**SAS codes**

**Import dataset in the SAS environment and check top 10 record of import dataset (2 Mark)**

FILENAME REFFILE '/home/u59032982/Life+Insurance+Dataset.csv';

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

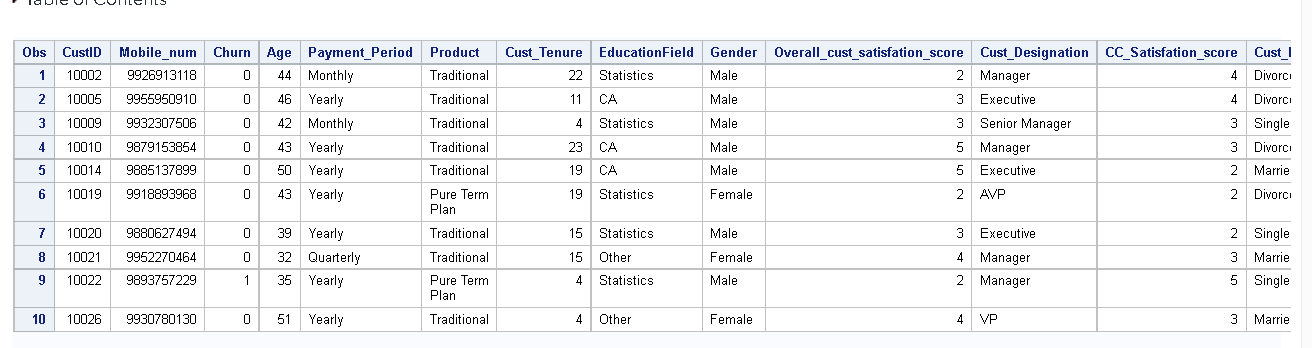
OUT=WORK.insurance;

GETNAMES=YES;

RUN;

proc print data=insurance(obs=10);

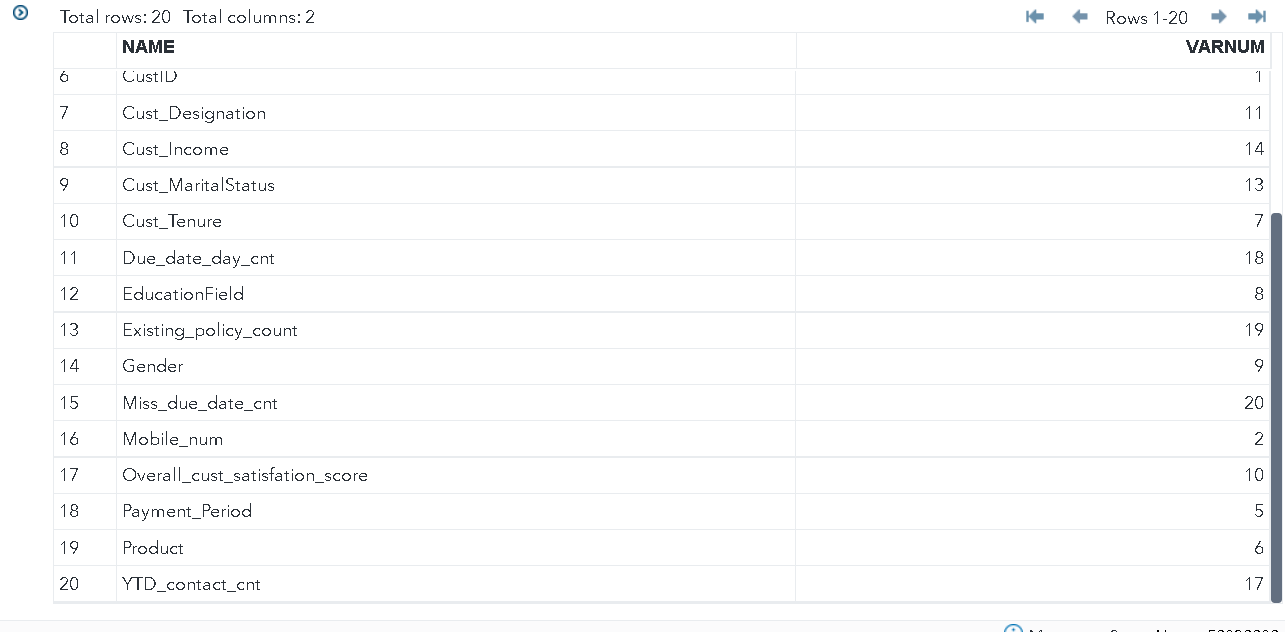
run;

Expected output: Top ten rows:

**Check variable type of the import dataset (2 Mark)**

PROC CONTENTS DATA=WORK.insurance noprint out=info(keep=name varnum);

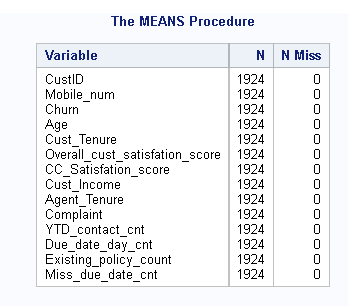
RUN;



**Checks if any variables have missing values, if yes then do treatment? (3 Mark)**

proc means data=insurance n nmiss;

run;

Expected: no missing values

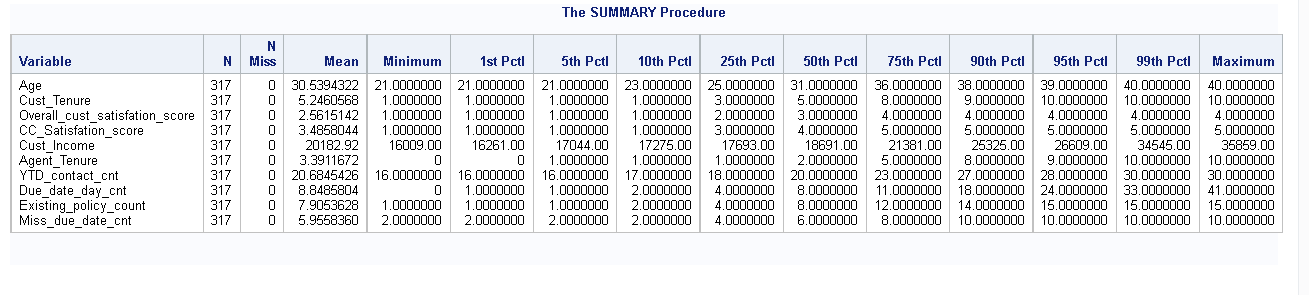
**Check summary and percentile distribution of all numerical variables for churners and non-churners? (5 Marks)**

proc summary data=insurance print n nmiss mean min p1 p5 p10 p25 p50 p75 p90 p95 p99 max;

var Age cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income agent\_Tenure YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt;

where churn=1;

run;

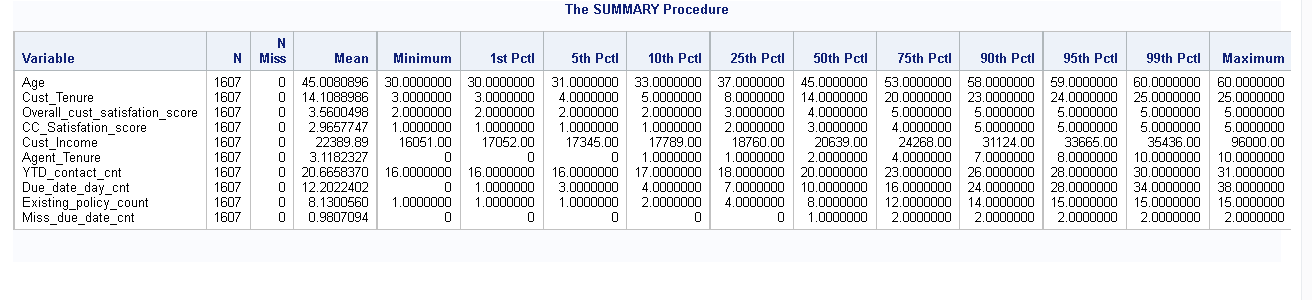


proc summary data=insurance print n nmiss mean min p1 p5 p10 p25 p50 p75 p90 p95 p99 max;

var Age cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income Agent\_Tenure

YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt;

where churn=0;

run;

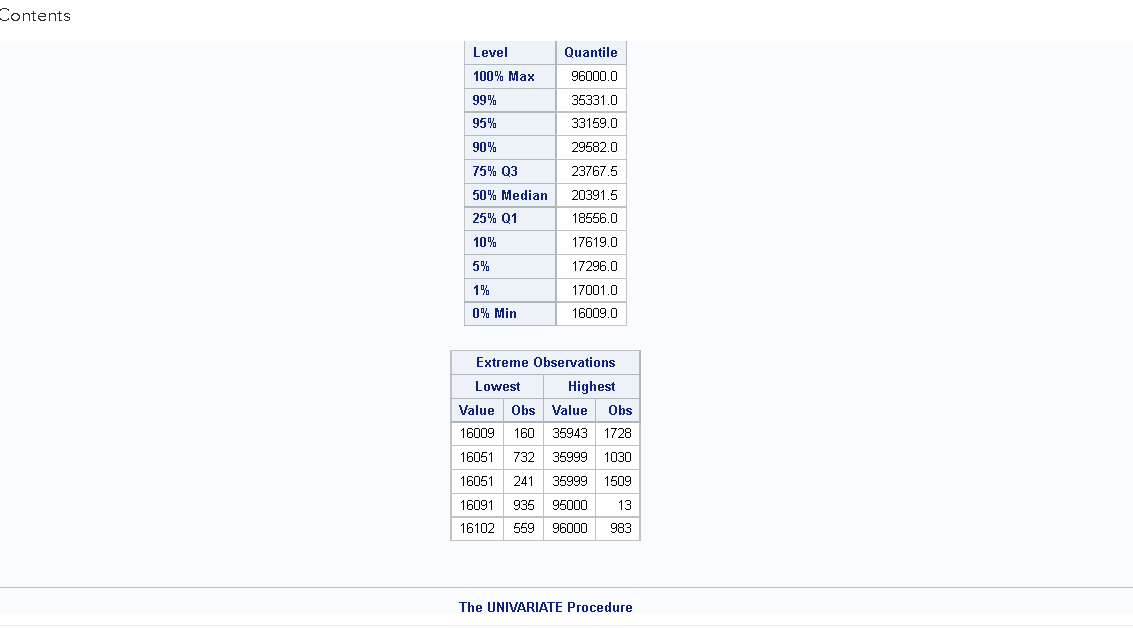
**Check for outlier, if yes then do treatment? (3 Mark)**

proc univariate data=insurance;

var Age cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income Agent\_Tenure

YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt;

run;

Expected: Customer income has 2 outliers

**assuming last 2 incomes in $90000 range**

data insurance;

set insurance;

if cust\_income >36000 then cust\_income = 36000;

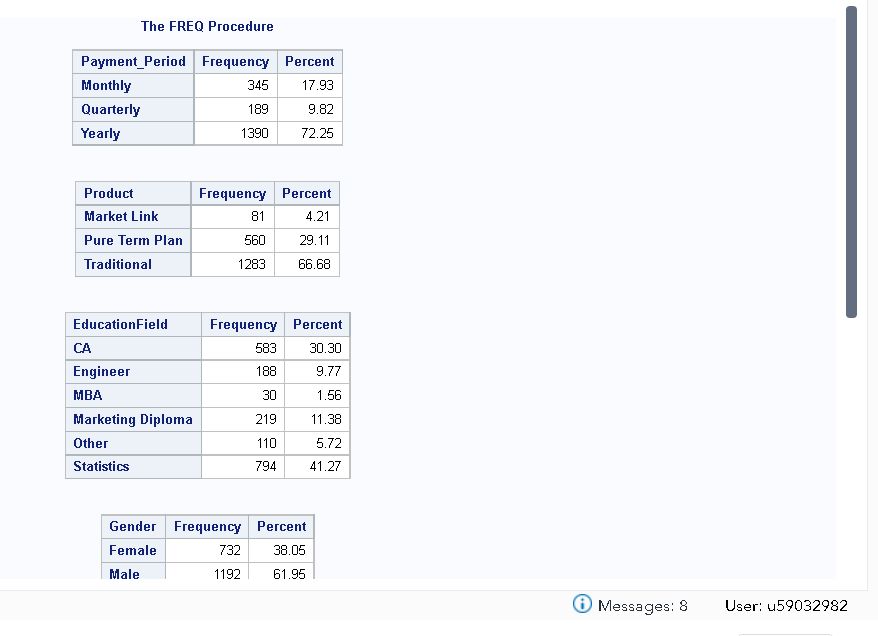
run;

**Check the proportion of all categorical variables and extract percentage contribution of each class in respective variables? (5 Marks)**

proc freq data=insurance;

table Payment\_Period Product EducationField Gender Cust\_Designation Cust\_MaritalStatus Complaint/ nocum;

run;

****

**Customer service management want you to create a macro where they will just put mobile number and they will get all the important information like Age, Education, Gender, Income and CustID (6 Marks)**

%MACRO cust\_info();

DATA output (keep = custid age mobile\_num educationfield gender cust\_income );

SET insurance;

where mobile\_num in (&mobile\_num.);

RUN;

proc print data=output;

run;

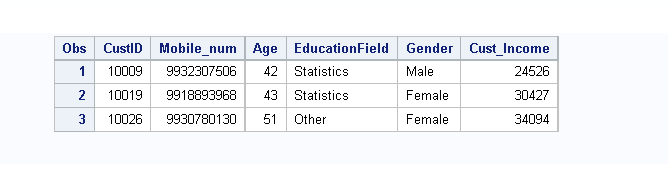
%MEND;

**Provided input mobile number**

%let mobile\_num = 9932307506, 9918893968, 9930780130 ;

**run macro for output**

%cust\_info;

Expected: displays relevant info of given numbers

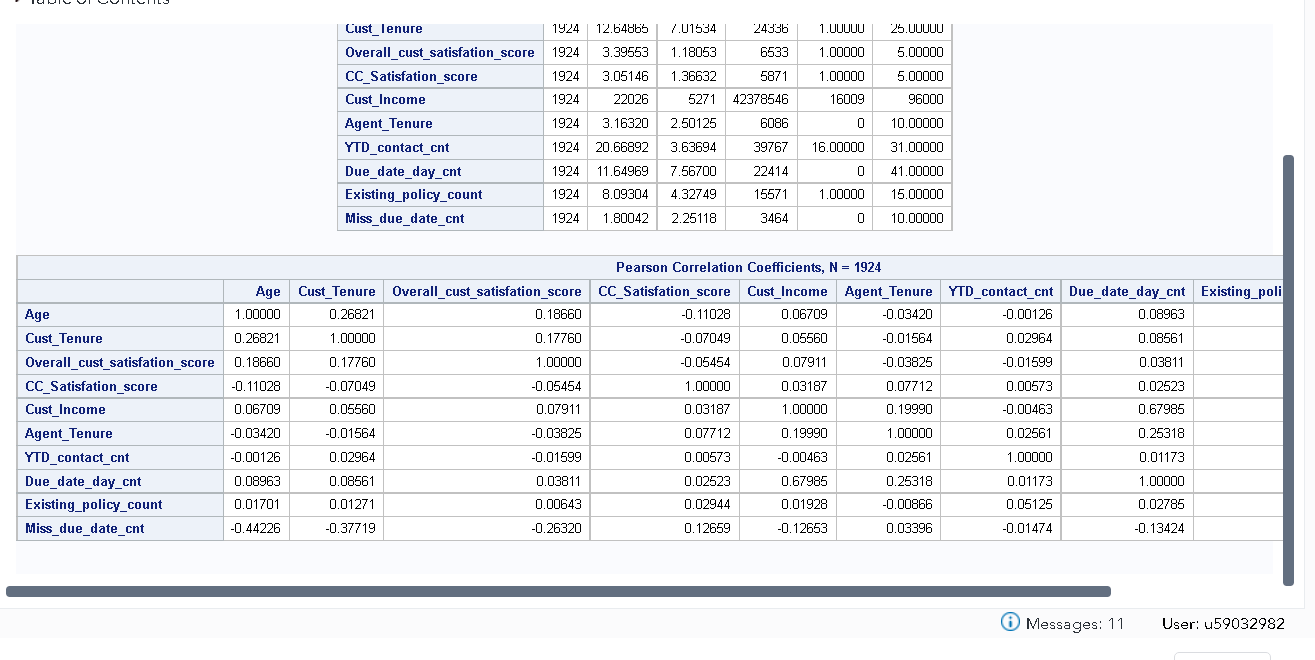
**Check correlation of all numerical variables before building model, because we cannot add correlated variables in model? (4 Marks)**

proc corr data=insurance NOPROB;

var Age cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income Agent\_Tenure

YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt;

run;

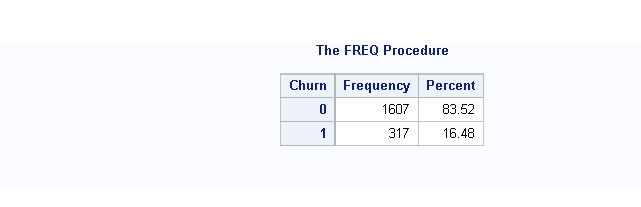
expected: no variables seem to be correlated

**Create train and test (70:30) dataset from the existing data set. Put seed 1234? (4 Marks)**

proc freq data=insurance;

table churn /nocum;

run;

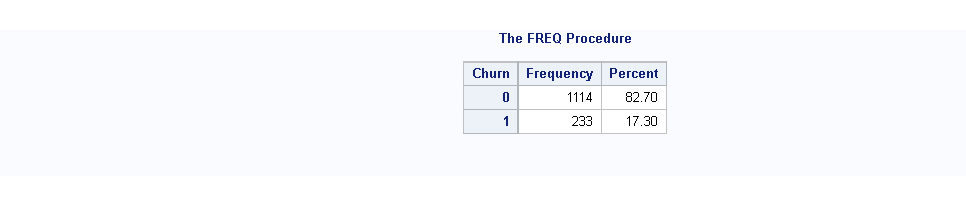


proc surveyselect data= insurance method = srs rat=0.7 seed = 1234 out =train\_set;

RUN;

proc freq data=train\_set;

table churn /nocum;

run;

proc sql;

create table test\_set as select t1.\* from insurance as t1

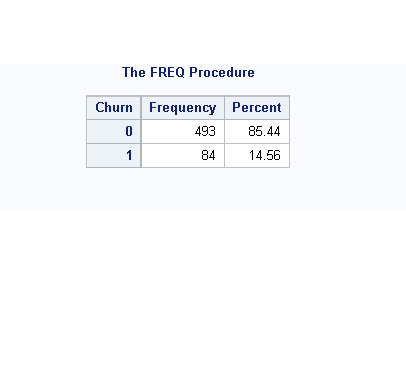
where custid not in (select custid from train\_set);

quit;

proc freq data=test\_set;

table churn /nocum;

run;

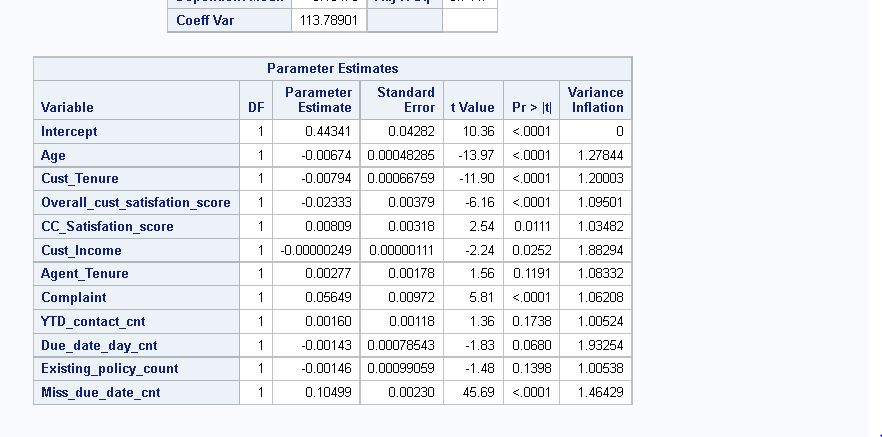


**Develop linear regression model first on the target variable to extract VIF information to check multicollinearity? (6 Marks)**

proc reg data=insurance;

model churn= Age cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income Agent\_Tenure complaint YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt/ vif;

run;



**Create clean logistic model on the target variables? (4 Marks)**

%let var = Age cust\_Tenure Overall\_cust\_satisfation\_score CC\_Satisfation\_score Cust\_Income Agent\_Tenure complaint YTD\_contact\_cnt Due\_date\_day\_cnt Existing\_policy\_count Miss\_due\_date\_cnt;

proc logistic data=train\_set descending outmodel=model;

model churn = &var / lackfit;

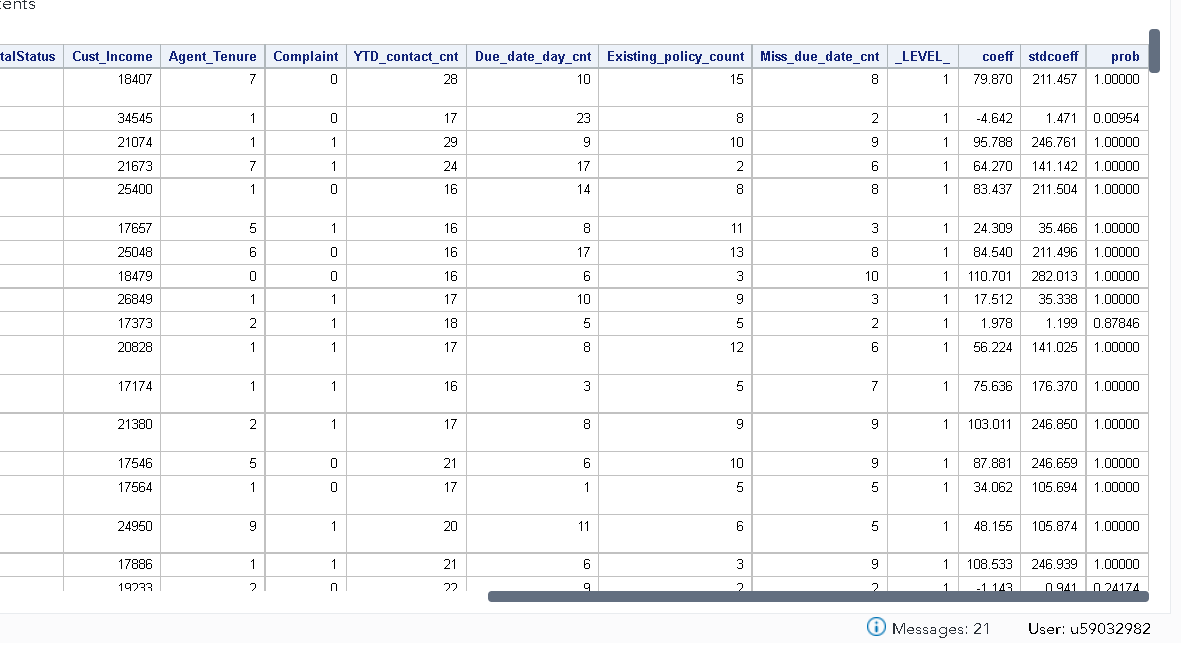
output out = train\_output xbeta = coeff stdxbeta = stdcoeff predicted = prob;

run;

proc print data=train\_output;

where churn=1;

run;



**12. Create a macro and take a KS approach to take a cut off on the calculated scores?**

**Predict test dataset using created model? (2 Marks)**

data test\_set;

set test\_set;

prob = 0.6509 - Age\*0.3712-Cust\_Tenure\*0.8291-Overall\_cust\_satisfation\_score\*2.1467+CC\_Satisfation\_score\*1.4308+ Cust\_Income\*0.000056 + agent\_Tenure\*0.1662+ Complaint\*5.6805-YTD\_contact\_cnt\*0.5071-Due\_date\_day\_cnt\*0.1113-Existing\_policy\_count\*0.3730+ Miss\_due\_date\_cnt\*13.8588;

score = exp(prob)/(1+exp(prob));

run;

proc print data=test\_set;

where churn=0;

run;

data train\_output;

set train\_output;

if prob>0.7 then churnpred = 1;

else churnpred = 0;

run;

data test\_set;

set test\_set;

if score>0.80 then churnpred = 1;

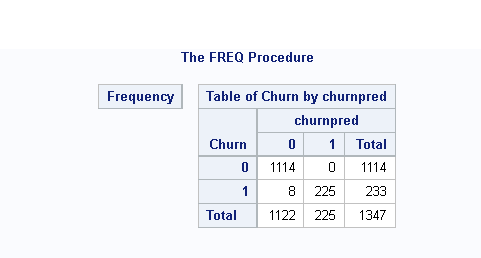
else churnpred = 0;

run;

proc freq data= train\_output;

table churn\*churnpred / nocol norow nopercent;

run;



proc freq data= test\_set;

table churn\*churnpred / nocol norow nopercent;

run;

