模型评估代码

libname r "D:\level2\book\data";

data cardraw;

set r.cardraw;

run;

proc sort data=cardraw out=r1

dupout=dup nodupkey;

by id;

run;

Proc freq过程步统计频次

proc freq data=r1

tables response;

run;

数据抽样代码

proc sort data=r1;

by response;

run;

proc surveyselect data=r1 out=r2 seed=1234 samprate=(0.01,1);

strata response;

run;

训练集合与验证集划分

proc surveyselect data=r2 out=r3 seed=456 samprate=0.6 outall;

run;

data card\_model card\_access;

set r3;

if selected=1then output card\_model;

else output card\_access;

drop selected SelectionProb SamplingWeight;

run;

连续变量间的相关系数

%let invar1=utilrate age balance confreq conp homein income points recent repay tmaddress tmbank;

proc corr data=card\_model pearson spearman rank;

var &invar1. ;

with response;

run;

计算分类变量间的相关系数

proc sort data=card\_model out=r1;

by descending response;

run;

%let inclass1=emp res otherp edu car lastact conab cardtyp;

proc freq data=r1 order=data;

tables response\*&inclass1./chisq expected measures;

run;

Proc test过程步

%let invar2=utilrate age balance confreq conp homein income points recent repay tmbank;

proct test data=card\_model;

class response;

var &invar2.;

run;

Proc anova方差检验

proc anova data=card\_model;

class res;

model income=res;

means res/hovtestwelch;

run;

侦测错误值和缺失值

%let inclass2=emp otherp edu lastact conab cardtyp;

\*错误值与缺失值侦测;

proc freq data=card\_model;

tables &inclass2./missing;

run;

错误值或缺失值甄别

proc means data=card\_model n nmiss max min mean median mode max dec=2;

var &invar2. ;

run;

错误值修正

\*错误修正;

proc print data=card\_model;

where age>80;

var id age;

run;

data card\_model;

set card\_model;

if id=222 then age=25;

if id =685 then age=49;

if id=686 then age=40;

run;

单值替换

\*单值替换;

data card\_model1;

set card\_model;

if income=. then income1=38867.65;

else income1=income;

run;

proc stdize data=card\_model out=card\_model1 outstat=inc\_miss

reponly sprefix=l\_ oprefix=m\_;

var income;

run;

类均值替换

proc means data=card\_model n nmiss max min mean median mode max dec=2;

var income homein;

class cardtyp;

run;

data card\_model2;

set card\_model1;

if income=.thendo;

if cardtyp=0 then m\_income=38753;

else m\_income=38409;

end;

else m\_income=income;

if homein=. then do;

if cardtyp=0 then m\_homein=48545;

else m\_homein=48246;

end;

else m\_homein=homein;

drop homein income;

run;

proc sort data=card\_model;

by cardtyp;

run;

proc stdize data=card\_model out=card\_model2 outstat=inc\_miss

reponly sprefix=m\_ method=median;

var income homein;

by cardtyp;

run;

回归替换

proc reg data=card\_model outest=r3;

miss\_reg: model income=utilrate age balance confreq conp

points recent repay tmbank/

selection=backward;

run;

proc score data=card\_model score=r3 out=card\_tem

type=parms predict;

var utilrate age balance confreq conp

points recent repay tmbank;

run;

data card\_model1;

set card\_tem;

if income=.then income1=miss\_reg;

else income1=income;

run;

异常值识别

proc means data=card\_model1 noprint;

var balance;

outputout=means(drop=\_type\_ \_freq\_)

mean=m\_bal

std=s\_bal;

run;

data w1;

set card\_model1(keep=id balance);

if \_n\_ = 1 then set means;

if balance lt m\_bal- 2\*s\_bal or

balance gt m\_bal + 2\*s\_bal ;

drop m\_bal s\_bal;

run;

极差计算

proc means data=card\_model1 noprint;

var balance;

output out=tmp q1=Lower q3=Upper range=Iqr;

run;

data w2;

set card\_model1(keep=id balance);

if \_n\_ = 1then set tmp;

if balance le Lower - 1.5\*Iqr or

balance ge Upper +1.5\*Iqr;

run;

快速聚类

%let invar3=utilrate age balance confreq conp m\_homein m\_income points recent repay tmbank;

proc fastclus data=card\_model1 maxc=8 maxiter=1 out=w3\_1 ;

var&invar3.;

run;

proc freq data=w3\_1;

table cluster;

run;

极端值的处理

%macro extre(indataset,invar,outvar,mapdata);

其中indataset为输入数据集

Invar为输入变量

Outvar是输出变量

Mapdata为离群值的maping关系

%extre(card\_model2,balance,balance\_e,extr\_balance);

%extre(card\_model2,m\_homein,m\_homein\_e,extr\_homein);

%extre(card\_model2,m\_income,m\_income\_e,extr\_income);

%extre(card\_model2,utilrate,utilrate\_e,extr\_utilrate);

%extre(card\_model2,age,age\_e,extr\_age);

%extre(card\_model2,confreq,confreq\_e,extr\_confreq);

%extre(card\_model2,conp,conp\_e,extr\_conp);

%extre(card\_model2,points,points\_e,extr\_points);

%extre(card\_model2,recent,recent\_e,extr\_recent);

%extre(card\_model2,repay,repay\_e,extr\_repay);

%extre(card\_model2,tmbank,tmbank\_e,extr\_tmbank);

保存宏到逻辑库

proc format lib=r;

invalue empf

"Unknown"=0

"Employed"=1

;

run;

options fmtsearch=(r work library);

data card\_model4;

set card\_model3;

emp\_f=input(emp,empf.);

drop emp;

run;

计算类别频数

proc freq data=card\_model5;

table response\*edu;

run;

计算WOE和IV

%macro CalcWOEIV(dsin, ivvar, dvvar, dsout);

%macro CalcWOEIV(dsin, ivvar, dvvar, newvar,woemap,dsout);

/\*\*dsin表示输入数据集;

/\*ivvar表示自变量\*/

/\*dvvar表示因变量\*/

newvar表示WOE转换后的变量名称

woemap表示ivvar变量的mapping关系

/\*dsout表示输出数据集\*/

计算edu的宏程序调用形式为：

data C1;

set card\_model4;

keep id response &inclass3.;

run;

%CalcWOEIV(C1, emp\_f, response,emp\_woe,woe\_emp,C2);

%CalcWOEIV(C2, otherp, response,otherp\_woe,woe\_otherp,C3);

%CalcWOEIV(C3, edu, response,edu\_woe,woe\_edu,C4);

多重共线性诊断

proc reg data=card\_model4;

model response=utilrate\_e age\_e balance\_e confreq\_e conp\_e m\_homein\_e m\_income\_e points\_e recent\_e repay\_e tmbank\_e/vif tol collin colli noint;

quit;

主成分分析

proc princomp data=card\_model4 out=r1;

var&invar4.;

run;

变量聚类

proc varclus data=card\_model4 maxeigen=0.5;

var&invar4.;

run;

变量最优分类

data w1;

set card\_model4;

keep id response &invar4.;

run;

以balance\_e为例

%BinContVar(w1, balance\_e, response, 1, 3, 0.01, bin\_balance);

%ApplyMap2(w1, balance\_e, balance\_e\_b, bin\_balance, w2);

%CalcWOEIV(w2,balance\_e\_b, response,balance\_woe,

woe\_balance,w3);

变量筛选

proc logistic data=card\_model5 ;

model response(event="1")=balance\_woe;

run;

proc logistic data=card\_model5desc;

model response=balance\_woe;

run;

Logistic建模

proc logistic data=card\_model5 des;

model response=balance\_woe cardtyp\_woe conab\_woe edu\_woe emp\_woe

income\_woe lastact\_woe otherp\_woe points\_woe homein\_woe age\_woe conp\_woe

tmbank\_woe/stbselection=stepwise slstay=0.01;

run;

模型打分-直接在Logistic过程步中打分

proc logistic data=card\_model5 des;

model response=balance\_woe cardtyp\_woe conab\_woe edu\_woe emp\_woe

income\_woe lastact\_woe otherp\_woe points\_woe homein\_woe age\_woe conp\_woe

tmbank\_woe/selection=backward slstay=0.01;

score data=card\_access5 out=r1;

run;

模型打分-单独打分

proc logistic data=card\_model5 des outtest=beta;

model response=balance\_woe cardtyp\_woe conab\_woe edu\_woe emp\_woe

income\_woe lastact\_woe otherp\_woe points\_woe homein\_woe age\_woe conp\_woe

tmbank\_woe/selection=backward slstay=0.01;

run;

proc score data=card\_access5 out=r1 score=beta

type=parms;

var balance\_woe cardtyp\_woe conab\_woe edu\_woe emp\_woe

income\_woe lastact\_woe otherp\_woe points\_woe homein\_woe age\_woe conp\_woe

tmbank\_woe;

run;

生成混淆矩阵

proc logistic data=card\_model5 des ;

model response=balance\_woe cardtyp\_woe edu\_woe emp\_woe

lastact\_woe otherp\_woe points\_woe homein\_woe age\_woe /ctablepprob=0.5;

run;

绘制ROC曲线

proc logistic data=card\_model5 desplots=(roc);

model response=balance\_woe cardtyp\_woe edu\_woe emp\_woe

lastact\_woe otherp\_woe points\_woe homein\_woe age\_woe ;

run;

proc logistic data=card\_model5 plots=roc(id=prob);

model response(event='1') = balance\_woe cardtyp\_woe edu\_woe / nofit;

roc'balance' balance\_woe;

roc'card\_type' cardtyp\_woe;

roc'edu' edu\_woe;

/\* roccontrast reference('K-G Score') / estimate e;\*/

run;

生成KS

proc npar1way data=r1 edf;

class response;

var p\_1;

run;

过度抽样调整方法1

proc logistic data=card\_model5 des;

model response=balance\_woe…tmbank\_woe/selection=forward slstay=0.01;

score data=card\_access5 out=r1 priorevent=;

run;

过度抽样调整方法2

data card\_model5;

set card\_model5;

off=log((1-p1)\*r1)/(p1\*(1-r1);

run;

proc logistic data=card\_model5 des outest=beta1;

model response=balance\_woe…tmbank\_woe/offset=off;

run;

proc score data=card\_access5 out=scored score=beta1

type=parms;

var balance\_woe…tmbank\_woe;

run;

data scored;

set scored;

p=1/(1+exp(-ins));

run;