCAR_TYPE_3	2.233	1.608	3.101
MCAR_TYPE_4	3.595	2.475	5.221
MCAR_TYPE_5	1.644	1.100	2.457
MCAR_TYPE_6	2.574	1.921	3.448
MCAR_USE_2	0.432	0.321	0.581
MEDUCATION_2	0.622	0.474	0.816
MJOB_1	4.053	2.641	6.219
MJOB_3	2.780	1.659	4.659
MJOB_4	1.771	1.114	2.816
MJOB_6	1.893	1.204	2.977
MJOB_7	3.544	2.143	5.862
MJOB_8	3.021	2.014	4.532
MREVOKED_N	0.490	0.366	0.656
MSTATUS_N	1.708	1.282	2.275
MURBANICITY_1	13.887	9.040	21.333
MVR_PTS_3	3.656	1.816	7.362
M_JOB_FLAG	1.825	1.074	3.102
MPARENT1_N	0.682	0.480	0.967

Prasanna Rao

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
TIF	0.951	0.925	0.977
TRAVTIME	1.011	1.004	1.018

The model generated above was used on testing data to generate

1) ROC curve :



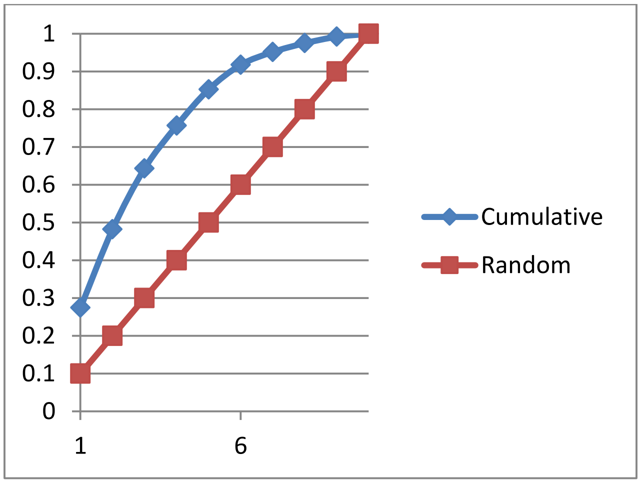
2) Ranked Data set

Rank for Variable phat	N Obs	Sum
0	238	177.0000000
1	238	134.0000000
2	239	104.0000000
3	238	73.0000000
4	239	62.0000000
5	238	42.0000000
6	239	22.0000000
7	238	15.0000000
8	239	11.0000000
9	238	5.0000000

Lift and KS statistics

GROUP	OBS	Responses	Total OBS	Total Responses	Theoretical Responses	LIFT	Cumulative	RANDOM	Difference
1	238	177	238	177	64.392	2.75	0.27442	0.1	17%
2	238	134	476	311	64.392	2.08	0.48217	0.2	28%
3	239	104	715	415	64.392	1.62	0.64341	0.3	34%
4	238	73	953	488	64.392	1.13	0.75659	0.4	36%
5	239	62	1192	550	64.392	0.96	0.85271	0.5	35%
6	238	42	1430	592	64.392	0.65	0.91783	0.6	32%
7	239	‘22	1669	614	64.392	0.34	0.95194	0.7	25%
8	238	15	1907	629	64.392	0.23	0.97519	0.8	18%
9	239	11	2146	640	64.392	0.17	0.99225	0.9	9%
10	238	5	2384	645	64.392	0.08	1	1	0%

KS statistic: 36 % and Gain: 2.75



Model 3: Model 3 was run without using buckets for the continuous variables using backward selection.

Output

Prasanna Rao

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	2785.600	2160.705
SC	2791.377	2287.789
-2 Log L	2783.600	2116.705

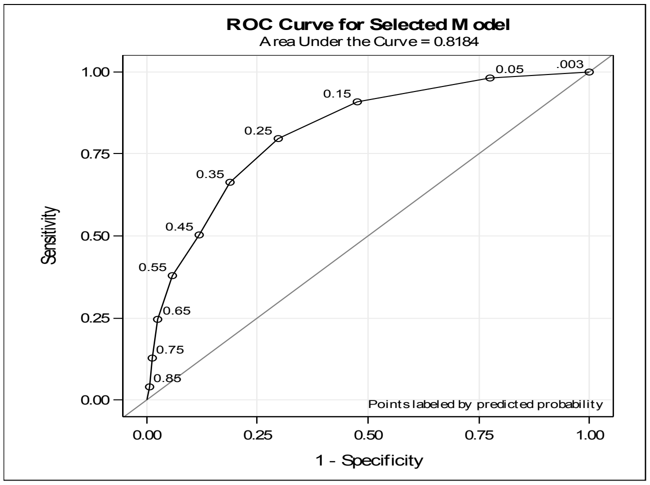
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	666.8951	21	<.0001
Score	578.7685	21	<.0001
Wald	425.5555	21	<.0001

ODDS Ratio for significant variables

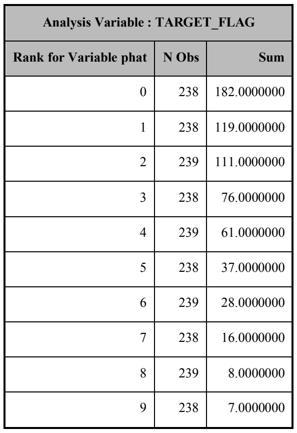
Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
BLUEBOOK	1.000	1.000	1.000
CLM_FREQ	1.149	1.045	1.264
IMPUTED_HOME_VAL	1.000	1.000	1.000
KIDSDRIV	1.598	1.297	1.970
MCAR_TYPE_1	0.535	0.399	0.717
MCAR_TYPE_4	1.500	1.068	2.107
MCAR_USE_1	2.572	1.972	3.353
MEDUCATION_2	0.551	0.421	0.720
MEDUCATION_3	0.647	0.459	0.912
MGENDER_M	0.738	0.574	0.948
MJOB_1	1.425	1.038	1.957
MJOB_2	0.257	0.128	0.513
MJOB_5	0.484	0.337	0.696
MREVOKED_N	0.486	0.364	0.649
MSTATUS_N	1.685	1.274	2.229
MURBANICITY_1	14.332	9.381	21.894
MVR_PTS	1.132	1.077	1.190
M_JOB_FLAG	2.077	1.184	3.644
MPARENT1_N	0.650	0.461	0.916
TIF	0.957	0.932	0.983
TRAVTIME	1.011	1.004	1.018

The model generated above was used on testing data to generate

1) ROC curve :



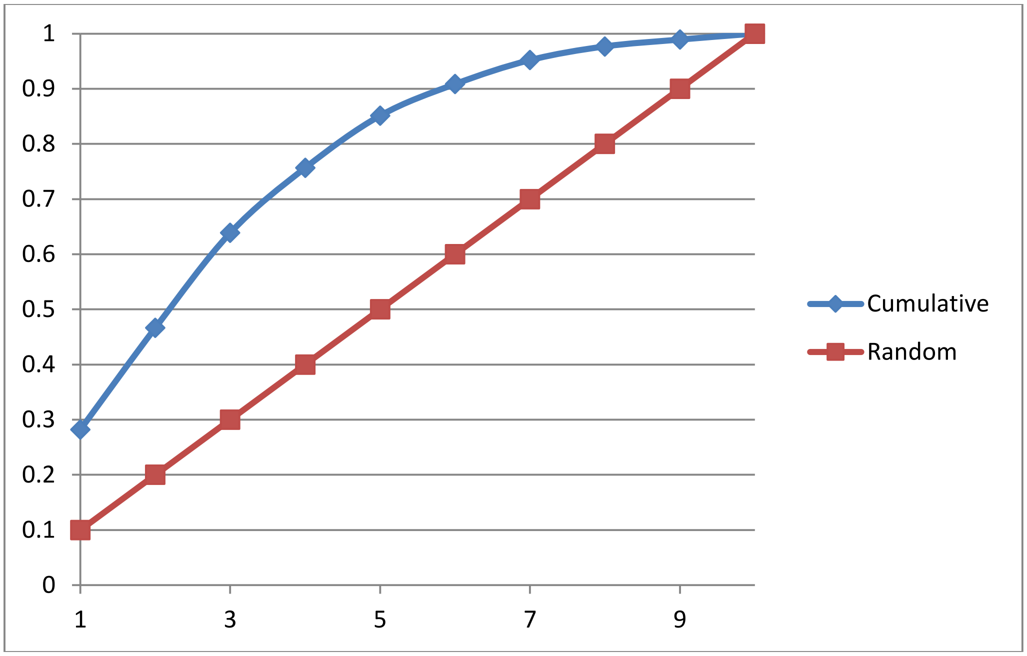
2) Ranked Data set



Analysis Variable : TARGET_FLAG		
Rank for Variable phat	N Obs	Sum
0	238	182.0000000
1	238	119.0000000
2	239	111.0000000
3	238	76.0000000
4	239	61.0000000
5	238	37.0000000
6	239	28.0000000
7	238	16.0000000
8	239	8.0000000
9	238	7.0000000

GAIN and KS statistic

GROUP	OBS	Responses	Total OBS	Total Responses	Theoretical Responses	LIFT	Cumulative	RANDOM	Difference
1	238	182	238	182	64.3918	**2.8264**	0.28217	0.1	0.18217
2	238	119	476	301	64.3918	1.8481	0.46667	0.2	0.26667
3	239	111	715	412	64.3918	1.7238	0.63876	0.3	0.33876
4	238	76	953	488	64.3918	1.1803	0.75659	0.4	0.35659
5	239	61	1192	549	64.3918	0.9473	0.85116	0.5	**0.35116**
6	238	37	1430	586	64.3918	0.5746	0.90853	0.6	0.30853
7	239	28	1669	614	64.3918	0.4348	0.95194	0.7	0.25194
8	238	16	1907	630	64.3918	0.2485	0.97674	0.8	0.17674
9	239	8	2146	638	64.3918	0.1242	0.98915	0.9	0.08915
10	238	7	2384	645	64.3918	0.1087	1	1	0



Model Selection

Model	AIC	Log Likelihood	LIFT	KS	ROC
1	2142	687.1951	2.84	37%	**81.90%**
2	2146	687.03	2.75	36%	81.85%
3	2160	666.8951	2.82	36%	81.84%

Model 1 was chosen based on

a) **Lower of AIC among the models**

b) **Higher Log likelihood ratio among the three models**

c) **Higher lift, higher KS and larger area under ROC**

Selected Model odds Analysis:

Y= 1.4559 -0.00003 * BLUEBOOK -0.4723 * CLM_FREQ_1 -.000001776 *IMPUTED_HOME_VAL
+0.4638 * KIDSDRIV - 0.6341 *MCAR_TYPE_1 + 0.3606 * MCAR_TYPE_4
+0.9406 *MCAR_USE_1 -0.5938*MEDUCATION_2 -0.4296*MEDUCATION_3
-0.2938*MGENDER_M + 0.3558 * MJOB_1 -1.3653*MJOB_2
-0.7305*MJOB_5 -0.7134*MREVOKED_N+ 0.5214*MSTATUS_N
+ 2.6197*MURBANICITY_1 -0.2652*MVR_PTS_1
+1.265*MVR_PTS_3+ 0.7339*M_JOB_FLAG -0.4474*MPARENT1_N

-0.0448 *TIF +0.0105*TRAVTIME;

Variable		**Reasoning**
BLUEBOOK	1	
CLM_FREQ_1	0.624	This makes sense as those who have had no claims would have 62.4 lesser chance of having a crash compared to those wo have claims(Having claims is reference)
IMPUTED_HOME_VAL	1	
KIDSDRIV	1.59	
MCAR_TYPE_1	0.53	Minivans have a lesser probability of having an accident compared to a SUV(SUV is reference)
MCAR_TYPE_4	1.434	Sports car has 43 percent more chance having a crash compared to a SUV.
MCAR_USE_1	2.562	Commercial car user has 156% chance having an accident compared to a private car (private car id reference)
MEDUCATION_2	0.552	Those having Bachelors have 44.8% lesser chance of an accident compared to senior High school
MEDUCATION_3	0.651	Those having Bachelors have 34.9% lesser chance of an accident compared to senior High school
MGENDER_M	0.745	Male's have a 25.5% lesser chance compared to females(reference is females)
MJOB_1	1.427	Those having clerical posts have 42.7% chances of having acident compared to Blue collared jobs(reference is blue collared)
MJOB_2	0.255	Doctors have 74.6% lesser chances of having accident compared to Blue collared jobs(reference is blue collared). This is in synch with conventional wisdom
MJOB_5	0.482	Home makers have 51.8% lesser chances of having accident compared to Blue collared jobs(reference is blue collared).This is in sync with conventional wisdom
MREVOKED_N	0.49	Those who have not their license revoked have 51 % lesser chance of crashing their car
MSTATUS_N	1.684	Those who are unmarried have 68.4% chances of crashing car compared to married
MURBANICITY_1	13.732	Those in highly urban area have a 1272% more chance of accidents compared to those in rural areas

MVR_PTS_1	0.767	Those having lesser tickets have 33.3% chance of lesser crashes.
MVR_PTS_3	3.543	Those having MORE tickets have 254% more chance of lesser crashes .
M_JOB_FLAG	2.083	
MPARENT1_N	0.639	Those who are not single parents have 36.1%of lesser chances of crashes compared to single parets
TIF	0.956	
TRAVTIME	1.011	

STAND ALONE SCORING PROGRAM

```
 libname p411 '/folders/myfolders/411' ;
%let INFILE            =  p411.logit_insurance_test ;
%let INFILE1       =  p411.logit_insurance;
%LET SCORETEST  =  P411.SCORETEST;
%LET  SCORED    =  P411.SCORED;
%LET  SCOREAMT    =  P411.SCOREAMT;

   data &SCORETEST.;
    SET &INFILE.;
   length IMPUTED_JOB  $ 13;
   IF KIDSDRIV > 3 THEN
      KIDSDRIV=3;
   IF AGE      > 64 THEN
      AGE=64;
   IF HOMEKIDS > 4 THEN
      HOMEKIDS=4;
   IF YOJ      > 19 THEN
      YOJ=19;
   IF INCOME   > 200000 THEN
      INCOME=200000;
   IF HOME_VAL > 500309 THEN
      HOME_VAL=500309;
   IF TRAVTIME > 75 THEN
      TRAVTIME=75;
   IF BLUEBOOK > 39090 THEN
      BLUEBOOK=39090;
```

```
  IF TIF     > 17 THEN
    TIF=17;
  IF MVR_PTS  >  8 THEN
    MVR_PTS=8;
  IF OLDCLAIM >  30000 THEN
    OLDCLAIM=27090;
   IMPUTED_AGE=AGE;
   M_AGE_FLAG=0;
  IF MISSING (AGE) THEN
   DO;       IMPUTED_AGE = 44.5 ;
          M_AGE_FLAG=1;
   END;
  * Missing value imputation ;
  ****Income*********************************** ;

   IMPUTED_HOME_VAL = HOME_VAL;
   IMPUTED_YOJ = YOJ;
   IMPUTED_JOB = JOB;
   IF MISSING (HOME_VAL) THEN  IMPUTED_HOME_VAL = 158000;
   IF MISSING (YOJ) THEN  IMPUTED_YOJ = 10.3;
   IF MISSING (JOB) THEN IMPUTED_JOB ='z_Blue Collar';
   IMPUTED_INCOME = INCOME;


  *IMPUTED_INCOME = INCOME;
  M_INCOME_FLAG=0;
  IF MISSING(INCOME) THEN DO;
  M_INCOME_FLAG=1;
  IF IMPUTED_HOME_VAL< 266569 AND IMPUTED_JOB IN ("Home Maker","Student")  THEN
   IMPUTED_INCOME = 8367 ;
  IF IMPUTED_HOME_VAL< 266569 AND IMPUTED_JOB IN ("Clerical")  THEN
   IMPUTED_INCOME = 33544 ;
  IF IMPUTED_HOME_VAL< 266569 AND IMPUTED_JOB IN
("Doctor","Lawyer","Manager","Professional","z_Blue Collar")
    AND EDUCATION IN ("<High School","z_High School")
    THEN  IMPUTED_INCOME = 47954.53 ;

  IF IMPUTED_HOME_VAL< 266569 AND
    IMPUTED_JOB IN ("Doctor","Lawyer","Manager","Professional","z_Blue Collar")
    AND    EDUCATION IN ("Bachelors","Masters")
    AND  IMPUTED_HOME_VAL >29182.5
      THEN
         IF IMPUTED_HOME_VAL < 206100 THEN
           IMPUTED_INCOME=46130 ;
         ELSE
            IMPUTED_INCOME=73734.400 ;


  IF IMPUTED_HOME_VAL< 266569 AND
    IMPUTED_JOB IN ("Doctor","Lawyer","Manager","Professional","z_Blue Collar") AND
    EDUCATION IN ("Bachelors","Masters")
    AND IMPUTED_HOME_VAL <=29182.5  THEN  IMPUTED_INCOME=84714.750 ;


  IF IMPUTED_HOME_VAL< 266569 AND
    IMPUTED_JOB IN ("Doctor","Lawyer","Manager","Professional","z_Blue Collar")
    AND EDUCATION IN ("PhD")  AND
    IMPUTED_HOME_VAL>=29159
    THEN  IMPUTED_INCOME = 74369.170 ;

  IF IMPUTED_HOME_VAL< 266569 AND
    IMPUTED_JOB IN ("Doctor","Lawyer","Manager","Professional","z_Blue Collar")
    AND EDUCATION IN ("PhD")  AND IMPUTED_HOME_VAL < 29159
    THEN  IMPUTED_INCOME = 132882.90 ;

  IF IMPUTED_HOME_VAL>=266569 AND IMPUTED_HOME_VAL < 324581   THEN
IMPUTED_INCOME=97729.140;
  IF IMPUTED_HOME_VAL>=324581 AND   IMPUTED_HOME_VAL< 400853 THEN
IMPUTED_INCOME=126747.50;
  IF IMPUTED_HOME_VAL>=400853 AND   IMPUTED_HOME_VAL< 509381 THEN
IMPUTED_INCOME=172639.10;
  IF IMPUTED_HOME_VAL>=509381 THEN IMPUTED_INCOME=172639.10 ;
  END;
  ******HOME VALUE******************;

   IMPUTED_HOME_VAL = HOME_VAL;
   IMPUTED_YOJ = YOJ;
   IMPUTED_JOB = JOB;

   IF MISSING (YOJ) THEN  IMPUTED_YOJ = 10.3;
   IF MISSING (JOB) THEN IMPUTED_JOB ='z_Blue Collar';


   M_HOME_VAL_FLAG=0;
  IF MISSING(HOME_VAL) THEN DO;
   M_HOME_VAL_FLAG=1;
  IF MSTATUS='z_No' AND  IMPUTED_INCOME< 74991.5 THEN IMPUTED_HOME_VAL = 58331.53;
  IF MSTATUS='z_No' AND  IMPUTED_INCOME >= 74991.5 THEN IMPUTED_HOME_VAL = 130901.50;
  IF MSTATUS="Yes"  AND  IMPUTED_INCOME < 63991 AND IMPUTED_JOB='Student' THEN
IMPUTED_HOME_VAL=139519.30;
  IF MSTATUS="Yes"  AND  IMPUTED_INCOME < 63991 AND
   IMPUTED_JOB IN ("Clerical","Doctor","Home Maker","Lawyer","Manager","Professional","z_Blue Collar")
    THEN  DO ; IF IMPUTED_INCOME <  28565

           THEN IMPUTED_HOME_VAL= 115548.30;
          ELSE
           IMPUTED_HOME_VAL= 181307.80 ;
       END;
 IF MSTATUS="Yes"  AND  IMPUTED_INCOME >= 63991     AND IMPUTED_INCOME  < 90699.5     THEN
IMPUTED_HOME_VAL=237957.90;
```

```
  IF MSTATUS="Yes"  AND  IMPUTED_INCOME >= 90699.5   AND IMPUTED_INCOME  < 130387  THEN
IMPUTED_HOME_VAL=303154.70;
  IF MSTATUS="Yes"  AND  IMPUTED_INCOME >= 130387      THEN IMPUTED_HOME_VAL=425040.60;
  END;

****************************CAR AGE**********************  ;
     M_CAR_AGE_FLAG = 0 ;
     IF MISSING(CAR_AGE) THEN
     M_CAR_AGE_FLAG = 1 ;
     DO;
     IF EDUCATION ='PhD' THEN IMPUTED_CAR_AGE = 13.74 ;
     ELSE
     IF EDUCATION = 'Masters' THEN IMPUTED_CAR_AGE =14.05;
     ELSE
     IF EDUCATION = 'Bachelors' THEN IMPUTED_CAR_AGE = 8.91;
     ELSE
     IF EDUCATION = '<High School' THEN  IMPUTED_CAR_AGE =3.48;
     ELSE
     IMPUTED_CAR_AGE = 4.16;
     END;
 ****************************** JOB PREDICTION BASED ON DECISION TREE***** ;
    IMPUTED_JOB = JOB;
     M_JOB_FLAG = 0;
   IF MISSING (JOB) THEN  DO;
    M_JOB_FLAG = 1;
    IF EDUCATION ="PhD" THEN IMPUTED_JOB = "Doctor" ;
    IF EDUCATION ="Masters" AND CAR_USE ="Private"  THEN IMPUTED_JOB = "Lawyer" ;
    IF EDUCATION ="Masters" AND CAR_USE ="Commercial"  THEN IMPUTED_JOB = "Manager" ;
    IF EDUCATION  IN ("<High School","Bachelors","z_High School") AND IMPUTED_INCOME< 12824.5
       AND  IMPUTED_HOME_VAL >= 25111.5  THEN IMPUTED_JOB ="Home Maker";
    IF EDUCATION  IN ("<High School","Bachelors","z_High School") AND IMPUTED_INCOME< 12824.5
       AND  IMPUTED_HOME_VAL < 25111.5   THEN IMPUTED_JOB ="Student";
    IF EDUCATION  IN ("<High School","Bachelors","z_High School") AND IMPUTED_INCOME>= 12824.5
       AND CAR_USE="Private" AND  IMPUTED_INCOME <  52373.5  THEN IMPUTED_JOB ="Clerical";
   IF EDUCATION  IN ("<High School","Bachelors","z_High School") AND IMPUTED_INCOME>= 12824.5
       AND CAR_USE="Private" AND  IMPUTED_INCOME >=52373.5  AND
       EDUCATION IN ("Bachelors","z_High School")  THEN IMPUTED_JOB = "Professional";
   IF EDUCATION  IN ("<High School","Bachelors","z_High School") AND IMPUTED_INCOME>= 12824.5
       AND CAR_USE="Private" AND  IMPUTED_INCOME >=52373.5  AND
       EDUCATION ="<High School"  THEN IMPUTED_JOB = "z_Blue Collar";
   IF EDUCATION  IN ("<High School","Bachelors","z_High School") AND IMPUTED_INCOME>= 12824.5
       AND CAR_USE="Commercial"  THEN IMPUTED_JOB = "z_Blue Collar";
   END;

   ************************YOJ YOJ ***8 ;
   M_YOJ_FLAG=0;
  IMPUTED_YOJ=YOJ ;
  IF MISSING(YOJ) THEN DO; M_YOJ_FLAG=1;
  IF IMPUTED_INCOME < 2.5 THEN IMPUTED_YOJ = 0;
  IF IMPUTED_INCOME >  2.5 AND HOMEKIDS < 2 AND MSTATUS = "Yes" THEN IMPUTED_YOJ=11.55 ;
  IF IMPUTED_INCOME >  2.5 AND HOMEKIDS < 2 AND MSTATUS = "z_No" THEN IMPUTED_YOJ=10.40 ;
  IF IMPUTED_INCOME >  2.5 AND HOMEKIDS >= 2   THEN IMPUTED_YOJ=10.40 ;
   END;
   RUN;


data &SCORETEST.;
    set &SCORETEST.;

    MSTATUS_Y   =           MSTATUS in ('Yes');
    MSTATUS_N   =           MSTATUS in ('z_No');

    MGENDER_M=              SEX in ('M');
    MGENDER_N =             SEX in ('z_F');

    MEDUCATION_1=           EDUCATION  in ('<High School');
    MEDUCATION_2=           EDUCATION  in ('Bachelors');
    MEDUCATION_3=           EDUCATION  in ('Masters');
    MEDUCATION_4=           EDUCATION  in ('PhD');
    MEDUCATION_5=           EDUCATION  in ('z_High School');

    MCAR_USE_1 =            CAR_USE  in ('Commercial');
    MCAR_USE_2 =            CAR_USE  in ('Private');

    MCAR_TYPE_1 =           CAR_TYPE  in ('Minivan');
    MCAR_TYPE_2 =           CAR_TYPE  in ('Panel Truck');
    MCAR_TYPE_3 =           CAR_TYPE  in ('Pickup');
    MCAR_TYPE_4 =           CAR_TYPE  in ('Sports Car');
    MCAR_TYPE_5 =           CAR_TYPE  in ('Van');
    MCAR_TYPE_6 =           CAR_TYPE  in ('z_SUV');

    MREVOKED_Y  =  REVOKED IN ('Yes');
    MREVOKED_N  =  REVOKED IN ('No');

    MURBANICITY_1 =URBANICITY IN ('Highly Urban/ Urban');
    MURBANICITY_2 =URBANICITY IN ('z_Highly Rural/ Rural');

    MJOB_1        = IMPUTED_JOB IN ('Clerical');
    MJOB_2        = IMPUTED_JOB IN ('Doctor');
    MJOB_3        = IMPUTED_JOB IN ('Home Maker');
    MJOB_4        = IMPUTED_JOB IN ('Lawyer');
    MJOB_5        = IMPUTED_JOB IN ('Manager');
    MJOB_6        = IMPUTED_JOB IN ('Professional');
    MJOB_7        = IMPUTED_JOB IN ('Student');
    MJOB_8        = IMPUTED_JOB IN ('z_Blue Collar');

    MPARENT1_N     = PARENT1 IN ('No');
    MPARENT1_Y    = PARENT1 IN ('Yes');

    RUN;
```

```
 DATA &SCORETEST.;
   SET &SCORETEST.;
      AGE_GRP_1 = 0;
      AGE_GRP_2 = 0 ;

      MVR_PTS_1 = 0;
      MVR_PTS_2=0;
      MVR_PTS_3=0;

      CLM_FREQ_2=0 ;
      CLM_FREQ_1=0;

      IMPUTED_INCOME_1 = 0;
      IMPUTED_INCOME_2 = 0;
      IMPUTED_INCOME_3 = 0;

      IF IMPUTED_INCOME <= 60000 THEN IMPUTED_INCOME_1=1 ;
      IF IMPUTED_INCOME > 60000 AND  IMPUTED_INCOME < 135000
       THEN IMPUTED_INCOME_2 =1 ;
      IF IMPUTED_INCOME >= 135000 THEN  IMPUTED_INCOME_3= 1;


      IF    IMPUTED_AGE <= 25    THEN   AGE_GRP_1 = 1 ;
      IF    IMPUTED_AGE > 25     AND    IMPUTED_AGE <= 65  THEN AGE_GRP_2= 1 ;
      IF    IMPUTED_AGE >65      THEN   AGE_GRP_1= 1 ;


     IF MVR_PTS <=2 THEN MVR_PTS_1 = 1   ;
     IF MVR_PTS > 2 AND  MVR_PTS < 7 THEN MVR_PTS_2 = 1   ;

     IF  MVR_PTS > 7   THEN MVR_PTS_3 = 1   ;


      IF  CLM_FREQ > 0 THEN CLM_FREQ_2 = 1 ;ELSE    CLM_FREQ_1=1;

      AGETRV = imputed_AGE*TRAVTIME;
      AGEPTS = IMPUTED_AGE*MVR_PTS;

      clmmulti=OLDCLAIM*CLM_FREQ;

RUN;


  DATA &SCORED. ;
      SET &SCORETEST.;

     TEMP= -1.4559 -0.00003 * BLUEBOOK -0.4723 * CLM_FREQ_1 -.000001776 *IMPUTED_HOME_VAL
         +0.4638 * KIDSDRIV - 0.6341 *MCAR_TYPE_1 + 0.3606 * MCAR_TYPE_4
         +0.9406 *MCAR_USE_1 -0.5938*MEDUCATION_2 -0.4296*MEDUCATION_3
        -0.2938*MGENDER_M + 0.3558 * MJOB_1 -1.3653*MJOB_2
        -0.7305*MJOB_5 -0.7134*MREVOKED_N+ 0.5214*MSTATUS_N
         + 2.6197*MURBANICITY_1 -0.2652*MVR_PTS_1
         +1.265*MVR_PTS_3+ 0.7339*M_JOB_FLAG -0.4474*MPARENT1_N
         -0.0448 *TIF +0.0105*TRAVTIME;

      AMT = 3710.08 +.11604*BLUEBOOK +140.9685*MVR_PTS;

      YHAT = exp(TEMP);
      PROB = YHAT / (1.0+YHAT);


      P_TARGET_FLAG = PROB;
      P_TARGET_AMT  = AMT;


      KEEP INDEX P_TARGET_FLAG P_TARGET_AMT ;
 RUN;
```

SCORED DATA FILE

Scored file sent along with the email : Scoredsas7.bdat