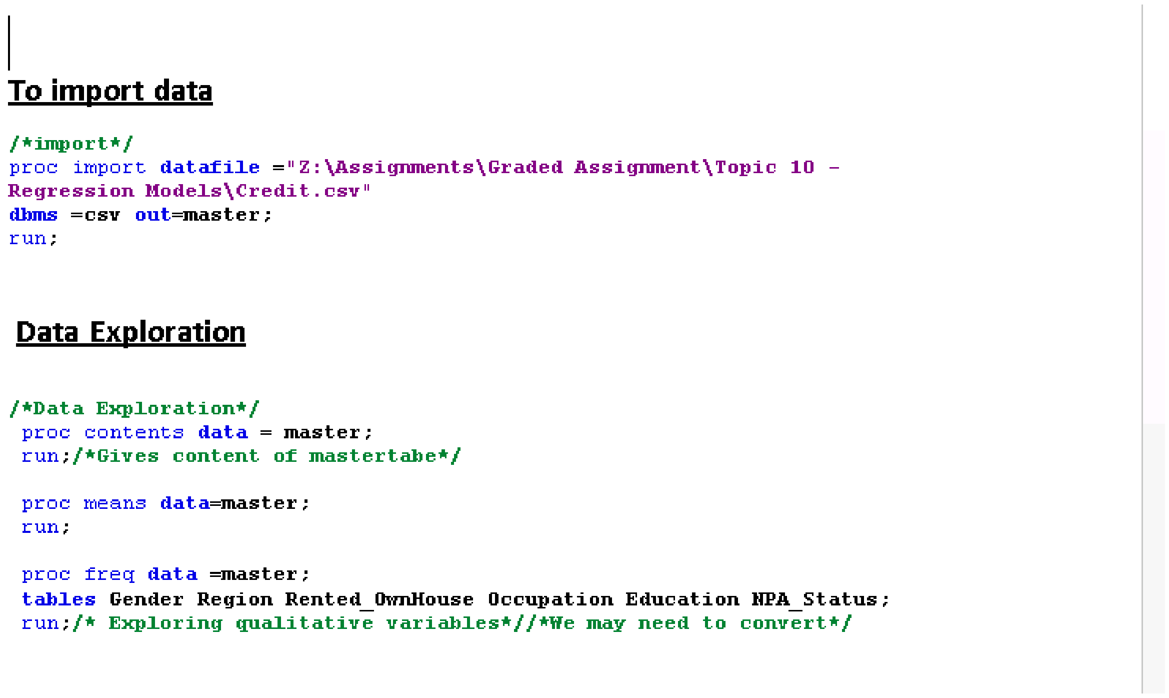
Aim:-

To analyse data of the customers of the credit bureu company . Create model to identify the probablity of customers to default on the loan



To import data

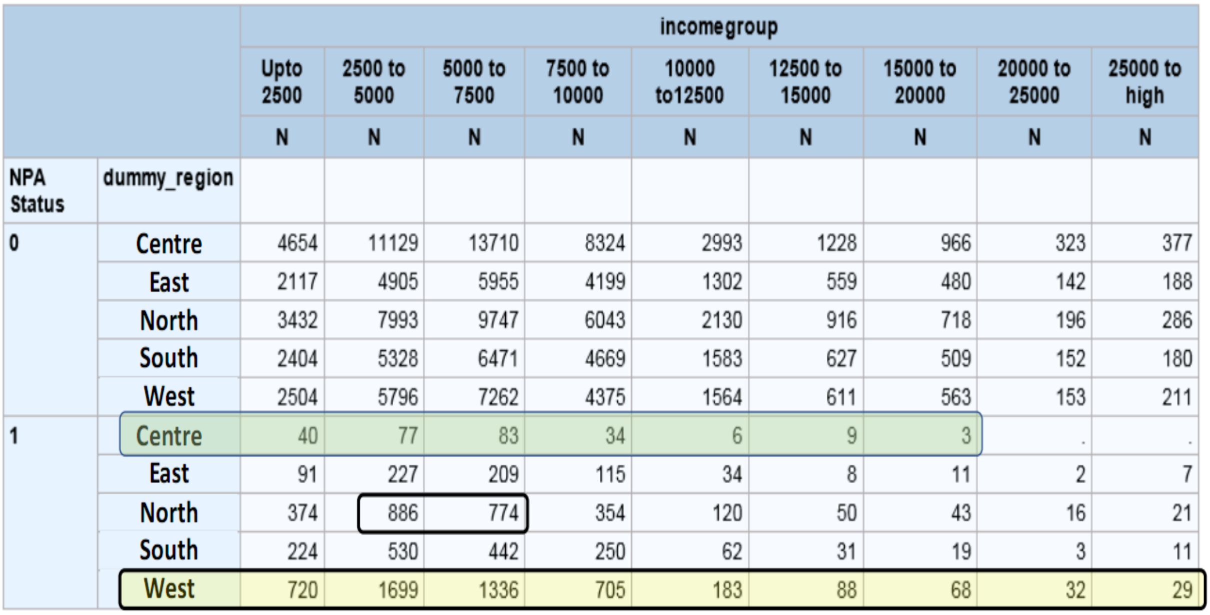
```
/*import*/
proc import datafile ="Z:\Assignments\Graded Assignment\Topic 10 -
Regression Models\Credit.csv"
dbms =csv out=master;
run;
```

Data Exploration

```
/*Data Exploration*/
 proc contents data = master;
 run;/*Gives content of mastertabe*/

 proc means data=master;
 run;

 proc freq data =master;
 tables Gender Region Rented_OwnHouse Occupation Education NPA_Status;
 run;/* Exploring qualitative variables*//*We may need to convert*/
```

- Age:-
  About 75% of Defaulting customers are between Age Group 30 to 60.
- Region:-
  Defaulted customers are concentrated in West (48%) and North (23%)
- Credit lines:-
  About 40% of defaulted customers have one credit line
- 60 -90 Days Past due :-
  Out of Default customers 28% customers defaulted on loan atlist 1nce for 60- 90 days 18% defaulted only 1nce
- 90+ Days Past Due:-
  35% of defaulted customers have not paid the loan for more than 1 time and 18% haven't paid 1 time for more than 90 days.
- Income:-
  93% percent of defaulters have income less than 1000. And 2500 to 5000 have 34% default rate
- Credit Limit:-
  45% of defauted customers used the entire credit limit.
- Debt to Income ratio:-
  Debt to income ratio is high for defaulters

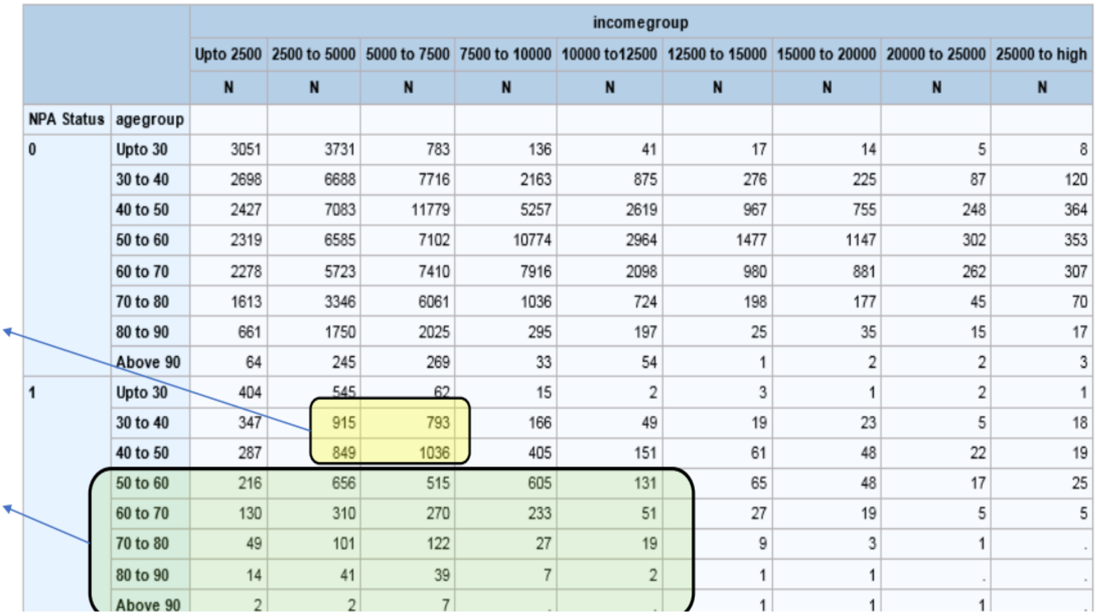
Cross table – Income, region and NPA Status



		incomegroup								
		Upto 2500	2500 to 5000	5000 to 7500	7500 to 10000	10000 to12500	12500 to 15000	15000 to 20000	20000 to 25000	25000 to high
		N	N	N	N	N	N	N	N	N
NPA Status	dummy_region									
0	Centre	4654	11129	13710	8324	2993	1228	966	323	377
	East	2117	4905	5955	4199	1302	559	480	142	188
	North	3432	7993	9747	6043	2130	916	718	196	286
	South	2404	5328	6471	4669	1583	627	509	152	180
	West	2504	5796	7262	4375	1564	611	563	153	211
1	Centre	40	77	83	34	6	9	3	.	.
	East	91	227	209	115	34	8	11	2	7
	North	374	886	774	354	120	50	43	16	21
	South	224	530	442	250	62	31	19	3	11
	West	720	1699	1336	705	183	88	68	32	29

- Default rate is high across income group in the West Region, the company must investigate if loans approval procedures are correctly followed.
- On the other loan default rate are relatively lower in the central region, company must learn how they are able to maintain this.
- The northern region also seem to have high loan defaults in the income groups of 2500 to 7500.

Cross table – Income, age group and NPA Status



		incomegroup								
		Upto 2500	2500 to 5000	5000 to 7500	7500 to 10000	10000 to12500	12500 to 15000	15000 to 20000	20000 to 25000	25000 to high
		N	N	N	N	N	N	N	N	N
NPA Status	agegroup									
0	Upto 30	3051	3731	783	136	41	17	14	5	8
	30 to 40	2698	6688	7716	2163	875	276	225	87	120
	40 to 50	2427	7083	11779	5257	2619	967	755	248	364
	50 to 60	2319	6585	7102	10774	2964	1477	1147	302	353
	60 to 70	2278	5723	7410	7916	2098	980	881	262	307
	70 to 80	1613	3346	6061	1036	724	198	177	45	70
	80 to 90	661	1750	2025	295	197	25	35	15	17
	Above 90	64	245	269	33	54	1	2	2	3
1	Upto 30	404	545	62	15	2	3	1	2	1
	30 to 40	347	915	793	166	49	19	23	5	18
	40 to 50	287	849	1036	405	151	61	48	22	19
	50 to 60	216	656	515	605	131	65	48	17	25
	60 to 70	130	310	270	233	51	27	19	5	5
	70 to 80	49	101	122	27	19	9	3	1	.
	80 to 90	14	41	39	7	2	1	1	.	.
	Above 90	2	2	7	.	.	1	1	1	.

Default in the age group of 30 to 50 with income 2500 to 7500 is on the higher side

Default in the age group of 50 and above must be checked and prevented, as repayment ability decrease post retirement

CODE

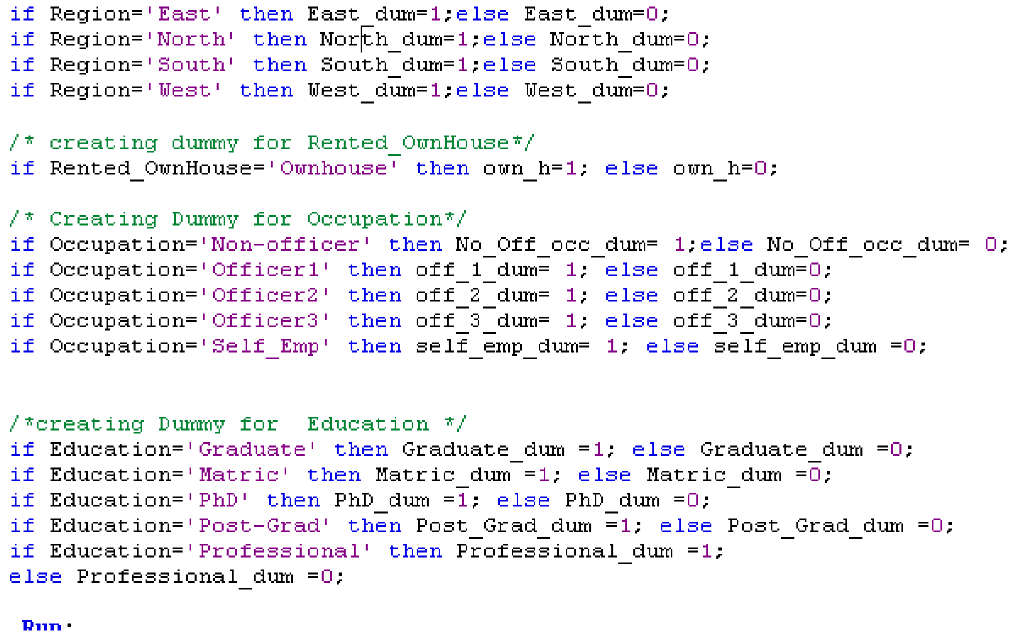
Data Preparation

1 Converting char variables to numeric

```
/*1 converting char variables to Numeric*/
  data master_prepared;/*This will contain data after preparation*/
  set master;

  month_income=input(MonthlyIncome,8.);/*Converting char to numeric*/
  drop MonthlyIncome;/*deleating previous variable which no longer needed*/

  No_of_Dependents= input(NumberOfDependents,8.);
  drop NumberOfDependents;
```

2 Creating Dummy Variables

```
  /*2 Creating dummy variables*/
  /*qualitative -> quantotative*/

  /* Dummy for Gender variable*/
  if Gender='Male' then Gender_dumm=1;else Gender_dumm=0;

  /* Dummy for  Region variables*/
  if Region='Central' then centr_dum=1;else centr_dum=0;
```



```
if Region='East' then East_dum=1;else East_dum=0;
if Region='North' then North_dum=1;else North_dum=0;
if Region='South' then South_dum=1;else South_dum=0;
if Region='West' then West_dum=1;else West_dum=0;

/* creating dummy for Rented_OwnHouse*/
if Rented_OwnHouse='Ownhouse' then own_h=1; else own_h=0;

/* Creating Dummy for Occupation*/
if Occupation='Non-officer' then No_Off_occ_dum= 1;else No_Off_occ_dum= 0;
if Occupation='Officer1' then off_1_dum= 1; else off_1_dum=0;
if Occupation='Officer2' then off_2_dum= 1; else off_2_dum=0;
if Occupation='Officer3' then off_3_dum= 1; else off_3_dum=0;
if Occupation='Self_Emp' then self_emp_dum= 1; else self_emp_dum =0;


/*creating Dummy for  Education */
if Education='Graduate' then Graduate_dum =1; else Graduate_dum =0;
if Education='Matric' then Matric_dum =1; else Matric_dum =0;
if Education='PhD' then PhD_dum =1; else PhD_dum =0;
if Education='Post-Grad' then Post_Grad_dum =1; else Post_Grad_dum =0;
if Education='Professional' then Professional_dum =1;
else Professional_dum =0;

 Run;
```

3 Finding Missing Values

```
/* 3 Finding Missing Values */
proc means data = master_prepared nmiss mean min max ;
run;
```

Obervation : nearly 24% data of month_income and 2% data of No_of_Dependents is missing some variables have .13% missing values .

4 Treating Missing Values

```
/*  4 treating Missing Values */
    data master_prepared;
    set master_prepared;
    if month_income ='.' then month_income= 6670;
    /* since month_income has large missing data we replace missing by mean*/
    if No_of_Dependents ='.' then No_of_Dependents = 1;
    /* whole no close to mean replaces large missing values*/
    if NPA_Status='.' then delete;
    /* very few records deleted*/
     run;
```

Obervation : After treatment no missing values left.

5 Finding Outliers

```
/* 5 finding Outliers */
     proc univariate data =  master_prepared;
     var age month_income No_of_Dependents DebtRatio;
        run;
```

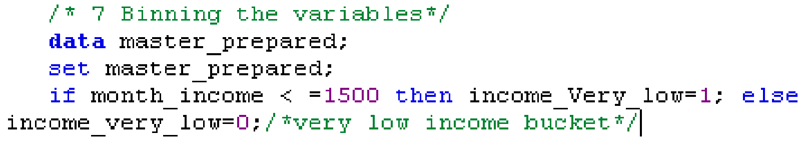
Observation :

1. **month_income has some extreame values but that can be due to random chance month_income can be binned.**
2. **Age can not be 0 some observations some of them have age more than 100 whcih can be due to random chance so it could be deleted.**
3. **DebtRatio has some extreme values which may be due to random chance so better we devide.**
4. **No_of_Dependents have some values more than 10 which is too high than 99% quantile so should be deleted.**

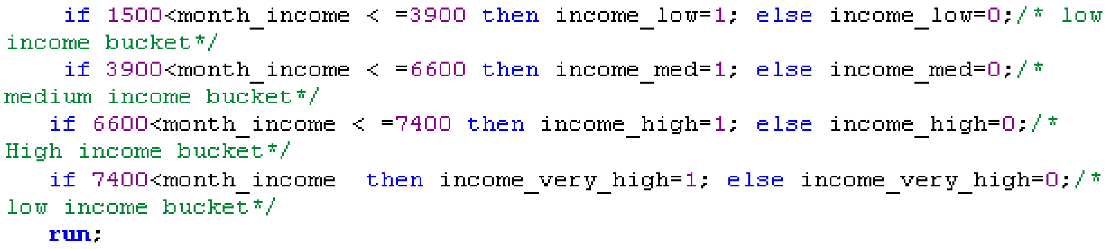
6 Treating The Outliers.

```
data master_prepared;
     set master_prepared;
     if age=0 then delete;
     if age>100 then delete;
     if no_of_dependants>10 then delete;/* no of dependant large is outlier*/

     if RevolvingUtilizationOfUnsecuredL >0.5 then debt=1;else
debt=0;/*devide debt ratio by 50% and credit utilization by 50%*/


     proc univariate data = master_prepared;
     var month_income debt;
     run;
```

7 Binning The variables



```
   /* 7 Binning the variables*/
   data master_prepared;
   set master_prepared;
   if month_income < =1500 then income_Very_low=1; else
income_very_low=0;/*very low income bucket*/
```



```
    if 1500<month_income < =3900 then income_low=1; else income_low=0;/* low
income bucket*/
    if 3900<month_income < =6600 then income_med=1; else income_med=0;/*
medium income bucket*/
   if 6600<month_income < =7400 then income_high=1; else income_high=0;/*
High income bucket*/
   if 7400<month_income  then income_very_high=1; else income_very_high=0;/*
low income bucket*/
   run;
```

8 Preparing Training and Validation dataset

```
proc surveyselect data = master_prepared
      method= SRS out= model_select samprate=0.5 outall;
     run; /*test data */

     data Model_train model_validation;
     set Model_select;

     if select =0 then output=model_train;else output=Model_validation;
     run;
```

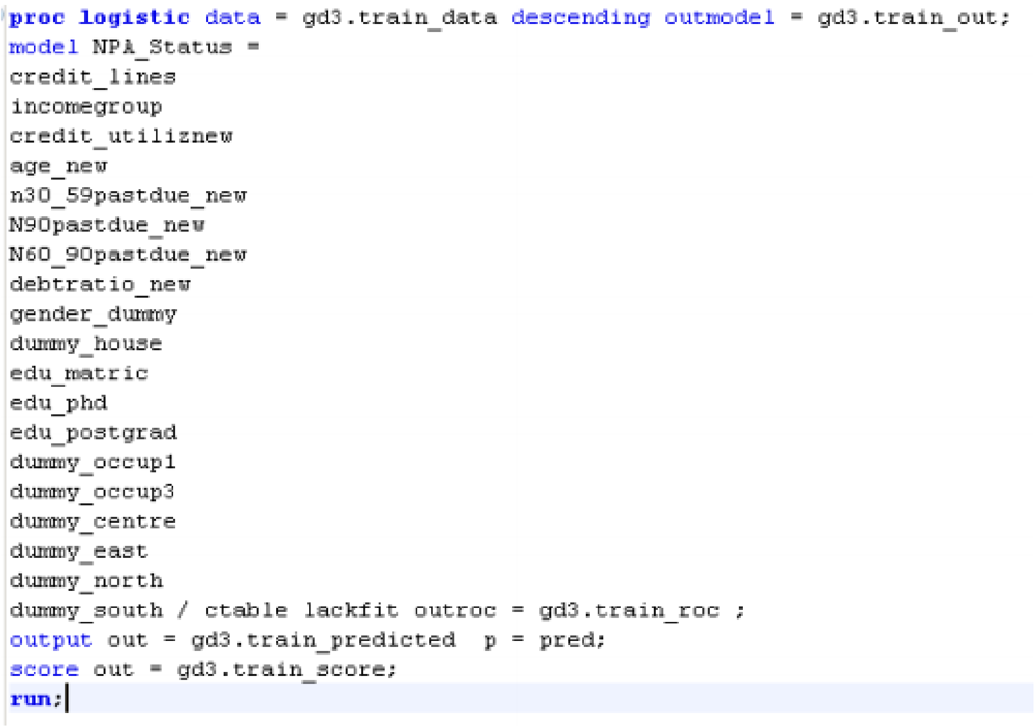
9 Running Logistic Regression

Iteration 1

```
 /* Logistic regression */

     /*Iteration 1*/

     Proc Logistic data =model_train descending;
     model NPA_Status = RevolvingUtilizationOfUnsecuredL  age 
NumberOfTime30_59DaysPastDueNotW DebtRatio NumberOfOpenCreditLinesAndLoans
     NumberOfTimes90DaysLate NumberRealEstateLoansOrLines 
NumberOfTime60_89DaysPastDueNotW month_income No_of_Dependents Gender_dumm
     centr_dum East_dum North_dum South_dum West_dum own_h No_Off_occ_dum
off_1_dum off_2_dum off_3_dum self_emp_dum
     Graduate_dum Matric_dum PhD_dum Post_Grad_dum Professional_dum  debt
income_Very_low income_low income_med income_high income_very_high ;
     /selection =backward ctable lackfit;

     Run;
```

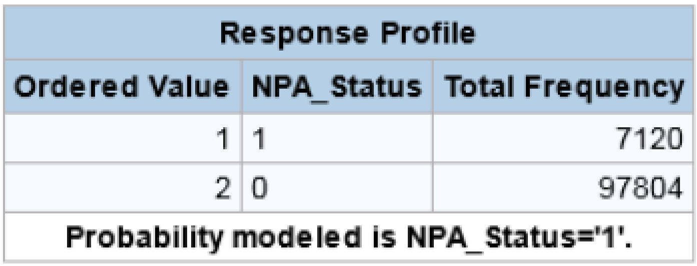
Multiple forward back ward iterations were done in order to generate the final model as below

FINAL MODEL



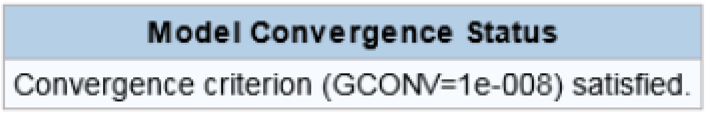
```
proc logistic data = gd3.train_data descending outmodel = gd3.train_out;
model NPA_Status =
credit_lines
incomegroup
credit_utiliznew
age_new
n30_59pastdue_new
N90pastdue_new
N60_90pastdue_new
debtratio_new
gender_dummy
dummy_house
edu_matric
edu_phd
edu_postgrad
dummy_occup1
dummy_occup3
dummy_centre
dummy_east
dummy_north
dummy_south / ctable lackfit outroc = gd3.train_roc ;
output out = gd3.train_predicted  p = pred;
score out = gd3.train_score;
run;
```

Model output and It’s meaning



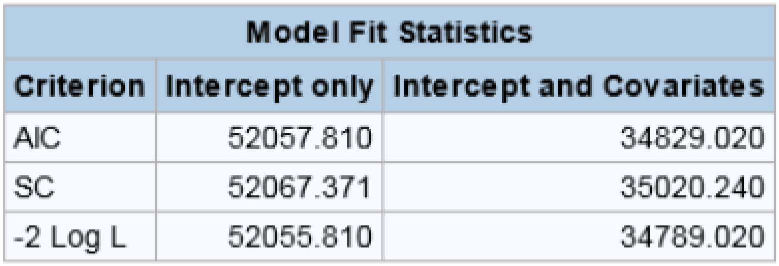
Response Profile		
Ordered Value	**NPA_Status**	**Total Frequency**
1	1	7120
2	0	97804
Probability modeled is NPA_Status='1'.		

- **Training set consist of 104924 observations out of which 7120 are defauted ie. 67.86 defaulted**



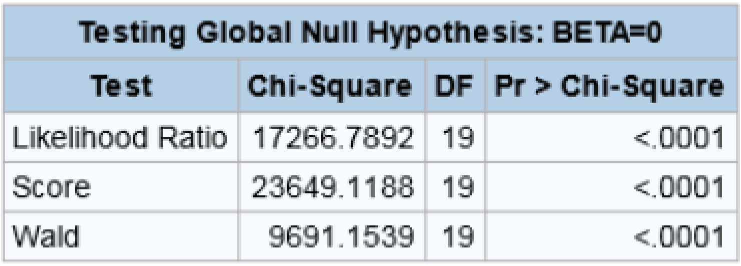
Model Convergence Status
Convergence criterion (GCONV=1e-008) satisfied.

- **Conversion Criteria is satisfied**



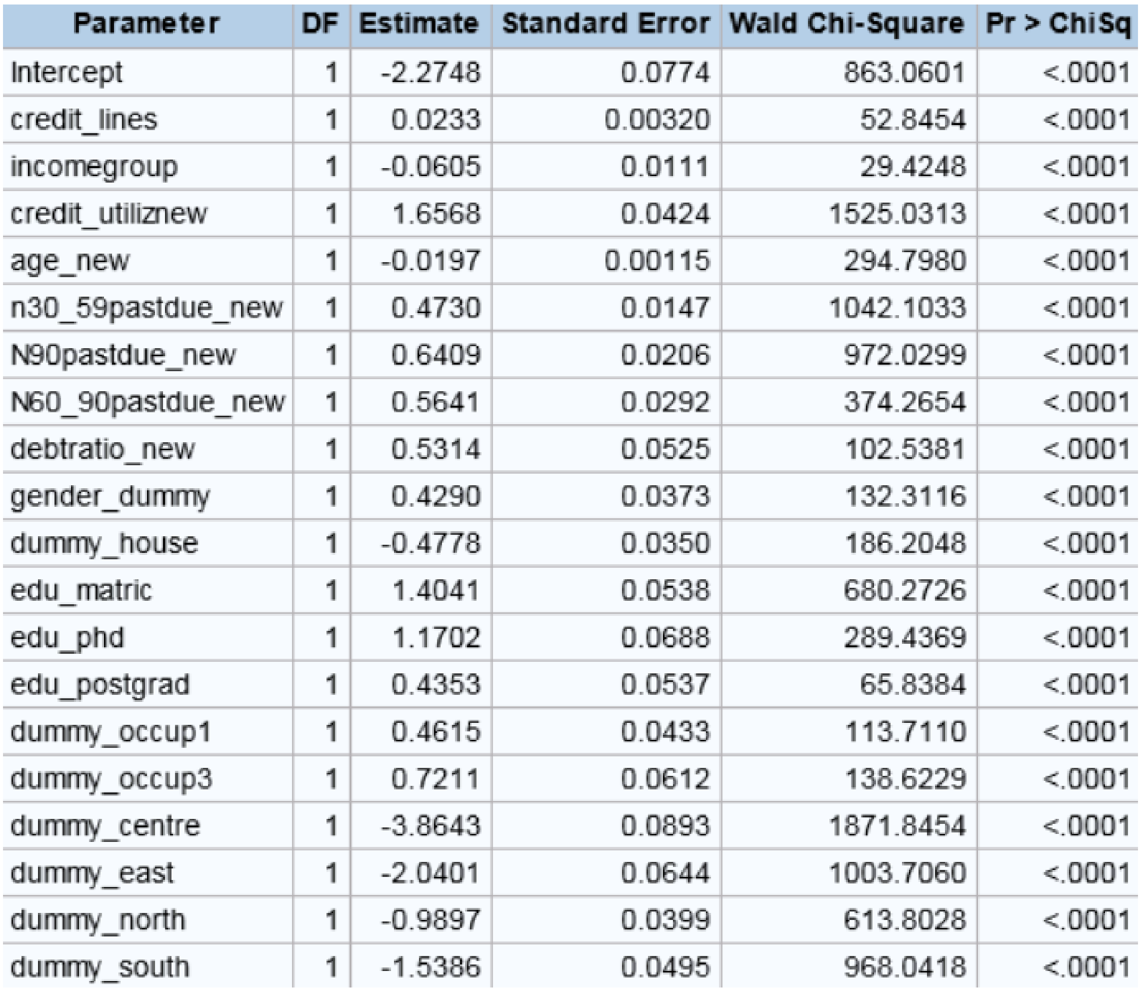
Model Fit Statistics		
Criterion	Intercept only	Intercept and Covariates
AIC	52057.810	34829.020
SC	52067.371	35020.240
-2 Log L	52055.810	34789.020

- **AIC , SC , -2 Log L has decreased over multiple iterations hence we can imply that this is model has least data lost out of all iterations run.**



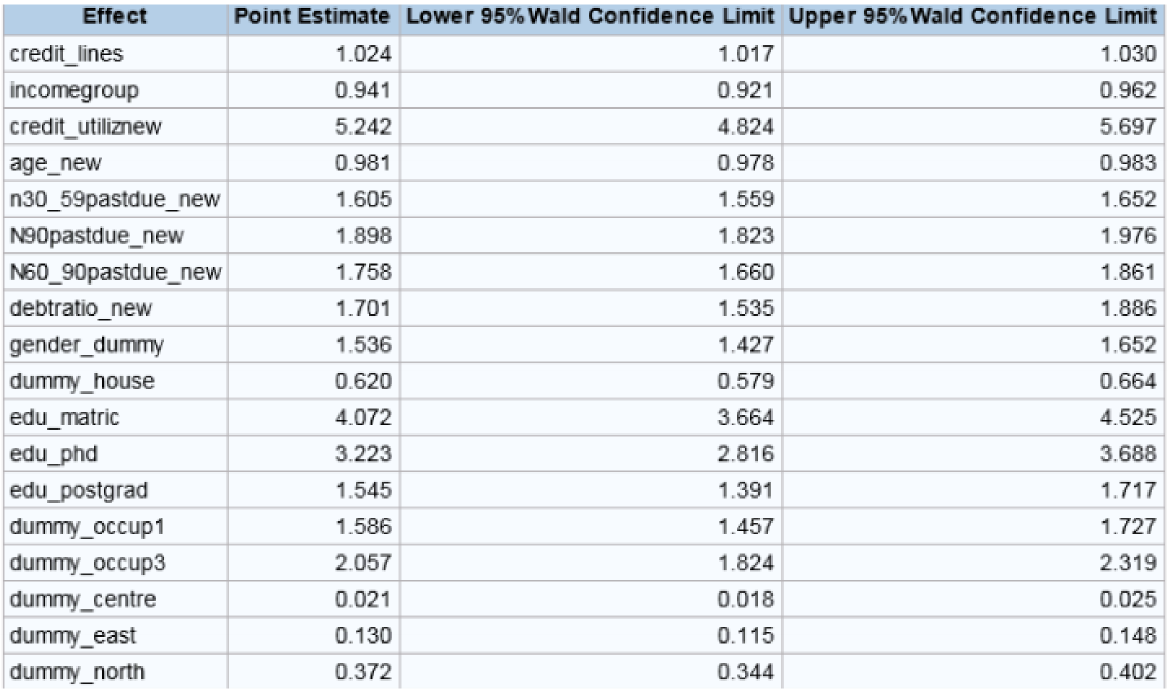
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > Chi-Square
Likelihood Ratio	17266.7892	19	<.0001
Score	23649.1188	19	<.0001
Wald	9691.1539	19	<.0001

- **Global Null hypothesis is rejected ie. At list one independent variable have significant impact on the dependant variable. Ie at list one variable in the model changes probability of default significantly.**



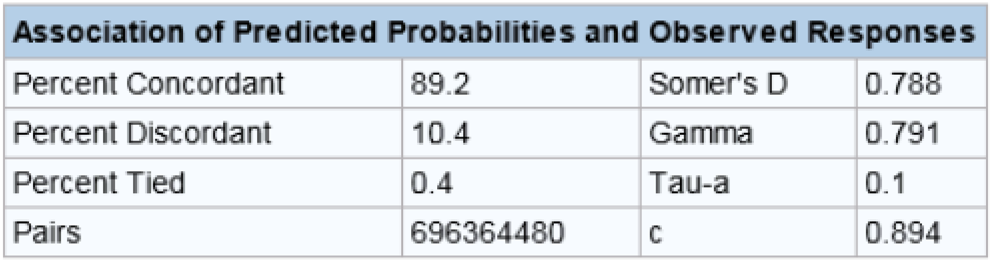
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.2748	0.0774	863.0601	<.0001
credit_lines	1	0.0233	0.00320	52.8454	<.0001
incomegroup	1	-0.0605	0.0111	29.4248	<.0001
credit_utiliznew	1	1.6568	0.0424	1525.0313	<.0001
age_new	1	-0.0197	0.00115	294.7980	<.0001
n30_59pastdue_new	1	0.4730	0.0147	1042.1033	<.0001
N90pastdue_new	1	0.6409	0.0206	972.0299	<.0001
N60_90pastdue_new	1	0.5641	0.0292	374.2654	<.0001
debtratio_new	1	0.5314	0.0525	102.5381	<.0001
gender_dummy	1	0.4290	0.0373	132.3116	<.0001
dummy_house	1	-0.4778	0.0350	186.2048	<.0001
edu_matric	1	1.4041	0.0538	680.2726	<.0001
edu_phd	1	1.1702	0.0688	289.4369	<.0001
edu_postgrad	1	0.4353	0.0537	65.8384	<.0001
dummy_occup1	1	0.4615	0.0433	113.7110	<.0001
dummy_occup3	1	0.7211	0.0612	138.6229	<.0001
dummy_centre	1	-3.8643	0.0893	1871.8454	<.0001
dummy_east	1	-2.0401	0.0644	1003.7060	<.0001
dummy_north	1	-0.9897	0.0399	613.8028	<.0001
dummy_south	1	-1.5386	0.0495	968.0418	<.0001

- **All independent variable in the model have significant impact on the dependant Variable**



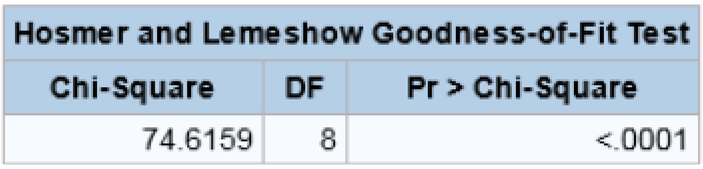
Effect	Point Estimate	Lower 95% Wald Confidence Limit	Upper 95% Wald Confidence Limit
credit_lines	1.024	1.017	1.030
incomegroup	0.941	0.921	0.962
credit_utiliznew	5.242	4.824	5.697
age_new	0.981	0.978	0.983
n30_59pastdue_new	1.605	1.559	1.652
N90pastdue_new	1.898	1.823	1.976
N60_90pastdue_new	1.758	1.660	1.861
debtratio_new	1.701	1.535	1.886
gender_dummy	1.536	1.427	1.652
dummy_house	0.620	0.579	0.664
edu_matric	4.072	3.664	4.525
edu_phd	3.223	2.816	3.688
edu_postgrad	1.545	1.391	1.717
dummy_occup1	1.586	1.457	1.727
dummy_occup3	2.057	1.824	2.319
dummy_centre	0.021	0.018	0.025
dummy_east	0.130	0.115	0.148
dummy_north	0.372	0.344	0.402

- **None of confidence interval of the odds ratio consist of 1 hence are significant.**
- **Also above table indicates percentage change in odds ratio for unit change (increase) in individual independent variable keeping all other IV constant.**



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	89.2	Somer's D	0.788
Percent Discordant	10.4	Gamma	0.791
Percent Tied	0.4	Tau-a	0.1
Pairs	696364480	c	0.894

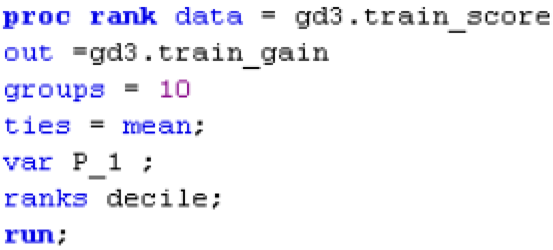
- **We achieve 89% concordance ie. Predicted probability of 1 and 0 is high in response.**



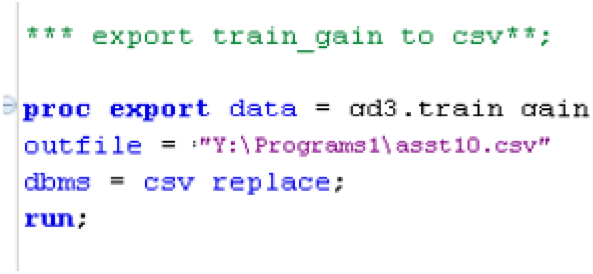
Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > Chi-Square
74.6159	8	<.0001

- **Model is not over fit.**

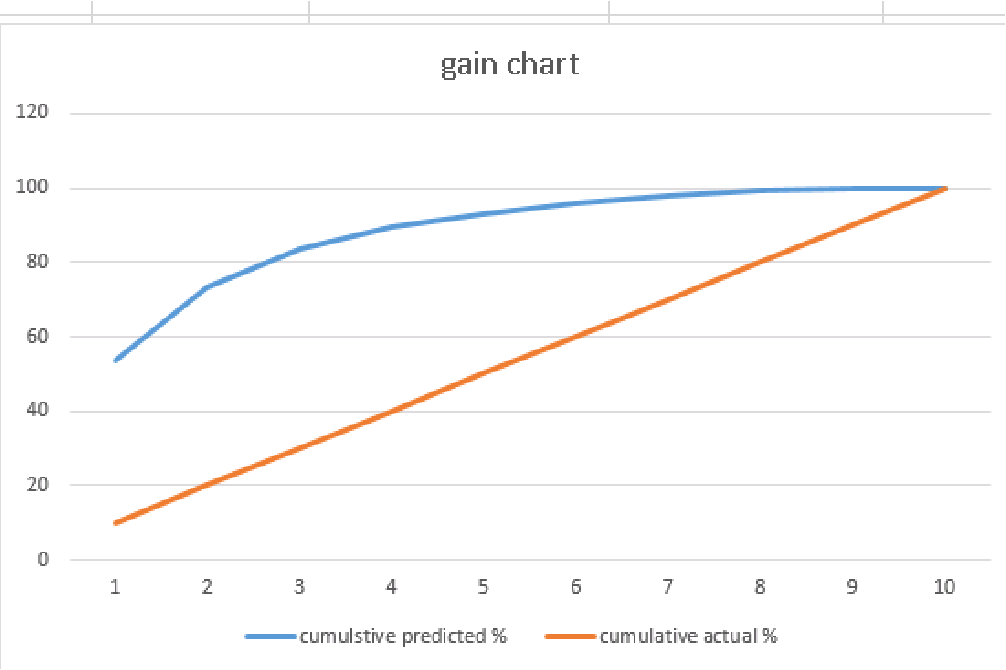
Scoring data



```
proc rank data = gd3.train_score
out =gd3.train_gain
groups = 10
ties = mean;
var P_1 ;
ranks decile;
run;
```



```
*** export train_gain to csv**;

proc export data = gd3.train_gain
outfile = "Y:\Programs1\asst10.csv"
dbms = csv replace;
run;
```

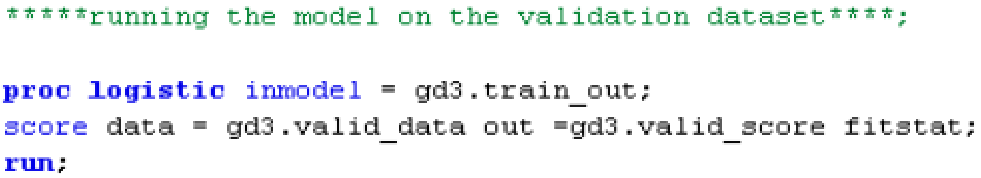


Observation:

from the above chart we can say that distance between the cumulative predicted percentage and cumulative atucal percentage is signifiantly high
So the predictive model proposed is good.

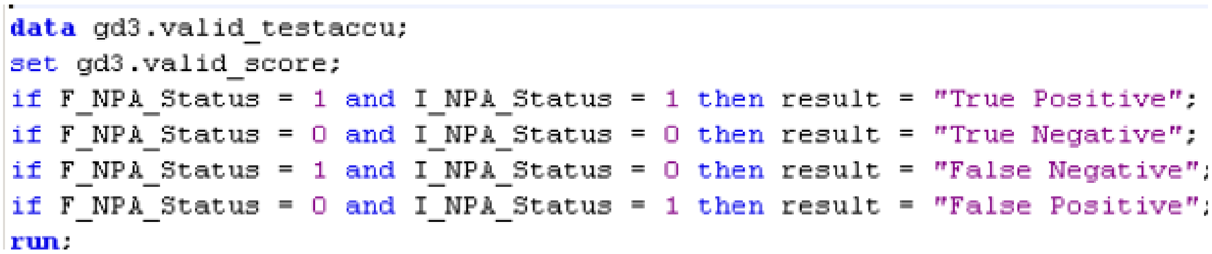
]

VALIDATION

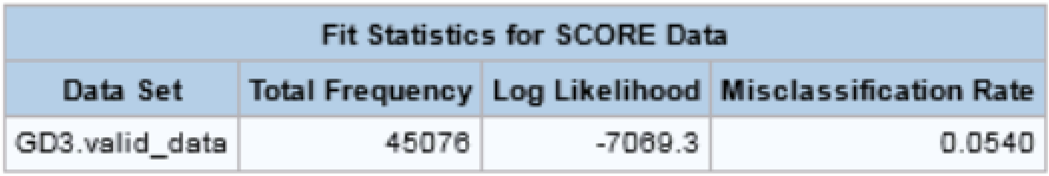


```
*****running the model on the validation dataset****;

proc logistic inmodel = gd3.train_out;
score data = gd3.valid_data out =gd3.valid_score fitstat;
run;
```

Accuracy



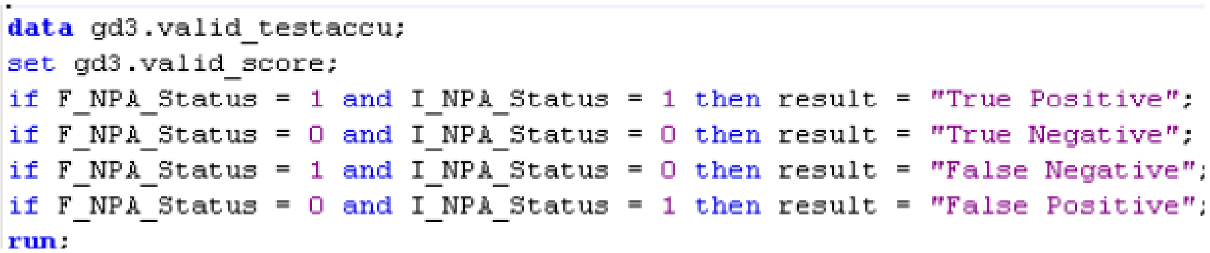
```
data gd3.valid_testaccu;
set gd3.valid_score;
if F_NPA_Status = 1 and I_NPA_Status = 1 then result = "True Positive";
if F_NPA_Status = 0 and I_NPA_Status = 0 then result = "True Negative";
if F_NPA_Status = 1 and I_NPA_Status = 0 then result = "False Negative";
if F_NPA_Status = 0 and I_NPA_Status = 1 then result = "False Positive";
run;
```



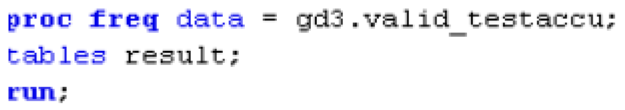
Fit Statistics for SCORE Data			
Data Set	Total Frequency	Log Likelihood	Misclassification Rate
GD3.valid_data	45076	-7069.3	0.0540

- **Misclassification rate is very low**
- **Model predicts correct event 94% of times**

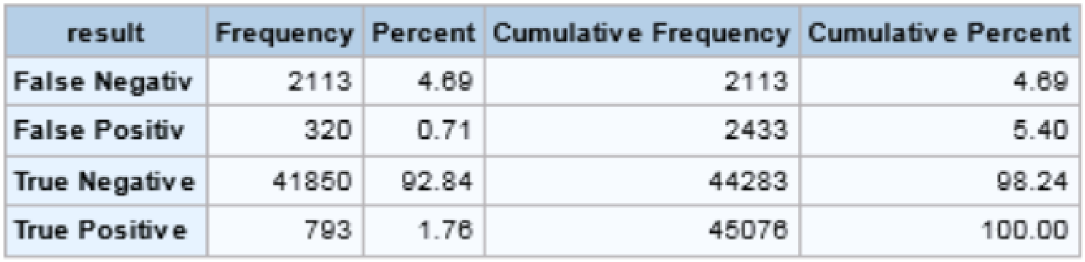
Accuracy



```
data gd3.valid_testaccu;
set gd3.valid_score;
if F_NPA_Status = 1 and I_NPA_Status = 1 then result = "True Positive";
if F_NPA_Status = 0 and I_NPA_Status = 0 then result = "True Negative";
if F_NPA_Status = 1 and I_NPA_Status = 0 then result = "False Negative";
if F_NPA_Status = 0 and I_NPA_Status = 1 then result = "False Positive";
run;
```



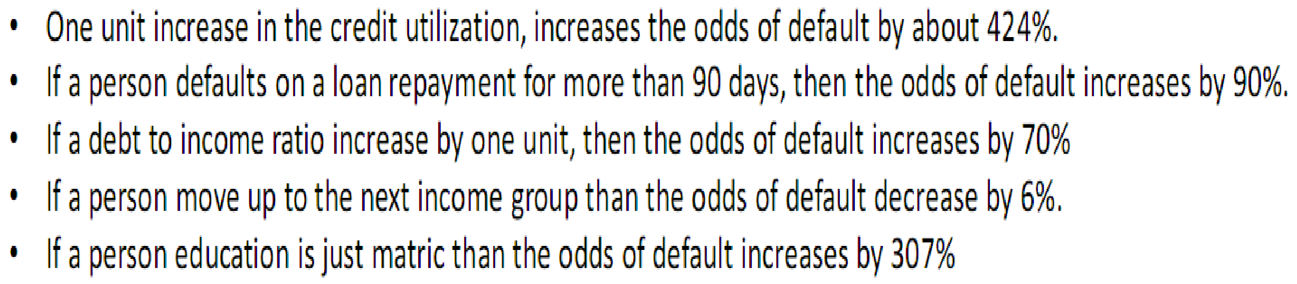
```
proc freq data = gd3.valid_testaccu;
tables result;
run;
```



result	Frequency	Percent	Cumulative Frequency	Cumulative Percent
False Negativ	2113	4.69	2113	4.69
False Positiv	320	0.71	2433	5.40
True Negative	41850	92.84	44283	98.24
True Positive	793	1.76	45076	100.00

- **.7%   False positive is**
- **4.69 % false –ve. (for alfa .32)**

OUTCOMES OF MODEL



- One unit increase in the credit utilization, increases the odds of default by about 424%.
- If a person defaults on a loan repayment for more than 90 days, then the odds of default increases by 90%.
- If a debt to income ratio increase by one unit, then the odds of default increases by 70%
- If a person move up to the next income group than the odds of default decrease by 6%.
- If a person education is just matric than the odds of default increases by 307%