## /\*深入解析sas-第18章-SAS数据挖掘的一般流程\*/

/\*creat a library\*/

libname ex 'D:\DATA ANALYSIS\sas documents\sas project\from wechat\Data&Code\Data';

**data** smbl;

set ex.smbl;

**run**;

/\*check the info of data set\*/

**proc** **contents** data=smbl;**run**;

/\* look at the first rows of data set\*/

**proc** **print** data=smbl(obs=**5**);**run**;

/\*drop-duplicate\*/

**proc** **sql**;

create table smbl1 as

select distinct\*

from smbl;

**quit**;

/\*check missing value and mean etc\*/

**proc** **means** data = ex.smbl N nmiss min mean median max std;

var \_numeric\_;

**run**;

/\*the situation of each var\*/

**%macro** FreqBar(ds, varname);/\*data set =ds,var=varname\*/

proc freq data=&ds;

tables &varname / plots(only)=freqplot;

run;

**%mend**;

%***FreqBar***(smbl, local)

%***FreqBar***(smbl,creditlevel )

%***FreqBar***(smbl, bad)

%***FreqBar***(smbl, education)

%***FreqBar***(smbl, indarea)

%***FreqBar***(smbl, reason)

/\*the 'reason' of missing value is higher \*/

/\*变量与目标变量的相关性\*/

/\*分类变量\*/

**proc** **freq** data = smbl;

tables (reason creditlevel education indarea local)\*bad/chisq nocol nopercent;/\*norow nocol nopercent：有时我们只需要频数，不需要各行各列的百分比\*/

**run**;

/\*'reason' have two styles , good style almost equal bad style,and missing value is higher,

so ,if create model,we will not think about it\*/

/\*数值变量\*/

**%macro** plottrend(ds, varname,y, obsingroup);

data temp; \*生成一个仅包含待分析变量的子数据集;

set &ds(keep = &varname &y.);

run;

proc sort data = temp out = tmp; \*对子数据集进行排序;

by &varname;

run;

data tmp; \*对排序完的子数据集进行分组标记;

set tmp;

group = ceil(\_N\_ /&obsingroup);

run;

data plot; \*根据分组标记，对每组内的观测求平均，将平均值存储为新的数据集plot;

set tmp;

by group;

if first.group then

sum =**0**;

if last.group then do;

avg = (sum/&obsingroup);

output;

end;

else sum + &y.;

run;

proc sgplot data = plot; \*画图;

title"&varname - Bad Trend";

series x= group y = avg / markers;

run;

**%mend**;

%***plottrend***(smbl, age,bad, **1200**)

%***plottrend***(smbl, yropen,bad, **600**)

%***plottrend***(smbl, revenue,bad, **600**)

%***plottrend***(smbl, rent,bad, **600**)

%***plottrend***(smbl, debtinc,bad, **600**)

%***plottrend***(smbl, delinq,bad, **600**)

%***plottrend***(smbl, profitrate,bad, **600**)

%***plottrend***(smbl, creditage,bad,**600**)

%***plottrend***(smbl, storearea,bad,**600**)

/\*数据加工\*/

/\*缺失值处理

连续型变量，指定默认值，平均数，中位数，最大值，最小值，不做处理；离散型变量，指定默认值，众数，不做处理

负债比和拖欠次数，一般来说是越大坏客户占比多，前几组坏账率高的原因是缺失值导致的，因此缺失值应修改为8~10组对应的值，

其他的缺失变量都修改为均值。\*/

**proc** **means** data = smbl n nmiss mean p90 ;

var debtinc delinq creditage age revenue;

**run**;

/\*负债比和拖欠次数取90分位数，信用史，年龄，营业额取均值\*/

**data** smbl;

set smbl;

if (debtinc = **.** )then debtinc = **41.78**;

if (delinq = **.**) then delinq = **2**;

if (creditage = **.**) then creditage =**6.9**;

if (age = **.**) then age = **39**;

if (revenue = **.**) then revenue = **100349**;

drop indarea local reason;

**run**;

/\*数据抽样\*/

/\*1、需要用到抽样的情况

①假设要对大规模数据(数亿条记录)集探索变量的分布，对原始变量进行操作是较为耗费时间，此时需要通过抽样来进行分析。

②当进行建模时，需要划分训练集、测试集，以验证模型的效果，此时需要用到抽样。

2、抽样的分类

①不放回简单抽样，按等概率原则，从总体N中抽取n个元素(不放回)。

②放回的无限制随机抽样，按等概率原则，从总体N中抽取n个元素(放回)。

③系统随机抽样，对数据集中的观测进行编号，1~N，假设样本容量为n，则抽样距离K = N/n，从1~K中随机选择i，接着选择i+K，i+2K,直到得到n个观测为止。

④逐次抽样，不断增加抽取的样本量，直到能够接受或拒绝原假设。

⑤伯努利抽样，等概率抽样，但每次的样本容量不固定。总体中一个个体是否被抽取，可以看做是一个伯努利试验。\*/

/\*拆分数据集\*/

/\*获取每个组的数量\*/

**proc** **sql** noprint; select count(ID) into: TotalObs from smbl ;**quit**;

%let train\_obs = %sysevalf(0.7\*&TotalObs); \*训练数据集观测数;

%let validate\_obs = %sysevalf(0.3\*&TotalObs); \*验证数据集观测数;

/\*开始抽样\*/

**proc** **surveyselect** data = smbl out = split seed = **9999**

group = (&train\_obs, &validate\_obs);

**run**;

/\*选项 \*/

/\*分割数据集\*/

**data** smbl\_train smbl\_validate ;

set split;

if GroupID =**1** then do;

drop GroupID;

output smbl\_train;

end;

else do;

drop GroupID;

output smbl\_validate;

end;

**run**;

/\*数据建模\*/

/\*因变量为0/1二分类，自变量为连续型变量和离散型变量，因此用逻辑回归进行建模。\*/

**proc** **logistic** data = smbl\_train desc ;

class creditlevel ;

model bad = creditlevel delinq debtinc profitrate revenue yropen /selection = none;

**run**;

/\*creditlevel2,pr 0.4334,对模型的贡献不显著，因些使用outdesign选项将信用等级为2的水平

排除在模型外\*/

**proc** **logistic** data = smbl\_train desc outdesign = work.design;

class creditlevel ;

model bad = creditlevel delinq debtinc profitrate revenue yropen /selection = none;

**run**;

/\*将删除分类变量信用等级，替换为信用等级1和信用等级3，等于剔除了原信用等级为2的水平\*/

**proc** **logistic** data = work.design desc outmodel = work.estimate1 ; /\*outmodel输出模型参数\*/

model bad = creditlevel1 creditlevel3 delinq debtinc profitrate revenue yropen /selection = none;

score data= work.design out=scores1; /\*score输出评分数据集\*/

**run**;

/\*模型和参数都是显著了\*/

/\*模型评估\*/

/\*对测试集的信用等级进行同样的转换\*/

**proc** **logistic** data = work.smbl\_validate desc outdesign = work.design2;

class creditlevel ;

model bad = creditlevel delinq debtinc profitrate revenue yropen/selection = none;

**run**;

/\*调用模型，对测试集进行评分\*/

**proc** **logistic** inmodel=work.estimate1 ;

score data= work.design2 out=work.scores2;

**run**;

/\*模型分类效果\*/

**proc** **freq** data = work.scores1;

title ‘data set smbl\_train\_impute prediction of classifiction situdation’;

table F\_bad\*I\_bad;

**run**;

**proc** **freq** data = work.scores2;

title ‘data set smbl\_validate \_impute prediction of classifiction situdation’;

table F\_bad\*I\_bad;

**run**;

/\*通过交叉表可以看到训练集和测试集中，分类错误的占比。\*/

/\*使用lift图评估模型\*/

/\*训练集\*/

/\*排序\*/

**proc** **sort** data = work.scores1 out = work.sorted\_s1; by descending P\_1; **run**;

/\*分组\*/

**data** work.temp;

set work.sorted\_s1;

group = ceil(\_N\_ /**420**);

**run**;

/\*根据分组标记，对每组内的观测求p值-坏概率;\*/

**data** work.plot;

set work.temp;

by group;

if first.group then sum =**0**;

if last.group then

do;

avg = (sum/**420**);

p = (**100**\*avg)/(**859**/**4200**);

output;

end;

else sum + bad;

**run**;

/\*画图\*/

**proc** **sgplot** data = work.plot;

title"testing data set of LIFT";

series x= group y = p / markers;

**run**;

/\*测试集\*/

/\*排序\*/

**proc** **sort** data = work.scores2 out = work.sorted\_s2; by descending P\_1; **run**;

/\*分组\*/

**data** temp;

set work.sorted\_s2;

group = ceil(\_N\_ /**180**);

**run**;

/\*根据分组标记，对每组内的观测求p值-坏概率;\*/

**data** plot;

set temp;

by group;

if first.group then sum =**0**;

if last.group then

do;

avg = (sum/**180**);

p = (**100**\*avg)/(**336**/**1800**);

output;

end;

else sum + 是否不良;

**run**;

**proc** **sgplot** data = plot; \*画图;

title"testing data set LIFT";

series x= group y = p / markers;

**run**;

/\*结论，模型从左到右，呈下降趋势，故随着p值的减小，坏客户的占比也较小，使用模型比不使用模型效果好；训练集和测试集的LIFT图趋势一致，

形状类似，说明模型同样适用于测试集\*/

/\*模型实施

根据模型中参数的最大似然估计分析，可以得到模型的数学表达式。

ln(p/(1-p)) = -2.3688 + 信用等级1 \* -1.1946 + 信用等级3 \* 0.2384 + 拖欠次数 \* 0.4188 + 负债比 \* 0.1384+ 资产收益率 \* -1.3416 + 营业额 \* -0.00003 + 经营时长 \* -0.2255

信用等级1的取值（信用等级为1,取1；信用等级为4，取-1；其他信用等级取0）

信用等级3的取值（信用等级为2,取1；信用等级为4，取-1；其他信用等级取0）

公式中，p为是否不良取1时的概率，可以根据p的值，决定是否发放贷款。\*/