```
PROC IMPORT DATAFILE='/home/pareshbits90/sasuser.v94/telecomfinal.csv'
        DBMS=CSV
        OUT=Cap2
        REPLACE;
        GETNAMES=YES;
RUN;

* Basic Data Exploration;
Proc contents data=cap2;
run;

Proc Means data=cap2 n nmiss min max;
run;

Proc univariate data=cap2;
run;

*Doing Bivariate Analysis for some combinations;
proc freq data=cap2;
tables crclscod*churn /norow nocol;
run;
*(AA class has high churn rate hence likely to be significant );

Proc Freq data=cap2;
tables hnd_webcap*churn /norow nocol;
run;
*(WCMB client has a high churn rate hence likely to be significant);

Proc Freq data=cap2;
tables ethnic*churn /norow nocol;;
run;

Proc Freq data=cap2;
tables area*churn;
run;

*Bivariate analysis;
Proc hpbin data=cap2 numbin=10 out=bucketed bucket;
var non_optimal_plan;
run;

Proc sort data=cap2;
by non_optimal_plan;
run;

Data usage;
merge cap2(in=a) bucketed(in=b);
if a and b;
run;

proc Freq data=usage;
tables bin_non_optimal_plan*churn / nocol;
run;

**Checking correlations for some variables;
Proc corr data=cap2 plots(maxpoints=none)=matrix(histogram);
var  non_optimal_plan churn;
run;

* Let us check if there are any missing/NA/unknown values left in excel as it's easy;
Proc Export data=cap2 outfile='/home/pareshbits90/sasuser.v94/Capstone Project.csv'
dbms=csv replace;
run;

*After doing some basic data exploration in Excel& SAS let us drop 14 variables having more than 25% missing
data;

Data Cap2;
Set Cap2;
drop income dwlltype dwllsize mailordr occu1 numbcars retdays wrkwoman solflag proptype mailresp cartype
car_buy children div_type csa ;
run;

*There are some other variables which have some missing values, but can be handled by deleting some
observations;

data cap2;
set cap2;
if avg6mou ne "NA";
if age1 ne "NA";
if hnd_price ne "NA";
```

```
if change_mou ne "NA";
if hnd_webcap ne "NA";
if prizm_social_one ne "NA";
if area ne "NA";
run;

**After bivariate analysis of credit class code vs attrition, we decide to reduce 49 different classes
 into 6 classes during data preparation and similarly reduce area into five zones and ethnic into 5 types;

**Data Preparation;
Data Cap2;
Set Cap2;

if crclscod="A" then creditclass1=1; else creditclass1=0;
if crclscod="AA" then creditclass2=1; else creditclass2=0;
if crclscod="BA" then creditclass3=1; else creditclass3=0;
if crclscod="CA" then creditclass4=1; else creditclass4=0;
if crclscod="EA" then creditclass5=1; else creditclass5=0;

if crclscod="A2" or crclscod="A3" or crclscod="B" or crclscod="B2" or crclscod="C" or crclscod="C2"
or crclscod="C5" or crclscod="CC" or crclscod="CY" or crclscod="D" or crclscod="D2" or crclscod="D4"
or crclscod="D5" or crclscod="DA" or crclscod="E" or crclscod="E2" or crclscod="E4" or crclscod="EC"
or crclscod="EM" or crclscod="G" or crclscod="GA" or crclscod="GY" or crclscod="H" or crclscod="I"
or crclscod="J" or crclscod="JF" or crclscod="K" or crclscod="L" or crclscod="M" or crclscod="O"
or crclscod="P1" or crclscod="TP" or crclscod="U" or crclscod="U1" or crclscod="V1" or crclscod="W"
or crclscod="Y" or crclscod="Z" or crclscod="Z1" or crclscod="Z2" or crclscod="Z4" or crclscod="Z5"
or crclscod="ZA" or crclscod="ZY" then creditclass6=1;else creditclass6=0;

if asl_flag="Y" then spending_limit1=1; else spending_limit1=0;
if asl_flag="N" then spending_limit2=1; else spending_limit2=0;

if prizm_social_one="C" then social_group1=1; else social_group1=0;
if prizm_social_one="R" then social_group2=1; else social_group2=0;
if prizm_social_one="S" then social_group3=1; else social_group3=0;
if prizm_social_one="T" then social_group4=1; else social_group4=0;
if prizm_social_one="U" then social_group5=1; else social_group5=0;

if area="ATLANTIC SOUTH AREA" or area="CENTRAL/SOUTH TEXAS AREA" or area="DALLAS AREA"
or area="DC/MARYLAND/VIRGINIA AREA" or area="HOUSTON AREA" or area="TENNESSEE AREA"
or area="NORTH FLORIDA AREA" or area="SOUTH FLORIDA AREA" then Area1=1; else Area1=0;
if area="CALIFORNIA NORTH AREA" or area="LOS ANGELES AREA" or area="SOUTHWEST AREA"
or area="NORTHWEST/ROCKY MOUNTAIN AR" then Area2=1; else Area2=0;
if area="GREAT LAKES AREA" then area3=1; else area3=0;
if area="CHICAGO AREA" or area="MIDWEST AREA" or area="OHIO AREA"
then Area4=1; else Area4=0;
if area="NEW ENGLAND AREA" or area="PHILADELPHIA AREA" or area="NEW YORK CITY AREA "
then Area5=1; else Area5=0;

if refurb_new="N" then refurb1=1; else refurb1=0;
if refurb_new="R" then refurb2=1; else refurb2=0;

if hnd_webcap="WC" then capacity1=1; else capacity1=0;
if hnd_webcap="WCMB" then capacity2=1; else capacity2=0;

if marital="A" then marital1=1; else marital1=0;
if marital="B" then marital2=1; else marital2=0;
if marital="M" then marital3=1; else marital3=0;
if marital="S" then marital4=1; else marital4=0;
if marital="U" then marital5=1; else marital5=0;

if ethnic="H" then ethnic1=1; else ethnic1=0;
if ethnic="N" then ethnic2=1; else ethnic2=0;
if ethnic="S" then ethnic3=1; else ethnic3=0;
if ethnic="U" then ethnic4=1; else ethnic4=0;
if ethnic="B" or ethnic="C" or ethnic="D" or ethnic="F" or ethnic="G" or ethnic="I" or ethnic="J"
or ethnic="M" or ethnic="O" or ethnic="P" or ethnic="R" or ethnic="X" or ethnic="Z" then
ethnic5=1; else ethnic5=0;

run;

*Let us add derived variables for non-optimal plan, recency, relative usage and billing error;

Data cap2;
set cap2;
non_optimal_plan=ovrrev_mean/avgrev;
recency=avg3mou/avg6mou;
relative_usage=mou_mean/avgmou;
billing_error_rev=adjrev/totrev;
run;

**Deleting Outliers(based on judgement, no specific rule);
```

```
Data cap2;
set cap2;
if 300>totmrc_mean>0;
if rev_range<321;
if mou_range<3000;
if -900 <change_mou <1200;
if drop_vce_range le 150;
if mou_opkv_range<1200;
if 0<mou_mean<4000;
if totcalls<30000;
if eqpdays>10;
if custcare_mean<50;
if ovrrev_mean<605;
if 9<rev_mean<650;
if ovrmou_mean<800;
if roam_mean<100;
if mou_pead_mean<100;
if datovr_range<300;
if datovr_mean<150;
if non_optimal_plan<1.8;
if drop_blk_mean<150;
if owylis_vce_Range<250;
if callwait_Mean<100;
if iwylis_vce_Mean<150;
if callwait_Range<70;
if ccrndmou_Range< 250;
if adjqty<35000;
if comp_vce_Mean<1000;
if plcd_vce_Mean<1200;
if avg3qty<1500;
if avgqty<1500;
if avg6qty<1500;
if models le 10;
if opk_dat_Mean<100;
if da_mean<25;
if adjmou<80000;
if avgrev<500;
if billing_error_rev>0.55;
if drop_blk_mean<125;
run;

**Dividing dataset into training and validation sample;
Data development_sample validation_sample;
set cap2;
if ranuni(100)<0.7 then output development_sample;
else output validation_sample;
run;

*Making use of decision tree to improve model by making clusters i.e let us make a model for 'high' churn;
Data Cap3;
Set cap2;
if eqpdays >300 and relative_usage le 0.4807;
run;

Data Cap4;
Set cap2;
if eqpdays le 300 and months>10;
run;

Proc sort data=cap3;
by customer_id;
run;

Proc sort data=cap4;
by customer_id;
run;

Data Cap_merged;
merge cap3(in=a) cap4(in=b);
if a or b;
by customer_id;
run;

Data cap5;
set cap2;
if avg3mou le 1 and change_mou le 0;
run;

Proc sort data=cap5;
by customer_id;
run;
```

```
Proc sort data=cap_merged;
by customer_id;
run;

Data model1;
merge cap_merged(in=a) cap5(in=b);
if a or b;
by customer_id;
run;

*First iteration;
Proc Logistic data=cap_merged descending;
model churn=totmrc_mean--avg6qty age1--mou_pead_mean da_mean--avgrev creditclass1-creditclass5
spending_limit1 social_group1-social_group4 area1-area4 refurb1 capacity1 marital1-marital4
ethnic1-ethnic4 non_optimal_plan--billing_error_rev /selection=stepwise slentry=0.1
slstay=0.05 stb ;
run;

*second Iteration with only significant variables;
Proc Logistic data=cap_merged descending outmodel=dmm plots(maxpoints=none)=all;
model churn= avg3mou non_optimal_plan comp_vce_Mean drop_blk_Mean eqpdays age2 spending_limit1 hnd_price change_mou months
recency relative_usage models  creditclass5 refurb1 Area1 Area2
billing_error_rev  social_group4 drop_vce_Range marital1 marital4  avgmou
 mou_range/stb ctable;
score out=dmp;
run;

**Let us generate deciles for Lift Chart by following code and then exporting the file to excel;
Proc sort data=dmp;
by P_1;
run;

proc rank data=dmp groups=10 ties=mean out=Lift_Curve;
Var P_1;
Ranks decile;
run;

Proc Export outfile='/home/pareshbits90/sasuser.v94/Lift_Curve.csv'
dbms=csv replace;
run;

**Let us try to make clusters to improve the result of log regression[THIS CODE WAS NOT USED FINALLY];
Proc standard data=cap2 mean=0 std=1 out=Standardized;
Var churn mou_Mean--avg6qty age1--mou_pead_mean da_mean--avgrev creditclass1--billing_error_rev;
run;

Data Standardized;
set standardized;
rev_mean3=rev_mean*3;
drop rev_mean;
run;

Proc fastclus data=standardized maxclusters=15 maxiter=100 out=Clustered_Data ;
Var churn rev_mean3 mou_Mean mou_range months eqpdays custcare_mean ovrrev_mean age1 models--uniqsubs
roam_mean drop_blk_mean creditclass1-creditclass5
spending_limit1 social_group1-social_group4 area1-area4 refurb1 capacity1 marital1-marital4
ethnic1-ethnic4 non_optimal_plan recency relative_usage billing_error_rev;
run;

Data Clustered_Data;
Set clustered_data;
Keep customer_id cluster;
run;

Proc sort data=clustered_data;
by customer_id;
run;

Proc sort data=cap2;
by customer_id;
run;

Data cluster_F;
merge cap2(in=a) clustered_data(in=b);
by customer_id;
if a and b;
run;

Data Cap_separate;
set cluster_F;
if cluster=7;
run;
```

```
Proc freq data=cap_separate;
tables churn;
run;

**Perform Logistic Regression on the following cluster;
Proc Logistic data=cap_separate descending plots=;
model churn=mou_Mean--avg6qty age1--mou_pead_mean da_mean--avgrev creditclass1-creditclass5
spending_limit1 social_group1-social_group4 area1-area18 refurb1 capacity1 marital1-marital4
ethnic1-ethnic4 non_optimal_plan /selection=forward;
run;

**Let us perform Survival Analysis on the dataset;
proc phreg data=cap_merged outest=temp;
model months*churn(0)=mou_Mean--mou_opkv_range totcalls--avg6qty age1--mou_pead_mean da_mean--avgrev creditclass1-creditc
spending_limit1 social_group1-social_group4 area1-area4 refurb1 capacity1 marital1-marital4
ethnic1-ethnic4 non_optimal_plan--billing_error_rev /selection=stepwise slentry=0.0025 slstay=0.0025
details;
run;

proc lifetest data=cap_merged outsurv=temp2
method=life plot=(S,H) width=1 graphics;
time months*churn(0);
run;

**significant variables from PHREG analysis::
mou_Mean totmrc_Mean rev_range mou_Range change_mou drop_blk_Mean totcalls eqpdays callwait_Mean adjqty ovrrev_Mean
rev_Mean avgmou age1 models actvsubs uniqsubs roam_Mean totrev adjrev avgrev creditclass1
creditclass2 creditclass3 creditclass4 spending_limit1 social_group2 social_group4 area4 refurb1
ethnic2 non_optimal_plan billing_error_rev;
```