REMOVING OUTLIERS

libname adijit'/sasdata3/SAS\_MINING/egusers/adijit';

/\*CHECKING FOR OUTLIERS 11867912 \*/

/\*ALL DATA POINTS WITH VALUES <P1 AND DATA POINTS WITH VALUES>P99 ARE REMOVED\*/

%macro bin(var,i);

proc univariate data=adijit.FINAL\_EKYC\_DATA\_UPW\_A;

var &var.;

output out=&var. pctlpts= 1 to 100 by 1 pctlpre=P\_ ;

run;

data \_null\_;

set &var.;

call symput('P\_1',compress(round((P\_1),0.001)));

call symput('P\_99',compress(round((P\_99),0.001)));

run;

data adijit.FINAL\_EKYC\_DATA\_UPW\_A;

set adijit.FINAL\_EKYC\_DATA\_UPW\_A;

if &var. <&P\_1. then flag\_&i.=1;

else if &var. > &P\_99. then flag\_&i.=1;

run;

%mend;

%bin(SUM\_TOT\_RCHG\_CNT,1);

%bin(SUM\_TOT\_RCHG\_MRP,2);

%bin(SUM\_TOTAL\_MOU,3);

%bin(SUM\_USAGE\_DAYS,4);

%bin(SUM\_TOT\_DATA\_USG,5);

%bin(unl\_tag,6);

%bin(HANDSET\_3G\_4G,7);

%bin(AON,8);

%bin(age,9);

%bin(count\_fam\_mem,10);

%bin(cat\_digital\_sms\_cnt,11);

%bin(tot\_dig\_sms\_cnt,12);

%bin(GENDER\_TAG,13);

data base;

set adijit.FINAL\_EKYC\_DATA\_UPW\_A;

outliers\_sum=sum(of flag\_1:flag\_13);

run;

/\*CLEAN DATA AFTER REMOVING OUTLIERS 11595515\*/

data adijit.FINAL\_EKYC\_DATA\_UPW2\_9;

set base;

where outliers\_sum eq .;

run;

/\*checkinggggggggggggggggggggggggggggggggggggggg\*/

data check1;

set base;

where outliers\_sum in (1,2,3);

run;

data check2;

set check1;

where age < 0; run;

data check;

set adijit.FINAL\_EKYC\_DATA\_UPW2\_5;

where input(DOB, date9.) > "1JAN2017"d ;run;

/\*proc freq data=base;tables outliers\_sum;\*/

/\*run;\*/

STANDARDIZE AND VARCLUS

/\*STANDARDIZING THE DATA POINTS\*/

data adijit.FINAL\_EKYC\_DATA\_UPW2\_9;

set adijit.FINAL\_EKYC\_DATA\_UPW2\_9;

age\_Y= int(age/365)\*12 ;

run;

proc stdize data=adijit.FINAL\_EKYC\_DATA\_UPW2\_9 out=adijit.stdardised\_data\_RANGE METHOD=RANGE ;

run;

/\*STANDARDIZE ONLY CONTINUOUS VARS\*/

DATA adijit.stdardised\_data1\_RANGE;

SET adijit.stdardised\_data\_RANGE

(KEEP=subs\_msisdn

SUM\_TOT\_RCHG\_CNT

SUM\_TOT\_RCHG\_MRP

SUM\_TOTAL\_MOU

SUM\_USAGE\_DAYS

SUM\_TOT\_DATA\_USG

AON

age\_y

count\_fam\_mem

cat\_digital\_sms\_cnt

tot\_dig\_sms\_cnt

)

; **RUN**;

**PROC** **SQL**;

CREATE TABLE adijit.FINAL\_EKYC\_DATA\_UPW2\_10

AS SELECT a.unl\_tag,a.HANDSET\_3G\_4G,a.gender\_tag, b.\*

from adijit.FINAL\_EKYC\_DATA\_UPW2\_9 as a

inner join adijit.stdardised\_data1\_RANGE as b

on a.subs\_msisdn=b.subs\_msisdn;

**quit**;

/\*proc freq data= adijit.FINAL\_EKYC\_DATA\_UPW2\_10 ;\*/

/\*TABLES HANDSET\_3G\_4G \*/

/\*;\*/

/\*RUN;\*/

**PROC** **MEANS** DATA=adijit.FINAL\_EKYC\_DATA\_UPW2\_10 N MEAN MIN MAX; **RUN**;

**PROC** **CORR** DATA=adijit.FINAL\_EKYC\_DATA\_UPW2\_10 OUT=CORR1;

VAR

SUM\_TOT\_RCHG\_CNT

SUM\_TOT\_RCHG\_MRP

SUM\_TOTAL\_MOU

SUM\_USAGE\_DAYS

SUM\_TOT\_DATA\_USG

unl\_tag

HANDSET\_3G\_4G

AON

age\_y

gender\_tag

count\_fam\_mem

cat\_digital\_sms\_cnt

tot\_dig\_sms\_cnt

;

**RUN**;

**proc** **varclus** data=adijit.FINAL\_EKYC\_DATA\_UPW2\_10 outtree=tree

centroid maxclusters=**15**; var

SUM\_TOT\_RCHG\_CNT

SUM\_TOT\_RCHG\_MRP

SUM\_TOTAL\_MOU

SUM\_USAGE\_DAYS

SUM\_TOT\_DATA\_USG

unl\_tag

HANDSET\_3G\_4G

AON

age\_y

gender\_tag

count\_fam\_mem

cat\_digital\_sms\_cnt

tot\_dig\_sms\_cnt

; **run**;

**proc** **stdize** data=adijit.FINAL\_EKYC\_DATA\_UPW2\_9 out=adijit.stdardised\_datarange method= range;

**run**;

/\*CREATING CLUSTERS USING THE LIST OF VARIABLES \*/

/\*MAXCLUSTERS=10......THE NUMBER OF CLUSTERS CAN BE SPECIFIED\*/

/\*MAXCLUSTERS SHOULD BE CHANGED TILL MEANINGFUL CLUSTERS ARE ARRIVED AT\*/

**data** adijit.temp\_stat;

input;

datalines;

**run**;

**data** adijit.final\_results;

input;

datalines;

**run**;

**%macro** ref\_macro(clus);

PROC FASTCLUS DATA=adijit.FINAL\_EKYC\_DATA\_UPW2\_10

MAXCLUSTERS =&clus.

OUT=adijit.cluster

OUTSEED=adijit.CENTRES

OUTSTAT=adijit.CLUSTER\_STAT\_&clus.

REPLACE=RANDOM RANDOM =**187413849**

CONVERGE =**0.001** MAXITER =**100**;

VAR

SUM\_TOT\_RCHG\_CNT

SUM\_TOT\_RCHG\_MRP

SUM\_TOTAL\_MOU

SUM\_USAGE\_DAYS

SUM\_TOT\_DATA\_USG

unl\_tag

HANDSET\_3G\_4G

AON

age\_y

gender\_tag

count\_fam\_mem

cat\_digital\_sms\_cnt

tot\_dig\_sms\_cnt

; RUN;

data adijit.CLUSTER\_STAT\_&clus.;

set adijit.CLUSTER\_STAT\_&clus.;

iteration\_final=&clus.;

run;

data adijit.temp\_stat;

set adijit.temp\_stat adijit.CLUSTER\_STAT\_&clus. ;

run;

proc sql;

create table cluster\_var

as select a.\*,b.cluster

from adijit.FINAL\_EKYC\_DATA\_UPW2\_9 a left join adijit.cluster b

on a.subs\_msisdn=b.subs\_msisdn;

quit;

proc sql;

create table final

as select cluster,

count(cluster) as cases,

mean(SUM\_TOT\_RCHG\_CNT) as A\_SUM\_TOT\_RCHG\_CNT,

mean(SUM\_TOT\_RCHG\_MRP) as A\_SUM\_TOT\_RCHG\_MRP,

mean(SUM\_TOTAL\_MOU) as A\_SUM\_TOTAL\_MOU,

mean(SUM\_USage\_DAYS) as A\_SUM\_USage\_DAYS,

mean(SUM\_TOT\_DATA\_USG) as A\_SUM\_TOT\_DATA\_USG,

mean(AON) as A\_AON,

mean(age\_Y) as A\_age\_Y,

mean(unl\_tag) as A\_unl\_tag,

mean(count\_fam\_mem) as A\_count\_fam\_mem,

mean(cat\_digital\_sms\_cnt) as A\_cat\_digital\_sms\_cnt,

mean(HANDSET\_3G\_4G) as A\_HANDSET\_3G\_4G,

mean(gender\_tag) as A\_gender\_tag,

mean(tot\_dig\_sms\_cnt) as A\_tot\_dig\_sms\_cnt

from cluster\_var

group by cluster;

quit;

data final;

set final;

iteration\_final=&clus.;

run;

data adijit.final\_results;

set adijit.final\_results final;

run;

**%mend**;

%***ref\_macro***(**3**);

%***ref\_macro***(**4**);

%***ref\_macro***(**5**);

%***ref\_macro***(**6**);

%***ref\_macro***(**7**);

%***ref\_macro***(**8**);

%***ref\_macro***(**9**);

%***ref\_macro***(**10**);

%***ref\_macro***(**11**);

%***ref\_macro***(**12**);

%***ref\_macro***(**13**);

%***ref\_macro***(**14**);

%***ref\_macro***(**15**);

**proc** **sql**;

create table final\_OVERALL\_SUMMARY

as select

count(cluster) as cases,

mean(SUM\_TOT\_RCHG\_CNT) as A\_SUM\_TOT\_RCHG\_CNT,

mean(SUM\_TOT\_RCHG\_MRP) as A\_SUM\_TOT\_RCHG\_MRP,

mean(SUM\_TOTAL\_MOU) as A\_SUM\_TOTAL\_MOU,

mean(SUM\_USage\_DAYS) as A\_SUM\_USage\_DAYS,

mean(SUM\_TOT\_DATA\_USG) as A\_SUM\_TOT\_DATA\_USG,

mean(AON) as A\_AON,

mean(age\_Y) as A\_age\_Y,

mean(count\_fam\_mem) as A\_count\_fam\_mem,

mean(unl\_tag) as A\_unl\_tag,

mean(gender\_tag) as A\_gender\_tag,

mean(cat\_digital\_sms\_cnt) as A\_cat\_digital\_sms\_cnt,

mean(HANDSET\_3G\_4G) as A\_HANDSET\_3G\_4G,

mean(tot\_dig\_sms\_cnt) as A\_tot\_dig\_sms\_cnt

from cluster\_var;

**quit**;

**proc** **contents** data= cluster\_var; **run**;

ADDITIONAL VARS

**proc** **delete** data=db2proj.MNP;**run**;

**PROC** **SQL**;

CREATE TABLE DB2PROJ.MNP (BULKLOAD=YES BL\_METHOD=CLILOAD)AS

SELECT DISTINCT t1.subs\_msisdn

FROM adijit.AAMART\_EKYC\_DATA t1;

**QUIT**;

**PROC** **SQL**;

connect to db2 (user=&user password=&pwd database=biprod);

CREATE TABLE adijit.ECOM1 AS

select \* from connection to db2

(

select distinct(subs\_msisdn) as msisdn

from

/\*select \* from\*/

AGGR\_V.AGGR\_DIGITAL\_SHORT\_CODE\_SUBS\_SUMM T4

where circle\_id=**22** and event\_date<='2017-12-31'

);

disconnect from db2;

**quit**;

/\* CHECK FOR DISCNCT\_DT\*/

**DATA** CHECK;

SET adijit.AAMART\_EKYC\_DATA ;

WHERE DISCNCTN\_DT<=31DEC2017;

**RUN**;

/\*6406496\*/

**PROC** **SQL**;

CREATE TABLE TEMP AS SELECT \* FROM adijit.AAMART\_EKYC\_DATA ;

**QUIT**;

/\*drawing sample of 7Lakh\*/

/\*700000\*/

**PROC** **SURVEYSELECT** DATA=TEMP METHOD=srs rep=**1**

SAMPSIZE=**2000000** seed = **12345** out = adijit.AAMART\_EKYC\_DATA1 ;

**RUN**;

**proc** **freq** data=adijit.FINAL\_EKYC\_DATA\_UPW1; tables (digital\_tag ecom\_tag family\_tag main\_bal\_inr\_last

prepost\_ind sum\_cust\_rev sum\_data\_usg\_days sum\_roaming\_days sum\_tot\_data\_usg sum\_tot\_rchg\_cnt

sum\_tot\_rchg\_mrp sum\_tot\_std\_mou sum\_total\_mou sum\_usage\_days sum\_vlr\_days unl\_tag); **run**;

/\*ECOMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMMM\*/

**proc** **delete** data=db2proj.MNP1;**run**;

**PROC** **SQL**;

CREATE TABLE DB2PROJ.MNP11 (BULKLOAD=YES BL\_METHOD=CLILOAD )

AS SELECT DISTINCT t1.subs\_msisdn

FROM adijit.FINAL\_EKYC\_DATA\_UPW2\_6 t1;

**QUIT**;

**PROC** **SQL**;

connect to db2 (user=&user password=&pwd database=biprod);

CREATE TABLE adijit.ECOM1 AS

select \* from connection to db2

(

select R.subs\_msisdn,event\_date,

SUM(COALESCE(BANKING,**0**)) As BANKING,

SUM(COALESCE(COMPETITIVE\_EXAM,**0**)) As COMPETITIVE\_EXAM,

SUM(COALESCE(CROSS\_CAT,**0** )) As CROSS\_CAT,

SUM(COALESCE(ENTERTAINMENT,**0** )) As ENTERTAINMENT,

SUM(COALESCE(FASHION,**0** )) As FASHION,

SUM(COALESCE(FINANCE,**0** )) As FINANCE,

SUM(COALESCE(FOOD,**0** )) As FOOD,

SUM(COALESCE(HEALTH,**0** )) As HEALTH,

SUM(COALESCE(JOB\_PORTALS,**0** )) As JOB\_PORTALS,

SUM(COALESCE(SATELLITE\_TV,**0** )) As SATELLITE\_TV,

SUM(COALESCE(SOCIAL\_MEDIA,**0** )) As SOCIAL\_MEDIA,

SUM(COALESCE(SPORTS,**0** )) As SPORTS,

SUM(COALESCE(TRAVEL,**0** )) As TRAVEL,

SUM(COALESCE(UTILITIES,**0**)) As UTILITIES

from db2proj.MNP11 as R

INNER JOIN

/\*select \* from\*/

AGGR\_V.AGGR\_DIGITAL\_SHORT\_CODE\_SUBS\_SUMM T4

/\*fetch first 10 rows only\*/

ON (T4.subs\_msisdn = R.subs\_MSISDN )

where circle\_id=**22** and event\_date between '2017-10-03' and '2017-12-31'

group by R.subs\_msisdn,event\_date

);

disconnect from db2;

**quit**;

**DATA** adijit.ECOM11;

SET adijit.ECOM1;

if sum(BANKING,COMPETITIVE\_EXAM,CROSS\_CAT,ENTERTAINMENT,FASHION,FINANCE,FOOD,HEALTH,JOB\_PORTALS

,SATELLITE\_TV,SOCIAL\_MEDIA,SPORTS,TRAVEL,UTILITIES,**0**)> **0** then ECOM\_TAG=**1**; else ECOM\_TAG=**0**;

**RUN**;

**PROC** **SQL**;

CREATE TABLE adijit.ECOM112

AS SELECT subs\_msisdn,max(ecom\_tag) AS ECOM\_TAG1

from adijit.ECOM11

group by subs\_msisdn;

**quit**;

**proc** **freq** data=adijit.ECOM112 ; tables ecom\_tag1; **run**;