第五届“泰迪杯”
数据分析技能赛

优
秀
报
告

作品名称：银行客户忠诚度分析

荣获奖项：一等奖并获泰迪杯

作品单位：南开大学

作品成员：徐妙杰　许梦娇　张馨悦

银行客户忠诚度分析

摘　　要：

改革开放以来，居民收入水平呈现快速增长的趋势。随着个人财富的积累，人们渐渐将注意力转移到自身财产管理上，各类理财产品层出不穷。在巨大的市场机遇面前，各家银行面临巨大的市场竞争。“以客户为中心”业务模式可以为客户带来极致体验和价值成长，形成路径依赖，进而实现价值共赢。因此，有效进行客户忠诚度分析，维护和提升客户忠诚度是当代银行发展历程中的重要课题。本文主要研究银行忠诚度的诸多问题，查阅相关文献并根据题中所给数据进行分析，综合运用XGBoost模型、随机森林预测模型等方法建立相关问题的数学模型，并利用Python、Excel、SPSS等软件进行模型的求解，得出合理的结论。

对于任务一，首先对短期数据进行缺失值和重复值处理，通过dropna()函数删除缺失值所在行，再通过drop.duplicates函数，删除user_id列重复值所在行数据。接着对于长期数据进行异常值处理，即取子数据框功能筛选并删除异常值所在行，对于需要保留的部分数据通过filter函数筛选并删除异常字符所在行。最后，对短期数据中的字符型数据进行特征编码。

对于任务二，首先选用 spearman 相关系数计算短期数据中所有指标之间的相关性，用 heatmap 函数绘制相关系数热力图。接着分别统计两种产品购买结果下不同年龄客户量占比，由 bar 函数绘制成分组柱状图。然后统计蓝领、学生与其他职业的产品购买情况，由 pie 函数绘制成饼状图。分别统计两种产品购买结果下的拜访客户通话时长，由 boxplot 函数绘制拜访客户的通话时长箱线图。

对于任务三，利用 Excel 统计两种流失情况下不同年龄客户量占比，根据 Excel 数据透视表，绘制相应折线图。统计两种流失情况下客户信用资格与年龄分布，绘制相应散点图。构造包含各账号户龄在不同流失情况下的客户量占比透视表，绘制相应堆叠柱状图。依照图表，对账号户龄和客户金融资产进行划分，利用 Excel 分别进行特征编码，作为新的客户特征。统计各资产阶段中新、老客户流失的客户量，由 heatmap 函数绘制热力图并设置上下限，对新老客户各资产阶段的客户流失情况进行分析。

对于任务四，根据给定图表，利用 Excel 统计各类特征，进行特征构建。

对于任务五，遵循代表性与不重复性原则，选取适当的客户特征，建立客户长期忠诚度预测模型。构建客户特征指标。基于任务 3 和任务 4 处理后的结果，在任务四构建的 IsActiveStatus，IsActiveAssetStage，CrCardAssetStage 指标的基础上，继续选取客户信用资格、性别、年龄、客户购买产品数量、个人年收入、新老用户活跃程度、不同金融资产客户活跃程度、不同金融资产信用卡持有状态 8 个特征考虑，分别利用 XGBoost 分类预测和随机森林分类预测两种模型进行预测，结果显示训练集的 F1 值分别为 98.4%和 84.4%，训练效果较好。分别使用混淆矩阵、F1 Score、准确率、召回率、精确率对预测模型进行评估，经过效果比较，最终选取 XGBoost 模型完成预测。

关键字：XGBoost模型、随机森林预测模型、用户忠诚度分析

目录

1 问题分析

1.1 问题重述

任务一：使用Python删除短期客户产品购买数据“short-customer-data.csv”中的缺失值和重复值，同时将字符型数据进行特征编码；使用Python对长期客户资源信息数据的训练集“long-customer-train.csv”中异常值进行处理。

任务二：使用Python基于短期数据分析所有指标的相关性并绘制相关系数热力图，挖掘短期客户对银行的忠诚度。

任务三：使用Excel基于长期数据进行客户流失因素的分析并可视化呈现结果。

任务四：基于长期数据提取影响客户流失的因素，构建与银行客户长期忠诚度相关特征。

任务五：银行客户长期忠诚度预测建模。

1.2 思路分析

围绕题目所给任务，本文解题思路与基本步骤如下：

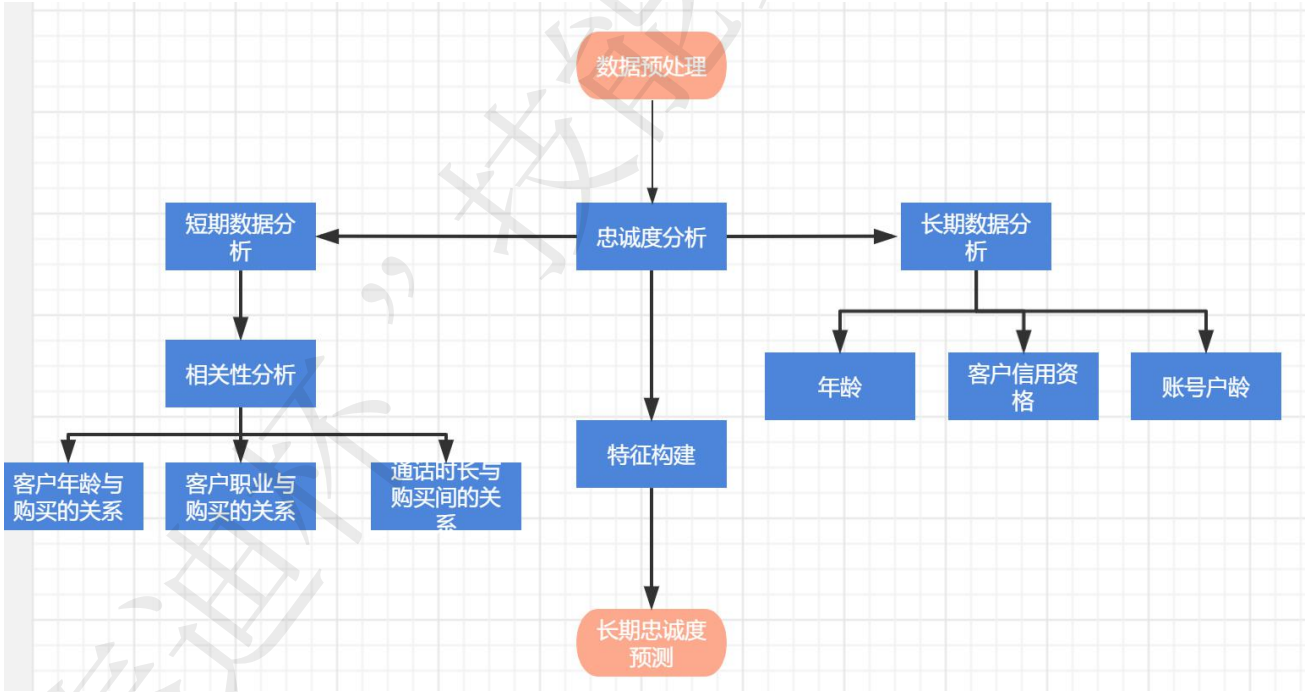

图 1-1 解题思维导图

2 任务一：数据探索与清洗

2.1 任务1.1 数据探索与预处理

2.1.1 数据缺失值及重复值定义及处理

根据题目所给数据特征，本文将缺失值定义为数据为空值，“use_id”重复值定义为在数据集中“use_id”字段相同。

根据上述定义，本文对其进行数据预处理，具体步骤如下：

Step 1 遍历查找各指标的缺失值数据行，即为空值的数据行并做删除处理。

Step 2 遍历查找“user_id”列重复值，并将重复值所在行数据做删除处理。

Step 3 将删除缺失值及重复值后的结果保存至“result1_1.xlsx”中。

针对上述过程叙述，主要代码如下：

```
In [92]: import pandas as pd
         import numpy as np

         #导入短期数据
         data1_1_duanqi=pd.read_csv("E:\\泰迪杯数据分析\\B题：银行客户忠诚度分析赛题数据\\B题：银行客户忠诚度分析赛题数据\\short-customer-data.csv",engi

In [133]: tezheng_list=list(data1_1_duanqi)
          for i in tezheng_list:
              data1_1_duanqi=data1_1_duanqi.dropna(subset=[i])

In [98]: #删除user_id重复的
         data1_1_duanqi=data1_1_duanqi.drop_duplicates(subset='user_id', keep='first', inplace=False)

In [99]: #保存到文件“result1_1.xlsx”中

         data1_1_duanqi.to_excel("E:\\teddy2\\result1_1.xlsx",index=False,header=True)
```

图 2-1 缺失值及重复值处理程序

处理结果如下：

表 2-1 缺失值处理结果描述表

指标名称	缺失值数量
user_id	0
age	0
job	330
marital	80
education	1730
default	8596
housing	990
loan	990
contact	0
month	0

day_of_week	0
duration	0
poutcome	0
y	0

表 2-2 重复值处理结果描述表

指标名称	重复值数量
age	33

2.1.2 异常值定义及处理

数据预处理过程中会针对进行过缺失值及重复值处理的数据集进一步做异常值处理。异常值定义为数据集中表现为特殊符号或极度不符合现实因素的数据，针对长期数据的数据特征本文将长期数据中“Age”列数据异常的情况总结为以下两类：

（1）数值为-1，0和“-”的异常值；

（2）数值中存在空格和“岁”等异常字符。

根据上述定义的两种数据异常情况分别进行处理，具体步骤如下：

Step 1 针对数值为-1，0和“-”的异常值的情况，删除存在该情况的行数据。

Step 2 针对数值中存在空格和“岁”等异常字符，删除异常字符并保存年龄数值。

Step 3 将删除异常值后的结果保存至“result1_2.xlsx”中。

针对上述过程叙述，主要代码如下：

```
#读取长期客户数据
data1_1_changqi=pd.read_csv("E:\泰迪杯数据分析\B题：银行客户忠诚度分析赛题数据\B题：银行客户忠诚度分析赛题数据\long-customer-train.csv",engine
```

```
#去掉Age为-1、0 和“-”的异常值
#注意Age是字符型
idx1=data1_1_changqi[data1_1_changqi.Age=='-1'].index
idx1=list(idx1)
idx2=data1_1_changqi[data1_1_changqi.Age=='0'].index
idx2=list(idx2)
idx3=data1_1_changqi[data1_1_changqi.Age=='-'].index
idx3=list(idx3)
idx=set(idx1+idx2+idx3)
data1_1_changqi.drop(index = idx,inplace = True)
```

```
#处理空格和“岁”等异常字符

import re

for i in data1_1_changqi.index:
    a=data1_1_changqi.loc[i,'Age']
    a="".join(filter(str.isdigit, a))
    data1_1_changqi.loc[i,'Age']=a
```

```
#结果保存到文件“result1_2.xlsx”中
data1_1_changqi.to_excel("E:\\teddy2\\result1_2.xlsx",index=False,header=True)
```

图 2-2 异常值处理程序

部分处理结果如下：

表 2-2 异常值部分处理结果

user_id	age
15553251	52
15553256	41
15553283	42
15553308	61
15553387	39
15553444	44
……	……
15815660	34
15815690	40

2.2 任务1.2 特征编码

2.2.1 字符型数据特征编码

为了便于分析，本文将短期数据中的字符型数据进行特征编码，具体规则如下：
（1）user_id：数据值中的“BA”删去同时保留其原数据值，例如将“BA2200001”改为“2200001”。
（2）job：分别将 admin., blue-collar, entrepreneur, housemaid, management, retired, self-employed, services, student, technician, unemployed 按照顺序编码为 1-11。
（3）marital：分别将 divorced, married, single 编码为 1, 2, 3。
（4）education：分别将 postgraduate, high school, illiterate, junior college, undergraduate 编码为 1, 2, 3, 4, 5。
（5）contact：将 cellular 编码为 1, telephone 编码为 2。
（6）month：将最后一次拜访客户的月份编码为 1-12。
（7）day_of_week：将最后一次拜访客户的星期编码为 1-5。
（8）poutcome：将失败（failure）编码为 1，不存在（nonexistent）编码为 2，成功（success）编码为 3。
（9）default, housing, loan, y：将否（no）编码为 0，是（yes）编码为 1。

针对上述过程叙述，主要代码如下：

```
#短期数据data1_1_duanqi
data1_1_duanqi

#先对job编码
list1=list(data1_1_duanqi.job)
list1=list(set(list1))
print(list1)
list1=['admin.','blue-collar','entrepreneur','housemaid','management','retired', 'self-employed','services','student','technician', 'unemploye
for i in data1_1_duanqi.index:
    a=data1_1_duanqi.loc[i,'job']
    idx=list1.index(a)
    data1_1_duanqi.loc[i,'job']=idx+1
```

```
['retired', 'housemaid', 'self-employed', 'services', 'admin.', 'entrepreneur', 'student', 'management', 'blue-collar', 'unemployed', 'techn
ician']
```

```
#对marital编码
list2=list(data1_1_duanqi.marital)
list2=list(set(list2))
print(list2)
list2=['divorced','married', 'single']
for i in data1_1_duanqi.index:
    a=data1_1_duanqi.loc[i,'marital']
    idx=list2.index(a)
    data1_1_duanqi.loc[i,'marital']=idx+1
```

```
['married', 'single', 'divorced']
```

```
#对education编码
list3=list(data1_1_duanqi.education)
list3=list(set(list3))
print(list3)
list3=['postgraduate','high school','illiterate', 'junior college', 'undergraduate']
for i in data1_1_duanqi.index:
    a=data1_1_duanqi.loc[i,'education']
    idx=list3.index(a)
    data1_1_duanqi.loc[i,'education']=idx+1
```

```
['high school', 'illiterate', 'junior college', 'undergraduate', 'postgraduate']
```

```
#对poutcome编码
list10=list(data1_1_duanqi.poutcome)
list10=list(set(list10))
print(list10)
for i in data1_1_duanqi.index:
    a=data1_1_duanqi.loc[i,'poutcome']
    idx=list10.index(a)
    data1_1_duanqi.loc[i,'poutcome']=idx+1
```

```
['failure', 'nonexistent', 'success']
```

```
#对y编码
list11=list(data1_1_duanqi.y)
list11=list(set(list11))
print(list11)
list11=['no','yes']
for i in data1_1_duanqi.index:
    a=data1_1_duanqi.loc[i,'y']
    idx=list11.index(a)
    data1_1_duanqi.loc[i,'y']=idx
```

```
['yes', 'no']
```

图 2-3 特征编码程序

2.2.2 字符型数据特征编码结果

部分处理结果如下：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	user_id	age	job	marital	ducatio	default	housing	loan	contact	month	ay_of_we	duration	poutcome	y
2	2200001	56	4	2	1	0	0	0	2	5	1	261	2	0
3	2200077	37	8	2	2	0	1	0	2	5	1	226	2	0
4	2200004	40	1	2	1	0	0	0	2	5	1	151	2	0
5	2200005	56	8	2	2	0	0	1	2	5	1	307	2	0
6	2200007	59	1	2	4	0	0	0	2	5	1	139	2	0
7	2200009	24	10	3	4	0	1	0	2	5	1	380	2	0
8	2200010	25	8	3	2	0	1	0	2	5	1	50	2	0
9	2200012	25	8	3	2	0	1	0	2	5	1	222	2	0
10	2200013	29	2	3	2	0	0	1	2	5	1	137	2	0
11	2200056	57	4	1	1	0	1	0	2	5	1	293	2	0
12	2200015	35	2	2	1	0	1	0	2	5	1	146	2	0
13	2200017	35	2	2	1	0	1	0	2	5	1	312	2	0
14	2200019	50	2	2	1	0	1	1	2	5	1	353	2	0
15	2200021	30	11	2	2	0	0	0	2	5	1	38	2	0
16	2200023	55	6	3	2	0	1	0	2	5	1	342	2	0
17	2200024	41	10	3	2	0	1	0	2	5	1	181	2	0
18	2200025	37	1	2	2	0	1	0	2	5	1	172	2	0
19	2200026	35	10	2	5	0	0	1	2	5	1	99	2	0
20	2200035	54	2	1	1	0	0	0	2	5	1	208	2	0
21	2200037	34	8	2	2	0	0	0	2	5	1	365	2	0
22	2200038	52	10	2	1	0	1	0	2	5	1	1666	2	0

图 2-4 特征编码部分处理结果

3 任务二：产品营销数据可视化分析

3.1 数据预处理

对短期数据进行缺失值、异常值、重复值处理，得到以下结果：

表 3-1 数据处理结果描述表

名称	user_id	age	job	marital	education
数量	30445	30445	30445	30445	30445

名称	default	housing	loan	contact	month
数量	30445	30445	30445	30445	30445

名称	day_of_week	duration	poutcome	y
数量	30445	30445	30445	30445

通过上表可知，经过数据预处理的清洗环节，可得题中所给数据规范，可以正常使用，为以下任务的完成提供了严谨的数据保障。

3.1 任务2.1：短期数据指标相关性分析

3.1.1 短期数据指标相关性

由于客户的不同特征向量往往具有一定的相关性，本文利用 Spearman 相关系数对上述变量进行相关性分析。

若设变量 M 和 N 的观测值分别为 m_i 和 $n_j(i=1,2,\cdots,n)$，且它们的样本均值分别为 $\overline{m}$ 和 $\overline{n}$，则 M 和 N 的 Person 相关性系数 R_p 如下所示：

$$R_p=\frac{\sum_{i=1}^{n}(m_i-\overline{m})(n_i-\overline{n})}{\sqrt{\sum_{i=1}^{n}(m_i-\overline{m})^2}\sqrt{\sum_{i=1}^{n}(n_i-\overline{n})^2}}$$

其中，$R_p\in[-1,1]$。若 $R_p>0$，则两变量为正相关关系；若 $R_p<0$，则两变量为负相关关系；若 $R_p=0$ 则两变量之间无相关性关系。若 $|R_p|$ 越接近 1 则两变量相关性越强，反之相关性则越弱。

斯皮尔曼相关性系数被定义成等级变量之间的 Person 相关性系数，通常也叫斯皮尔曼秩相关系数。“秩”可以理解成是一种顺序或者排序，那么它就是根据原始数据的排序位置进行求解，具体公式如下：

$$R_s=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}$$

具体计算过程如下：

Step 1 对两变量 M 和 N 进行排序，并记下排序后的秩次；

Step 2 计算两变量秩次差值的平方 d_i^2；

Step 3 代入 R_s 公式。

```
import pandas as pd
import numpy as np
#导入数据
data2_1=pd.read_excel("E:\\teddy2\\result1_3.xlsx")
```

```
data_tezheng=data2_1
data_tezheng=data_tezheng.drop('user_id',axis=1)

#采用spearman进行相关性分析
corr_matrix=data_tezheng.corr(method='spearman')
```

```
import pandas as pd
import seaborn as sns
from matplotlib import pyplot as plt
plt.rcParams['font.sans-serif']=['SimHei']  #图片显示中文
plt.rcParams['axes.unicode_minus'] =False #减号unicode编码

plt.figure(figsize=(12.5,10))
ax=sns.heatmap(corr_matrix, annot=True)
# 设置刻度字体大小
plt.xticks(fontsize=18)
plt.yticks(fontsize=18)
plt.show()
#保存为指定文件
figure = ax.get_figure()
figure.savefig('E:\\teddy2\\相关系数热力图.jpg')
```

图 3-1 相关性分析程序

3.1.2 指标相关性结果分析

利用 Python 求解可得：

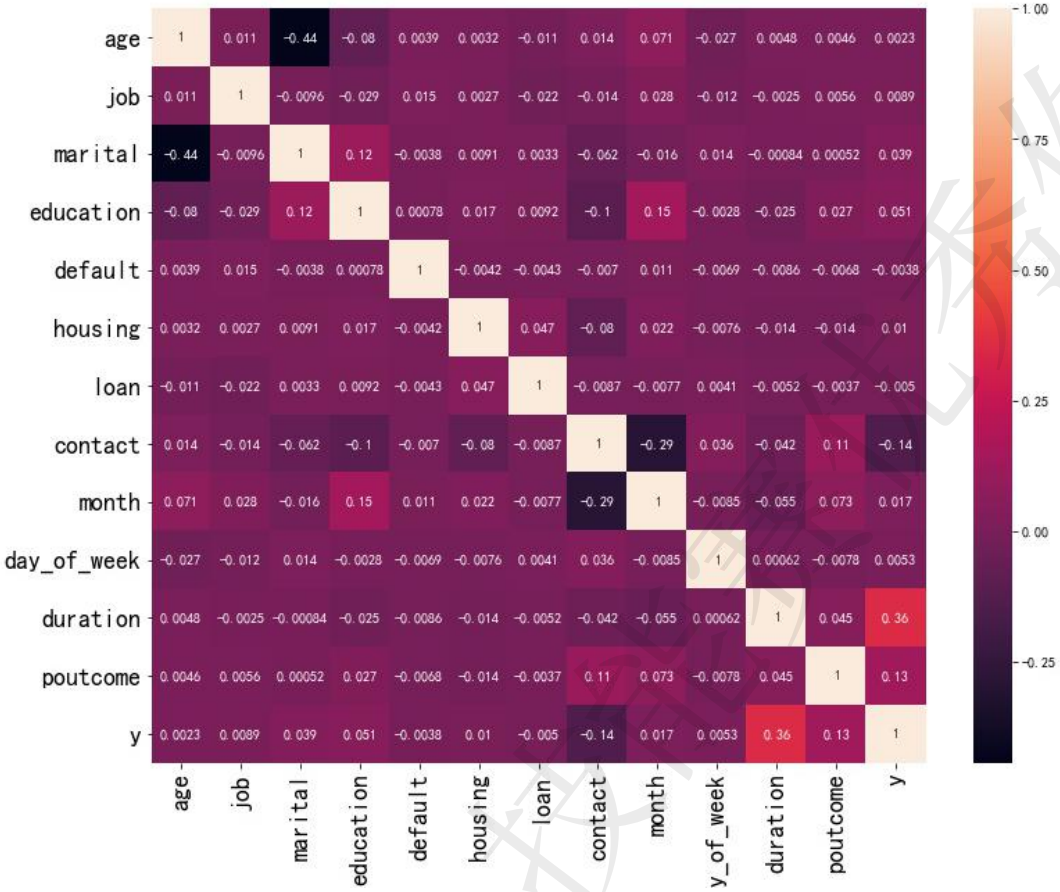

图 3-2 指标相关系数热力图

表 3-1 相关系数判别表

相关系数绝对值取值	相关程度
$[0.0,0.3)$	不相关
$[0.3,0.5)$	弱相关
$[0.5,0.8)$	中度相关
$[0.8,1.0]$	强相关

最近一次拜访客户的星期（duration）与产品购买结果（y）的相关系数为0.36，根据上述判别准则，两者属于弱正相关关系；婚姻状况（marital）与年龄（age）的相关系数为-0.44，两者为弱负相关关系；其余特征指标之间的相关性系数均小于0.3，无显著相关关系。

根据本文对duration与marital的编码准则可知，拜访客户时间越接近周五，该客户越可能购买此产品，这是由于大多数客户一般在工作日的结尾时间较为充裕，所以拜访客户的时间多选取为工作日后半周，从而产品成交情况与其弱正相关。年龄越大客户为单身的可能性越小，

故呈现负相关关系，尽管它的负相关性是由于分类变量的赋值规则导致的，但其表征的现实情况是符合常理的，从而印证了本文结果的正确性。

3.2 任务2.2：不同产品购买情况下年龄结构分布

在客户画像的刻画中，客户的年龄预期对产品的购买状况也有所相关，因此对不同产品购买情况下客户年龄结构分布是必要的。

将客户年龄以 10 岁为一组进行分类，统计各年龄阶段的分布数量并计算其占比情况，利用 bar 函数绘制成分组柱状图，具体程序如下：

```
#产品购买结果即特征y，根据分析，将年龄按十年为一组划分，分为11~20，21~30等共9组
a=data2_1.age
print(min(a))
print(max(a))
```

```
17
95
```

```
#分别统计未买用户（y=0）和购买用户（y=1）中各年龄段的分布人数
#初始化list1,list2
list1=[]
for i in range(1,10):
    list1.append(0)
list2=[]
for i in range(1,10):
    list2.append(0)

for i in data2_1.index:
    x=data2_1.loc[i,'y']
    if x==0:
        a=data2_1.loc[i,'age']
        idx=int((a-1)/10)-1
        list1[idx]+=1
    if x==1:
        a=data2_1.loc[i,'age']
        idx=int((a-1)/10)-1
        list2[idx]+=1

#分别统计未买用户（y=0）和购买用户（y=1）中各年龄段占比
a=sum(list1)
b=sum(list2)
for i in range(0,9):
    list1[i]=list1[i]/a
    list2[i]=list2[i]/b

print(list1)
print(list2)
```

图 3-3 不同产品购买情况下年龄结构分布程序

可得如下结果：

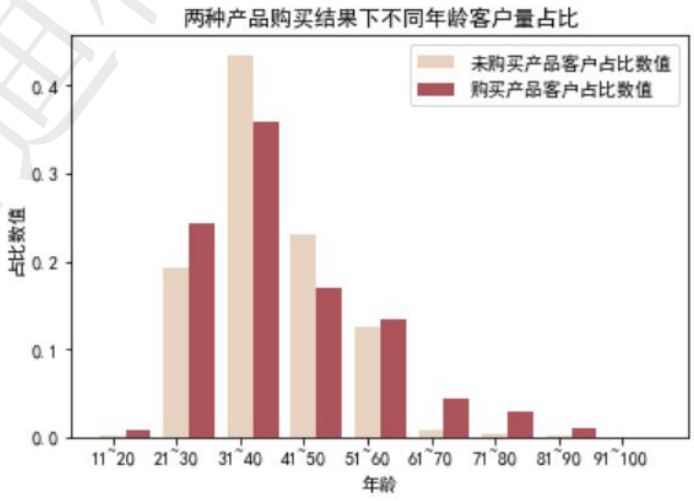

图 3-3 不同产品购买情况下年龄占比图

由上图所示，购买产品的客户年龄大多集中在 21-60 岁，其中 21-30 岁客户所占比例为 19.35%，31-40 岁为客户主要集中年龄段，比例高达 35.89%，41-50 岁客户占比 22.98%，51-60 岁客户占比 12.51%，61-70 岁客户占比为 0.82%；未购买产品的客户年龄集中分布比购买商品的客户年龄分布更加集中，未购买商品客户在 31-40 岁中所占比例高达 43.47%，其中 31-40 岁、41-50 岁、21-30 岁分别为占比前三名，11-20 岁、61-100 岁的客户数量较少。

此外，在 31-50 年龄段中未购买产品客户占比大于购买产品客户占比，剩余年龄阶段下，未购买产品客户占比均小于购买产品客户占比。因此，银行在设计产品时，应主要关注 21-60 岁群体客户的偏爱和需求，对于已购买产品的客户继续维护客户关系，完善自身服务；对于未购买产品的客户应集中调研其未购买的原因，对已有产品或未来推出的产品进行相应改进，以提高自身产品的竞争能力，从而赢得更多客户的青睐。

3.3 任务2.3：蓝领与学生产品购买情况饼图

此外，客户的职业也与客户是否选择购买产品有所联系，接下来具体对蓝领与学生的产品购买情况进行分析。

计算不同职业客户的产品购买情况，并用pie函数绘制画出饼状图，具体结果如下：

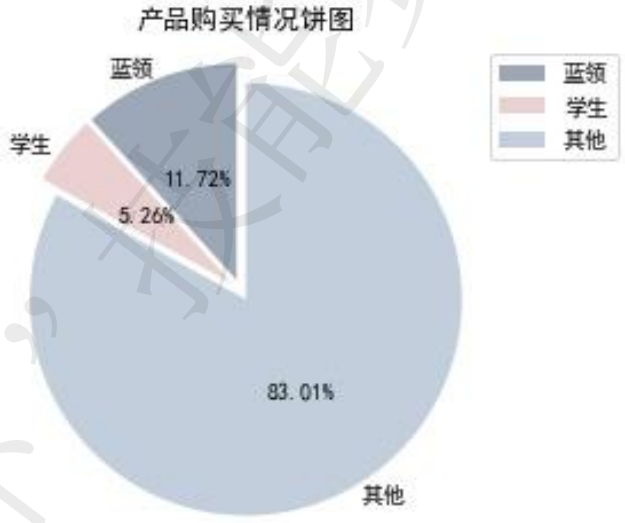

图 3-4 蓝领与学生产品购买情况饼图

如上图所示，蓝领和学生购买产品所占比例分别为7.97%和33.33%，两者共计41.3%，接近一半，说明该产品的主要购买人员职业为学生和蓝领，其中学生近乎占据购买人数总量的1/3。

因此，银行此类产品的目标客户群体为学生和蓝领，在后续的产品优化和更新中，应着重关注这两类群体的需求，从而维持较好的销售情况。此外，对于其他职业的群体，银行也应适当收集相关偏好以扩大该产品的受众范围，从而达到更好的销售水平。

3.4 任务2.4：拜访客户通话时长箱线图

由于与客户通话时长直接决定了产品的推销效果，一般而言，与客户通话时间越长，客户越了解推荐的产品，客户对于产品的意向了解度越高，从而越有可能购买该产品，因此对于拜访客户通话时长与产品购买结果关系的分析是必要的。

利用 boxplot 函数绘制拜进行画图，主要程序如下：

```
#提取通话时长
notbuyers=data2_1[data2_1.y==0]
duration_notbuy=notbuyers.duration
buyers=data2_1[data2_1.y==1]
duration_buy=buyers.duration

import matplotlib.pyplot as plt
x = duration_notbuy  #数据集
y = duration_buy
plt.boxplot((x,y),labels=('未购买产品','已购买产品'))
plt.savefig("E:\\teddy2\\boxplot.jpg")
plt.show()
```

图 3-5 箱线图程序

运行上述程序，所得结果如下：

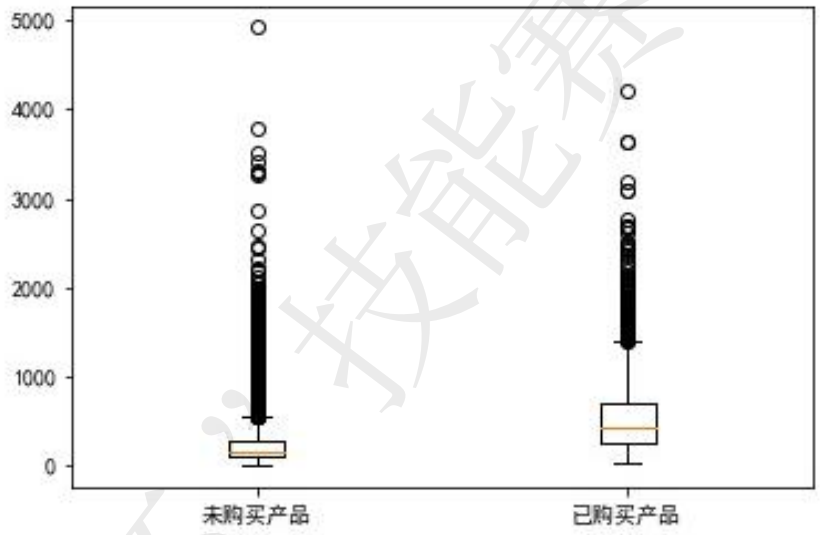

图 3-6 拜访客户通话时长箱线图

如上图所示，当客户购买产品时，拜访客户的通话时间集中在 1000-2500 分钟之间；当客户未购买产品时，拜访客户的通话时间集中在 500-1800 分钟之间。由此可以看出，与客户通话时间越长，客户越了解推荐的产品，客户对于产品的意向了解度越高，从而越有可能购买该产品。

然而，客户未购买产品时也有打了将近 5000 分钟的情况，这可能是由于推销时间过长客户产生反感，因而拒绝购买产品，因此在银行实际推销业务中，应当合理把握推销时间，切忌使客户产生反感，从而影响下次营销活动。

因此，在实际银行向潜在客户推销产品时，要在一定限度内与客户保持联系，既不能过分打扰客户，也不能不与客户通话，从而有充足机会向客户介绍产品，让客户更加了解产品，并对产品产生购买欲望。

4 任务三：客户流失因素可视化分析

4.1 数据预处理

4.1.1 缺失值处理

按照上文缺失值处理方法对长期数据遍历，发现数据表“Exited”列有 48 个空值，并将空值所在数据行删除。

4.1.2 重复值处理

按照上文重复值处理方法对长期数据“CustomerID”遍历并没有发现重复值存在。

4.1.3 异常值处理

上文已将“Age”列中的异常值处理完成，下对其他剩余指标进行处理。

“Tenure”表示以年为单位的账号户龄，故数据中存在特别大的数值不符合常理，视作异常值；“Balance”表示客户的金融资产，0 表示该客户在此银行没有任何金融资产即为空户，符合实际情况因此不做处理；“IsActiveMember”为分类变量，但数据表中有过大的数值，为异常值，故做删除处理。

4.1.4 数据处理结果

经过上述操作，清洗后部分数据如下：

	A	B	C	D	E	F	G	H	I	J	K
1	CustomerIc	CreditScore	Gender	Age	Tenure	Balance	NumOfPro	HasCrCard	IsActiveMe	EstimatedS	Exited
2	15553251	713	1	52	0	185891.5	1	1	1	46369.57	1
3	15553256	619	1	41	8	0	3	1	1	79866.73	1
4	15553283	603	1	42	8	91611.12	1	0	0	144675.3	1
5	15553308	589	1	61	1	0	1	1	0	61108.56	1
6	15553387	687	1	39	2	0	3	0	0	188150.6	1
7	15553444	480	0	44	10	129608.6	1	1	0	5472.7	1
8	15553496	717	1	42	5	190305.8	1	1	0	99347.8	1
9	15553610	556	1	52	9	0	1	1	0	175149.2	1
10	15553627	545	0	53	5	114421.6	1	1	0	180598.3	1
11	15553653	623	1	48	1	108076.3	1	1	0	118855.3	1
12	15553659	748	0	60	0	152335.7	1	1	0	126743.3	1
13	15553670	635	1	32	8	0	2	1	1	19367.98	1
14	15553719	580	1	38	1	128218.5	1	1	0	125953.8	1
15	15553727	613	1	51	7	147262.1	1	1	1	53630.9	1
16	15553866	850	0	56	1	169743.8	1	0	0	155850.4	1
17	15553901	678	1	28	4	0	2	1	1	144423.2	1
18	15553913	753	0	57	7	0	1	1	0	159475.1	1
19	15553935	446	1	45	10	125191.7	1	1	1	128260.9	1
20	15553999	587	0	46	9	107850.8	1	1	0	139431	1
21	15554025	490	1	40	1	0	1	1	1	49594.19	1
22	15554053	689	1	30	5	136650.9	1	1	1	41865.72	1

4.1 任务3.1：两种流失情况下不同年龄客户量占比折线图

使用 Excel 可视化工具可得如下结果：

图 4-1 两种流失情况下不同年龄客户量占比折线图

表 4-1 两种流失情况下不同年龄客户量占比

年龄	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
已流失	0	0.27%	7.29%	26.29%	38.60%	22.21%	4.95%	0.33%	0.05%	0
未流失	0.25%	1.04%	22.04%	48.09%	18.76%	4.54%	2.90%	1.44%	0.11%	0.03%

根据上图可知，银行主要的客户群体年龄分布在 21-60 岁，其中在已流失客户群体中年龄 41-50 岁占比最大，达到了 38.60%，未流失客户群体年龄集中分布在 31-40 岁，达到 48.09% 接近一半。1-10 岁与 81-100 岁年龄段的客户在已流失和未流失客户群体中占比较少，其中 1-10 岁与 91-100 岁年龄的客户在已流失客户中占比为 0。

已流失客户情况下，年龄段占比最高的前三名分别是 41-50 岁、31-40 岁、51-60 岁；未流失客户情况下，年龄段占比最高的前三名分别是 31-40 岁、21-30 岁、41-50 岁。未流失与已流失两种情况下，1-20 岁与 81-100 岁占比趋势基本一致，没有太大区别，这说明该年龄段的客户流失率较低，相比于其他年龄段的客户具有较高的忠诚度。

因此，银行应着重关注 31-60 岁的客户群体，他们对于产品及服务的要求较高，若产品服务达到客户预期即可解决该类客户流失问题。此外，客户群体中 1-10 岁以及 81-100 岁人数较少，因此银行可适当减少维护此类客户群体的行为，将更多精力放到维护易流失客户群体中。

4.2 任务3.2：两种流失情况下客户信用资格与年龄分布散点图

使用 Excel 可视化工具可得如下结果：

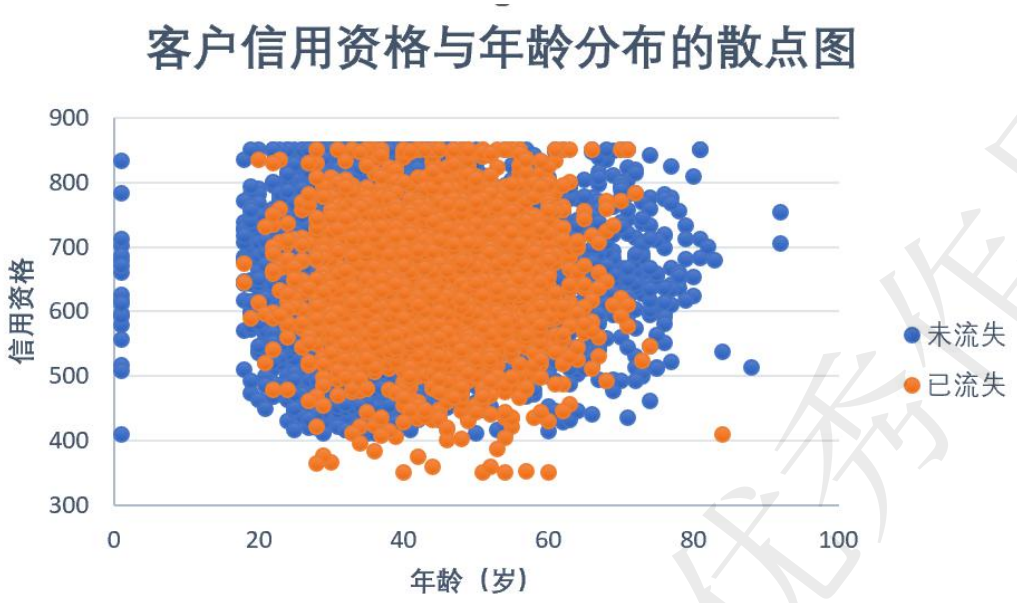

图 4-1 两种流失情况下客户信用资格与年龄分布图

由上图可知，客户年龄与其信用资格并无线性关系，已流失客户群体年龄集中度较高，且信用资格分布呈中心集中态势；未流失客户群体集中态势略弱，且部分 65-80 岁高信用值客户偶有流失。

因此，银行应重点关注中高信用值且年龄为 31-60 岁的客户群体，了解客户偏好，制定相关营销策略，提升自身服务质量以满足客户需求，从而提高客户对银行的忠诚度。

4.3 任务3.3：两种流失情况下账号户龄占比情况

使用 Excel 可视化工具可得如下结果：

表 4-2 透视表

Tenure / Exited	0	1	2	3	4	5	6	7	8	9	10
0	4.05%	10.09%	10.58%	9.94%	9.91%	10.06%	9.71%	10.67%	10.42%	9.67%	4.90%
1	4.63%	11.54%	9.80%	10.18%	10.13%	10.67%	9.96%	8.11%	9.69%	10.72%	4.84%

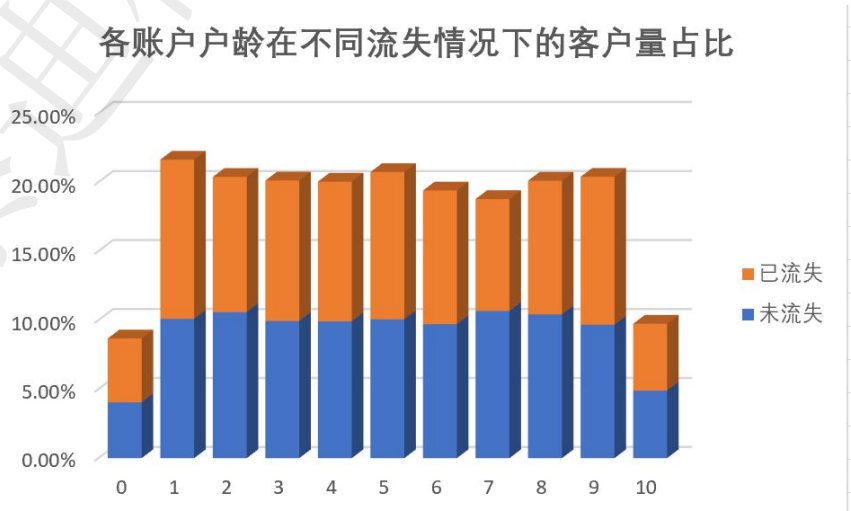

图 4-2 两种流失情况下各账户户龄占比

由上图可知，已流失与未流失客户在不同户龄所占比例较为均衡，其中户龄 10 年已流失与未流失客户占比分别为 4.9%，4.84%，相差仅为 0.06%。户龄为 0、1、3、4、5、6、9 年的已流失客户占比高于未流失客户，其中户龄为 1 年的已流失客户占比为 11.54%，高于未流失客户 1.45%。总体而言，银行客户的户龄大多集中在 1-9 年，短期户龄账户（0 年）与超长期户龄账户（10 年）所占比例较少。

客户并未随着户龄的增长流失有所降低，比如户龄为 9 年的客户已流失占比为 10.72%，未流失占比为 9.67%。因此，在户龄指标上已流失客户与未流失客户并无显著不同，银行应广泛关注各个户龄的客户，了解客户需求、精准营销，通过客户的喜好与其建立紧密联系，使得客户对于银行的某种产品和服务产生一定的依赖性。

4.4 任务3.4：新老客户各资产阶段的情况分析

4.4.1 户龄及金融资产特征编码

利用 Excel 的排序功能分别将户龄和金融资产区间从小到大进行排序，按照题目要求进行特征编码作为新的客户特征，部分结果如下：

CustomerId	Status	AssetStage
15553251	新客户	高资产
15553256	老客户	低资产
15553283	老客户	中上资产
15553308	新客户	低资产
15553387	新客户	低资产
15553444	老客户	高资产
15553496	稳定客户	高资产
15553610	老客户	低资产
15553627	稳定客户	中上资产
15553653	新客户	中上资产
15553659	新客户	高资产
15553670	老客户	低资产
15553719	新客户	高资产
15553727	老客户	高资产
15553866	新客户	高资产
15553901	稳定客户	低资产
15553913	老客户	低资产
15553935	老客户	高资产
15553999	老客户	中上资产
15554025	新客户	低资产
15554053	稳定客户	高资产

图 4-3 户龄及金融资产特征编码部分结果

4.2.2 新、老客户各资产阶段流失客户量热力图

利用 4.2.1 中编码好的特征分别对新、老客户进行数量统计，并绘制热力图，具体程序如下：

```python
#绘制新客户在各资产阶段中流失热力图
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

pd.set_option('expand_frame_repr',False)

#自定义横纵轴标签
x_ticks=['低资产','中下资产','中上资产','高资产']
y_ticks = ['新客户','老客户']

#可以正常显示中文
plt.rcParams['font.sans-serif'] = ['SimHei']

sns.heatmap(df,annot=True,fmt='.0f',cmap="Blues",vmax=1300, vmin=100,xticklabels=x_ticks, yticklabels=y_ticks)

plt.title('新老客户在各资产阶段中流失热力图')  # 图标题
plt.xlabel('客户金融资产划分情况')  # x轴标题
plt.ylabel('账号户龄划分情况')  # y轴标题

#保存为指定文件
plt.savefig("E:\\teddy2\\heatmap2.jpg")

plt.show()
```

图 4-4 新、老客户各资产阶段流失客户量部分程序

具体运行结果如下：

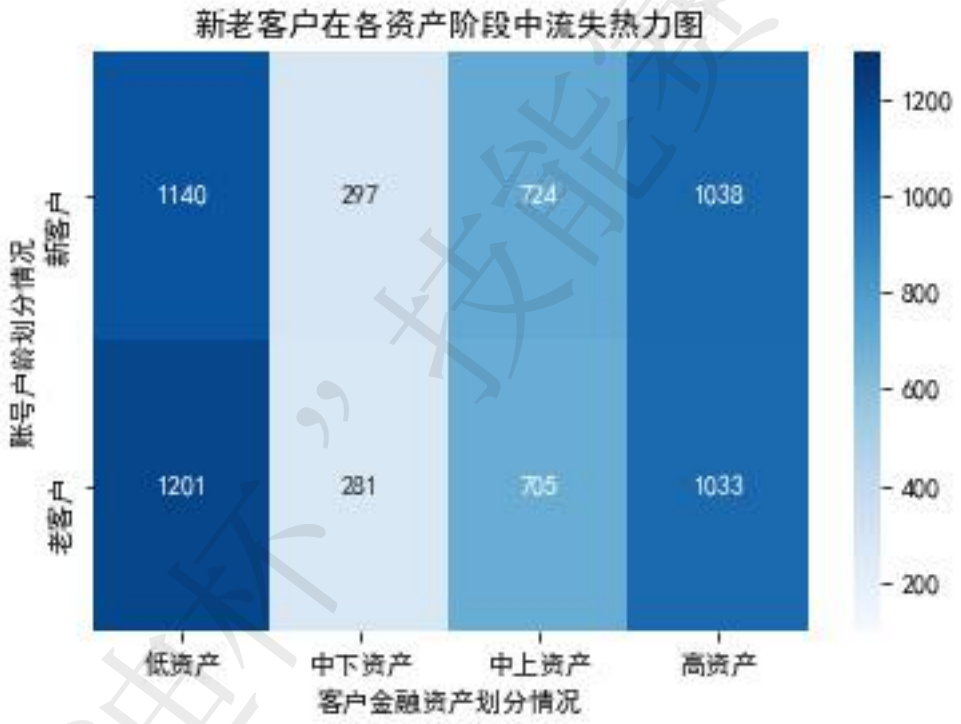

图 4-5 新、老客户各资产阶段流失客户量热力图

由上图可知，新客户在低资产、中下资产、中上资产、高资产的流失量分别为 1140、297、724、1038；老客户在低资产、中下资产、中上资产、高资产的流失量分别为 1201、281、705、1033，其中新、老客户均在低资产、高资产阶段流失量较大，两类客户在不同资产阶段流失量趋势接近一致。老客户在低资产阶段的流失量略大于新客户，新客户在中下资产、中上资产、高资产阶段流失量略大于老客户。

因此，银行在维护客户过程中要重视金融资产处于低资产、高资产和中上资产阶段的客户，其中老客户中的低资产阶段客户相较于对应阶段的新客户更易流失。对于银行而言，高资产客户的流失要比低资产客户流失影响更加严重，因而不论新老客户，银行都应加强与高资产客户的联系，接受客户的建议，维护好客户关系。

5 任务4：特征构建

5.1 新老客户活跃度特征构建

根据题目所给新老客户活跃程度特征构建规则，利用 Excel 操作得到如下部分结果：

CustomerIc	IsActiveStatus
15553251	3
15553256	5
15553283	2
15553308	0
15553387	0
15553444	2
15553496	1
15553610	2
15553627	1
15553653	0
15553659	0
15553670	5
15553719	0
15553727	5
15553866	0
15553901	4
15553913	2
15553935	5
15553999	2
15554025	3
15554053	4

图 5-1 新老客户活跃程度特征构建结果

5.2 不同金融资产客户活跃程度特征构建

根据题目所给不同金融资产客户活跃程度特征构建规则，利用 Excel 操作得到部分结果：

CustomerIc	IsActiveAssetStage
15553251	9
15553256	6
15553283	2
15553308	0
15553387	0
15553444	3
15553496	3
15553610	0
15553627	2
15553653	2
15553659	3
15553670	6
15553719	3
15553727	9
15553866	3
15553901	6

图 5-2 不同金融资产客户活跃程度特征构建结果

5.3 不同金融资产信用卡持有状态特征构建

根据题目所给不同金融资产信用卡持有状态特征构建规则，利用 Excel 操作得到如下部分结果：

CustomerIc	CrCardAssetStage
15553251	9
15553256	6
15553283	5
15553308	6
15553387	0
15553444	9
15553496	9
15553610	6
15553627	9
15553653	9
15553659	9
15553670	6
15553719	9
15553727	9
15553866	5
15553901	6
15553913	6
15553935	9
15553999	9

图 5-3 不同金融资产信用卡持有状态特征构建结果

6 银行客户长期忠诚度预测建模

6.1 模型的建立

6.1.1 特征的选取

由于 5.1、5.2 与 5.3 所构建的新的特征“IsActiveStatus”、“IsActiveAssetStage”、“CrCardAssetStage”包含原特征中的账号户龄、金融资产、持有信用卡状态、活动状态、购买产品数量，故直接选取为所需特征。接着，将剩余未利用的指标包括客户信用资格、性别、年龄、个人年收入纳入所需特征集，客户流失情况为因变量。

CustomerId	CreditScore	Gender	Age	EstimatedSala	Exited	IsActiveStatus	IsActiveAsset	CrCardAssetStage
15553251	713	1	52	46369.57	1	3	9	9
15553256	619	1	41	79866.73	1	5	6	6
15553283	603	1	42	144675.3	1	2	2	5
15553308	589	1	61	61108.56	1	0	0	6
15553387	687	1	39	188150.6	1	0	0	0
15553444	480	0	44	5472.7	1	2	3	9
15553496	717	1	42	99347.8	1	1	3	9
15553610	556	1	52	175149.2	1	2	0	6
15553627	545	0	53	180598.28	1	1	2	9
15553653	623	1	48	118855.26	1	0	2	9
15553659	748	0	60	126743.33	1	0	3	9

6.1.2 XGBoost模型

XGBoost 是在 GBDT 的基础上对 Boosting 算法进行的改进，其内部决策树使用的是回归树。XGBoost 是由 k 个基模型组成的一个加法模型，假设第 t 次迭代要训练的数模型是 $f_t(x_t)$，则有

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

其中，$\hat{y}_i^{(t)}$ 表示第 t 次迭代后样本 i 的预测结果，$\hat{y}_i^{(t-1)}$ 表示前 t-1 课数的预测结果，$f_t(x_i)$ 表示第 t 棵数的模型，所有的模型求和即为预测结果值。XGBoost 的损失函数由预测值和真实值进行表示，其目标函数分别由损失函数和正则化项构成，该项将全部树的复杂度进行求和，用于抑制模型的复杂度，从而避免过拟合发生，公式为

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_1) + \sum_{i=1}^{t} \Omega(f_i)$$

通过上述 XGBoost 目标函数构建过程，可以看到 XGBoost 使用二阶展开、引入正则化项等方式提高预测精度，因此该模型具有较好的回归预测功能。XGBoost 算法的核心过程总结如下：

Step 1 不断地添加树，不断地进行特征分裂来生成一棵新树，每次添加一个树，其实是以上次的预测数据为基础学习一个新的函数 $f_t(x_i)$ 拟合上次预测学习的残差；

Step 2 模型训练完成后，XGBoost 模型由 k 个树模型组成，要预测一个样本的特征，即根据该样本的输入特征，在每棵树中找到一个对应的叶子节点，每个叶子节点对应一个值；

Step 3 将 k 棵树对应的叶子节点的值加起来就得到了模型的输出值，即样本某个特征的预测值。

6.1.3 随机森林模型

随机森林模型将决策树的简单性与集成模型的灵活性和强大功能相结合。在一片森林中，可以忘记一棵特定树木的高变异性，不必担心每个元素，因此可以种植更好的，更大的树木，比修剪的树木具有更大的预测能力。严格的随机森林模型没有提供像单棵树那么多的解释能力，它们的性能要好得多，而且不必像单独树一样担心调整森林的参数。

具体步骤如下：

Step 1 为每棵树创建一个引导数据集，训练 n 个决策树；

Step 2 使用其对应的数据集创建决策树，但在每个节点上使用随机的子变量或特征子样本进行分解；

Step 3 对 n 棵树重复进行以上操作以创建森林。

使用随机森林进行预测非常容易。 我们只需要获取我们每棵单独的树，通过我们想要对其进行预测的观测值，从每棵树中获取一个预测(总计 N 个预测)，然后获得总体的，汇总的预测。引导数据，然后使用聚合进行预测称为 Bagging，对于分类问题，最终预测是森林所做的就是预测出频数最高的结果。

6.2 模型的求解与评估

6.2.1 模型评价指标

1、混淆矩阵

混淆矩阵		真实值	
		正样本	负样本
预测值	正样本	TP	FP
	负样本	FN	TN

TP（真正）：被模型预测为正的正样本。
FP（假正）：被模型预测为正的负样本。
FN（假负）：被模型预测为负的正样本。
TN（真负）：被模型预测为负的负样本。

2、准确率（Accuracy）：预测正确样本占总样本的比例，准确率越大越好。

$$ACC = (TP+TN)/(TP+TN+FP+FN)$$

3、召回率（Recall）：也是灵敏度(Sensitivity)，衡量额检测系统的查全率，实际为正样本的结果中，预测为正样本的比例，召回率越大越好。

$$Recall = (TP)/(TP+FN)$$

4、精确率（Precision）：衡量的检测系统的查准率，预测出来为正样本的结果中，实际为正样本的比例，精确率越大越好。

$$Precision = (TP)/(TP + FP)$$

5、F1 值：为了能够评价不同算法的优劣，在 Precision 和 Recall 的基础上提出了 F1 值的概念。精确率和召回率的调和平均，精确率和召回率是互相影响的，虽然两者都高是一种期望的理想情况，然而实际中常常是精确率高、召回率就低，或者召回率低、但精确率高。若需要兼顾两者，那么就可以用 F1 指标。F1 的定义如下：

$$F1 = (\text{精确率} * \text{召回率} * 2)/(\text{精确率} + \text{召回率})$$

F1 的取值在 0-1 之间，数值越大表示模型效果越好。

6.2.2 XGBoost求解与评估

利用 R 语言进行求解可得如下结果：

表 6-1 指定的 5 个客户 ID 的预测结果 1

CustomerID	Excited
15579131	0
15674442	0
15719508	1

15730076	1
15792228	1

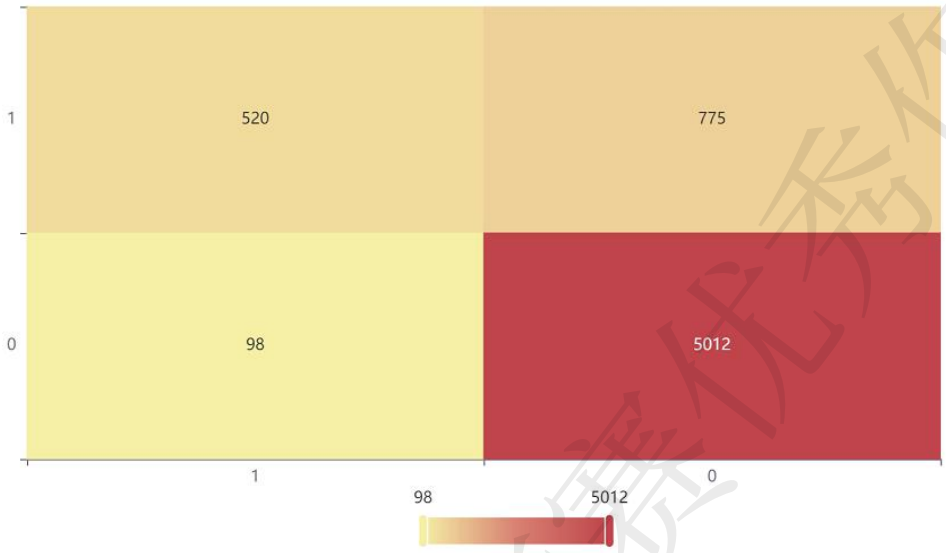

图 6-1　XGBoost 混淆矩阵

表 6-2 各项评估指标

	准确率	召回率	精确率	F1
训练集	0.864	0.864	0.861	0.844
测试集	0.856	0.856	0.845	0.838

根据上述结果可得，模型的测试集准确率、召回率、精确率、F1 分别达到了 86%，86%，85%，84%；训练集准确率、召回率、精确率、F1 分别达到了 86%，86%，86%，84%，拟合结果较好。

6.2.3 随机森林模型求解与评估

利用 R 语言进行求解可得如下结果：

表 6-2 指定的 5 个客户 ID 的预测结果 2

CustomerID	Excited
15579131	0
15674442	0
15719508	1
15730076	1

15792228	1

随机森林模型的结果与 XGBoost 所预测的结果完全相同。

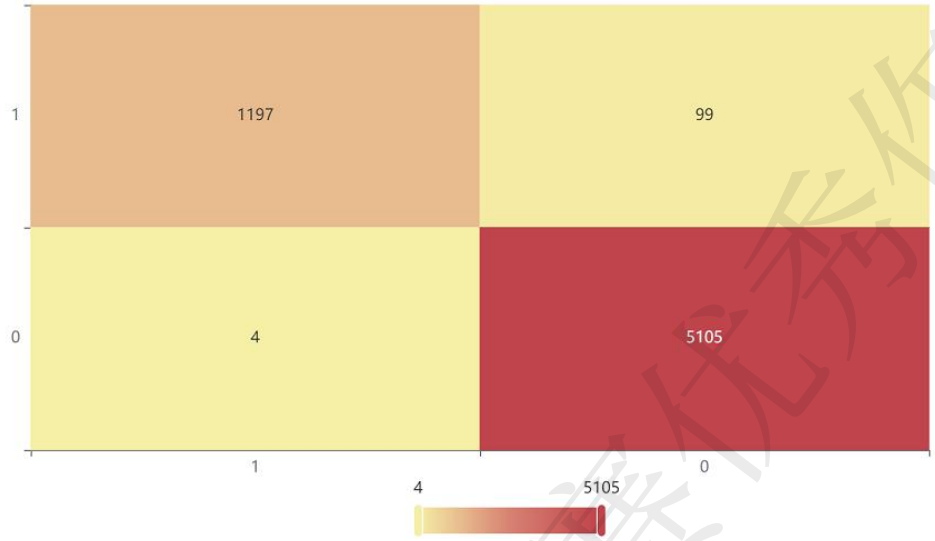

图 6-2　随机森林混淆矩阵

表 6-2 随机森林各项评估指标

	准确率	召回率	精确率	F1
训练集	0.984	0.984	0.984	0.984
测试集	0.844	0.844	0.828	0.827

根据上述结果可得，随机森林模型的测试集准确率、召回率、精确率、F1 分别达到了 85%，85%，83%，83%，在测试集的拟合结果略差于 XGBoost；训练集准确率、召回率、精确率、F1 均达到了 98%，在训练集上结果拟合结果优于 XGBoost。

参考文献

[1] 刘学方,潘丽丽,孙世重. 商业银行客户满意度与忠诚度关系实证研究[J]. 价格理论与实践, 2015(04):110-112.

[2] 孙莹莹. 中国建设银行个人客户忠诚度影响因素研究[D]. 哈尔滨工业大学, 2014.

[3] 国刚,杨青. 基于数据挖掘的客户忠诚度分析[J]. 价值工程, 2013,32(06):140-143.

[4] 吴载斌,王斌会. 数据挖掘中的预测及其应用[J]. 统计与预测, 2002(01):34-36.

[5] 王文贤,金阳,陈道斌. 基于RFM模型的个人客户忠诚度研究[J]. 金融论坛, 2012, 17(03): 75-80.

[6] 胡瑜慧. 论商业银行顾客忠诚度的提高[J]. 商业研究, 2006(06):74-77.

[7] 王建军,张勇,池宏. 我国商业银行客户忠诚度研究[J]. 南开管理评论, 2006(04):29-34.

[8] 刘敦利. 基于栅格尺度的土地沙漠化预警模式研究[D]. 乌鲁木齐:新疆大学, 2010.