随机森林

代码清单2-1 模型训练代码

\*指定模型训练后打分代码存放的目录;

%let pth=C:\Users;

\*以下模型训练;

%macro rftrain(indat,p);

proc contents noprint data=&indat. out=train\_data\_name;

run;

data train\_data\_name;

set train\_data\_name;

where type=1 and upcase(compress(name)) not in ( 'CSR\_ID','TARGET') ;

keep name;

run;

%do i=1 %to &p.;

\*每次随机筛选50个变量;

proc surveyselect data=train\_data\_name out=var\_01 sampsize=50;quit;

proc sql ;

select distinct name into:var separated by ' ' from var\_01;

quit;

\*每次随机筛选20%的观测;

data step01;

set &indat.;

x=ranuni(0);

if x<=0.2;

run;

\*构建决策树, criterion=entropy指定信息熵作为决策树分割的依据，可以选择信息增益、Gini系数等其他指标;

Proc split data=step01 outleaf=leaf

outimportance=importance outtree=tree outmatrix=matrix outseq=seq

criterion=entropy

assess=impurity

maxbranch=3

maxdepth=5

exhaustive=100

leafsize=30

splitsize=30

subtree=assessment;

code file="&pth.\sas\_rule&i..txt";

describe file="&pth.\rulefinal&i..txt";

input &var./ level=interval;

target target/level=binary;

run;

%end;

%mend;

\* fraud\_train\_samp是建模数据，训练1000棵决策树;

%rftrain(fraud\_train\_samp1,1000);

代码清单2-2 模型打分代码

%macro rfscore(indat,p,outdat);

%do i=1 %to &p.;

data score\_&i.;

set &indat.;

%include "&pth.\sas\_rule&i..txt";

p\_&i.=p\_target1;

keep csr\_id p\_&i.;

run;

proc sort data=score\_&i.;

by csr\_id;

run;

%end;

proc sort data=&indat.(keep=csr\_id target) out=tmp1;

by csr\_id;

run;

data &outdat.;

merge tmp1 %do i=1 %to &p.;score\_&i. %end;;

by csr\_id;

pr=sum(of p\_1-p\_&p.)/&p.;

keep csr\_id target pr;

run;

%mend;

\*对验证数据打分，结果输出到表score\_valdt;

%rfscore(fraud\_valdt\_samp1,1000,score\_valdt);

代码清单2-3 模型评估代码

%macro Fit(in,out,grp\_cnt,pred\_var,act\_var);

data work.tt1;

set &in;

run;

data \_null\_;

set work.tt1 nobs=obs ;

call symput("Base",obs/&grp\_cnt);

stop;

run;

proc sort data=work.tt1;

by descending &pred\_var;

run;

data work.tt1;

N=\_N\_;

set work.tt1;

format Grp2 4.0;

Grp2=INT((N-1)/&base);

run;

proc means data=work.tt1 nway noprint;

class Grp2 ;

output out=&out mean(&pred\_var &act\_var)=pred\_evt actual\_evt;

run;

%mend;

\*fit\_valdt就是表2-5前五列;

%Fit(score\_valdt,fit\_valdt,20,pr,target);