

**PROJECT REPORT**

**ON**

**BANK CUTOMER PROFILE ANALYSIS**

**USING**

**LOGISTIC REGRESSION MODEL**

**INSTRUCTOR: -**

ARKAR MIN

**SUBMITTED BY: -**

RIMA NIGAM

|  |  |
| --- | --- |
| S.No. | TABLE OF CONTENTS |
| 1. | Title |
| 2. | Introduction of the Dataset |
| 3. | Executive Summary |
| 4. | Background and Business Questions |
| 5. | Objectives of the Project |
| 6. | Methodology |
| 7. | Data Exploration |
| 8. | Data Validation – Missing values and Outliers Treatment |
| 9. | Study framework of Y and Xs variables |
| 10. | Descriptive Analysis – Univariate, Bivariate and Multivariate Analysis |
| 11. | Hypothesis Testing |
| 12. | Model Building |
| 13. | Model Findings |
| 14. | Conclusions |
| 15. | Recommendations |
| 16. | Appendix |

**TITLE**

Project Report on Bank Customer Profile Analysis dataset using Logistic Regression model.

**INTRODUCTION**

The bank customers data were collected between Jan 2014-December 2014. It provided the information on active and inactive customer profiles.

Key Attributes of the Dataset:

* Active/ Inactive Customers
  + Income/Age/Credit Score
  + Products (Checking, Credit Cards, Insurance, Loan, Mortgage, Money Market, Line of Credit)
* Banking habits: Checks, Telephone, Teller visit, Point of sales
* Account Balance/NSF

**EXECUTIVE SUMMARY**

This report is part of a SAS Business project, as a capstone for the Data Science and Application. The final goal is to predict status of the customers as Active and Inactive based on the Customer’s details. The aim of this report is to explain how to explore, validate and prepare the data and build a significant model. All steps are going to be explained and supported by SAS codes including SAS Macros.

**BACKGROUND AND BUSINESS QUESTIONS**

**Time Period**: The bank customers data was collected between Jan 2014-December 2014. It provided the information on active and inactive customer profiles.

**PROBLEM STATEMENTS**

* Do the Banking habits like the Telephone Banking, Teller visits, Number of checks, Point of Sales let the customers to become active or inactive?
* Will the complimentary Direct Deposit Facility help to retain the customers?
* Do the card users’ customers stay with the bank for a longer period of time?
* Does the bank prefer checking account customers over the saving account customers?
* Is it possible that customers with Insufficient funds in their accounts still remains active with the banks?
* Do the customers having loan with the bank comes out to be the loyal customers?
* What age group of customers are the most active customers?

**OBJECTIVES OF THE PROJECT**

* The foremost aim of Profiling bank customer’s data is that it will cognize the issuer's decisions about whom to give banking facilities and what a credit limit to provide.
* Through this project, the issuers will be able to get a better understanding of their potential and current customers.
* This project will help in finding the best technique, which will lead to improvement in accuracy and helps banks to get higher profitability by customer satisfaction through a focus on the valuable customer (companies) which are considered as the main engine in the bank's profitability.

**METHODOLOGY**

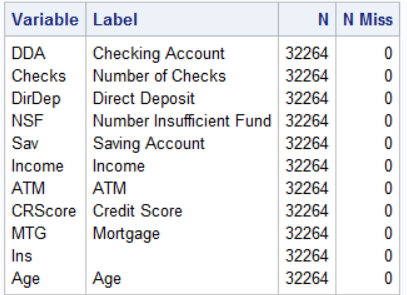
* **Data Collection:** The Bank data about the Customers is collected through various means like questionnaire, interview etc. It is primarily collected and consolidated directly from the customers. By directly analyzing the feedback of its customers, financial banking institutions can provide personalized products and services tailored to their customer needs.
* **Data Definition:** After gathering the necessary data, the data has been analyzed based on the behavior of the customers. In this project, we are going to predict the Active/Inactive customers. This research has been done on the bank data collected between Jan 2014-Dec 2014. This Methodology contains six phases of business understanding, data understanding, data preparation, modelling, evaluation and scoring. In the Business understanding phase business goals of this project were set which include receiving a general perspective of bank’s business and using that for marketing and targeting actual and potential customers. In the Understanding data phase, necessary information and concepts for achieving business goal has been achieved. In the phase of Database preparing, data were analyzed in terms of accuracy, validation of data and the volume of missing data and the necessary solutions for estimating missing data that were used. In the phase modeling, a Logistic Regression model is created to predict the binary class target variable using the significant predictors. Then created models according to the significant into two parts of training and test.
* SAS 9.4 Software is used to conduct this research using SAS Macros with around 18 SAS Procedures.

**DATASET EXPLORATION**

The foremost step to begin with the journey of analysis of data is to explore the data then various analysis and visualization methods and information acquisition will be done with SAS. This dataset was explored using various inbuilt procedure like proc contents to know the variable/ column names, values and observations in the dataset. This gave us a quick insight into the dataset. Accessing the column names gave a quick idea as from where the analysis should be start.

**DATA VALIDATION**

**HANDLING MISSING VALUES: -** Handling missing values is an important part of the EDA process. The variables which are having very few missing values as compared to the size of the dataset, we may choose to drop the rows that have missing values. Otherwise, we planned to impute missing values based on the non-missing values in a column with the mean or median value. So, by using Stdize procedure we have replaced the missing values.



**OUTLIER TREATMENT: -**An outlier is that pattern which is dissimilar with respect to all the remaining patterns in the dataset. The discovery of outliers or extreme values is useful in detection of unpredicted and unidentified data in various areas.

**Steps to Treat the outliers: -**

1. Find the Q1, Q2 and IQR using PROC MEANS.
2. If a value is higher than 3 times of IQR above the Q3,It is considered as extreme outlier. Similarly, if it is lower than that then also it will be considered as extreme outlier.
3. A single macro program can detect and treat the outliers at the same time and remove the observations that are out of range.

**STUDY FRAMEWORK OF Y AND X VARIABLES**

In this project, my Target variable is Checking Account where I will be predicting the Active and the Inactive customers based on my Xs variables and they are -Age, Branch, Bank Area, ATM, Credit Card, Phone, Teller, NSF, Credit Score, Checks, Direct Deposit, Mortgage.

**DATA ANALYSIS**

After exploring the dataset, there was a need to analyze the data. In this study, I aimed to find a relation between Checking Account and different variables, and compare them.

**DESCRIPTIVE ANALYSIS**

Data Visualization is the most important step because it gives the visual summary of the data which makes it easier for a layman as well to understand and identify the data patterns and trends than looking through thousands of rows on a spreadsheet.

So, to gain insights, we have visualized our Bank Customer Profile Analysis dataset using various visualization methods. In this project, we have used various graphs like bar plot, boxplot, waterfall plot, butterfly plot, histogram and mosaic plot to give a clear picture of the dataset.

By analysing all the plots created in the project code, we can see that,

* Customers are preferring to keep the checking account over saving account for their day-to-day transactions.
* Maximum branches are not offering the direct deposit facility to deposit the funds electronically into a bank account or the customers are not preferring may be the bank is charging a certain amount for the deposits to take place between banks.
* Average income group customers are the most Active customers. The Branch B4 has the maximum number of Average Income customers.
* The most active customers fall under the ‘Teen’ and ‘Adult’ age group as they seem to do much transactions.
* Maximum customers are having very good score between 600-700 as they are the most active customers because may be they are doing transactions on the regular basis and paying bills on time as well.
* Maximum customers in Urban Areas are preferring to maintain the Checking Accounts over the Saving Accounts for daily transactions.
* The more the customers are using their phone banking and the teller visits, the more they are becoming inactive, may be because the customers are not satisfied with the services or they may not be able to solve their queries well neither through phone or in personal.
* Customers are only opening the loan account to apply for the mortgage but seems not to be an active customer. The reason could be either they are not liking the bank services but due to low rate of interest on loan they have opened the accounts.
* B4 and B2 are the oldest branches as per the age of the Account.
* The branch B4 and B2 are having the maximum inactive customers. The bank must find the reason for the customer’s churn in order to find the way to retain the customers.
* Most of the ATM users are the Active customers as they prefer to escape from long queues of Teller visits for withdrawing, deposit and transferring money.

**HYPOTHESIS TESTING**

The **hypothesis testing** is a statistical test used to determine whether the hypothesis assumed for the sample of data stands true for the entire population or not.

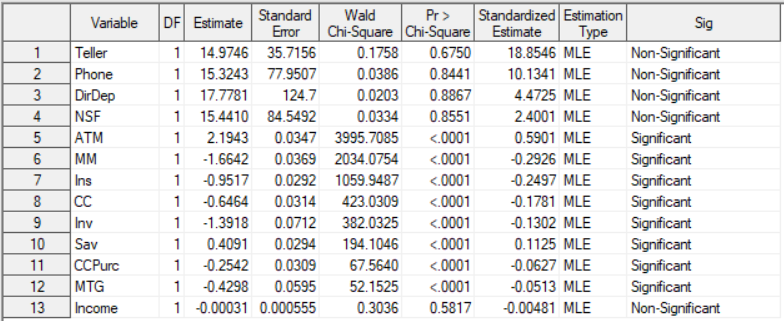
**Setting the Null Hypothesis (Ho) and the Alternative Hypothesis (Ha) at 5% Significance level**

Ho: The Target variable is not having a relationship with the Independent Variables.

Ha: The Target variable is having a relationship with the Independent Variables.

**Based on this dataset, I have prepared some hypothesis questions to answer in the next slide.**

* Does the customers having mortgage and loan installments turned out to the Active customers?
* Are the banking habits – checks, phone, teller etc. creating any impact on the status of the customers?
* Will the Card users customers are the active customers?
* Which variables are having relationship with the Target Variable?

****

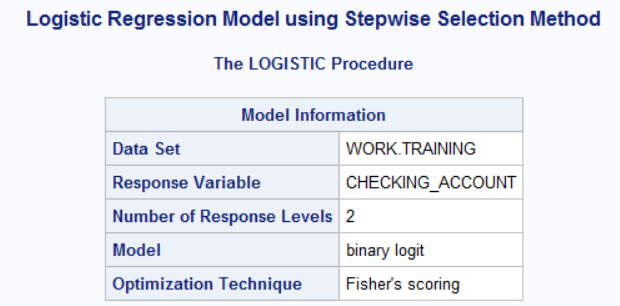
* As the P-values of Mortgage and Instalment variable is less than 0.05. That shows these variables are helping the customers to remain Active.
* The banking habits variable comes out to insignificant that means there is no relationship of them with the Target Variable.
* Card user customers also turned out to be the Active customers.
* Out of all variables selected, ATM, CC, Ccpurc, MTG, Ins shows the significant relationship with the target variable.

**MODEL BUILDING**

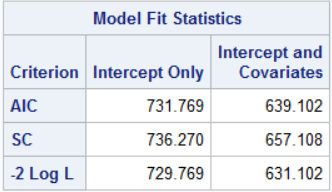
Logistic regression describes the relationship between a categorical response variable and a set of predictor variables. A categorical response variable can be a binary variable, an ordinal variable or a nominal variable. Each type of categorical variables requires different techniques to model its relationship with the predictor variables.

In this dataset, I am having a binary response variable as Active/Inactive.

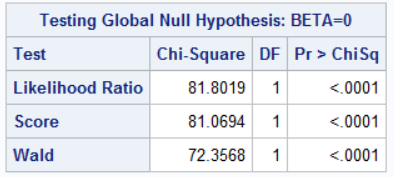
So, I have built a Logistic Regression Model using a Stepwise Selection Method.

****

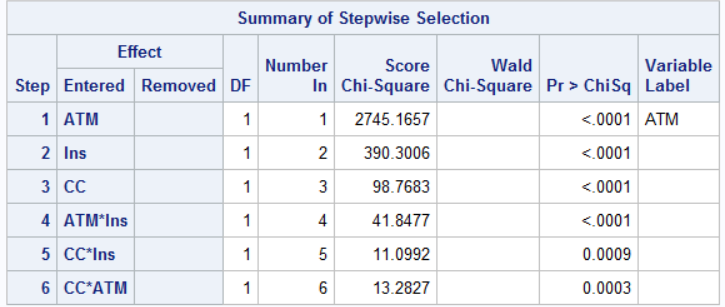
**MODEL FINDINGS**



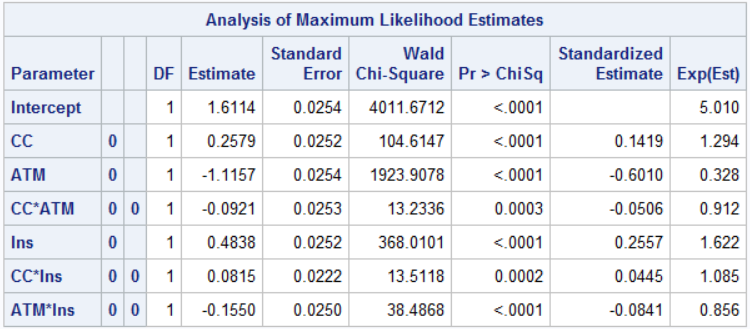
* Ultimately, the model with the smallest AIC is considered the best.
* SC penalizes for the number of predictors in the model and the smallest SC is most desirable.
* The -2 Log L is used in hypothesis tests for nested models.

****

* Likelihood Ratio – This is the Likelihood Ratio (LR) Chi-Square test that at least one of the predictors’ regression coefficient is not equal to zero in the model.
* Score – This is the Score Chi-Square Test that at least one of the predictors’ regression coefficient is not equal to zero in the model.
* Wald – This is the Wald Chi-Square Test that at least one of the predictors’ regression coefficient is not equal to zero in the model.

****

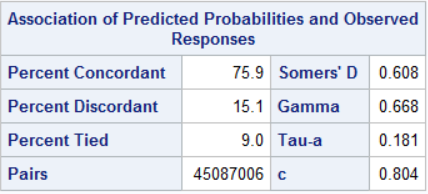
After addition of all the selected variables into the model, the model selected the following variables and they are all significant as their p values are less than 0.05. As the model is predicting whether the customer is active or not so the card users and the customers having loan instalment sounds more active.



The logistic regression model models the log odds of a positive response as a linear combination the predictor variables. This is written as

**log[ p / (1-p) ] = b0 + b1\*x1 + b2\*x2 + b3 \*x3..**,

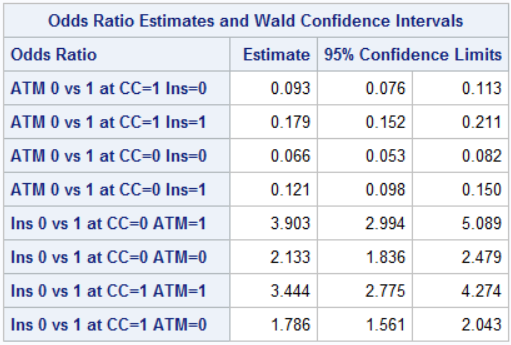
**Here all the variables are significant as their P-values are less than 0.05.**



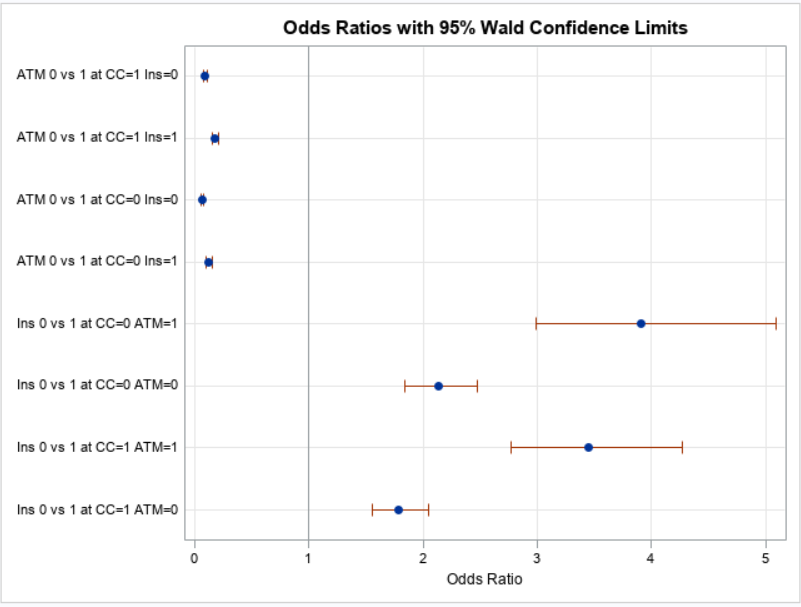
**Percent Concordant** – A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response.

**Percent Discordant** – If the observation with the lower ordered response value has a higher predicted mean score than the observation with the higher ordered response value, then the pair is discordant.

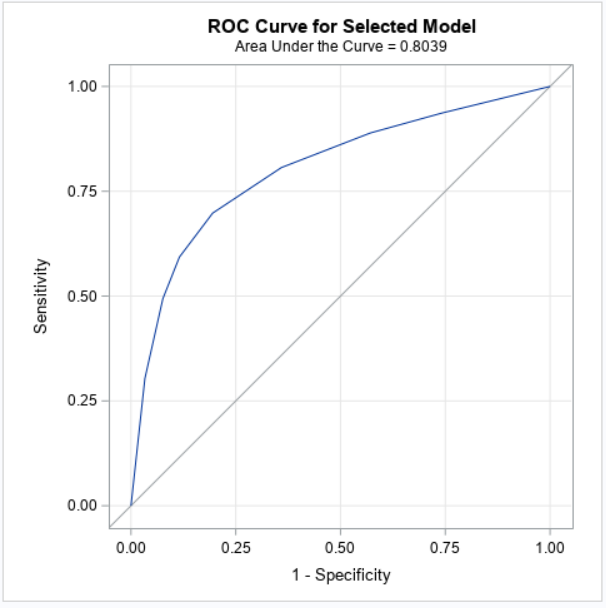
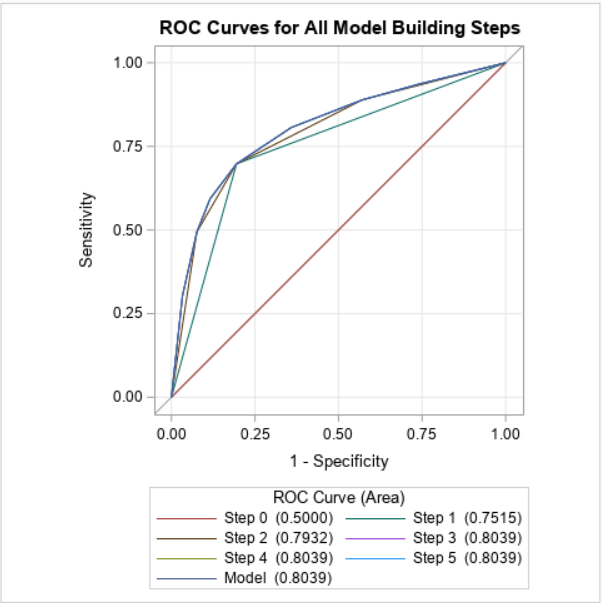
**Percent Tied** – If a pair of observations with different responses is neither concordant nor discordant, it is a tie.



* A confidence limit that contains 1 will not be significant.



The default odds ratio plot is shown. Four estimates are less than 1 and four are greater than 1. These four confidence intervals in contains 1, which indicates ratios that are not significant.

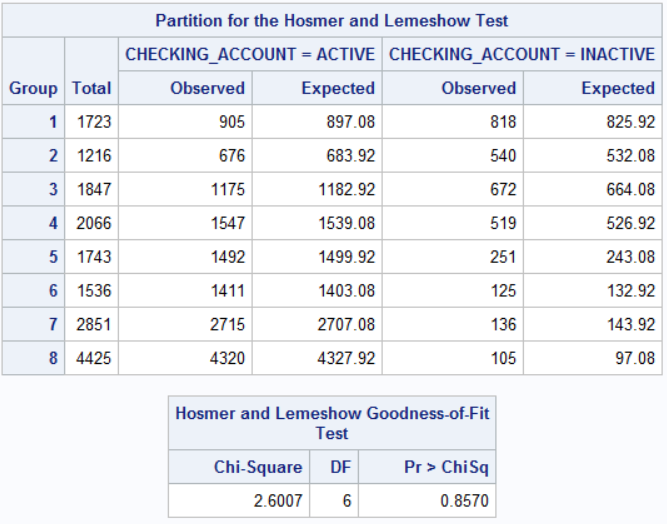
****

The AUC value lies between 0.5 to 1 where 0.5 denotes a bad classifier and 1 denotes an excellent classifier.

Here AUC = 0.8039, which is near to 1 that means it is a good model.

**Hosmer and Lemeshow Test**

The [Hosmer and Lemeshow goodness of fit (GOF) test](http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm) is a way to assess whether there is evidence for lack of fit in a logistic regression model. Simply put, the test compares the expected and observed number of events in bins defined by the predicted probability of the outcome

****

The partition table shows the observed and expected count or events in each decile of the predicted probabilities.

A large value of Chi-squared (with small p-value < 0.05) indicates poor fit and small Chi-squared values (with larger p-value closer to 1) indicate a good logistic regression model fit.

**FINAL RESULT OF THE ANALYSIS**

* Using the stepwise selection method, the best model came out was having three variables ATM, INSTALLMENT AND CREDIT CARD and the result was 80.4% which indicates a good-fit model.
* Customer using Credit cards are the most Active customers and banks should give them some offers so that they can retain them long either by increasing the Credit Limit of the most valuable customers or give them some offers on the shopping through credit cards. Interest is the source of handsome revenue for the banks.
* Those customers who are having loan installments with the banks also turned out to be the Active customers. In that case, Bank should try to convert or retain them as a long-lasting customer by offering them good services as may be, they are with the bank till their loan period.
* ATM helps the customers to avoid long teller lines to withdraw and deposit cash which make them Active customers as well but on the other hand it is a benefit for the banks too as they charge Certain fees for doing the transactions which is also a major revenue for them.
* So, this is not a bad model though, however, we still have a larger scope to improve the model by using other variables.

**RECOMMENDATIONS**

* Bank should concentrate more on the Card users’ customers as they are the one who are the most Active customers and bringing revenue to the banks.
* In order to pay the loan instalments on time, the customers having loan with the bank comes out to be an active customer. Bank should try them to make their Potential customers.
* Bank should try to improve the Teller Services and Phone Banking services in order to retain more customers.
* Bank should provide complementary Direct Deposit Facility to entertain more customers.
* As per the Age group, Bank should concentrate more on Young and Adult Age Group as they are not that much satisfied with the bank services.
* In terms of Area, the branches in Rural and Semi-urban are not doing well, bank should taken this into consideration.
* The Branches B4 and B2 are the oldest branches but as they are having maximum inactive customers so bank should find out the reason behind that. May be these branches are using old style of Banking habits.
* Bank should also try to convert the Loan customers to their potential customers.

**APPENDIX**

/\*MACRO PROGRAM TO ASSIGN THE LIBRARY\*/

%LET DIR =C:\Users\Veena Nigam\Desktop\SAS Documents\SAS Business Project ;

%PUT &DIR.;

%LET DSN = RIMA.PROJECT4;

LIBNAME RIMA "&DIR.";

/\*Describing the properties of the project data\*/

**PROC** **CONTENTS** DATA = &DSN. OUT =CONTENTS VARNUM SHORT;

**RUN**;

/\*COLUMN NAMES\*/

/\*AcctAge DDA DDABal CashBk Checks DirDep NSF NSFAmt Phone Teller Sav SavBal ATM ATMAmt POS POSAmt CD CDBal IRA IRABal LOC LOCBal ILS ILSBal

MM MMBal MMCred MTG MTGBal CC CCBal CCPurc SDB Income HMOwn LORes HMVal Age CRScore Moved InArea Ins Branch Res Dep DepAmt Inv InvBal\*/

/\*SPLIT INTO NUMERIC AND CATEGORICAL\*/

**PROC** **SQL**;

SELECT NAME INTO : NUM\_ONLY SEPARATED BY " "

FROM CONTENTS

WHERE TYPE EQ **1**

;

SELECT NAME INTO : CHAR\_ONLY SEPARATED BY " "

FROM CONTENTS

WHERE TYPE EQ **2**

;

**QUIT**;

%PUT &NUM\_ONLY;

%PUT &CHAR\_ONLY;

%LET X = %SCAN(&NUM\_ONLY.,1);

%PUT &X;

%LET X = %SCAN(&NUM\_ONLY.,2);

%PUT &X;

%LET X = %SCAN(&NUM\_ONLY.,3);

%PUT &X;

/\*DATA PROFILING\*/

**%MACRO** PROF\_NUMERIC(DSN = , VAR= );

%LET N = %SYSFUNC(COUNTW(&VAR.));

%DO I = **1** %TO &N;

%LET X = %SCAN (&VAR, &I);

TITLE "DISTRIBUTION OF ALL NUMERIC VARIABLES:SUMMARY";

PROC MEANS DATA = RIMA.RISK N NMISS Q1 Q3 MEAN MEDIAN MAX;

VAR &X.;

RUN;

TITLE "DISTRIBUTION OF &X. : HISTOGRAM AND DENSITY CURVE";

PROC SGPLOT DATA = &DSN.;

HISTOGRAM &X.;

DENSITY &X./ TYPE=KERNEL;

KEYLEGEND/LOCATION=INSIDE POSITION = TOPRIGHT ACROSS=**1** NOBORDER;

RUN;

QUIT;

TITLE "DISTRIBUTION OF &X. :VERTICLE BOX PLOT";

PROC SGPLOT DATA = &DSN.;

VBOX &X.**.**;

YAXIS GRID;

XAXIS DISPLAY=(NOLABEL);

RUN;

QUIT;

%END;

**%MEND**;

ODS PDF FILE = "&DIR.\NUMERIC\_SUMMARY\_&SYSDATE9..PDF";

%***PROF\_NUMERIC*** (DSN = &DSN., VAR = &NUM\_ONLY);

ODS PDF CLOSE;

/\*CHARACTER VARIABLES\*/

%LET X = %SCAN(&CHAR\_ONLY.,1);

%PUT &X;

%LET X = %SCAN(&CHAR\_ONLY.,2);

%PUT &X;

**%MACRO** PROF\_CHAR(DSN = ,CVAR = , COLOR= );

TITLE "COUNT OF ALL CATEGORICAL VARIABLES:SUMMARY";

PROC FREQ DATA= &DSN. ORDER=FREQ;

TABLE &CVAR./MISSING;

RUN;

%LET N = %SYSFUNC(COUNTW(&CVAR.));

%DO I = **1** %TO &N;

%LET X = %SCAN(&CVAR.,&I);

TITLE "COUNT BY %UPCASE(&X)";

PROC FREQ DATA = &DSN. ORDER=FREQ;

TABLE &X./MISSING;

RUN;

TITLE "COUNT BY %UPCASE(&X.)";

PROC SGPLOT DATA = &DSN.;

VBAR &X/categoryorder=respasc barwidth=**0.6** fillattrs= &COLOR.;

xaxis display=(nolabel);

RUN;

QUIT;

PROC TEMPLATE;

DEFINE STATGRAPH PIE;

BEGINGRAPH;

ENTRYTITLE "COUNT BY %UPCASE(&X.)";

LAYOUT REGION;

PIECHART CATEGORY=&X / DATALABELLOCATION=OUTSIDE DATASKIN = CRISP DATALABELCONTENT = ALL CATEGORYDIRECTION = CLOCKWISE START = **180** NAME = 'PIE' ;

DISCRETELEGEND 'PIE';

ENDLAYOUT;

ENDGRAPH;

END;

RUN;

PROC SGRENDER DATA = &DSN. TEMPLATE = PIE;

RUN;

%END;

**%MEND**;

%***PROF\_CHAR***(DSN = &DSN.,CVAR = &CHAR\_ONLY , COLOR = GRAPHDATA9 );

/\*Select Y(DDA/Checking Account) and Xs variables (Using Proc Sql)\*/

**PROC** **SQL**;

CREATE TABLE RIMA.BANK\_PROFILE AS

SELECT DDA, Checks, DirDep, NSF,Phone, Teller, Sav,Income, ATM, CC, CCPurc, CRScore, Res, Branch, MTG,INS,AGE

FROM RIMA.PROJECT4;

**QUIT**;

/\*Selecting a new data based on variable identification process\*/

%LET DSN\_1 = RIMA.BANK\_PROFILE;

/\*Missing Values Detection\*/

/\* create a format to group missing and nonmissing \*/

**proc** **format**;

value $charmiss ' '='Missing' other='Not Missing';

value nummiss **.** ='Missing' other='Not Missing';

**run**;

**proc** **freq** data= &DSN\_1.;

format \_CHAR\_ $charmiss.;

tables \_CHAR\_ / missing missprint nocum nopercent;

format \_NUMERIC\_ nummiss.;

tables \_NUMERIC\_ / missing missprint nocum nopercent;

**run**;

/\*Other way to find the missing values\*/

**PROC** **MEANS** DATA = &DSN\_1. N NMISS;

**RUN**;

/\*UNIVARIATE ANALYSIS\*/

/\*Analysing Numerical Column\*/

**%MACRO** UNIVARIATE\_NUM(DSN= , VAR= );

%LET N = %SYSFUNC(COUNTW(&VAR.));

%DO I = **1** %TO &N;

%LET X = %SCAN (&VAR, &I);

PROC UNIVARIATE DATA = &DSN.;

TITLE "COMPREHENSIVE UNIVARIATE ANALYSIS OF %UPCASE(&X.) ";

VAR &X.;

RUN;

TITLE "DISTRIBUTION OF &X. :BOX PLOT";

PROC SGPLOT DATA = &DSN.;

VBOX &X.;

YAXIS GRID;

XAXIS DISPLAY=(NOLABEL);

RUN;

TITLE "CORRELATION OF &X. WITH CHECKING ACCOUNT";

PROC CORR DATA=&DSN. PLOTS(MAXPOINTS=NONE) = MATRIX(HISTOGRAM);

VAR &X.;

WITH DDA;

RUN;

%END;

**%MEND**;

ODS PDF FILE = "&DIR.\Images.PDF";

%***UNIVARIATE\_NUM***(DSN = &DSN\_1.,VAR = &NUM\_ONLY.);

ODS PDF CLOSE;

/\*AFTER LOOKING INTO THE BOX PLOTS- THESE ARE THE COLUMNS IN WHICH OUTLIERS WERE DETECTED --

TELLER, PHONE,NSF, MTG, DDA, CHECKS, CRSCORE, CCPURC, AGE

I will treat these outliers after handling the missing values\*/

/\*Analysing Categorical Variables\*/

**%MACRO** UNIVARIATE\_CAT(DSN = , CVAR = ,DEP = ,COLOR = );

TITLE "COUNT OF ALL CATEGORICAL VARIABLES:SUMMARY";

PROC FREQ DATA= &DSN. ORDER=FREQ;

TABLE &CVAR./MISSING;

RUN;

%LET N = %SYSFUNC(COUNTW(&CVAR.));

%DO I = **1** %TO &N;

%LET X = %SCAN(&CVAR.,&I);

TITLE "COUNT BY %UPCASE(&X)";

PROC FREQ DATA = &DSN. ORDER=FREQ;

TABLE &X./MISSING;

RUN;

TITLE "COUNT BY %UPCASE(&X.)";

PROC SGPLOT DATA = &DSN.;

VBAR &X/categoryorder=respasc barwidth=**0.6** fillattrs= &COLOR.;

xaxis display=(nolabel);

RUN;

QUIT;

PROC ANOVA DATA = &DSN.;

TITLE "RELATIONSHIP BETWEEN DDA AND &X.";

CLASS &X.;

MODEL &DEP. = &X.;

/\*MEANS &DEP./SCHEFFE;\*/

RUN;

QUIT;

%END;

**%MEND**;

%***UNIVARIATE\_CAT***(DSN = &DSN\_1.,CVAR = &CHAR\_ONLY ,DEP = DDA, COLOR = GRAPHDATA9 );

/\*Only the Res column is more significant and shows a strong relationship with the DDA as its P value is less than 0.05 \*/

/\*BIVARIATE ANALYSIS\*/

**%MACRO** BIVARIATE (DSN = ,VAR1 = , VAR2 = ,VAR3 = ,VAR4 = );

PROC SGPLOT DATA = &DSN\_1.;

VBAR &VAR1. / GROUP = &VAR2. GROUPDISPLAY = CLUSTER;

TITLE 'DISTRIBUTION OF CUSTOMERS HAVING THE CHECKING ACCOUNT AS PER THE AREA';

RUN;

PROC SGPLOT DATA = &DSN\_1.;

VBAR &VAR3. / GROUP = &VAR2. GROUPDISPLAY = CLUSTER;

TITLE 'DISTRIBUTION OF CUSTOMERS HAVING THE SAVING ACCOUNT AS PER THE AREA';

RUN;

PROC SGPLOT DATA = &DSN\_1.;

TITLE'Customers having Insufficient Funds';

VBAR &VAR4./GROUP = &VAR1.;

RUN;

QUIT;

**%MEND**;

ODS PDF FILE = "&DIR.\Images.PDF";

%***BIVARIATE***(DSN = &DSN\_1.,VAR1 = DDA, VAR2= RES, VAR3 = SAV, VAR4 = NSF);

ODS PDF CLOSE;

/\*As it is a normal distribution but having outliers so we can replace the missing values with median\*/

/\*Missing Values Treatment\*/

**%MACRO** MISSING (DSN = , OUT = );

PROC MEANS DATA = &DSN. N NMISS;

RUN;

PROC STDIZE DATA = &DSN. OUT= &OUT. METHOD= MEDIAN REPONLY;

VAR Age Income CRScore;

RUN;

/\*Checking if the missing values are replaced by median value\*/

PROC MEANS DATA = &OUT. MAXDEC=**2** N NMISS;

RUN;

**%MEND**;

%***MISSING***(DSN= &DSN\_1., OUT = RIMA.BANK\_NEW);

/\*Replacing the missing values of Binary Class Variable with the Word 'Missing'\*/

/\* Recode by using IF-THEN \*/

**%MACRO** RECODE (DATA = , VAR = , OUTPUT = );

data &OUTPUT.;

set &DATA.;

/\* use IF-THEN logic to recode gender \*/

length &VAR.\_N $**8**;

if &VAR.= **0** then &VAR.\_N = "NO";

else if &VAR.=**1** then &VAR.\_N = "YES";

else &VAR.\_N = "MISSING ";

DROP &VAR.;

run;

**%MEND**;

%***RECODE*** (DATA = RIMA.BANK\_NEW, VAR = CC, OUTPUT = RIMA.BNK\_001);

%***RECODE*** (DATA = RIMA.BNK\_001, VAR = TELLER, OUTPUT = RIMA.BNK\_02);

%***RECODE*** (DATA = RIMA.BNK\_02, VAR =CCPurc , OUTPUT = RIMA.BNK\_03);

%***RECODE*** (DATA = RIMA.BNK\_03, VAR = Phone, OUTPUT = RIMA.BNK\_04);

/\*CHECKING IF ALL THE MISSING VALUES HAS BEEN TREATED\*/

**PROC** **MEANS** DATA=RIMA.BNK\_04 N NMISS;

**RUN**;

/\*Other way to do the Missing Value Treatment of Binary class variable (with mode)\*/

**%MACRO** REPLACE (INPUT= ,STATS= ,VARS= ,OUTPUT= );

\* GENERATE ANALYSIS RESULTS ;

PROC UNIVARIATE DATA=&INPUT NOPRINT;

VAR &VARS;

OUTPUT OUT=DUMMY &STATS= &VARS;

RUN;

\* CONVERT TO VERTICAL ;

PROC TRANSPOSE DATA=DUMMY OUT=DUMMY;

RUN;

\* REPLACE MISSING WITH ANALYSIS RESULTS ;

DATA &OUTPUT;

SET &INPUT;

ARRAY VARS &VARS ;

DO I =**1** TO DIM(VARS);

SET DUMMY(KEEP=COL1) POINT= I ;

VARS(I)=COALESCE(VARS(I),COL1);

DROP COL1 ;

END;

RUN;

**%MEND**;

/\*Variables with outliers

TELLER, PHONE,NSF, MTG, DDA, CHECKS, CRSCORE, CCPURC, AGE\*/

/\*TREATING THE OUTLIERS\*/

**%MACRO** OUTLIER (DATA= , VARNAME = , THRESHOLD= , OUTPUT = );

PROC MEANS DATA = &DATA. MAXDEC=**2** N P25 P75 QRANGE;

VAR &VARNAME.;

RUN;

PROC MEANS DATA = &DATA. MAXDEC = **2** N P25 P75 QRANGE;

VAR &VARNAME.;

OUTPUT OUT = RIMA.DEL P25 = Q1 P75 = Q3 QRANGE=IQR;

RUN;

DATA RIMA.TEMP1;

SET RIMA.DEL ;

LOWER\_LIMIT = Q1 - (&THRESHOLD.\*IQR);

UPPER\_LIMIT = Q1 + (&THRESHOLD.\*IQR);

RUN;

/\*CARTESIAN PRODUCT\*/

PROC SQL;

CREATE TABLE RIMA.DATA\_01 AS

SELECT A.\*,B.LOWER\_LIMIT, B.UPPER\_LIMIT

FROM &DATA. AS A, RIMA.TEMP1 AS B

;

QUIT;

DATA RIMA.DATA\_02;

SET RIMA.DATA\_01;

IF &VARNAME. LE LOWER\_LIMIT THEN RANGE = "BELOW LOWER LIMIT";

ELSE IF &VARNAME. GE UPPER\_LIMIT THEN RANGE = "ABOVE UPPER LIMIT";

ELSE RANGE = "WITHIN RANGE";

RUN;

QUIT;

/\*PRINTING WITHIN RANGE DATA\*/

PROC SQL;

CREATE TABLE &OUTPUT.(DROP= LOWER\_LIMIT UPPER\_LIMIT RANGE) AS

SELECT \*

FROM RIMA.DATA\_02

WHERE RANGE = "WITHIN RANGE";

QUIT;

PROC PRINT DATA=&OUTPUT.;

RUN;

/\*DROP THE UNNECESSARY VARIABLES AND THE DATASETS \*/

/\* PROC DELETE DATA= &DATA.;\*/

/\* RUN;\*/

/\* DATA &OUTPUT.;\*/

/\* DROP = LOWER\_LIMIT UPPER\_LIMIT RANGE);\*/

/\* RUN;\*/

/\* PROC DELETE DATA= RIMA.DEL; RUN;\*/

/\* PROC DELETE DATA= RIMA.TEMP1; RUN;\*/

/\* PROC DELETE DATA= RIMA.DATA\_01; RUN;\*/

/\* PROC DELETE DATA= RIMA.DATA\_02; RUN;\*/

**%MEND** OUTLIER;

%***OUTLIER***(DATA = RIMA.BNK\_04, VARNAME = CRSCORE, THRESHOLD = **1.5**, OUTPUT = RIMA.BANK\_000);

/\*DATA TRANSFORMATION: CONTINUOUS TO CATEGORICAL VARIABLES\*/

/\*RECODING THE DDA and SAV VARIABLE AS ACTIVE AND INACTIVE CUSTOMERS\*/

**%MACRO** RECODE\_SAV (DATA = , VAR = , OUTPUT = );

data &OUTPUT.;

set &DATA.;

/\* use IF-THEN logic to recode gender \*/

length SAVING\_ACCOUNT $**8**;

if &VAR.= **1** then SAVING\_ACCOUNT = "ACTIVE";

else SAVING\_ACCOUNT = "INACTIVE ";

DROP &VAR.;

run;

PROC PRINT DATA = &OUTPUT.;

RUN;

**%MEND**;

%***RECODE\_SAV***(DATA = RIMA.BANK\_FINAL , VAR = SAV , OUTPUT = RIMA.BANK\_UPDATED\_0);

**%MACRO** RECODE\_DDA (DATA = , VAR = , OUTPUT = );

data &OUTPUT.;

set &DATA.;

/\* use IF-THEN logic to recode gender \*/

length CHECKING\_ACCOUNT $**8**;

if &VAR.= **1** then CHECKING\_ACCOUNT = "ACTIVE";

else CHECKING\_ACCOUNT = "INACTIVE ";

DROP &VAR.;

run;

PROC PRINT DATA = &OUTPUT.;

RUN;

**%MEND**;

%***RECODE\_DDA*** (DATA = RIMA.BANK\_UPDATED\_0 , VAR = DDA , OUTPUT = RIMA.BANK\_UPDATED);

/\*GROUPING THE DATA USING HPBIN PROCEDURE\*/

**%MACRO** BIN (DATA = , VAR1 = ,BINS = , OUT =);

PROC HPBIN DATA = &DATA. OUTPUT = &OUT. NUMBIN = &BINS.;

INPUT &VAR1.;

RUN;

**%MEND**;

%***BIN*** (DATA = RIMA.BANK\_UPDATED , VAR1 = AGE , BINS = **3** , OUT = RIMA.BANK1);

/\*Creating the final dataset after treating the missing values and the outliers\*/

**PROC** **SQL**;

CREATE TABLE RIMA.BANK\_FINALISED AS

SELECT \*

FROM RIMA.BANK\_UPDATED;

**QUIT**;

%LET DSN\_2 = RIMA.BANK\_FINALISED;

/\*The other way of grouping data using Proc Format if we know the range\*/

/\*Transforming Age column into groups\*/

**PROC** **MEANS** DATA = &DSN\_2. MAXDEC=**1** N NMISS MIN MAX RANGE;

VAR AGE;

**RUN**;

**PROC** **FORMAT** ;

VALUE AGE\_GROUP LOW- **30** = "YOUNG"

**31**- **60** = " ADULT"

**61**- HIGH = "SENIOR";

**RUN**;

%LET DSN = &DSN\_2.;

%LET VAR1 = AGE;

%LET VAR2 = CHECKING\_ACCOUNT;

**PROC** **FREQ** DATA = &DSN\_2.;

TITLE "RELATIONSHIP BETWEEN &VAR1. AND &VAR2.";

TABLE &VAR1. \* &VAR2. /CHISQ NOROW NOCOL ;

FORMAT &VAR1. AGE\_GROUP.;

**RUN**;

**PROC** **SGPLOT** DATA = &DSN\_2.;

TITLE "ACTIVE/INACTIVE CUSTOMERS AS PER THE AGE GROUP";

VBAR &VAR1./ GROUP = &VAR2. ;

FORMAT &VAR1. AGE\_GROUP.;

**RUN**;

**QUIT**;

/\*Transforming Credit Score column into Categories\*/

**PROC** **MEANS** DATA = &DSN\_2. MAXDEC=**1** N NMISS MIN MAX RANGE;

VAR CRSCORE;

**RUN**;

**PROC** **FORMAT** ;

VALUE CRS LOW- **600** = "GOOD"

**601**- **700** = " VERY GOOD"

**701**- HIGH = "EXCELLENT";

**RUN**;

%LET DSN = &DSN\_2.;

%LET VAR1 = CRSCORE;

%LET VAR2 = CHECKING\_ACCOUNT;

**PROC** **FREQ** DATA = &DSN\_2.;

TITLE "RELATIONSHIP BETWEEN &VAR1. AND &VAR2.";

TABLE &VAR1. \* &VAR2. /CHISQ NOROW NOCOL ;

FORMAT &VAR1. CRS.;

**RUN**;

**PROC** **SGPLOT** DATA = &DSN\_2.;

TITLE "ACTIVE/INACTIVE CUSTOMERS AS PER THE CREDIT SCORE DISTRIBUTION";

VBAR &VAR1./ GROUP = &VAR2. ;

FORMAT &VAR1. CRS.;

**RUN**;

**QUIT**;

/\*BUSINESS QUESTIONS\*/

ODS PDF FILE = "C:\Users\Veena Nigam\Desktop\SAS Documents\SAS Business Project\Images.PDF";

/\*Which type of Account customers are preferring to keep with the bank?\*/

/\*Distribution of Checking Account\*/

**PROC** **TEMPLATE**;

DEFINE STATGRAPH PIE;

BEGINGRAPH;

ENTRYTITLE "DISTRIBUTION OF CHECKING ACCOUNT";

LAYOUT REGION;

PIECHART CATEGORY= CHECKING\_ACCOUNT / DATALABELLOCATION=CALLOUT DATASKIN = SHEEN DATALABELCONTENT = ALL CATEGORYDIRECTION = CLOCKWISE START = **180** NAME = 'PIE' ;

DISCRETELEGEND 'PIE';

ENDLAYOUT;

ENDGRAPH;

END;

**RUN**;

**PROC** **SGRENDER** DATA = &DSN\_2. TEMPLATE = PIE;

**RUN**;

/\*Distribution of the Saving Account\*/

goptions cback=black;

**PROC** **TEMPLATE**;

DEFINE STATGRAPH PIE;

BEGINGRAPH;

ENTRYTITLE "DISTRIBUTION OF SAVING ACCOUNT";

LAYOUT REGION;

PIECHART CATEGORY= SAVING\_ACCOUNT / DATALABELLOCATION=CALLOUT DATASKIN = SHEEN DATALABELCONTENT = ALL CATEGORYDIRECTION = CLOCKWISE START = **180** NAME = 'PIE' ;

DISCRETELEGEND 'PIE';

ENDLAYOUT;

ENDGRAPH;

END;

**RUN**;

**PROC** **SGRENDER** DATA = &DSN\_2. TEMPLATE = PIE;

**RUN**;

/\*What type of Account are liked by the customers as per the income\*/

**PROC** **MEANS** DATA = RIMA.BANK\_FINAL MAXDEC=**1** N NMISS MIN MAX RANGE;

VAR INCOME;

**RUN**;

**PROC** **FORMAT**;

VALUE INC LOW- **50**= "LOW"

**51**- **100** = "AVERAGE"

**100**- HIGH = "EXCELLENT";

**RUN**;

/\*Is there any effect of Banking habits on customers becoming active or inactive?\*/

/\*Creating a table of non-missing values\*/

**PROC** **SQL**;

CREATE TABLE RIMA.HABIT AS

SELECT \*

FROM &DSN\_2.

WHERE PHONE\_N NE "MISSING" AND TELLER\_N NE "MISSING";

**QUIT**;

/\*CUSTOMER SATISFACTION COUNT USING PHONE BANKING\*/

%let graphs='C:\Users\Veena Nigam\Desktop\SAS Documents\SAS Business Project\IMAGES';

%let dpi=100;

%let w=8in;

%let h=4.5in;

title;

footnote;

ods html close;

/\*--Macro Program to find the effect of banking habits on the customers\*/

**%macro** RGBHex(rr,gg,bb);

%sysfunc(compress(CX%sysfunc(putn(&rr,hex2.))

%sysfunc(putn(&gg,hex2.))

%sysfunc(putn(&bb,hex2.))))

**%mend** RGBHex;

**proc** **template**;

define statgraph BarPie;

dynamic \_tsize \_lsize;

begingraph;

entrytitle "Customers Satisfaction Count using Phone Banking as per the Area" / textattrs=(size=\_tsize);

/\*--Define an attribute map to use specific colors by type--\*/

discreteattrmap name='PHONE\_N' / ignorecase=true;

value 'YES' / fillattrs=(color=%***rgbhex***(**100**, **150**, **40**));

value 'NO' / fillattrs=(color=%***rgbhex***(**180**, **110**, **50**));

enddiscreteattrmap;

/\*--Associate the Attribute Map to the TYPE variable--\*/

discreteattrvar attrvar=type var=type attrmap='PHONE\_N';

/\*--Define a one row x two columns layout --\*/

layout lattice / columns=**2** columnweights=(**0.6** **0.4**) columngutter=**20**;

/\*--First cell has a Bar Chart--\*/

layout overlay / xaxisopts=(display=(tickvalues)) walldisplay=none

yaxisopts=(display=(tickvalues) linearopts=(tickvalueformat=percent.)

griddisplay=on offsetmax=**0.2**);

barchart category=PHONE\_N / name='Y' barlabel=true stat=proportion

dataskin=pressed group=RES groupdisplay=cluster barlabelformat=percent6.1

baselineattrs=(thickness=**0**) barlabelattrs=(size=\_lsize) grouporder=descending

/\*--FILLTYPE= requires SAS 9.4 (TS1M2) or later--\*/

filltype=gradient

;

discretelegend 'Y' / location=inside halign=center valign=top autoitemsize=true valueattrs=(size=\_lsize);

endlayout;

/\*--Second cell has a Pie Chart--\*/

layout region / pad=(bottom=**30**);

piechart category=RES / dataskin=sheen centerfirstslice=true start=**270** stat=pct

datalabelattrs=(size=\_lsize);;

endlayout;

endlayout;

endgraph;

end;

**run**;

/\*--Render the graph--\*/

ods listing style=listing gpath=&graphs image\_dpi=&dpi;

ods graphics / reset width=&w height=&h imagename='BarPie';

**proc** **sgrender** data=RIMA.HABIT template=BarPie;

dynamic \_tsize=**16** \_lsize=**13**;

**run**;

/\*CUSTOMER SATISFACTION COUNT USING TELLER SERVICES\*/

%let graphs='C:\Users\Veena Nigam\Desktop\SAS Documents\SAS Business Project\IMAGES';

%let dpi=100;

%let w=8in;

%let h=4.5in;

title;

footnote;

ods html close;

/\*--Macro Program to find the effect of banking habits on the customers\*/

**%macro** RGBHex(rr,gg,bb);

%sysfunc(compress(CX%sysfunc(putn(&rr,hex2.))

%sysfunc(putn(&gg,hex2.))

%sysfunc(putn(&bb,hex2.))))

**%mend** RGBHex;

**proc** **template**;

define statgraph BarPie;

dynamic \_tsize \_lsize;

begingraph;

entrytitle "Customers Satisfaction Count visiting Teller as per the Area" / textattrs=(size=\_tsize);

/\*--Define an attribute map to use specific colors by type--\*/

discreteattrmap name='TELLER\_N' / ignorecase=true;

value 'YES' / fillattrs=(color=%***rgbhex***(**100**, **150**, **40**));

value 'NO' / fillattrs=(color=%***rgbhex***(**180**, **110**, **50**));

enddiscreteattrmap;

/\*--Associate the Attribute Map to the TYPE variable--\*/

discreteattrvar attrvar=type var=type attrmap='TELLER\_N';

/\*--Define a one row x two columns layout --\*/

layout lattice / columns=**2** columnweights=(**0.6** **0.4**) columngutter=**20**;

/\*--First cell has a Bar Chart--\*/

layout overlay / xaxisopts=(display=(tickvalues)) walldisplay=none

yaxisopts=(display=(tickvalues) linearopts=(tickvalueformat=percent.)

griddisplay=on offsetmax=**0.2**);

barchart category=TELLER\_N / name='Y' barlabel=true stat=proportion

dataskin=pressed group=RES groupdisplay=cluster barlabelformat=percent6.1

baselineattrs=(thickness=**0**) barlabelattrs=(size=\_lsize) grouporder=descending

/\*--FILLTYPE= requires SAS 9.4 (TS1M2) or later--\*/

filltype=gradient

;

discretelegend 'Y' / location=inside halign=center valign=top autoitemsize=true valueattrs=(size=\_lsize);

endlayout;

/\* --Second cell has a Pie Chart--\*/

layout region / pad=(bottom=**30**);

piechart category=RES / dataskin=sheen centerfirstslice=true start=**270** stat=pct

datalabelattrs=(size=\_lsize);;

endlayout;

endlayout;

endgraph;

end;

**run**;

/\*--Render the graph--\*/

ods listing style=listing gpath=&graphs image\_dpi=&dpi;

ods graphics / reset width=&w height=&h imagename='BarPie1';

**proc** **sgrender** data=RIMA.HABIT template=BarPie;

dynamic \_tsize=**16** \_lsize=**13**;

**run**;

/\*What age group of customers are the most active customers?\*/

**data** RIMA.AGE;

set &DSN\_2.;

if CHECKING\_ACCOUNT = "ACTIVE" then ACTIVE=**1**;

else INACTIVE=**0**;

/\*ACTIVE=round(-2500\*(1+ranuni(2))); INACTIVE=round(2400\*(1+ranuni(2)));\*/

CHECKING\_ACCOUNT=**0**;

LENGTH AGEGROUP $**12**;

IF AGE LE **20** THEN AGEGROUP="TEEN";

IF **21**<AGE<=**30** THEN AGEGROUP="YOUNG ADULT";

IF **31**<AGE<=**50** THEN AGEGROUP="ADULT";

IF AGE GE **50** THEN AGEGROUP="SENIOR";

**run**;

**data** RIMA.AGE;

length Age $**12**;

do Age='Teen', 'Young Adult', 'Adult', 'Senior';

ACTIVE=round(-**500**\*(**1**+ranuni(**2**))); INACTIVE=round(**400**\*(**1**+ranuni(**2**))); CHECKING\_ACCOUNT=**0**; output;

end;

**proc** **format**;

picture positive low-<**0**='0000' **0**<-high='0000';

title 'ACTIVE/INACTIVE CUSTOMERS AS PER THE AGE';

**proc** **sgplot** data= RIMA.AGE noautolegend;

/\* format A B positive.;\*/

hbarparm category=age response=ACTIVE / dataskin=sheen name='m'

fillattrs=graphdata1 datalabel=ACTIVE datalabelattrs=(size=**10**) transparency=**0.2**;

hbarparm category=age response=INACTIVE / dataskin=sheen name='f'

fillattrs=graphdata2 datalabel=INACTIVE datalabelattrs=(size=**10**) transparency=**0.2**;

scatter x=CHECKING\_ACCOUNT y=age / markerchar=age markercharattrs=(size=**11** weight=bold);

keylegend 'm' 'f';

xaxis values=(-**1000** to **1000** by **200**) display=(nolabel) grid offsetmin=**0.05** offsetmax=**0.05**;

yaxis display=(noticks novalues nolabel);

**run**;

/\*Which branches are offering the direct deposit facilities\*/

title"COMPLEMENTORY DIRECT DEPOSIT OFFERED BY VARIOUS BRANCHES";

**proc** **sgplot** data=&DSN\_2. noautolegend;

waterfall category=BRANCH response=DIRDEP/ colorgroup=checking\_account dataskin=sheen datalabel name='a';

keylegend 'a' / location=outside position=topright across=**1**;

xaxis display=(nolabel);

yaxis grid display=(nolabel) offsetmin=**0**;

**run**;

/\*Which branch is having the maximum number of Inactive customers?\*/

**proc** **template**;

define statgraph barchart;

begingraph / attrpriority=none;

entrytitle "ACTIVE/INACTIVE CUSTOMERS AS PER THE BRANCHES";

layout overlay;

barchart x=CHECKING\_ACCOUNT / name="BRANCHES"

stat=pct display=all

group=BRANCH groupdisplay=cluster group100=positive

barlabel=true;

discretelegend "BRANCHES";

endlayout;

endgraph;

end;

**run**;

**proc** **sgrender** data=&dsn\_2. template=barchart;

**run**;

/\*Which is the oldest Branch as per the age of the Account?\*/

**proc** **template**;

define statgraph barchart;

begingraph;

entrytitle "Oldest Branch as per the Account Age";

layout overlay;

barchart category=branch response=acctage/ name="bar"

stat=mean orient=horizontal

colorbyfreq=true colorstat=pct;

continuouslegend "bar" /

title="Percent of age of the Account";

endlayout;

endgraph;

end;

**run**;

**proc** **sgrender** data=rima.project4 template=barchart;

**run**;

/\*What type of Income people are the most active customers in the different branches of the bank ?\*/

**DATA** RIMA.INCOME;

SET &DSN\_2.;

IF INCOME LE **50** THEN INCOME\_1= **0**;

ELSE IF **51**< INCOME <=**100** THEN INCOME\_2 = **1**;

ELSE INCOME\_3 =**2**;

**RUN**;

**proc** **template**;

define statgraph barchart;

begingraph;

entrytitle "INCOME GROUP AS PER THE AREA";

layout overlay / cycleattrs=true

xaxisopts=(display=(tickvalues))

yaxisopts=(label="INCOME" offsetmax=**0.2**);

barchart category=BRANCH response=INCOME\_1 / stat=sum name="POOR INCOME"

legendlabel="POOR INCOME" datatransparency=**0.2**

discreteoffset=-**0.2** barwidth=**0.5** ;

barchart category=BRANCH response=INCOME\_2 / stat=sum name="AVERAGE INCOME"

legendlabel="AVERAGE INCOME" datatransparency=**0.2**

discreteoffset=**0** barwidth=**0.5** ;

barchart category=BRANCH response= INCOME\_3 / stat=sum name="HIGHER INCOME"

legendlabel="HIGHER INCOME" datatransparency=**0.2**

discreteoffset=+**0.2** barwidth=**0.5** ;

discretelegend "POOR INCOME" "AVERAGE INCOME" "HIGHER INCOME" / title="INCOME:"

location=inside halign=right valign=top;

endlayout;

endgraph;

end;

**run**;

**proc** **sgrender** data= RIMA.INCOME template=barchart;

**run**;

/\*\*/

**data** rima.atm;

set &dsn\_2.;

if atm = **1** then atm\_n = "YES";

else if atm =**0** then atm\_n = "NO";

**run**;

**proc** **freq** data=rima.atm;

tables atm\_n\*checking\_account / norow chisq plots=MOSAIC; /\* alias for MOSAICPLOT \*/

**run**;

/\*Does people also prefer to keep Money Market accounts instead of other type of accounts?\*/

**proc** **sgpanel** data=rima.project4;

panelby dda;

vbar sav / response=MM group= Res groupdisplay=cluster stat=mean;

title "Distribution of Accounts as per the Area";

**run**;

title;

/\*Do the customers with mortgage are the most loyal customers?\*/

**proc** **sgpanel** data=rima.project4;

panelby dda;

vbar MTG / response= INS group= DDA groupdisplay=cluster stat=mean;

title "Customers keeping the loan account but not the checking account";

**run**;

title;

ODS PDF CLOSE;

/\*Hypothesis Testing\*/

/\* Perform the t-test \*/

title 'Two Sample T-Test';

**proc** **ttest** data= &dsn\_2.;

class checking\_account; /\* defines the grouping variable \*/

var age income; /\* variable whose means will be compared \*/

**run**;

/\*FINAL VARIABLE SELECTION\*/

/\*SAS MACRO : VARIABLE SELECTION BASED ON WALD CHI-SQUARE\*/

/\*Variable Selection based on Univariate Analysis (Wald Chi-Square and Standardized Coefficient)\*/

/\*PROC LOGISTIC is run on each of the variables and tracking p-value of wald chi-square and standardized coefficient\*/

**%macro** perf(data=,targetvar=,vars=,output=);

%let n=%sysfunc(countw(&vars));

%do i=**1** %to &n;

%let val = %scan(&vars,&i);

ods select none;

ods output ParameterEstimates=Estimate&i;

proc logistic data=&data;

model &targetvar(event='1')=&val / stb;

run;

data Estimate&i;

set Estimate&i;

length Sig $**15**;

where Variable NE 'Intercept';

if ProbChiSq < **.05** then Sig ='Significant';

else if ProbChiSq >= **.05** then Sig = 'Non-Significant';

help = abs(StandardizedEst);

run;

%end;

data &output;

set Estimate1 - Estimate&n;

run;

proc datasets library=work

nodetails nolist;

delete Estimate1 - Estimate&n;

run;

quit;

proc sort data = &output;

by descending help;

run;

data &output;

set &output(DROP=HELP);

run;

**%mend** perf;

options symbolgen mlogic;

%***perf***(data=rima.project4 ,targetvar= dda ,vars= ATM INS MTG CC CCPURC TELLER PHONE DIRDEP INV INCOME MM SAV NSF ,output= rima.result1);

/\*MODEL BUILDING\*/

/\*LOGISTIC REGRESSION MODEL\*/

/\*Y - DDA

X- CRSCORE, DIRDEP, NSF, RES, ATM ,CC, MTG\*/

/\*SPLITING THE DATA INTO TRAINING AND TESTING\*/

**PROC** **SURVEYSELECT** DATA= RIMA.BANK\_DATA OUT= RIMA.MODEL\_DATA RATE=**0.7** OUTALL;

**RUN**;

**PROC** **FREQ** DATA = RIMA.MODEL\_DATA;

TABLE SELECTED/MISSING;

**RUN**;

**DATA** TRAINING TESTING;

SET RIMA.MODEL\_DATA;

IF SELECTED EQ **1** THEN OUTPUT TRAINING;

ELSE OUTPUT TESTING;

**RUN**;

/\*MODEL 1\*/

**PROC** **LOGISTIC** DATA= TRAINING plots(only MAXPOINTS=NONE)=(roc ODDSRATIO);

class CC(param=ref ref="1") CCPURC(param=ref ref="1") mtg(param=ref ref="1") ins(param=ref ref="1") atm(param=ref ref="1");

model checking\_account(event="active") = cc|ccpurc|mtg|ins|atm/CLODDS=PL;

oddsratio CC;

oddsratio CCPURC;

oddsratio mtg;

oddsratio ins;

oddsratio atm;

OUTPUT OUT = BANK\_PROB P = PRED\_PROB;

**RUN**;

**QUIT**;

/\*INTERACTIONS\*/

TITLE "RUNNING A LOGISTIC REGRESSION MODEL WITH INTERACTIONS";

**PROC** **LOGISTIC** DATA = TRAINING plots(only MAXPOINTS=NONE)=(roc ODDSRATIO) ;

class CC(param=ref ref="1") CCPURC(param=ref ref="1") mtg(param=ref ref="1") ins(param=ref ref="1") atm(param=ref ref="1");

model checking\_account(event="active") = cc|ccpurc|mtg|ins|atm @**2**/CLODDS=PL SELECTION= BACKWARD SLSTAY=**0.1**;;

oddsratio CC;

oddsratio CCPURC;

oddsratio mtg;

oddsratio ins;

oddsratio atm;

OUTPUT OUT = BANK\_PROB P = PRED\_PROB;

**RUN**;

**QUIT**;

/\* Creating Logistic regression model using stepwise selection method \*/

Title"Logistic Regression Model using Stepwise Selection Method";

**proc** **logistic** data=training plots(only MAXPOINTS=NONE)=(roc ODDSRATIO);

class CC CCPURC atm mtg ins;

model checking\_account(event="active") = cc|CCPURC|atm|mtg|ins/selection=stepwise expb stb lackfit;

oddsratio CC ;

oddsratio CCPURC;

oddsratio atm;

oddsratio mtg;

oddsratio ins;

output out = temp p=new;

store bank\_logistic;

**run**;

\*MODEL SCORING: TRAINING AND TESTING;

\*1. ;

TITLE "MODEL SCORING : SCORING OPTION";

**PROC** **LOGISTIC** DATA = TRAINING plots(only MAXPOINTS=NONE) ;

class CC atm ins;

model checking\_account(event="active") = cc|atm|ins @**2**/CLODDS=PL SELECTION= BACKWARD SLSTAY=**0.1** ;

OUTPUT OUT= TEST P=PROB;

SCORE DATA = TESTING OUT=TESTING\_SCORE\_FINAL;

**RUN**;

**QUIT**;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* END \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*