



**题目: 银行理财产品客户定位分析**

**姓 名 杜泽宣**

**学 院 理学院**

**专 业 数学与应用数学**

**班 级 2021214101**

**学 号 2021212487**

**指导教师 徐文哲**

# **2024年 4月**

**摘 要**

本论文针对银行理财产品的客户定位问题进行了深入研究。随着当前经济环境的复杂变化，某商业银行为了更有效地推广理财产品并减少客户消极反应，通过精准定位潜在客户群体，采用了2009年至2016年间来自UCI数据库的部分实际客户数据（共54个特征变量和1个二分类标签，样本量达到354293条）。这些数据涵盖了客户的个人信息（如性别、婚姻状况、教育程度、年龄、子女数量等）、财务状况（如资产管理规模、基金投资、借记卡消费记录、存款情况、理财产品投资额度等）以及客户在银行的各项服务使用情况。

研究过程中首先将数据集划分为训练集和测试集，采用随机抽样的方式确保数据分布合理性，并根据目标变量“是否购买理财产品”对数据进行平衡处理，保证不同类别样本的相对比例适中。随后对数据进行了预处理，包括删除冗余记录、填补缺失值和处理异常值。对于缺失的连续性变量，采取中位数填充法；分类变量则采用众数填充。此外，对明显无统计意义的变量予以剔除，并对极端离群值运用3σ原则进行识别和修正，用中位数替换超出3σ范围的数据。

通过SAS工具对数据进行深入挖掘和分析，提取出有助于市场决策的关键信息，并构建模型预测新用户是否会购买理财产品，从而帮助银行实现更为精准的市场营销策略和资源投放。整个研究过程充分体现了对数据分析方法的综合运用，以及在解决现实金融问题中的实践价值。

**关键词** ：客户定位、数据分析、SAS、变量筛选、缺失值处理、模型预测、分类算法、客户行为分析、定量研究、投资偏好、历史交易信息

1. **引言**

由于目前经济形势不容乐观，为推销自己的理财产品，精确瞄准目标客户进行推荐，减少储户的负面感受，某商业银行进行客户定位分析，以求较为精准的筛选出目标客户。故选取了部分（2009-2016）储户数据进行分析。基于SAS Base分析,提取有价值信息供市场决策以及预测哪些新用户会购买该理财产品。

# **第二章 数据集介绍**

**模型需求**

数据集（数据来源 UCI） 2009-2016年商业银行部分客户的历史数据（54个自变量，1个二分类因变量，354293条数据），主要包括用户基本信息，资产状况，历史交易信息，是否购买了该理财产品等信息，数据集来源于UCI,为2009-2016年某商业银行部分客户数据,主要包括用户基本信息，资产状况，历史交易信息，是否购买了该理财产品等信息,根据现有信息，基于SAS Base分析，提取有价值信息供市场决策以及预测哪些新用户会购买该理财产品。

**变量说明**

现对变量进行说明

CUST\_ID 客户ID

GENDER 性别

MARR 婚姻状况

EDUCATION 教育程度

AGE 年龄

CHILDREN 子女数量

F\_VIP 是否是VIP客户

F\_STAFF 是否是员工

F\_PAYROLL 是否有工资代发

F\_YLJ 是否有养老金

F\_CC 是否有信用卡

F\_WEB 是否使用网银

F\_MOBILE 是否使用手机银行

F\_TEL 是否使用电话银行

F\_FUND 是否有基金

DOB 出生日期

AUM\_3 3个月的资产管理规模

FUND\_3 3个月的基金投资额

DEBIT\_3 3个月的借记卡消费额

AUM\_6 6个月的资产管理规模

FUND\_6 6个月的基金投资额

DEBIT\_6 6个月的借记卡消费额

DEPOSIT\_3 3个月的存款额

DEPOSIT\_6 6个月的存款额

FIX\_3 3个月的定期存款额

FIX\_6 6个月的定期存款额

FINACE\_3 3个月的理财产品投资额

FINACE\_6 6个月的理财产品投资额

YJL\_3 3个月的银行业务收入

YJL\_6 6个月的银行业务收入

PAYROLL\_3 3个月的工资代发额

PAYROLL\_6 6个月的工资代发额

C\_1W\_D\_3 3个月内的每周借记卡消费天数

C\_1W\_D\_6 6个月内的每周借记卡消费天数

C\_1W\_TR\_3 3个月内的每周借记卡交易次数

C\_1W\_TR\_6 6个月内的每周借记卡交易次数

C\_FIX\_3 3个月内的定期存款次数

C\_FIX\_6 6个月内的定期存款次数

C\_FUND\_3 3个月内的基金交易次数

C\_FUND\_6 6个月内的基金交易次数

C\_YJL\_3 3个月内的银行业务收入次数

C\_YJL\_6 6个月内的银行业务收入次数

C\_DEBIT\_3 3个月内的借记卡交易次数

C\_DEBIT\_6 6个月内的借记卡交易次数

C\_FIANCE\_3 3个月内的理财产品投资次数

C\_FIANCE\_6 6个月内的理财产品投资次数

GAP\_FINACE\_3 3个月内理财产品投资额与平均投资额的差值

GAP\_FIANCE\_6 6个月内理财产品投资额与平均投资额的差值

DT\_L\_FINACE 理财产品最后一次投资距今的天数

CHANNEL\_PRE 上次交易渠道

A\_L\_FIANCE 累计理财产品投资额

TARGET 是否购买理财产品 1购，0未购买

# **第三章 运行结果及分析**

##### 3.1 数据概览

# **3.1.1 导入数据**

上传数据集至sas服务器，使用proc import语句导入

1. proc import out=finance
2. datafile="/home/u63802491/1(UTF-8).csv" dbms=csv replace;
3. run;

# **3.1.2 查看数据**

查看数据情况

1. proc contents data=finance;
2. run;

结果：[3.1.2-1.html](3.1.2.html)

可以看出数据集有大量缺失，同时存在多组重复数据，需要调整。

##### 3.2数据清洗

筛选数值/字符变量

1. */\*数值变量\*/*
2. %let var=AGE AUM\_3 AUM\_6 A\_L\_FIANCE CHILDREN CUST\_ID C\_1W\_D\_3 C\_1W\_D\_6  C\_1W\_TR\_3 C\_1W\_TR\_6 C\_FIANCE\_6 C\_FIX\_3 C\_FIX\_6 C\_FUND\_3 C\_FUND\_6   DEBIT\_3 DEBIT\_6 DEPOSIT\_3 DEPOSIT\_6 DOB DT\_L\_FINACE FINACE\_3 FINACE\_6  FIX\_3 FIX\_6 FUND\_3 FUND\_6 GAP\_FIANCE\_6 PAYROLL\_3 PAYROLL\_6 TARGET  YJL\_6;
4. */\*字符变量\*/*
5. %let char=CHANNEL\_PRE C\_DEBIT\_3 C\_DEBIT\_6 C\_FIANCE\_3 C\_YJL\_3 C\_YJL\_6   EDUCATION F\_CC F\_CLOAN F\_FUND F\_HLOAN F\_MOBILE F\_PAYROLL F\_STAFF  F\_TEL  F\_VIP F\_WEB F\_YJL F\_YLJ GAP\_FINACE\_3 GENDER MARR YJL\_3;

统计数据

1. \*\*统计数值型变量\*\*;
2. proc means data=finance n nmiss mean median min max; */\* 在输出中包含观测数 (n)、缺失值数量 (nmiss)、均值 (mean)、中位数 (median)、最小值 (min)、最大值 (max) \*/*
3. var &var; */\* 使用 VAR 语句指定要进行统计分析的变量 \*/*
4. run;
5. \*\*统计字符型变量\*\*;
6. proc freq data=finance;
7. table &char */\* 使用 TABLE 语句指定要进行频率分析的字符型变量 \*/*
8. /plots(only)=freqplot; */\* 使用 PLOTS 选项生成频率图 \*/*
9. run;

结果：<3.1.2-2(1).html> <3.1.2-2(2).html>

发现1.目标变量观测严重不均

2.存在大量分类变量

3.存在重复变量

删除重复行，调整数据变量类型

1. */\*去重\*/*
2. proc SQL;
3. create table fin\_dis as
4. select distinct \* from finance order by CUST\_ID;
5. quit;
6. */\*数据清洗\*/*
7. proc sql;
8. delete from finance  where CUST\_ID not in (select min(CUST\_ID) from finance group by \_numerie\_, \_character\_ );
9. run;

平衡数据集

1. */\*根据因变量target 0 1 占比判断数据是否均衡\*/*
2. proc freq data=fin\_dis;
3. table target
4. /missing;
5. run;
6. */\*0/1 占比34:1 数据不均衡,平衡数据集\*/*
7. data fin\_1 fin\_0;
8. set fin\_dis;
9. if target=1 then output fin\_1;
10. else output fin\_0;
11. run;

结果<3.1.2-3.html>

1. */\*保留fin\_1中所有数据 抽样使1 0占比 1:3 \*/*
2. */\*随机抽样\*/*
3. proc surveyselect data=fin\_0
4. out=fin\_slt\_0 method=srs seed=12345  n=22228;
5. run;
6. */\*合成均衡数据集\*/*
7. data fin\_bal;
8. set fin\_slt\_0 fin\_1;
9. run;

结果：<3.1.2-4.html>

1. */\*查看数据缺失情况，数据缺失量过多有可能是由数据不均衡导致, 要先抽取均衡数据集\*/*
2. proc means data=fin\_bal N Nmiss min max;
3. var &var;
4. run;
5. proc freq data=fin\_bal;
6. table &char
7. /missing;
8. run;
9. \*筛除缺失占比达85%以上的变量;
10. %let miss=C\_DEBIT\_3 C\_DEBIT\_6 C\_FUND\_3 C\_YJL\_3 C\_FUND\_6 YJL\_3 PAYROLL\_3
11. C\_1W\_D\_3 C\_YJL\_6 C\_FIX\_3 YJL\_6 PAYROLL\_6 EDUCATION C\_1W\_D\_6 C\_FIX\_6 MARR;

结果：<3.1.2-5.html>

现在是一个平衡数据集。

定义数值/字符变量集

1. */\*除客户ID, 因变量TARGET外, 需要筛选的变量还有： 39-2=37  \*/*
2. */\*数值型变量：21 \*/*
3. %let var1=C\_1W\_TR\_3  FIX\_3  FIX\_6  FINACE\_3  C\_FIANCE\_6  GAP\_FIANCE\_6  FINACE\_6 C\_1W\_TR\_6  A\_L\_FIANCE  DT\_L\_FINACE  DEPOSIT\_3  DEPOSIT\_6  AGE  AUM\_3 DEBIT\_3  DEBIT\_6  FUND\_3  AUM\_6  FUND\_6  CHILDREN  DOB ;
4. */\*字符型变量: 16 \*/*
5. %let char1=F\_CC  F\_CLOAN  F\_FUND  F\_HLOAN  F\_MOBILE  F\_PAYROLL   F\_STAFF  F\_TEL  F\_VIP F\_WEB  F\_YJL  F\_YLJ  CHANNEL\_PRE  C\_FIANCE\_3   GAP\_FINACE\_3  GENDER;

将分类字符变量转化为分类数值变量

1. */\*使用数组,循环, 对字符型变量重编码\*/*
2. %macro recod(in\_data,out\_data);
3. data &out\_data(drop=i);
4. set &in\_data;
5. array cha{\*}\_character\_;
6. do i=1 to dim(cha);
7. if cha{i}='YES' then cha{i}=1;
8. else if cha{i}='NO' then cha{i}=0;
9. else if cha{i}='男性' then cha{i}=1;
10. else if cha{i}='女性' then cha{i}=0;
11. else if cha{i}='未知' then cha{i}='.';
12. else if cha{i}='柜面' then cha{i}=0;
13. else if cha{i}='网银' then cha{i}=1;
14. else if cha{i}='手机银行' then cha{i}=2;
15. else if cha{i}=' ' then cha{i}='.';
16. end;
17. run;
18. %mend;
19. %recod(fin\_bal,fin\_bal\_rec);
20. */\* 将字符型变量转换为数值型变量 \*/*
21. data fin\_bal\_REC;
22. set fin\_bal\_REC;
23. F\_CC\_n = input(F\_CC, best12.); */\* 转换格式 \*/*
24. drop F\_CC; */\* 删除原字符型变量 \*/*
25. run;
26. */\*...\*/*
27. data fin\_bal\_REC;
28. set fin\_bal\_REC;
29. F\_CLOAN\_n = input(F\_CLOAN, best12.); */\* 转换格式 \*/*
30. drop F\_CLOAN; */\* 删除原字符型变量 \*/*
31. run;
32. data fin\_bal\_REC;
33. set fin\_bal\_REC(rename=(F\_CC\_n=F\_CC));
34. */\*...\*/*
35. set fin\_bal\_REC(rename=(GENDER\_n=GENDER));
36. run;

检查数据集fin\_bal\_REC发现需要的变量已经改为数值型。

定义连续/分类变量

1. */\*连续变量\*/*
2. %let var\_list1=C\_1W\_TR\_3  FIX\_3  FIX\_6  FINACE\_3  C\_FIANCE\_6   GAP\_FIANCE\_6  FINACE\_6 C\_1W\_TR\_6  A\_L\_FIANCE  DT\_L\_FINACE  DEPOSIT\_3 DEPOSIT\_6  AGE  AUM\_3 DEBIT\_3  DEBIT\_6 FUND\_3  AUM\_6  FUND\_6   CHILDREN  DOB   C\_FIANCE\_3  GAP\_FINACE\_3;
3. */\*分类变量\*/*
4. %let cla\_list1=F\_CC  F\_CLOAN   F\_FUND   F\_HLOAN  F\_MOBILE   F\_PAYROLL F\_STAFF   F\_TEL   F\_VIP F\_WEB  F\_YJL  F\_YLJ  CHANNEL\_PRE  GENDER;

抽取训练集/测试集

1. */\*抽取训练集、测试集\*/*
2. data fin\_train\_rec fin\_valid\_rec;
3. set fin\_bal\_REC;
4. ran=ranuni(12345);
5. if ran<0.7 then output fin\_train\_rec;
6. else output fin\_valid\_rec;
7. drop ran &miss;
8. run;

检查数据

1. */\* 使用 PROC CONTENTS 过程查看数据集的结构和属性 \*/*
2. proc contents data=fin\_train\_rec;
3. run;
4. proc contents data=fin\_valid\_rec;
5. run;

结果：<3.1.2-6.html>

删除除ID不同外，其他信息均相同的冗余信息后，剩余352487条数据.

抽取均衡数据集，随机抽样，保留target=1中所有数据（11114），抽样target=0 （33342）使1：0占比1:3， 得到新的数据集包含44456条观测值，55个变量.

划分训练数据集和验证数据集，训练数据：验证数据 =7:3

填补缺失值，异常值删除或替换

连续变量缺失值填补：中位数填补

分类变量缺失值填补：众数填补。

1. /\*尝试补齐变量\*/
2. proc means data=fin\_train\_rec n nmiss mode mean median min  max std;
3. var &var\_list1;
4. run;
5. %macro miss(in\_data,out\_data);
6. data &out\_data;
7. set &in\_data;
8. if C\_1W\_TR\_3=.    then  C\_1W\_TR\_3 =1.0000000 ;
9. if FIX\_3=.        then  FIX\_3   =70000.00 ;
10. if FIX\_6=.        then  FIX\_6   =60000.00 ;
11. if FINACE\_3=.     then  FINACE\_3   =74725.27 ;
12. if C\_FIANCE\_6=.   then  C\_FIANCE\_6   =2.0000000 ;
13. if GAP\_FIANCE\_6=. then  GAP\_FIANCE\_6  =91.0000000 ;
14. if FINACE\_6=.     then  FINACE\_6   =69617.49 ;
15. if C\_1W\_TR\_6=.    then  C\_1W\_TR\_6   =1.0000000 ;
16. if A\_L\_FIANCE=.   then  A\_L\_FIANCE  =100000.00 ;
17. if DT\_L\_FINACE=.  then  DT\_L\_FINACE  =20171.00 ;
18. if DEPOSIT\_3=.    then  DEPOSIT\_3   =9368.34 ;
19. if DEPOSIT\_6=.    then  DEPOSIT\_6   =10863.97 ;
20. if AGE=.          then  AGE    =59.0000000 ;
21. if AUM\_3=.        then  AUM\_3    =22193.60 ;
22. if DEBIT\_3=.      then  DEBIT\_3  =0 ;
23. if DEBIT\_6=.      then  DEBIT\_6   =0 ;
24. if FUND\_3=.       then  FUND\_3     =0 ;
25. if AUM\_6=.        then  AUM\_6     =17229.62 ;
26. if FUND\_6=.       then  FUND\_6     =0 ;
27. if CHILDREN=.     then  CHILDREN     =0 ;
28. if DOB=.          then  DOB     =3425.00 ;
29. if C\_FIANCE\_3=.   then  C\_FIANCE\_3    =1.0000000 ;
30. if GAP\_FINACE\_3=. then  GAP\_FINACE\_3  =9.0000000 ;
31. if CHANNEL\_PRE=.  then CHANNEL\_PRE     =3;
32. if GENDER=.       then GENDER    =0;
33. run;
34. %mend;
35. %miss(fin\_train\_rec, train\_nomiss);
36. %miss(fin\_valid\_rec, valid\_nomiss);

检查数据集train\_nomiss

1. */\*分析数据集 train\_nomiss\*/*
2. proc means data=train\_nomiss n nmiss mode mean median min  max std;
3. var &var\_list1 ;
4. run;

结果：[3.1.2-7.html](file:///F:\Desktop\sas\3.1.2-7.html)

我们可以明显的看出CHILDREN变量没有统计意义，故剔除。

同时某些数据过分偏离均值，对离群值进行替换：超出3sigma范围的数据视为离群值，使用中位数替换；

使用去除量纲

1. */\*异常值处理 3sigma\*/*
2. proc means data=train\_nomiss n nmiss mode mean median min p1 p95 p99 max std;
3. var &var\_list1;
4. run;
5. %macro normal(in\_data,out\_data);
6. data &out\_data;
7. set &in\_data;
8. C\_1W\_TR\_3=max(min(C\_1W\_TR\_3, 2.0000000), 1.0000000);
9. FIX\_3=max(min(FIX\_3,  500000.00),4709.01);
10. FIX\_6=max(min(FIX\_6, 500811.16),2837.16);
11. FINACE\_3=max(min(FINACE\_3,701538.46),7142.86);
12. C\_FIANCE\_6=max(min(C\_FIANCE\_6,23.0000000),1.0000000);
13. GAP\_FIANCE\_6=max(min(GAP\_FIANCE\_6,183.0000000),7.0000000);
14. FINACE\_6=max(min(FINACE\_6,826775.96),6885.25);
15. C\_1W\_TR\_6=max(min(C\_1W\_TR\_6,3.0000000),1.0000000);
16. A\_L\_FIANCE=max(min(A\_L\_FIANCE,1200000.00),48200.00);
17. DT\_L\_FINACE=max(min(DT\_L\_FINACE,20632.00),17734.00);
18. DEPOSIT\_3=max(min(DEPOSIT\_3,624592.67),0.1800000);
19. DEPOSIT\_6=max(min(DEPOSIT\_6,600228.74),0.2500000);
20. AGE=max(min(AGE,87.0000000),24.0000000);
21. AUM\_3=max(min(AUM\_3,1521085.68),0);
22. DEBIT\_3=max(min(DEBIT\_3,350000.00),0);
23. DEBIT\_6=max(min(DEBIT\_6,266666.67),0);
24. FUND\_3=max(min(FUND\_3,322893.50),0);
25. AUM\_6=max(min(AUM\_6,1032059.36),0);
26. FUND\_6=max(min(FUND\_6,228829.39),0);
27. CHILDREN=max(min(CHILDREN,0),0);
28. DOB=max(min(DOB,5939.00),313.0000000);
29. C\_FIANCE\_3=max(min(C\_FIANCE\_3,5.0000000),1.0000000);
30. GAP\_FINACE\_3=max(min(GAP\_FINACE\_3,9.0000000),1.0000000);
31. run;
32. %mend;
34. %normal(train\_nomiss,train\_nomal);
35. %normal(valid\_nomiss,valid\_nomal);
36. proc means data=train\_nomal n nmiss mode mean median min p1 p99 max std;
37. var &var\_list1 ;
38. run;

结果：<3.1.2-8.html>

这样就减少了许多偏差。

查看数据分布，标准化数据

1. proc means data=train\_nomal n nmiss mean median min p1 p99 max std;
2. var &var\_list1;
3. run;
4. /\*直方图\*/
5. %macro plt(var\_name);
6. proc sgplot data=train\_nomal;
7. histogram &var\_name;
8. density &var\_name;
9. density &var\_name
10. /type=kernel;
11. run;
12. %mend;
13. %plt(AUM\_6);

结果<3.1.2-9.html>

可以看出数据右偏严重，对连续变量做log变换，对右偏的连续变量做sqrt变换

1. */\*对右偏的连续变量做log变换\*/*
2. %let log=FIX\_3 FIX\_6 FINACE\_3  FINACE\_6
3. A\_L\_FIANCE DT\_L\_FINACE DEPOSIT\_3 DEPOSIT\_6 AGE AUM\_3
4. DEBIT\_3 DEBIT\_6 FUND\_3 AUM\_6 FUND\_6 DOB;
5. */\*对右偏的计数型变量做 sqrt 变换 \*/*
6. %let sqrt=C\_1W\_TR\_3 C\_FIANCE\_6 C\_1W\_TR\_6  CHILDREN C\_FIANCE\_3 GAP\_FINACE\_3 GAP\_FIANCE\_6;
7. %macro std(in\_data,out\_data);
8. data &out\_data(drop=i j);
9. set &in\_data;
10. array var1{\*} &log;
11. array var2{\*} &sqrt;
12. do i=1 to dim(var1);
13. if var1{i} not in ( .,0) then var1{i}=log10(var1{i});
14. end;
15. do j=1 to dim(var2);
16. if var2{j} ^= . then var2{j}=sqrt(var2{j});
17. end;
18. run;
19. %mend;
20. %std(train\_nomal,train\_std);
21. %std(valid\_nomal,valid\_std);

检查数据集

1. proc means data=train\_std n nmiss mode mean median min p1 p95 p99 max std;
2. var &var\_list1;
3. run;
4. %plt(C\_1W\_TR\_3);
5. %plt(FIX\_6);
6. %plt(FINACE\_3);
7. %plt(C\_FIANCE\_6);
8. %plt(GAP\_FIANCE\_6);
9. %plt(FINACE\_6);
10. %plt(C\_1W\_TR\_6);
11. %plt(A\_L\_FIANCE);
12. %plt(DT\_L\_FINACE);
13. %plt(DEPOSIT\_3);
14. %plt(DEPOSIT\_6);
15. %plt(AGE);
16. %plt(AUM\_3);
17. %plt(DEBIT\_3);
18. %plt(DEBIT\_6);
19. %plt(FUND\_3);
20. %plt(AUM\_6);
21. %plt(FUND\_6);
22. %plt(DOB);
23. %plt(C\_FIANCE\_3);
24. %plt(GAP\_FINACE\_3);

结果：[3.1.2-10（1）.html](3.1.2-10(1).html) <3.1.2-2(2).html>

筛选变量

1. */\*需要筛选的变量\*/*
2. */\*连续变量 22 \*/*
3. %let var\_list1=FIX\_3  FIX\_6  FINACE\_3  FINACE\_6  C\_FIANCE\_3   C\_FIANCE\_6  GAP\_FINACE\_3  GAP\_FIANCE\_6 C\_1W\_TR\_3  C\_1W\_TR\_6   DEPOSIT\_3  DEPOSIT\_6  AUM\_3  AUM\_6  DEBIT\_3  DEBIT\_6  FUND\_3   FUND\_6 AGE  DOB  A\_L\_FIANCE  DT\_L\_FINACE;
4. */\*分类变量 14 \*/*
5. %let cla\_list1=F\_CC  F\_CLOAN  F\_FUND  F\_HLOAN  F\_MOBILE  F\_PAYROLL   F\_STAFF  F\_TEL  F\_VIP F\_WEB  F\_YJL  F\_YLJ  CHANNEL\_PRE  GENDER;

进行相关性分析，画热力图

1. */\*根据共线性筛除变量\*/*
2. */\*相关分析\*/*
3. */\*连续与分类间相关性\*/*
4. proc corr data=train\_std spearman;
5. var target &cla\_list1;
6. with &var\_list1;
7. run;
8. */\* 将结果输出到CSV文件，以便于在excel中绘制热力图 只能本地运行 \*/*
9. ods csv file="corr\_spearman.csv";
10. proc print data=\_LAST\_ (obs=max);
11. run;
12. ods csv close;
13. */\*连续与连续间相关性\*/*
14. proc corr data=train\_std pearson;
15. var &var\_list1;
16. run;
17. */\* 将Pearson相关系数结果输出到CSV文件 只能本地运行 \*/*
18. ods csv file="corr\_pearson.csv";
19. proc print data=\_LAST\_ (obs=max);
20. run;
21. ods csv close;

结果：<3.1.2-11(1).html> <3.1.2-11(2).jpg>

VIF检验

1. */\*VIF检验\*/*
2. proc reg data=train\_std;
3. model target=&var\_list1 &cla\_list1
4. /vif;
5. run;

结果：<3.1.2-12.html>

删去VIF>10的变量

删除共线性变量（13）：FIX\_3 CFINACE\_3 C\_FINACE\_6 GAP\_FINACE\_3 DEPOSIT\_3 DEPOSIT\_6 AUM\_ 3 DEBIT\_3 DT\_L\_ FINACE FUND\_3 FUND\_6 FINACE\_3 C\_ 1W\_TR\_6 AUM\_6 DEBIT\_6 AGE DOB A\_L\_FIANCE

剩余连续变量（9）：FIX\_6 FINACE\_6 GAP\_FIANCE\_6 C\_1W\_TR\_6 AUM\_6 DEBIT\_6 AGE DOB A\_L\_ FIANCE.

定义新变量集

1. */\*需要筛选的变量数: 23\*/*
2. */\*连续变量 9 \*/*
3. %let var\_list2= FIX\_6  FINACE\_6 GAP\_FIANCE\_6 C\_1W\_TR\_6  AUM\_6 DEBIT\_6 AGE DOB A\_L\_FIANCE ;
4. */\*分类变量 14 \*/*
5. %let cla\_list2=F\_CC F\_CLOAN F\_FUND F\_HLOAN F\_MOBILE F\_PAYROLL F\_STAFF F\_TEL F\_VIP F\_WEB F\_YJL F\_YLJ CHANNEL\_PRE GENDER;

逐步回归方法进一步压缩变量

1. */\* 多方法组合对比筛选变量 \*/*
2. */\*相关分析 corr\*/*
3. \*target--&var\_list2;
4. proc corr data=train\_std spearman;
5. var target;
6. with &var\_list2;
7. run;
8. \*target--&cla\_list2;
9. proc freq data=train\_std;
10. table (&cla\_list2)\*target
11. /chisq nocol nopercent;
12. run;
13. */\*逐步回归法筛选变量\*/*
14. proc logistic data=train\_std descending namelen=50;
15. model target(event='1')=&var\_list2 &cla\_list2
16. /selection=stepwise
17. sls=0.05  sle=0.05
18. stb  lackfit  parmlabel;
19. run;
20. */\*4、根据ROC曲线各变量解释信息百分比 筛选变量\*/*
21. proc logistic data=train\_std plots=roc;
22. class &cla\_list2;
23. model target(event='1')=&var\_list2 &cla\_list2/ctable;
24. ROC  "FIX\_6"         FIX\_6;
25. ROC  "FINACE\_6"      FINACE\_6;
26. */\*…\*/*
27. ROC  "CHANNEL\_PRE"   CHANNEL\_PRE;
28. ROC  "GENDER"        GENDER;
29. run;
30. */\*根据以上 相关系数 逐步回归 ROC这三种筛选结果对比选出共同认为重要的变量, 对有异议的变量依次放入模型查看模型精度是否有显著变化，有则进，否则剔除\*/*
31. ods graphics on;
32. proc logistic data=train\_std plots(MAXPOINTS=5000)=roc;
33. class &cla\_list2;
34. model target(event='1')=&var1;
35. run;
36. ods graphics off;

变量逐一进入模型方程，在引入变量时利用残差平方，当方程加入了该变量后，一旦该变量对残差的变化不显著时即可删除该变量，如此下去，逐一删除变量。

剩余影响显著的变量有AUM\_6 AGE GAP\_FIANCE\_6 CHANNEL\_PRE F\_WEB DOB GENDER

##### 3.3 建立模型

定义最终模型使用的7个变量

1. */\*精度80.74%，确定最终模型中的变量为: 7\*/*
2. %let var2=AUM\_6  CHANNEL\_PRE  AGE  GAP\_FIANCE\_6  F\_WEB  DOB  GENDER;

训练模型

1. ods graphics on;
2. */\*使用训练集（train\_std）建立逻辑回归模型，并输出模型至fin\_model。同时绘制ROC曲线以进行模型诊断，其中CHANNEL\_PRE、F\_WEB、GENDER为分类变量，&var2为待确定的连续或分类预测变量列表。pprob=0.33表示设定概率阈值为0.33，ctable生成列联表。同时，对训练集进行打分，输出到train\_score数据集中。\*/*
3. proc logistic data=train\_std outmodel=fin\_model plots(MAXPOINTS=5000)=roc;
4. class CHANNEL\_PRE F\_WEB GENDER;
5. model target(event='1')=&var2 / pprob=0.33 ctable;
6. score out=train\_score;
7. run;
8. */\*类似地，用验证集（valid\_std）运行逻辑回归过程，但不进行模型训练，仅用于生成ROC曲线以进一步验证模型性能。\*/*
9. proc logistic data=valid\_std plots(MAXPOINTS=5000)=roc;
10. class CHANNEL\_PRE F\_WEB GENDER;
11. model target(event='1')=&var2;
12. run;
13. ods graphics off;
14. */\*调用之前训练好的fin\_model，对验证集（valid\_std）进行打分，得到预测结果存储在valid\_score中，设定先验概率为0.33\*/*
15. proc logistic inmodel=fin\_model;
16. score data=valid\_std out=valid\_score
17. priorevent=0.33;
18. run;
19. */\*定义宏f\_i，用于计算训练集和验证集上预测结果（f\_target）与实际标签（i\_target）的交叉频数表。\*/*
20. %macro f\_i(data);
21. proc freq data=&data;
22. table f\_target\*i\_target;
23. run;
24. %mend;
25. %f\_i(train\_score);
26. %f\_i(valid\_score);
27. */\*保存模型根据最大似然估计分析表 确定各变量系数,将模型具体化。再次利用训练集（train\_std）建立逻辑回归模型，并输出至fin\_model，此步是为了确认模型系数或更新模型。\*/*
28. proc logistic data=train\_std outmodel=fin\_model;
29. class CHANNEL\_PRE F\_WEB GENDER;
30. model target(event='1')=&var2;
31. run;

结果：<3.1.3-1.html>

代码主要进行了模型训练、验证、评估以及基于业务需求调整阈值的过程，主要使用了proc logistic过程进行逻辑回归模型构建和评分，以及proc freq过程来比较预测结果与真实标签的吻合度。

模型的回归方程：



其中



注：先验概率 priorevent 为原始数据集（按均衡比例抽取出的均衡数据集）中事件 target=1 的占比，可根据具体业务、公司情况，以先验概率为基础调节阈值（这儿的阈值指的是预测结果中 target=1 的占比），若需扩大业务 扩张客户量 可调高该阈值，反之降低。

# **第四章 模型预测**

##### 4.1 验证集测试

使用验证集对模型进行评估

1. */\*重复操作，再次调用fin\_model对验证集评分，并生成验证集上的预测结果与实际标签的交叉频数表，以便进行效果评估和阈值选择。\*/*
2. */\*根据业务需求 以先验概率为基础 合理灵活的调整判断客户是否响应的阈值\*/*
3. proc logistic inmodel=fin\_model;
4. score data=valid\_std out=valid\_score
5. priorevent=0.33;
6. run;
7. proc freq data=valid\_score;
8. table f\_target\*i\_target;
9. run;

结果：<4.1-1.html>

##### 4.2模型解释

具体化解释模型结构中各变量对模型预测结果的影响

连续变量可直接通过正负相关性解释相应的影响，也可以将部分连续变量 如：age 离散化之后更易于解释，

分类变量可通过分析对比各分类水平上target=1和0 的占比 进一步解释哪一类水平的客户更可能响应或不响应，

也可以对分类变量做哑变量变换，将每一类视为一个独立的变量，进而比较各类哑变量对模型的贡献

1. */\*使用proc freq过程统计训练数据集中目标变量(target)与分类变量(CHANNEL\_PRE, gender, F\_WEB)之间的关系，并计算卡方检验统计量(chisq)，以探究各类别下购买理财产品的可能性是否存在显著差异。\*/*
2. proc freq data=train\_std;
3. table target\*(CHANNEL\_PRE gender F\_WEB )/chisq;
4. run;
5. */\*定义了一个名为lift\_var的宏，用于对指定的连续变量进行等频或等宽离散化，并绘制Lift图观察响应概率随变量变化的趋势。\*/*
6. */\*为了更易于解释连续变量对模型的影响，对部分连续变量（做离散化处理，探究随着连续变量的变化，响应概率的变化趋势）\*/*
7. %macro lift\_var(in\_data,Var\_group,n\_group);
8. proc sort data=&in\_data out=fin\_sort;
9. by &var\_group;
10. run;
11. */\*添加分组列 一般设置 10组\*/*
12. data fin\_group;
13. set fin\_sort;
14. group=ceil(\_N\_/&n\_group);
15. run;
16. */\*根据分组 求出每组的p值   p值=每组中实际target为1的占比除以整个数据集中target为1的占比\*/*
17. data fin\_plt\_lift;
18. set fin\_group;
19. by group;
20. if first.group then sum=0;
21. sum + target;
22. avg=sum/&n\_group;
23. if last.group;
24. run;
25. */\*绘制lift图\*/*
26. proc sgplot data=fin\_plt\_lift;
27. series x=group y=avg/markers;
28. run;
29. %mend;
30. %let var2=AUM\_6 CHANNEL\_PRE F\_WEB DOB AGE GAP\_FIANCE\_6 GENDER ;
31. %lift\_var(train\_std,age,4628);  %lift\_var(valid\_std,age,2042);
32. %lift\_var(train\_std,DOB,4628);  %lift\_var(valid\_std,DOB,2042);
33. %lift\_var(train\_std,GAP\_FIANCE\_6,4628);  %lift\_var(valid\_std,GAP\_FIANCE\_6,2042);
34. %lift\_var(train\_std,AUM\_6,4628);  %lift\_var(valid\_std,AUM\_6,2042);

结果：<4.2-1.html>

# **第五章 分析与结论**

根据上面的分析，可以看出下面这类人群购买银行理财产品的一员比较强

1. 年龄偏高的用户群体

2. 使用网络渠道

3. 关注行龄比较久的客户，特别是这类客户购买意向较高

4. 女性

目标用户预测：

如果要向新用户推荐该理财产品，可着重关注近六月在我行的月均流水，该用户的年龄，之前该买别的理财产品的频率，购买渠道偏好，是否已开通网银及手机银行，账龄，性别等维度信息，通过模型进行预测该用户会够买该理财产品的概率。从而让理财产品经理在筛选目标客户有所放矢，提高工作效率及成功营销概率。

附：<附.pdf>