A CASE STUDY:
LOGISTIC REGRESSION

Methodology Training Document (Module 7)

YEAR 2015

If '**A**' Is Client, '**B**' Is To Be You…

I. COMPLICATIONS

*'**A**' witnesses losses in business*

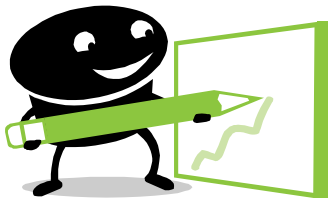
*'**A**' plans to investigate*

Things look like a black box !

II. SOLUTIONS

There strikes an idea !

*'**A**' seeks help from '**B**'*



*'**B**' provides a solution*

III. BLISS POINTS

*'**A**' gets desired results*

*'**B**' becomes a super-hero*

*How '**B**' comes up with a solution is now the area of focus*

Problem Statement

A departmental store owner experiences decline in total sales over a period of time. He wants to identify the segment of customers who are more likely to buy a product from his store in the next month. He seeks a modeler's help.

Two datasets are being provided by store-owner for the purpose of analysis and validation of results.

Modeling Dataset

(400 customers, 17 variables)

Validation Dataset

(100 customers, 17 variables)

S. NO.	VARIABLE	DESCRIPTION
1	CUSTOMER_ID	Customer Identification Number
2	IND_BUY	Takes value 1 if customer has purchased a product in current month, else takes value 0
3	IND_LOW_INCOME	Takes value 1 if customer belongs to low income group, else takes value 0
4	IND_ACTIVE_1M	Takes value 1 if customer has been tagged as 'active' in the last month, else takes value 0
5	IND_CAR_OWNER	Takes value 1 if customer owns a car, else takes value 0
6	NUM_POOR_FEEDBACK_1M	Number of times customer has filled up 'poor' feedback in the last one month
7	NUM_PURCHASES_1M	Number of purchases made by the customer in past one month
8	CUMM_DISC_AMT_1M	Cumulative amount of discount till last month as received by the customer
9	PREV_BUY_BILL_AMT	Bill amount of last purchase
10	NUM_FREEBIES_RECD_1M	Number of freebies received by customer in last one month
11	NUM_VISITS_2M	Total number of visits by customer in last 2 months
12	NUM_VISITS_3M	Total number of visits by customer in last 3 months
13	EST_SPENDING_LIMIT	Estimated spending limit of customer
14	DAYS_TO_NEXT_BDAY	Number of days left in nearest birthday in customer household
15	NUM_KIDS	Number of kids present in customer household
16	MARITAL_STATUS	M=Married, U=Unmarried, Blank=No Information Available
17	IND_CASH_PURCHASE	Takes value 1 if customer has made a purchase in current month by cash payment, else takes value 0

EDD Analysis

EDD Macro Syntax

Note: EDD macro will not work for variables with name greater than 28 bytes

```
%EDD (
      INLIB             = <Location of the input dataset>,
      INPUTDATA         = <Name of input dataset>,
      EDD_OUT_LOC_XLS   = <Location and name of output XLS file>,
      OUTLIB            = <Location of the output dataset>,
      OUTDATA           = <Name of output dataset>,
      NUM_UNIQ          = <Option*>
      );
*NUM_UNIQ can either be Y or N depending on whether the # of unique
values column is desired.
```

After a glance at EDD Output . . .

- *Character format*
- *Possibly an indicator (**Note:** Num. of unique values = 2, Min. = 0 and Max. = 1)*
- *Num. of unique values = Num. of obs.*
- *Missing values present (**Note**: NMISS > 0)*
- *Outliers present (**Note**: Max. / P99 ratio is significantly high)*

EDD Output of Modeling Dataset

Event Rate = 5%

VARIABLE	TYPE	# OBS.	NMISS	UNIQUE	MEAN	STD DEV	MIN	P1	P5	P25	MEDIAN	P75	P95	P99	MAX
CUSTOMER_ID	NUM	400	0	400	495	299.69	1	14	54	216	493	772	950.5	993.5	999
IND_BUY	NUM	400	0	2	0.05	0.22	0	0	0	0	0	0	0.5	1	1
IND_LOW_INCOME	NUM	400	0	2	0.55	0.50	0	0	0	0	1	1	1	1	1
IND_ACTIVE_1M	NUM	400	0	2	0.08	0.27	0	0	0	0	0	0	1	1	1
IND_CAR_OWNER	NUM	400	0	2	0.01	0.11	0	0	0	0	0	0	0	1	1
NUM_POOR_FEEDBACK_1M	NUM	400	0	6	0.52	0.96	0	0	0	0	0	1	3	4	5
NUM_PURCHASES_1M	NUM	400	0	27	3.35	6.51	0	0	0	0	1	4	14.5	26	80
CUMM_DISC_AMT_1M	NUM	400	6	56	1805	883.70	0	0	0	1250	1825	2500	3250	3750	4250
PREV_BUY_BILL_AMT	NUM	400	0	54	752	731.23	300	350	443.75	625	725	816.5	1000	1050	15000
NUM_FREEBIES_RECD_1M	NUM	400	0	7	0.39	0.77	0	0	0	0	0	1	2	3.5	6
NUM_VISITS_2M	NUM	400	0	22	3.12	4.23	0	0	0	1	2	4	11.5	19	31
NUM_VISITS_3M	NUM	400	0	24	3.66	4.75	0	0	0	1	2	4	13.5	23	35
EST_SPENDING_LIMIT	NUM	400	0	278	1432	1361.10	20	47.5	157.5	610	1137.5	1737.5	3675	7670	11970
DAYS_TO_NEXT_BDAY	NUM	400	0	62	19	14.39	0	0	4	10	16	24	49	69	79
NUM_KIDS	NUM	400	0	4	1.13	0.43	1	1	1	1	1	1	2	3	4
MARITAL_STATUS	CHAR	400	72	3	M::231	U::97	::72								
IND_CASH_PURCHASE	NUM	400	0	2	0.05	0.22	0	0	0	0	0	0	0.5	1	1

Variable Classification

PRIMARY KEY

CUSTOMER_ID

Necessary Condition: It's number of unique values equals the total number of records in the dataset.

Sufficient Condition: Dataset contains customer level information and as per the label, this is customer's identification number.

TARGET VARIABLE

IND_BUY

It takes value 1 if customer has purchased a product in the current month, else takes value 0. This is what is to be predicted for next month according to given problem statement. Here, ***1 is event and 0 is non-event***.

INELIGIBLE VARIABLES

IND_CASH_PURCHASE

It takes value 1 if customer has made a purchase in current month by cash payment, else takes value 0. Apparently, ***this variable contains post-event information and hence it can not be used as an input in the modeling process***. *(**Note**: Variables with single unique value are also ineligible as they do not distinguish between event and non-event)*

INDEPENDENT VARIABLES (ALSO CALLED PREDICTORS)

NUMERIC

INDICATORS

IND_LOW_INCOME

IND_ACTIVE_1M

IND_CAR_OWNER

Necessary Condition:

EDD output indicates

- These variables have only 2 unique values
- Minimum value is 0
- Maximum value is 1

Sufficient Condition:

Labels of these variables clearly mention that these are the binary indicators taking value 0 or 1

Note: **IND_BUY** and **IND_CASH_PURCHASE** are also indicator variables but they do not belong to the set of predictor variables.

DISCRETE / CONTINUOUS

NUM_POOR_FEEDBACK_1M

NUM_PURCHASES_1M

CUMM_DISC_AMT_1M

PREV_BUY_BILL_AMT

NUM_FREEBIES_RECD_1M

NUM_VISITS_2M

NUM_VISITS_3M

EST_SPENDING_LIMIT

DAYS_TO_NEXT_BDAY

NUM_KIDS

Both EDD output and the variable descriptions point out that these all are discrete or continuous variables.

CHARACTER

MARITAL_STATUS

It has character format. It takes three values:

- *M implies that the customer is married*
- *U implies that the customer is unmarried*
- *<Blank> implies that no information is available about the marital status of customer*

Outlier Treatment

TECHNIQUES FOR OUTLIER DETECTION / TREATMENT:

- **Capping and Flooring Technique**: The outliers are identified and treated based upon the values of P99 and P1
- **Exponential Smoothing Technique**: The curve between P95 to P99 is extrapolated beyond P99, to identify the values falling above the curve. The values falling outside the curve are outliers and are treated according to some functions depending upon the boundary conditions
- **Sigma Approach**: The outliers are identified and treated based upon the values of mean and standard deviation
- **Robust Regression Technique**: This technique involves running regression repeatedly to identify outliers by assigning weights to the observations. The weights are on the basis of the prediction error (residual) in different iterations. Higher residual means lower weight.
- **Mahalanobis Distance Technique**: The outliers are identified by the magnitude of 'Mahalanobis' or statistical distance from the origin. Weights are given to each observation as the inverse of 'Mahalanobis' distance.

Note: Detailed discussion on each of the outlier treatment techniques is beyond the scope of current exercise. Keeping in mind beginners' perspective, the most commonly used and the most easily implementable technique (**Capping and Flooring Technique**) is now being discussed.

FLOORING

CASES	LOGIC				ILLUSTRATION (Let X = 5)		
	MIN	P1	OUTLIER	TREATMENT	MIN	P1	TREATMENT
CASE I	< 0	< 0	Any value < X * P1	Floor at X * P1	- 200	- 10	Floor at -50
CASE II	< 0	= 0	Any value < - X	Floor at - X	- 200	0	Floor at -5
CASE III	< 0	> 0	Any value < P1 – (X * P1)	Floor at P1 – (X * P1)	- 200	10	Floor at -40
CASE IV	> 0	> 0	Any value < P1 / X	Floor at P1 / X	1	10	Floor at 2

CAPPING

CASES	LOGIC				ILLUSTRATION (Let X = 5)		
	P99	MAX	OUTLIER	TREATMENT	P99	MAX	TREATMENT
CASE I	> 0	> 0	Any value > X * P99	Cap at X * P99	10	200	Cap at 50
CASE II	= 0	> 0	Any value > X	Cap at X	0	200	Cap at 5
CASE III	< 0	> 0	Any value > P99 - (X * P99)	Cap at P99 - (X * P99)	- 10	200	Cap at 40
CASE IV	< 0	< 0	Any value > P99 / X	Cap at P99 / X	- 10	- 1	Cap at -2

Outlier Treatment Continued…

HOW TO DECIDE UPON THE VALUE OF X ?

As a rule of thumb, X may be assumed as 1 or 2 or 5 or 10.

X = 1 is as good as flooring and capping at P1 and P99 respectively. This is a very strict treatment.

X = 10 is too lenient.

X = 5 is commonly used.

X = 2 may be used based on data analysis.

An Excerpt from Modeling Dataset EDD

VARIABLE	TYPE	# OBS.	NMISS	UNIQUE	MEAN	STD DEV	MIN	P1	P5	P25	MEDIAN	P75	P95	P99	MAX
PREV_BUY_BILL_AMT	NUM	400	0	54	752	731.23	300	350	443.75	625	725	816.5	1000	1050	15000

Modeling dataset EDD indicates that PREV_BUY_BILL_AMT has outliers.

An Excerpt from Validation Dataset EDD

VARIABLE	TYPE	# OBS.	NMISS	UNIQUE	MEAN	STD DEV	MIN	P1	P5	P25	MEDIAN	P75	P95	P99	MAX
PREV_BUY_BILL_AMT	NUM	100	0	34	698	176.44	300	325	416.75	587.5	668.75	800	1000	1100	1200

Validation dataset EDD, however, shows that maximum value of PREV_BUY_BILL_AMT is not very far away from modeling data P99 value.

Taking X = 2 is likely to work for outlier treatment.

Outlier Treatment: *Cap PREV_BUY_BILL_AMT at 2100*

```
PREV_BUY_BILL_AMT = min(PREV_BUY_BILL_AMT , 2100);
```


Missing Value Imputation

An Excerpt from Modeling Dataset EDD

VARIABLE	TYPE	# OBS.	NMISS	UNIQUE	MEAN	STD DEV	MIN	P1	P5	P25	MEDIAN	P75	P95	P99	MAX
CUMM_DISC_AMT_1M	NUM	400	6	56	1805	883.70	0	0	0	1250	1825	2500	3250	3750	4250
MARITAL_STATUS	CHAR	400	72	3	M::231	U::97	::72								

Modeling dataset EDD indicates that CUMM_DISC_AMT_1M and MARITAL_STATUS have missing values.

CUMM_DISC_AMT_1M:

- *A continuous numeric variable*
- *Missing values are not too many; hence no need to create a separate indicator for missing values*
- *Missing values implying no information available; Imputation with ZERO is not meaningful*
- *Imputation with median value looks apt*
- *No outliers or extreme values to distort mean; mean value is in fact close to median; hence imputation with mean is also a good option*

MARITAL_STATUS

- *A character variable with three categories*
  - *M: Customer is married*
  - *U: Customer is unmarried*
  - *Missing Values: No information is available about customer's marital status*
- *To treat missing values, create indicators for each of the remaining categories*

An illustrative table showing MVI by indicator creation

MARITAL_STATUS	IND_MARRIED	IND_UNMARRIED
M	1	0
M	1	0
U	0	1
U	0	1
	0	0
	0	0

MVI Treatment: *Impute CUMM_DISC_AMT_1M with its median value and create dummy codes for MARITAL_STATUS*

```
If CUMM_DISC_AMT_1M   = .       then CUMM_DISC_AMT_1M = 1825     ;
If MARITAL_STATUS     = "M"     then IND_MARRIED      = 1        ; else IND_MARRIED    = 0;
If MARITAL_STATUS     = "U"     then IND_UNMARRIED    = 1        ; else IND_UNMARRIED  = 0;
```


Correlations

Syntax for computing Pearson's correlation coefficients

```
PROC CORR  DATA = <Modeling dataset after data prep>
           OUTP = <Output dataset containing correlation matrix>;
VAR _numeric_;
WITH _numeric_;
RUN;
```

For Correlation Matrix Analysis . . .

- *Perfect Positive Correlation (100%)*
- *40% ≤ Absolute Correlation < 100%*
- *20% ≤ Absolute Correlation < 40%*

Note: The primary key 'CUSTOMER_ID' is numeric too. But it should be dropped from this analysis as its correlation with variables would have no meaning.

VARIABLES		A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
IND_BUY	A	100.0%																
IND_LOW_INCOME	B	-7.3%	100.0%															
IND_ACTIVE_1M	C	10.1%	11.4%	100.0%														
IND_CAR_OWNER	D	7.7%	-8.1%	5.0%	100.0%													
NUM_POOR_FEEDBACK_1M	E	-3.2%	10.8%	3.3%	-6.1%	100.0%												
NUM_PURCHASES_1M	F	22.6%	-11.3%	10.8%	0.1%	-3.3%	100.0%											
CUMM_DISC_AMT_1M	G	7.4%	-5.7%	0.4%	4.3%	1.5%	16.4%	100.0%										
PREV_BUY_BILL_AMT	H	-8.8%	2.3%	-4.3%	-2.5%	-5.8%	-17.8%	-79.8%	100.0%									
NUM_FREEBIES_RECD_1M	I	7.9%	5.4%	10.4%	3.2%	14.1%	14.9%	56.0%	-57.4%	100.0%								
NUM_VISITS_2M	J	16.2%	-11.8%	-1.9%	3.9%	-12.8%	28.8%	24.0%	-24.5%	13.1%	100.0%							
NUM_VISITS_3M	K	18.6%	-14.1%	-1.2%	3.2%	-12.5%	26.6%	23.7%	-25.9%	14.6%	96.4%	100.0%						
EST_SPENDING_LIMIT	L	12.5%	-11.4%	2.2%	-2.2%	-7.5%	29.7%	23.0%	-29.2%	15.7%	52.8%	53.4%	100.0%					
DAYS_TO_NEXT_BDAY	M	-6.9%	9.1%	1.2%	-0.2%	1.6%	-15.9%	-23.5%	23.8%	-13.0%	-24.7%	-25.4%	-34.7%	100.0%				
NUM_KIDS	N	0.9%	-4.2%	3.8%	12.2%	14.3%	10.0%	19.3%	-17.6%	24.7%	5.2%	6.2%	12.1%	-5.2%	100.0%			
IND_CASH_PURCHASE	O	100.0%	-7.3%	10.1%	7.7%	3.2%	22.6%	7.4%	-8.8%	7.9%	16.2%	18.6%	12.5%	-6.9%	0.9%	100.0%		
IND_MARRIED	P	-1.3%	43.0%	8.4%	-8.6%	8.2%	-9.8%	-7.9%	5.9%	-3.9%	-7.9%	-7.7%	-6.6%	2.5%	-0.7%	-1.3%	100.0%	
IND_UNMARRIED	Q	0.4%	-4.8%	0.5%	-6.4%	-0.1%	-1.4%	3.4%	-6.1%	3.5%	7.1%	6.5%	6.0%	1.5%	-3.9%	0.4%	-66.1%	100.0%

Variable Clustering

PROC VARCLUS Syntax

```
PROC VARCLUS DATA = <Input Data>
MAXEIGEN = 0.7 MAXCLUSTERS = 100* SHORT HI;
VAR         IND_LOW_INCOME
            IND_ACTIVE_1M
            IND_CAR_OWNER
            NUM_POOR_FEEDBACK_1M
            NUM_PURCHASES_1M
            CUMM_DISC_AMT_1M
            PREV_BUY_BILL_AMT
            NUM_FREEBIES_RECD_1M
            NUM_VISITS_2M
            NUM_VISITS_3M
            EST_SPENDING_LIMIT
            DAYS_TO_NEXT_BDAY
            NUM_KIDS
            IND_MARRIED
            IND_UNMARRIED
            ;
ODS OUTPUT RSQUARE = <Output Data>;
RUN;

*Since number of variables listed in VAR
statement is 15, MAXCLUSTERS = 100 is not
putting any effective condition.
```

A variable selected from each cluster should have a high correlation with its own cluster and a low correlation with the other clusters

$$\text{R Square Ratio} = \frac{1 - \text{R Square Own Cluster}}{1 - \text{R Square Next Closest}}$$

Output *Cluster's best representative (Variable with minimum R Square Ratio)*

Num Clusters	Cluster	Variable	Own Cluster	Next Closest	R Square Ratio
2	Cluster 1	IND_LOW_INCOME	0.099	0.0003	0.9012
2		NUM_POOR_FEEDBACK_1M	0.0401	0.0093	0.9689
2		NUM_PURCHASES_1M	0.2227	0.0391	0.8089
2		NUM_VISITS_2M	0.782	0.0542	0.2305
2		NUM_VISITS_3M	0.7828	0.0595	0.2309
2		EST_SPENDING_LIMIT	0.5391	0.0705	0.4958
2		DAYS_TO_NEXT_BDAY	0.2031	0.0516	0.8403
2		IND_MARRIED	0.0866	0.0043	0.9173
2		IND_UNMARRIED	0.0403	0.0015	0.9611
2	Cluster 2	IND_ACTIVE_1M	0.0108	0.0001	0.9893
2		IND_CAR_OWNER	0.0098	0.0012	0.9915
2		CUMM_DISC_AMT_1M	0.7882	0.0858	0.2317
2		PREV_BUY_BILL_AMT	0.7936	0.0982	0.2289
2		NUM_FREEBIES_RECD_1M	0.6423	0.0273	0.3678
2		NUM_KIDS	0.1508	0.0073	0.8554
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
8	Cluster 1	NUM_VISITS_2M	0.9124	0.1236	0.1
8		NUM_VISITS_3M	0.9156	0.1167	0.0956
8		EST_SPENDING_LIMIT	0.5461	0.1791	0.5529
8	Cluster 2	CUMM_DISC_AMT_1M	0.8242	0.0694	0.1889
8		PREV_BUY_BILL_AMT	0.8338	0.0862	0.1819
8		NUM_FREEBIES_RECD_1M	0.6362	0.0347	0.3768
8	Cluster 3	IND_MARRIED	0.8307	0.1851	0.2077
8		IND_UNMARRIED	0.8307	0.0054	0.1702
8	Cluster 4	IND_CAR_OWNER	0.5612	0.0066	0.4417
8		NUM_KIDS	0.5612	0.0539	0.4639
8	Cluster 5	NUM_POOR_FEEDBACK_1M	1	0.0156	0
8	Cluster 6	IND_ACTIVE_1M	1	0.0131	0
8	Cluster 7	IND_LOW_INCOME	1	0.0688	0
8	Cluster 8	NUM_PURCHASES_1M	0.5793	0.0996	0.4672

Double Check Multicollinearity

There arises a problem of multicollinearity when predictors are highly correlated among themselves. Variable clustering does away with this to a large extent. To double check, variance inflation test is recommended for use.

VIF Macro Syntax

```
%VIF    (
        MOD_DAT         = <Input dataset with library name>,
        OUT_DAT         = <Output dataset containing stats of short-listed variables>,
        ELIM_SUM        = <Output dataset containing summary of variables  eliminated>,
        VAR_LIST        = <List of variables>,
        DP_VAR          = <Dependent variable>,
        MAX_VIF_LIMIT   = <Maximum value of VIF permitted>,
        IF_CORR         = <Option*>
        );

*IF_CORR can either be Y or N depending on whether correlation technique should be applied for variable
reduction or VIF only is sufficient.
```

As a rule of thumb, MAX_VIF_LIMIT is generally used as 2 or 5 or 10. However, MAX_VIF_LIMIT = 10 is a bit too lenient on variable elimination by variance inflation factor.

VIF Output . . .

VIF value is under 2

Dependent	Variable	DF	Estimate	StdErr	tValue	Probt	Variance Inflation
IND_BUY	NUM_VISITS_3M	1	0.0080	0.0024	3.2761	0.0011	1.1596
IND_BUY	PREV_BUY_BILL_AMT	1	0.0000	0.0001	-0.5625	0.5741	1.1258
IND_BUY	DAYS_TO_NEXT_BDAY	1	-0.0002	0.0008	-0.2720	0.7857	1.1137
IND_BUY	IND_LOW_INCOME	1	-0.0263	0.0221	-1.1870	0.2359	1.0550
IND_BUY	NUM_POOR_FEEDBACK_1M	1	0.0134	0.0114	1.1774	0.2398	1.0368
IND_BUY	IND_ACTIVE_1M	1	0.0837	0.0398	2.1016	0.0362	1.0197
IND_BUY	IND_CAR_OWNER	1	0.1260	0.0972	1.2956	0.1959	1.0191
IND_BUY	IND_UNMARRIED	1	-0.0041	0.0252	-0.1631	0.8705	1.0152
IND_BUY	INTERCEPT	1	0.0523	0.0549	0.9530	0.3412	0.0000

Logistic Regression

Logistic Regression Syntax

```
PROC LOGISTIC DATA = <Modeling dataset> NAMELEN = 32 DESCENDING
OUTEST = <Dataset containing estimated parameters (View 1)>;
MODEL <Dependent Variable>       = <List of independent variables>
           /          SELECTION  = <Selection method>
                      SLE        = <SLE Criterion>
                      SLS        = <SLS Criterion>
                      LACKFIT
                      RSQ
                      STB
                      CLPARM     = WALD;
OUTPUT OUT = <Scored modeling dataset> P = PRED;
ODS OUTPUT PARAMETERESTIMATES = <Dataset containing estimated parameters (View 2)>;
RUN;
```

Note: Importance of a variable in a model should never be deduced from magnitude of estimates. Different variables may have different units and scales. Standardized estimates provide more meaningful indications

Sample Output . . .

Variable	DF	Estimate	Std. Err.	Wald Chi Sq.	Prob. Chi Sq.	Standardized Est.
Intercept	1	-3.6669	0.3475	111.336	4.99E-26	.
NUM_VISITS_3M	1	0.1087	0.0319	11.61226	0.000655	0.2845
IND_ACTIVE_1M	1	1.2982	0.6061	4.588731	0.032183	0.1944

STATISTICAL SIGNIFICANCE:

As p-values are less than 0.05, both variables are statistically significant at 5% level

CHECK VARIABLE HYPOTHESIS:

NUM_VISITS_3M: More visits to the departmental store by customer in the past 3 months, more likely she is to visit again and buy some product

IND_ACTIVE_1M: Active customers in last 1 month are more probable to remain active and purchase a product.

Scoring

Before scoring, it should be ensured that all data prep has been done on validation dataset too.

Validation Dataset Preparation: *Replicate all steps (MVI / Outlier Treatment / New Variable Creation)*

```
PREV_BUY_BILL_AMT     = min(PREV_BUY_BILL_AMT , 2100);
If CUMM_DISC_AMT_1M   = .         then CUMM_DISC_AMT_1M = 1825     ;
If MARITAL_STATUS     = "M"       then IND_MARRIED      = 1        ; else IND_MARRIED    = 0;
If MARITAL_STATUS     = "U"       then IND_UNMARRIED    = 1        ; else IND_UNMARRIED  = 0;
```

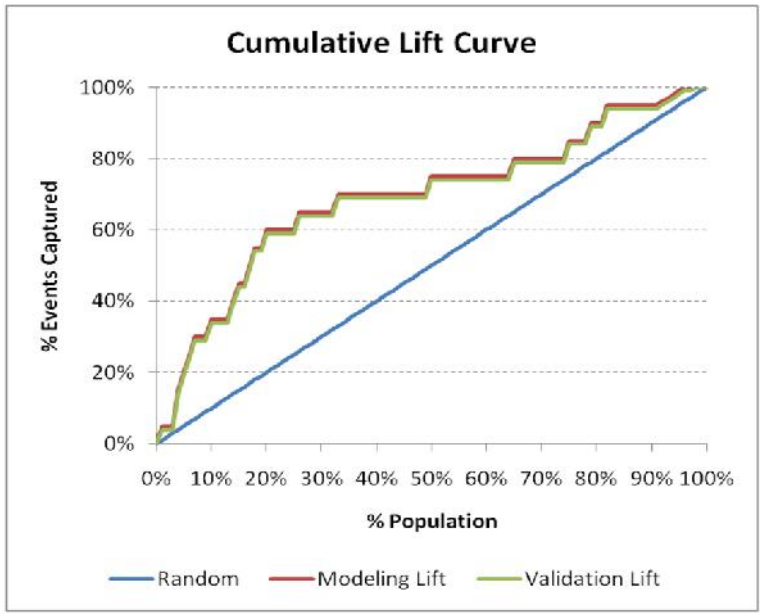
Syntax for Macro used for scoring a Logistic Regression Model

```
%LGTSCORE  (
           INDATA     = <Name of the Input dataset to be scored>,
           OUTDATA    = <Output dataset>,
           REGEST     = <Dataset having the estimated coefficients of the model*>,
           DEPVAR     = <Name of the dependent variable>,
           DEPVARH    = <New variable name for predicted values>
           );
*This should be the same dataset as obtained from PROC LOGISTIC STEP –
<Dataset containing estimated parameters (View 1)>
```


Lift Chart

Lorenz-curve (Lift) and KS statistic Macro Syntax

```
%KS_LZ_GB  (
           INDATA      = <name of  the input dataset>,
           OUTPUT_KS   = <Output location for KS stat dataset>,
           OUTPUT_LZ   = <Output location for Lorenz dataset>,
           NUM_BIN     = <Number of bins to be created>,
           DEP_VAR     = <Dependent variable Name>,
           SCORE       = <Scoring Variable name>,
           VAR_KEY     = <Key used for binning>,
           LIFT        = <Value at which lift is to be calculated>,
           ODS_OUT     = <Name of file containing KS stat and Lift with '.xls' as the extension>
           );
```



Model Performance	
Concordance	66.6
Modeling Lift at 5% Event Rate	20%
Validation Lift at 5% Event Rate	19%

Note: The lift charts are not smooth because of very few records in the current exercise, However, in LIVE projects, modeling population is likely to be much more, yielding smooth lift charts.

Appendix

For running any SAS Toolkit Macro, user must define the toolkit catalogue as follows:

```
LIBNAME CATLOG "Z:\MacroToolkit";
OPTIONS MSTORED SASMSTORE = CATLOG;
```

Thanks

For queries, contact Varun Aggarwal at Varun.Aggarwal@exlservice.com