/\* Generated Code (IMPORT) \*/

/\* Source File: telecomsample.csv \*/

/\* Source Path: /home/kvandanamba0/Projrct/Final \*/

/\* Code generated on: 6/24/18, 5:14 PM \*/

%web\_drop\_table(temp);

FILENAME REFFILE '/home/kvandanamba0/Projrct/Final/telecomsample.csv';

ROC IMPORT DATAFILE=REFFILE

DBMS=CSV replace

OUT=temp;

GETNAMES=YES;

RUN;

%web\_open\_table(temp);

proc contents data = temp varnum;

run;

/\*Checking Missing values of all Numeric variables\*/

proc means n nmiss mean data = temp;

run;

/\*Checking Missing values of all String variables\*/

proc freq data = temp;

tables income crclscod asl\_flag prizm\_social\_one area refurb\_new hnd\_webcap marital ethnic dwlltype dwllsize mailordr occu1 numbcars retdays proptype mailresp cartype car\_buy children csa div\_type;

run;

/\*We have chosen 'churn' as our target variable; checking 0's and 1's freq\*/

proc freq data = temp;

tables churn;

run;

/\*Treating missing values: Deleting values = 'NA' from variables

whose missing value % is <2%\*/

data temp1;

set temp;

if mou\_mean = . or totmrc\_Mean = . or rev\_Range = . or mou\_Range = . or

change\_mou = . or ovrrev\_Mean = . or rev\_Mean=. or ovrmou\_Mean=. or age1=. Or age2=. or

hnd\_price=. or forgntvl=. or mtrcycle=. or truck=. or roam\_Mean=. or

da\_Mean=. or da\_Range=. or datovr\_Mean=. or datovr\_Range=. then delete ;

run;

/\*Imputing mean to income,retdays,avg6mou,avg6qty\*/

data temp2;

set temp1;

income1 = input(income,10.);

retdays1 = input(retdays,5.);

run;

data temp3(drop = income retdays);

set temp2;

run;

data temp4;

set temp3;

rename income1=income retdays1=retdays;

run;

data temp5;

set temp4;

test=1;

run;

proc means mean data = temp5;

var avg6mou avg6qty income retdays;

output out = temp5\_mean mean(avg6mou) = mean\_avg6mou mean(avg6qty)=mean\_avg6qty

mean(income) = mean\_income mean(retdays) = mean\_retdays;

run;

data temp5\_mean(drop = \_TYPE\_ \_FREQ\_);

set temp5\_mean;

test =1;

run;

data temp\_merge;

merge temp5 temp5\_mean ;

by test;

run;

data temp\_merge1;

set temp\_merge;

if avg6mou=. then avg6mou\_1 = round(mean\_avg6mou,0.01);

else avg6mou\_1 = avg6mou;

if avg6qty =. then avg6qty\_1 = round(mean\_avg6qty,0.01);

else avg6qty\_1 = avg6qty;

if income = . then income\_1 = round(mean\_income,0.01);

else income\_1 = income;

if retdays = . then retdays\_1 = round(mean\_retdays,0.01);

else retdays\_1 = retdays;

run;

data temp\_merge2(drop = avg6qty avg6mou test mean\_avg6qty mean\_avg6mou income retdays mean\_income mean\_retdays);

set temp\_merge1;

run;

data temp\_merge3;

set temp\_merge2;

rename income\_1 = income retdays\_1 = retdays avg6mou\_1 =avg6mou avg6qty\_1=avg6qty;

run;

proc means n nmiss data = temp\_merge3;

run;

proc freq data =temp\_merge3;

tables income crclscod asl\_flag prizm\_social\_one area refurb\_new hnd\_webcap marital ethnic dwlltype dwllsize mailordr occu1 numbcars retdays proptype mailresp cartype car\_buy children csa div\_type;

run;

/\*Taking backup\*/

data temp\_backup1;

set temp\_merge3;

run;

/\*Data Preparation : Dummy Variables\*/

\*1.crclscod;

data temp\_dummy;

set temp\_backup1;

if substr(crclscod,1,1) = "A" or substr(crclscod,1,1) = "B" or substr(crclscod,1,1) = "C" or substr(crclscod,1,1) = "D" or substr(crclscod,1,1) = "E"

then best\_crclscod = 1; else best\_crclscod = 0;

if substr(crclscod,1,1) = "F" or substr(crclscod,1,1) = "G" or substr(crclscod,1,1) = "H" or substr(crclscod,1,1) = "I" or substr(crclscod,1,1) = "J"

then good\_crclscod = 1; else good\_crclscod = 0;

if substr(crclscod,1,1) = "K" or substr(crclscod,1,1) = "L" or substr(crclscod,1,1) = "M" or substr(crclscod,1,1) = "N" or substr(crclscod,1,1) = "O"

then average\_crclscod = 1; else average\_crclscod = 0;

if substr(crclscod,1,1) = "P" or substr(crclscod,1,1) = "Q" or substr(crclscod,1,1) = "R" or substr(crclscod,1,1) = "S" or substr(crclscod,1,1) = "T"

then bad\_crclscod = 1; else bad\_crclscod = 0;

if substr(crclscod,1,1) = "U" or substr(crclscod,1,1) = "V" or substr(crclscod,1,1) = "W" or substr(crclscod,1,1) = "X" or substr(crclscod,1,1) = "Y"

or substr(crclscod,1,1) = "Z" then worst\_crclscod = 1; else worst\_crclscod = 0;

\*2.asl\_flag;

if asl\_flag = 'Y' then asl\_flag\_yes = 1;

else asl\_flag\_yes = 0;

if asl\_flag = 'N' then asl\_flag\_no = 1;

else asl\_flag\_no = 1;

\*3.prizm\_social\_one;

if prizm\_social\_one = 'C' then area\_city = 1;

else area\_city = 0;

if prizm\_social\_one = 'R' then area\_rural = 1;

else area\_rural = 0;

if prizm\_social\_one = 'S' then area\_suburban = 1;

else area\_suburban = 0;

if prizm\_social\_one = 'T' then area\_town = 1;

else area\_town = 0;

if prizm\_social\_one = 'U' then area\_urban = 1;

else area\_urban = 0;

if prizm\_social\_one = 'NA' then area\_unknown = 1;

else area\_unknown = 0;

\*4.refurb\_new;

if refurb\_new = 'N' then handset\_new = 1;

else handset\_new =0;

if refurb\_new = 'R' then handset\_refurb = 1;

else handset\_refurb = 0;

\*5.hnd\_webcap;

if hnd\_webcap = 'WC' then handset\_wc = 1;

else handset\_wc = 0;

if hnd\_webcap = 'WCMB' then handset\_wcmb = 1;

else handset\_wcmb = 0;

if hnd\_webcap = 'NA' then handset\_na = 1;

else handset\_na = 0;

if hnd\_webcap = 'UNKW' then handset\_unkw = 1;

else handset\_unkw = 0;

\*6.marital;

if marital = 'M' then status\_married = 1;

else status\_married = 0;

if marital = 'A' then status\_infermarried =1;

else status\_infermarried =0;

if marital = 'B' then status\_infersingle =1;

else status\_infersingle =0;

if marital = 'S' or marital = 'B' then status\_single = 1;

else status\_single = 0;

if marital = 'U' then status\_unknown = 1;

else status\_unknown = 0;

\*7.dwlltype;

if index(dwlltype,'M')>0 then dwell\_multiple = 1;

else dwell\_multiple = 1;

if index(dwlltype,'S')>0 then dwell\_single = 1;

else dwell\_single = 1;

if index(dwlltype,'NA')>0 then dwell\_unknown = 1;

else dwell\_unknown = 1;

\*8.dwllsize;

if index(dwllsize,'A') > 0 then dsize\_1 = 1;

else dsize\_1 = 0;

if index(dwllsize,'B') > 0 then dsize\_2 = 1;

else dsize\_2 = 0;

if index(dwllsize,'C') > 0 then dsize\_3 = 1;

else dsize\_3 = 0;

if index(dwllsize,'D') > 0 then dsize\_4 = 1;

else dsize\_4 = 0;

if index(dwllsize,'E') > 0 then dsize\_5 = 1;

else dsize\_5 = 0;

if index(dwllsize,'F') > 0 then dsize\_6 = 1;

else dsize\_6 = 0;

if index(dwllsize,'G') > 0 then dsize\_7 = 1;

else dsize\_7 = 0;

if index(dwllsize,'H') > 0 then dsize\_8 = 1;

else dsize\_8 = 0;

if index(dwllsize,'I') > 0 then dsize\_9 = 1;

else dsize\_9 = 0;

if index(dwllsize,'J') > 0 then dsize\_10to19 = 1;

else dsize\_10to19 = 0;

if index(dwllsize,'K') > 0 then dsize\_20to29 = 1;

else dsize\_20to29 = 0;

if index(dwllsize,'L') > 0 then dsize\_30to39 = 1;

else dsize\_30to39 = 0;

if index(dwllsize,'M') > 0 then dsize\_40to49 = 1;

else dsize\_40to49 = 0;

if index(dwllsize,'N') > 0 then dsize\_50to99 = 1;

else dsize\_50to99 = 0;

if index(dwllsize,'O') > 0 then dsize\_100 = 1;

else dsize\_100 = 0;

if index(dwllsize,'NA') > 0 then dsize\_unknown = 1;

else dsize\_unknown = 0;

\*9.mailordr;

if index(mailordr,'B')>0 then mailorder\_buyer = 1;

else mailorder\_buyer = 0;

if index(mailordr,'NA')>0 then mailorder\_unknown = 1;

else mailorder\_unknown = 0;

\*10.occu1;

if index(occu1,'1')>0 then occu\_technical = 1;

else occu\_technical = 0;

if index(occu1,'2')>0 then occu\_admin = 1;

else occu\_admin = 0;

if index(occu1,'3')>0 then occu\_sales = 1;

else occu\_sales = 0;

if index(occu1,'4')>0 then occu\_wc = 1;

else occu\_wc = 0;

if index(occu1,'5')>0 then occu\_bc = 1;

else occu\_bc = 0;

if index(occu1,'6')>0 then occu\_student = 1;

else occu\_student = 0;

if index(occu1,'7')>0 then occu\_homemaker = 1;

else occu\_homemaker = 0;

if index(occu1,'8')>0 then occu\_retires = 1;

else occu\_retires = 0;

if index(occu1,'9')>0 then occu\_farmer = 1;

else occu\_farmer = 0;

if index(occu1,'A')>0 then occu\_military = 1;

else occu\_military = 0;

if index(occu1,'B')>0 then occu\_religious = 1;

else occu\_religious = 0;

if index(occu1,'C')>0 or index(occu1,'D')>0 or index(occu1,'E')>0 or

index(occu1,'F')>0 or index(occu1,'G')>0 or index(occu1,'H')>0 or

index(occu1,'I')>0 or index(occu1,'J')>0 or index(occu1,'K')>0

then occu\_selfemp = 1;else occu\_selfemp = 0;

\*11.numbcars;

if index(numbcars,'1')>0 then numcars\_1 = 1;

else numcars\_1 = 0;

if index(numbcars,'2')>0 then numcars\_2 = 1;

else numcars\_2 = 0;

if index(numbcars,'3')>0 then numcars\_3 = 1;

else numcars\_3 = 0;

if index(numbcars,'NA')>0 then numcars\_unknown = 1;

else numcars\_unknown = 0;

\*12.proptype;

if index(proptype,'A')>0 then proptype\_a = 1;

else proptype\_a = 0;

if index(proptype,'B')>0 then proptype\_b = 1;

else proptype\_b = 0;

if index(proptype,'D')>0 then proptype\_d = 1;

else proptype\_d = 0;

if index(proptype,'E')>0 then proptype\_e = 1;

else proptype\_e = 0;

if index(proptype,'G')>0 then proptype\_g = 1;

else proptype\_g = 0;

if index(proptype,'M')>0 then proptype\_m = 1;

else proptype\_m = 0;

if index(proptype,'NA')>0 then proptype\_unknown = 1;

else proptype\_unknown = 0;

\*13.mailresp;

if index(mailresp,'R')>0 then mailresp\_yes=1;

else mailresp\_yes = 0;

if index(mailresp,'NA')>0 then mailresp\_unknown=1;

else mailresp\_unknown = 0;

\*14.cartype;

if index(cartype,'A')>0 then cartype\_luxury = 1;

else cartype\_luxury = 0;

if index(cartype,'B')>0 then cartype\_truck = 1;

else cartype\_truck = 0;

if index(cartype,'C')>0 then cartype\_SUV = 1;

else cartype\_SUV = 0;

if index(cartype,'D')>0 then cartype\_minivan = 1;

else cartype\_minivan = 0;

if index(cartype,'E')>0 then cartype\_regular = 1;

else cartype\_regular = 0;

if index(cartype,'F')>0 then cartype\_upper = 1;

else cartype\_upper = 0;

if index(cartype,'G')>0 then cartype\_basic = 1;

else cartype\_basic = 0;

if index(cartype,'NA')>0 then cartype\_unknown = 1;

else cartype\_unknown = 0;

\*15.div\_type;

if index(div\_type,'BTH')>0 then division\_bth = 1;

else division\_bth = 0;

if index(div\_type,'LDD')>0 then division\_LDD = 1;

else division\_LDD = 0;

if index(div\_type,'LTD')>0 then division\_LTD = 1;

else division\_LTD = 0;

if index(div\_type,'NA')>0 then division\_unknown = 1;

else division\_unknown = 0;

\*16.ethnic;

if index(ethnic,'B')>0 then ethnic\_asian\_nor = 1;

else ethnic\_asian\_nor = 0;

if index(ethnic,'D')>0 then ethnic\_south\_euro = 1;

else ethnic\_south\_euro = 0;

if index(ethnic,'F')>0 then ethnic\_french = 1;

else ethnic\_french = 0;

if index(ethnic,'G')>0 then ethnic\_german = 1;

else ethnic\_german = 0;

if index(ethnic,'H')>0 then ethnic\_hispanic = 1;

else ethnic\_hispanic = 0;

if index(ethnic,'I')>0 then ethnic\_italian = 1;

else ethnic\_italian = 0;

if index(ethnic,'J')>0 then ethnic\_jewish = 1;

else ethnic\_jewish = 0;

if index(ethnic,'M')>0 then ethnic\_misc = 1;

else ethnic\_misc = 0;

if index(ethnic,'N')>0 then ethnic\_north\_euro = 1;

else ethnic\_north\_euro = 0;

if index(ethnic,'O')>0 then ethnic\_asian = 1;

else ethnic\_asian = 0;

if index(ethnic,'P')>0 then ethnic\_polynesia = 1;

else ethnic\_polynesia = 0;

if index(ethnic,'R')>0 then ethnic\_arab = 1;

else ethnic\_arab = 0;

if index(ethnic,'S')>0 then ethnic\_scot\_iris = 1;

else ethnic\_scot\_iris = 0;

if index(ethnic,'U')>0 or index(ethnic,'NA')then ethnic\_unknown = 1;

else ethnic\_unknown = 0;

if index(ethnic,'Z')>0 then ethnic\_afro\_amer = 1;

else ethnic\_afro\_amer = 0;

/\*17.children\*/

if index(children,'Y')>0 then children\_yes = 1;

else children\_yes= 0;

if index(children,'N')>0 then children\_no = 1;

else children\_no= 0;

if index(children,'NA')>0 then children\_unknown = 1;

else children\_unknown= 0;

\*18.car\_buy;

if index(car\_buy,'NEW')>0 then car\_buy\_new = 1;

else car\_buy\_new=0;

if index(car\_buy,'UNKNOWN')>0 or index(car\_buy,'NA') then car\_buy\_unknown = 1;

else car\_buy\_unknown=0;

\*run;

\*19.Area;

if index(area,'ATLANTIC SOUTH AREA')>0 then area\_atlantic = 1;

else area\_atlantic = 0;

if index(area,'CALIFORNIA NORTH AREA')>0 then area\_cali = 1;

else area\_cali = 0;

if index(area,'CENTRAL/SOUTH TEXAS AREA')>0 then area\_texas = 1;

else area\_texas = 0;

if index(area,'CHICAGO AREA')>0 then area\_chicago = 1;

else area\_chicago = 0;

if index(area,'DALLAS AREA')>0 then area\_dallas = 1;

else area\_dallas = 0;

if index(area,'DC/MARYLAND/VIRGINIA AREA')>0 then area\_dcmvir = 1;

else area\_dcmvir = 0;

if index(area,'GREAT LAKES AREA')>0 then area\_gla = 1;

else area\_gla = 0;

if index(area,'HOUSTON AREA')>0 then area\_houston = 1;

else area\_houston = 0;

if index(area,'LOS ANGELES AREA')>0 then area\_la = 1;

else area\_la = 0;

if index(area,'MIDWEST AREA')>0 then area\_mdwest = 1;

else area\_mdwest = 0;

if index(area,'NA')>0 then area\_unknown = 1;

else area\_unknown = 0;

if index(area,'NEW ENGLAND AREA')>0 then area\_neweng = 1;

else area\_neweng = 0;

if index(area,'NEW YORK CITY AREA')>0 then area\_nyc = 1;

else area\_nyc = 0;

if index(area,'NORTH FLORIDA AREA')>0 then area\_nfl = 1;

else area\_nfl = 0;

if index(area,'NORTHWEST/ROCKY MOUNTAIN AREA')>0 then area\_nrocky = 1;

else area\_nrocky = 0;

if index(area,'OHIO AREA')>0 then area\_ohio = 1;

else area\_ohio = 0;

if index(area,'PHILADELPHIA AREA')>0 then area\_phily = 1;

else area\_phily = 0;

if index(area,'SOUTH FLORIDA AREA')>0 then area\_sfl = 1;

else area\_sfl = 0;

if index(area,'SOUTHWEST AREA')>0 then area\_swest = 1;

else area\_swest = 0;

if index(area,'TENNESSEE AREA')>0 then area\_tenese = 1;

else area\_tenese = 0;

\*20.csa;

if substr(csa,1,1) = "A" or substr(csa,1,1) = "B" or substr(csa,1,1) = "C" or substr(csa,1,1) = "D"

or substr(csa,1,1) = "F" or substr(csa,1,1) = "G" then csa\_city1 = 1;

else csa\_city1 = 0;

if substr(csa,1,1) = "H" or substr(csa,1,1) = "I" or substr(csa,1,1) = "K"

or substr(csa,1,1) = "L" or substr(csa,1,1) = "M" then csa\_city2 = 1;

else csa\_city2 = 0;

if substr(csa,1,1) = "N" or substr(csa,1,1) = "O" or substr(csa,1,1) = "P" or substr(csa,1,1) = "S"

or substr(csa,1,1) = "V" then csa\_city3 = 1;

else csa\_city3 = 0;

run;

data temp\_backp2;

set temp\_dummy;

run;

data temp\_backup3(drop = crclscod asl\_flag prizm\_social\_one refurb\_new hnd\_webcap marital ethnic dwlltype dwllsize mailordr occu1 numbcars proptype mailresp cartype car\_buy children div\_type area csa);

set temp\_backp2;

run;

proc means n nmiss data = temp\_backup3;

run;

proc freq data =temp\_merge3;

tables income crclscod asl\_flag prizm\_social\_one area refurb\_new hnd\_webcap marital ethnic dwlltype dwllsize mailordr occu1 numbcars retdays proptype mailresp cartype car\_buy children csa div\_type;

run;

/\*outlier detection for variables:

actvsubs

adjmou

adjqty

adjrev

age1

age2

avg3mou

avg3qty

avg6mou

avg6qty

avgmou

avgqty

avgrev

callwait\_Mean

callwait\_Range

ccrndmou\_Range

change\_mou

comp\_vce\_Mean

custcare\_Mean

da\_Mean

da\_Range

datovr\_Mean

datovr\_Range

drop\_blk\_Mean

drop\_dat\_Mean

drop\_vce\_Mean

drop\_vce\_Range

eqpdays

iwylis\_vce\_Mean

months

mou\_mean

mou\_opkv\_Range

mou\_pead\_Mean

mou\_Range

opk\_dat\_Mean

ovrmou\_Mean

ovrrev\_Mean

owylis\_vce\_Range

plcd\_vce\_Mean

recv\_sms\_Mean

rev\_Mean

rev\_Range

roam\_Mean

totcalls

totmrc\_mean

totrev

uniqsubs

blck\_dat\_Mean\*/

proc univariate data = temp\_backup3;

var mou\_mean;

output out=percentile1\_99 pctlpts=99.9 pctlpre=percentile;

histogram mou\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e. 3765\*/

data temp\_out;

set temp\_backup3;

if mou\_mean ge 3765 then delete;

run;

proc univariate data = temp\_out;

var totmrc\_mean;

output out=percentile2\_99 pctlpts=99.9 pctlpre=percentile;

histogram totmrc\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e. 189.99\*/

data temp\_out;

set temp\_out;

if totmrc\_mean > 189.99 then delete;

run;

proc univariate data = temp\_out;

var rev\_Range;

output out=percentile3\_99 pctlpts=99.9 pctlpre=percentile;

histogram rev\_range;

run;

/\*Above 99.9 percentile values to be deleted i.e. 671.05\*/

data temp\_out;

set temp\_out;

if rev\_Range > 671.05 then delete;

run;

proc univariate data = temp\_out;

var mou\_Range;

output out=percentile4\_99 pctlpts=99.9 pctlpre=percentile;

Histogram mou\_range;

run;

/\*Above 99.9 percentile values to be deleted i.e. 3317\*/

data temp\_out;

set temp\_out;

if mou\_Range >3317 then delete;

run;

proc univariate data = temp\_out;

var change\_mou;

output out=percentile5\_99 pctlpts=99.9 pctlpre=percentile;

Histogram change\_mou;

run;

/\*Above 99.9 percentile values to be deleted i.e. 1387.25\*/

data temp\_out;

set temp\_out;

if change\_mou >1387.25 then delete;

run;

proc univariate data = temp\_out;

var drop\_blk\_Mean;

output out=percentile6\_99 pctlpts=99.9 pctlpre=percentile;

Histogram drop\_blk\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e. 157\*/

data temp\_out;

set temp\_out;

if drop\_blk\_Mean >157 then delete;

run;

proc univariate data = temp\_out;

var drop\_vce\_Range;

output out=percentile7\_99 pctlpts=99.9 pctlpre=percentile;

\*Histogram drop\_vce\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e. 79\*/

data temp\_out;

set temp\_out;

if drop\_vce\_Range > 79 then delete;

run;

proc univariate data = temp\_out;

var owylis\_vce\_Range;

output out=percentile8\_99 pctlpts=99.9 pctlpre=percentile;

histogram owylis\_vce\_range;

run;

/\*Above 99.9 percentile values to be deleted i.e. 214\*/

data temp\_out;

set temp\_out;

if owylis\_vce\_Range >214 then delete;

run;

proc univariate data = temp\_out;

var mou\_opkv\_Range;

output out=percentile9\_99 pctlpts=99.9 pctlpre=percentile;

\*Histogram mou\_pokv\_range;

run;

/\*Above 99.9 percentile values to be deleted i.e. 1505.12\*/

data temp\_out;

set temp\_out;

if mou\_opkv\_Range >1505.12 then delete;

run;

proc univariate data = temp\_out;

var months;

output out=percentile10\_99 pctlpts=99.9 pctlpre=percentile;

Histogram months;

run;

/\*Above 99.9 percentile values to be deleted i.e. 58\*/

data temp\_out;

set temp\_out;

if months>58 then delete;

run;

proc univariate data = temp\_out;

var totcalls;

output out=percentile11\_99 pctlpts=99.9 pctlpre=percentile;

Histogram totcalls;

run;

/\*Above 99.9 percentile values to be deleted i.e. 39755\*/

data temp\_out;

set temp\_out;

if totcalls>39755 then delete;

run;

proc univariate data = temp\_out;

var eqpdays;

output out=percentile12\_99 pctlpts=99.9 pctlpre=percentile;

Histogram eqpdays;

run;

/\*Above 99.9 percentile values to be deleted i.e. 1506\*/

data temp\_out;

set temp\_out;

if eqpdays>1506 then delete;

if eqpdays <= 0 then delete;

run;

proc univariate data = temp\_out;

var custcare\_Mean;

output out=percentile13\_99 pctlpts=99.9 pctlpre=percentile;

Histogram custcare\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e. 47.33333333\*/

data temp\_out;

set temp\_out;

if custcare\_Mean>47.33333333 then delete;

run;

proc univariate data = temp\_out;

var callwait\_Mean;

output out=percentile14\_99 pctlpts=99.9 pctlpre=percentile;

Histogram callwait\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e. 48.66666667\*/

data temp\_out;

set temp\_out;

if callwait\_Mean>48.66666667 then delete;

run;

proc univariate data = temp\_out;

var iwylis\_vce\_Mean;

output out=percentile15\_99 pctlpts=99.9 pctlpre=percentile;

Histogram iwylis\_vce\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.137\*/

data temp\_out;

set temp\_out;

if iwylis\_vce\_Mean>137 then delete;

run;

proc univariate data = temp\_out;

var callwait\_Range;

output out=percentile16\_99 pctlpts=99.9 pctlpre=percentile;

HIstogram callwait\_range;

run;

/\*Above 99.9 percentile values to be deleted i.e.36\*/

data temp\_out;

set temp\_out;

if callwait\_Range>36 then delete;

run;

proc univariate data = temp\_out;

var ccrndmou\_Range;

output out=percentile17\_99 pctlpts=99.9 pctlpre=percentile;

Histogram ccrndmou\_range;

run;

/\*Above 99.9 percentile values to be deleted i.e.171\*/

data temp\_out;

set temp\_out;

if ccrndmou\_Range>171 then delete;

run;

proc univariate data = temp\_out;

var adjqty;

output out=percentile18\_99 pctlpts=99.9 pctlpre=percentile;

Histogram adjqty;

run;

/\*Above 99.9 percentile values to be deleted i.e.31034\*/

data temp\_out;

set temp\_out;

if adjqty>31034 then delete;

run;

proc univariate data = temp\_out;

var ovrrev\_Mean;

output out=percentile19\_99 pctlpts=99.9 pctlpre=percentile;

Histogram ovrrev\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.245.35\*/

data temp\_out;

set temp\_out;

if ovrrev\_Mean>245.35 then delete;

run;

proc univariate data = temp\_out;

var rev\_Mean;

output out=percentile20\_99 pctlpts=99.9 pctlpre=percentile;

Histogram rev\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.322.38\*/

data temp\_out;

set temp\_out;

if rev\_Mean>322.38 then delete;

run;

proc univariate data = temp\_out;

var ovrmou\_Mean;

output out=percentile21\_99 pctlpts=99.9 pctlpre=percentile;

Histogram ovrmou\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.701\*/

data temp\_out;

set temp\_out;

if ovrmou\_Mean>701 then delete;

run;

proc univariate data = temp\_out;

var comp\_vce\_Mean;

output out=percentile22\_99 pctlpts=99.9 pctlpre=percentile;

Histogram comp\_vce\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.878\*/

data temp\_out;

set temp\_out;

if comp\_vce\_Mean>878 then delete;

run;

proc univariate data = temp\_out;

var plcd\_vce\_Mean;

output out=percentile23\_99 pctlpts=99.9 pctlpre=percentile;

Histogram plcd\_vce\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.1040\*/

data temp\_out;

set temp\_out;

if plcd\_vce\_Mean>1040 then delete;

run;

proc univariate data = temp\_out;

var avg3mou;

output out=percentile24\_99 pctlpts=99.9 pctlpre=percentile;

Histogram avg3mou;

run;

/\*Above 99.9 percentile values to be deleted i.e.3273\*/

data temp\_out;

set temp\_out;

if avg3mou>3273 then delete;

run;

proc univariate data = temp\_out;

var avgmou;

output out=percentile25\_99 pctlpts=99.9 pctlpre=percentile;

Histogram avgmou;

run;

/\*Above 99.9 percentile values to be deleted i.e.2694\*/

data temp\_out;

set temp\_out;

if avgmou>2694 then delete;

run;

proc univariate data = temp\_out;

var avg3qty;

output out=percentile26\_99 pctlpts=99.9 pctlpre=percentile;

Histogram avg3qty;

run;

/\*Above 99.9 percentile values to be deleted i.e.2694\*/

data temp\_out;

set temp\_out;

if avg3qty>2694 then delete;

run;

proc univariate data = temp\_out;

var avgqty;

output out=percentile27\_99 pctlpts=99.9 pctlpre=percentile;

Histogram avgqty;

run;

/\*Above 99.9 percentile values to be deleted i.e.1191\*/

data temp\_out;

set temp\_out;

if avgqty>1191 then delete;

run;

proc univariate data = temp\_out;

var actvsubs;

output out=percentile28\_99 pctlpts=99.9 pctlpre=percentile;

Histogram actvsubs;

run;

/\*Above 99.9 percentile values to be deleted i.e.5\*/

data temp\_out;

set temp\_out;

if actvsubs>5 then delete;

run;

proc univariate data = temp\_out;

var uniqsubs;

output out=percentile29\_99 pctlpts=99.9 pctlpre=percentile;

Histogram uniqsubs;

run;

/\*Above 99.9 percentile values to be deleted i.e.7\*/

data temp\_out;

set temp\_out;

if uniqsubs>7 then delete;

run;

proc univariate data = temp\_out;

var opk\_dat\_Mean;

output out=percentile30\_99 pctlpts=99.9 pctlpre=percentile;

HIstogram opk\_dat\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.59\*/

data temp\_out;

set temp\_out;

if opk\_dat\_Mean>59 then delete;

run;

proc univariate data = temp\_out;

var roam\_Mean;

output out=percentile31\_99 pctlpts=99.9 pctlpre=percentile;

Histogram roam\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.76\*/

data temp\_out;

set temp\_out;

if roam\_Mean>76 then delete;

run;

proc univariate data = temp\_out;

var recv\_sms\_Mean;

output out=percentile32\_99 pctlpts=99.9 pctlpre=percentile;

Histogram recv\_sms\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.9\*/

data temp\_out;

set temp\_out;

if recv\_sms\_Mean>9 then delete;

run;

proc univariate data = temp\_out;

var mou\_pead\_Mean;

output out=percentile33\_99 pctlpts=99.9 pctlpre=percentile;

Histogram mou\_pead\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.82\*/

data temp\_out;

set temp\_out;

if mou\_pead\_Mean>82 then delete;

run;

proc univariate data = temp\_out;

var da\_Mean;

output out=percentile34\_99 pctlpts=99.9 pctlpre=percentile;

Histogram da\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.19\*/

data temp\_out;

set temp\_out;

if da\_Mean>19 then delete;

run;

proc univariate data = temp\_out;

var da\_Range;

output out=percentile35\_99 pctlpts=99.9 pctlpre=percentile;

\*HISTOGRAM da-range;

run;

/\*Above 99.9 percentile values to be deleted i.e.22\*/

data temp\_out;

set temp\_out;

if da\_Range>22 then delete;

run;

proc univariate data = temp\_out;

var datovr\_Mean;

output out=percentile36\_99 pctlpts=99.9 pctlpre=percentile;

Histogram datovr\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.24\*/

data temp\_out;

set temp\_out;

if datovr\_Mean>24 then delete;

run;

proc univariate data = temp\_out;

var datovr\_Range ;

output out=percentile37\_99 pctlpts=99.9 pctlpre=percentile;

Histogram datovr\_range;

run;

/\*Above 99.9 percentile values to be deleted i.e.46\*/

data temp\_out;

set temp\_out;

if datovr\_Range>46 then delete;

run;

proc univariate data = temp\_out;

var drop\_dat\_Mean;

output out=percentile38\_99 pctlpts=99.9 pctlpre=percentile;

Histogram drop\_dat\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.4\*/

data temp\_out;

set temp\_out;

if drop\_dat\_Mean>4 then delete;

run;

proc univariate data = temp\_out;

var drop\_vce\_Mean;

output out=percentile39\_99 pctlpts=99.9 pctlpre=percentile;

Histogram drop\_vce\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.77\*/

data tepm\_out;

set temp\_out;

if drop\_vce\_Mean>77 then delete;

run;

proc univariate data = temp\_out;

var adjmou;

output out=percentile40\_99 pctlpts=99.9 pctlpre=percentile;

Histogram adjmou;

run;

/\*Above 99.9 percentile values to be deleted i.e.75256\*/

data temp\_out;

set temp\_out;

if adjmou>75256 then delete;

run;

proc univariate data = temp\_out;

var totrev;

output out=percentile41\_99 pctlpts=99.9 pctlpre=percentile;

Histogram totrev;

run;

/\*Above 99.9 percentile values to be deleted i.e.6687\*/

data temp\_out;

set temp\_out;

if totrev>6687 then delete;

run;

proc univariate data = temp\_out;

var adjrev;

output out=percentile42\_99 pctlpts=99.9 pctlpre=percentile;

Histogram adjrev;

run;

/\*Above 99.9 percentile values to be deleted i.e.5739\*/

data temp\_out;

set temp\_out;

if adjrev>5739 then delete;

run;

proc univariate data = temp\_out;

var avgrev;

output out=percentile43\_99 pctlpts=99.9 pctlpre=percentile;

Histogram avgrev;

run;

/\*Above 99.9 percentile values to be deleted i.e.246\*/

data temp\_out;

set temp\_out;

if avgrev>246 then delete;

run;

proc univariate data = temp\_out;

var avg6mou;

output out=percentile44\_99 pctlpts=99.9 pctlpre=percentile;

HISTOGRAM avg6mou;

run;

/\*Above 99.9 percentile values to be deleted i.e.2648\*/

data temp\_out;

set temp\_out;

if avg6mou>2648 then delete;

run;

proc univariate data = temp\_out;

var avg6qty;

output out=percentile45\_99 pctlpts=99.9 pctlpre=percentile;

Histogram avg6qty;

run;

/\*Above 99.9 percentile values to be deleted i.e.1055\*/

data temp\_out;

set temp\_out;

if avg6qty>1055 then delete;

run;

proc univariate data = temp\_out;

var blck\_dat\_Mean;

output out=percentile46\_99 pctlpts=99.9 pctlpre=percentile;

Histogram blck\_dat\_mean;

run;

/\*Above 99.9 percentile values to be deleted i.e.2\*/

data temp\_out;

set temp\_out;

if blck\_dat\_Mean>2 then delete;

run;

/\*Outliers detection in age1 and age2\*/

proc univariate data = temp\_out;

var age1;

output out=percentile47\_99 pctlpts=99.9 pctlpre=percentile;

Histogram age1;

run;

/\*Above 99.9 percentile values to be deleted i.e.86\*/

data temp\_out;

set temp\_out;

if age1>86 then delete;

run;

/\*Imputing age1=0 with mean of age1 i.e.31\*/

data temp\_out1;

set temp\_out;

if age1 = 0 then age1 = 31;

run;

proc univariate data = temp\_out;

var age2;

output out=percentile48\_99 pctlpts=99.9 pctlpre=percentile;

Histogram age2;

run;

/\*Above 99.9 percentile values to be deleted i.e.90\*/

data temp\_out;

set temp\_out;

if age2>90 then delete;

run;

data temp\_out2;

set temp\_out1;

if age2 = 0 then age2 = 21;

run;

proc means n nmiss min max mean data= temp\_out2;

var age1 age2 income;

run;

proc contents data = temp\_out2 varnum;

run;

proc freq data = temp\_dummy;

tables area csa;

run;

/\*No missing values in area and csa\*/

/\*\*\*\*\*\*\*\*\*\*\*End of Data Preparation\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Start of Modelling\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*Partitioning data into Development and Validation Dataset\*/

data development validation;

set temp\_out2;

if ranuni (100) <0.70 then output development;

else output validation;

run;

proc freq data = development;

tables churn;

run;

proc freq data = validation;

tables churn;

run;

/\*Running Logistic Regression on development dataset: Iteration 1\*/

proc logistic data = development descending;

model churn =

actvsubs

adjmou

adjqty

adjrev

age1

age2

area\_atlantic

area\_cali

area\_chicago

area\_city

area\_dallas

area\_dcmvir

area\_gla

area\_houston

area\_la

area\_mdwest

area\_neweng

area\_nfl

area\_nrocky

area\_nyc

area\_ohio

area\_phily

area\_rural

area\_sfl

area\_suburban

area\_swest

area\_tenese

area\_texas

area\_town

area\_unknown

area\_urban

asl\_flag\_no

asl\_flag\_yes

average\_crclscod

avg3mou

avg3qty

avg6mou

avg6qty

avgmou

avgqty

avgrev

bad\_crclscod

best\_crclscod

blck\_dat\_Mean

callwait\_Mean

callwait\_Range

car\_buy\_new

car\_buy\_unknown

cartype\_basic

cartype\_luxury

cartype\_minivan

cartype\_regular

cartype\_SUV

cartype\_truck

cartype\_unknown

cartype\_upper

ccrndmou\_Range

change\_mou

children\_no

children\_unknown

children\_yes

comp\_vce\_Mean

csa\_city1

csa\_city2

csa\_city3

custcare\_Mean

da\_Mean

da\_Range

datovr\_Mean

datovr\_Range

division\_bth

division\_LDD

division\_LTD

division\_unknown

drop\_blk\_Mean

drop\_dat\_Mean

drop\_vce\_Mean

drop\_vce\_Range

dsize\_1

dsize\_100

dsize\_10to19

dsize\_2

dsize\_20to29

dsize\_3

dsize\_30to39

dsize\_4

dsize\_40to49

dsize\_5

dsize\_50to99

dsize\_6

dsize\_7

dsize\_8

dsize\_9

dsize\_unknown

dwell\_multiple

dwell\_single

dwell\_unknown

eqpdays

ethnic\_afro\_amer

ethnic\_arab

ethnic\_asian

ethnic\_asian\_nor

ethnic\_french

ethnic\_german

ethnic\_hispanic

ethnic\_italian

ethnic\_jewish

ethnic\_misc

ethnic\_north\_euro

ethnic\_polynesia

ethnic\_scot\_iris

ethnic\_south\_euro

ethnic\_unknown

forgntvl

good\_crclscod

handset\_na

handset\_new

handset\_refurb

handset\_unkw

handset\_wc

handset\_wcmb

hnd\_price

income

iwylis\_vce\_Mean

mailorder\_buyer

mailorder\_unknown

mailresp\_unknown

mailresp\_yes

models

months

mou\_Mean

mou\_opkv\_Range

mou\_pead\_Mean

mou\_Range

mtrcycle

numcars\_1

numcars\_2

numcars\_3

numcars\_unknown

occu\_admin

occu\_bc

occu\_farmer

occu\_homemaker

occu\_military

occu\_religious

occu\_retires

occu\_sales

occu\_selfemp

occu\_student

occu\_technical

occu\_wc

opk\_dat\_Mean

ovrmou\_Mean

ovrrev\_Mean

owylis\_vce\_Range

plcd\_vce\_Mean

proptype\_a

proptype\_b

proptype\_d

proptype\_e

proptype\_g

proptype\_m

proptype\_unknown

recv\_sms\_Mean

rev\_Mean

rev\_Range

roam\_Mean

status\_infermarried

status\_infersingle

status\_married

status\_single

status\_unknown

totcalls

totmrc\_Mean

totrev

truck

uniqsubs

worst\_crclscod

;

output out = churn\_model1 predicted = pred\_prob;

run;

/\*iteration2\*/

proc logistic data = development descending;

model churn =

actvsubs

age1

area\_rural

area\_town

asl\_flag\_yes

best\_crclscod

avgmou

mou\_mean

change\_mou

custcare\_Mean

division\_bth

division\_LDD

drop\_blk\_Mean

dsize\_4

dsize\_40to49

eqpdays

handset\_new

hnd\_price

mailorder\_buyer

mailresp\_unknown

models

months

mou\_Range

opk\_dat\_Mean

ovrmou\_Mean

status\_single

uniqsubs/ctable lackfit rsquare

;

output out = churn\_model2 predicted = pred\_prob;

run;

proc sort data churn\_model2 out = churn\_sorted;

by descending pred\_prob ;

run;

proc rank data = churn\_sorted out = decile groups = 10 ties=mean;

var pred\_prob;

ranks decile;

run;

proc export data = decile

outfile ='C:\Users\Vandana Vaanya\Desktop\Jigsaw\SAS\14.final project'

dbms = xlsx replace ;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*End of Development Dataset Modelling\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*Validation Modeling\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

proc logistic data = Validation descending;

model churn =

actvsubs

adjmou

adjqty

adjrev

age1

age2

area\_atlantic

area\_cali

area\_chicago

area\_city

area\_dallas

area\_dcmvir

area\_gla

area\_houston

area\_la

area\_mdwest

area\_neweng

area\_nfl

area\_nrocky

area\_nyc

area\_ohio

area\_phily

area\_rural

area\_sfl

area\_suburban

area\_swest

area\_tenese

area\_texas

area\_town

area\_unknown

area\_urban

asl\_flag\_no

asl\_flag\_yes

average\_crclscod

avg3mou

avg3qty

avg6mou

avg6qty

avgmou

avgqty

avgrev

bad\_crclscod

best\_crclscod

blck\_dat\_Mean

callwait\_Mean

callwait\_Range

car\_buy\_new

car\_buy\_unknown

cartype\_basic

cartype\_luxury

cartype\_minivan

cartype\_regular

cartype\_SUV

cartype\_truck

cartype\_unknown

cartype\_upper

ccrndmou\_Range

change\_mou

children\_no

children\_unknown

children\_yes

comp\_vce\_Mean

csa\_city1

csa\_city2

csa\_city3

custcare\_Mean

da\_Mean

da\_Range

datovr\_Mean

datovr\_Range

division\_bth

division\_LDD

division\_LTD

division\_unknown

drop\_blk\_Mean

drop\_dat\_Mean

drop\_vce\_Mean

drop\_vce\_Range

dsize\_1

dsize\_100

dsize\_10to19

dsize\_2

dsize\_20to29

dsize\_3

dsize\_30to39

dsize\_4

dsize\_40to49

dsize\_5

dsize\_50to99

dsize\_6

dsize\_7

dsize\_8

dsize\_9

dsize\_unknown

dwell\_multiple

dwell\_single

dwell\_unknown

eqpdays

ethnic\_afro\_amer

ethnic\_arab

ethnic\_asian

ethnic\_asian\_nor

ethnic\_french

ethnic\_german

ethnic\_hispanic

ethnic\_italian

ethnic\_jewish

ethnic\_misc

ethnic\_north\_euro

ethnic\_polynesia

ethnic\_scot\_iris

ethnic\_south\_euro

ethnic\_unknown

forgntvl

good\_crclscod

handset\_na

handset\_new

handset\_refurb

handset\_unkw

handset\_wc

handset\_wcmb

hnd\_price

income

iwylis\_vce\_Mean

mailorder\_buyer

mailorder\_unknown

mailresp\_unknown

mailresp\_yes

models

months

mou\_Mean

mou\_opkv\_Range

mou\_pead\_Mean

mou\_Range

mtrcycle

numcars\_1

numcars\_2

numcars\_3

numcars\_unknown

occu\_admin

occu\_bc

occu\_farmer

occu\_homemaker

occu\_military

occu\_religious

occu\_retires

occu\_sales

occu\_selfemp

occu\_student

occu\_technical

occu\_wc

opk\_dat\_Mean

ovrmou\_Mean

ovrrev\_Mean

owylis\_vce\_Range

plcd\_vce\_Mean

proptype\_a

proptype\_b

proptype\_d

proptype\_e

proptype\_g

proptype\_m

proptype\_unknown

recv\_sms\_Mean

rev\_Mean

rev\_Range

roam\_Mean

status\_infermarried

status\_infersingle

status\_married

status\_single

status\_unknown

totcalls

totmrc\_Mean

totrev

truck

uniqsubs

worst\_crclscod;

output out = churn\_model3 predicted = pred\_prob1;

run;

/\*iteration2\*/

proc logistic data = validation descending;

model churn =

actvsubs

age1

area\_rural

area\_town

asl\_flag\_yes

best\_crclscod

avgmou

mou\_mean

change\_mou

custcare\_Mean

division\_bth

division\_LDD

drop\_blk\_Mean

dsize\_4

dsize\_40to49

eqpdays

handset\_new

hnd\_price

mailorder\_buyer

mailresp\_unknown

models

months

mou\_Range

opk\_dat\_Mean

ovrmou\_Mean

status\_single

uniqsubs/ctable lackfit rsquare

;

output out = churn\_model4 predicted = pred\_prob1;

run;

proc sort data churn\_model4 out = churn\_sorted1;

by descending pred\_prob1 ;

run;

proc rank data = churn\_sorted1 out = decile1 groups = 10 ties=mean;

var pred\_prob1;

ranks decile1;

run;

proc export data = decile1

outfile ='C:\Users\Vandana Vaanya\Desktop\Jigsaw\SAS\14.final project'

dbms = xlsx replace ;

run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*End of Validation Data Modeling\*\*\*\*\*\*\*\*\*\*\*\*\*/