基于随机森林分类模型的银行客户忠诚度分析

摘要

银行作为我国现代经济结构的重要组成部分，为国家的经济建设作出了巨大的贡献。随着金融市场的不断健全，银行产品也存在同质化现象，使得客户选择渠道增多，客户产品忠诚度下降，是当前银行业面临的难题之一。为此利用**随机森林分类模型**来分析银行客户忠诚度的相关因素，预测客户忠诚度趋势，对银行有效提高客户忠诚度做出相应策略有重要的作用。

**针对问题一：**本文先对所有数据进行预处理分析，筛选异常，重复，缺失的数据。对于重复和缺失的数据，删除缺失值和重复值的所在行数据。而对于异常数据，去除存在特殊数值为-1、0 和“-”的异常值的行数据，对存在特殊字符和空格的行数据，则去除特殊字符和空格，保留行数据。最后再进行特征编码，对不同的指标进行**权重划分编码**。

**针对问题二：从时间和空间维度**，基于短期数据分析不同指标客户与购买银行产品的关联性，从空间维度出发，由**Kappa一致性检验**对定类与定量做相关性分析，得出poutcome（上一次银行活动，客户购买产品的结果）与本次银行活动，客户购买产品有较大关联，同时绘制两种产品购买结果下不同年龄客户量占比的分组柱状图和蓝领（blue-collar）与学生（student）的产品购买情况饼图，分析不同客户群体的产品购买情况，最后从时间维度绘制拜访客户的通话时长箱线图，得出拜访客户的通话时长越长约有可能购买本次银行产品，进一步挖掘出短期客户对银行的忠诚度。

**针对问题三：**即对得到的数据进行可视化分析并撰写分析报告，根据所得到的可视化图表所示，从年龄与流失状况和年龄与信用分数来看：主要流失的客户和信用资格较低的人群为中年人群；从户龄与流失状况来看：主要流失客户户龄集中在前五年和第九年；从新、老客户在各资产阶段中流失的客户量来看：中上资产和低资产人群更易流失，高资产和中下资产人群较为稳定。而是否为新客户或者是老客户则对流失情况影响不大。

**针对问题四：**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的特征。主要使用Pandas库读入表格，根据题目所给表对各项指标值进行编码。与银行客户长期忠诚度相关的特征主要为：账号户龄，资产阶级，活跃状态和信用卡持有状态。

**针对问题五：**利用随机森林分类模型，选取了年龄（age）和不同金融资产信用卡持有状态（CrCardAssetStage）、不同金融资产客户活跃程度的特征（IsActiveAssetStage）、新老客户活跃程度（IsActiveStatus）的特征作为四条特征依据来进行预测，建立随机森林分类模型来对数据进行预测，运用SPSSPRO的随机森林分类进行预测得到预测结果，报告结果数据预测为(user\_id)15579131：0、15674442：0、15719508：0、15730076：0、15792228：1。再利用混淆矩阵、召回率和F1-score值来评估模型的准确率，最终得到（针对测试集）准确率为0.83，召回率为0.83，精确率为0.99以及F1-score值为0.903。

关键词 随机森林分类模型 权重划分编码 Kappa一致性检验 混淆矩阵

Abstract

As an important part of China's modern economic structure, banks have made great contributions to the country's economic construction. With the continuous improvement of the financial market, the homogenization of banking products also exists, which leads to the increase of customers' choice channels and the decrease of customers' product loyalty, which is one of the difficulties faced by the current banking industry. Therefore, the random forest classification model is used to analyze the related factors of bank customer loyalty, predict the trend of customer loyalty, and make corresponding strategies for banks to effectively improve customer loyalty. In view of question one: this paper first preprocesses and analyzes all data, and screens out abnormal, duplicate and missing data. For duplicate and missing data, delete the data in the row where the missing value and duplicate value are located. For the abnormal data, the line data with abnormal values of -1, 0 and "-"are removed, and for the line data with special characters and spaces, the special characters and spaces are removed and the line data is kept. Finally, feature coding is carried out, and different indexes are coded by weight division. In view of the second question: from the time and space dimensions, based on short-term data, this paper analyzes the correlation between customers with different indicators and the purchase of banking products. From the space dimension, Kappa consistency test makes correlation analysis between classification and quantification, and concludes that poutcome (the result of the last banking activity, customers' purchase of products) has a great correlation with this banking activity and customers' purchase of products. At the same time, we draw the grouping histogram of customers of different ages and the pie chart of blue-collar and student's product purchases under the two product purchase results, and analyze the product purchases of different customer groups. Finally, we draw the box chart of the call duration of visiting customers from the time dimension, and dig out the loyalty of short-term customers to the bank. In view of the third question: that is, visual analysis of the obtained data and writing an analysis report, according to the obtained visual chart, from the age and loss status, age and credit score, the main lost customers and people with low credit qualifications are middle-aged people; From the perspective of household age and loss status: the age of the main lost customers is concentrated in the first five years and the ninth year; Judging from the number of customers lost by new and old customers in each asset stage, the upper-middle assets and low-assets groups are more likely to lose, while the high-assets and middle-lower assets groups are more stable. Whether it is a new customer or an old customer has little effect on the loss. Question 4: Based on long-term data, extract the factors that affect customer churn, and construct the characteristics related to long-term loyalty of bank customers. The main Pandas library is used to read the table, and the index values are coded according to the table given by the title. The characteristics related to long-term loyalty of bank customers are: account age, bourgeoisie, active status and credit card holding status. Aiming at the fifth problem, using the random forest classification model, the age, credit card holding status of different financial assets, the characteristics of active degree of customers of different financial assets, and the characteristics of active degree of new and old customers are selected as four characteristics to predict. Establish a random forest classification model to predict the data, and use SPSSPRO's random forest classification to predict the results. The reported results are (user \_ id) 15579131: 0, 15674442: 0, 15719508: 0, 15730076: 0, 1579228: 1. Then, the confusion matrix, recall rate and F1-score value are used to evaluate the accuracy of the model. Finally, the accuracy rate (for the test set) is 0.83, recall rate is 0.83, accuracy rate is 0.99 and F1-score value is 0.903.

目录

[第一章 问题分析 5](#_Toc31764)

[1.1 问题重述 5](#_Toc954)

[1.2 思路分析 5](#_Toc30987)

[第二章 任务1数据探索与清洗 6](#_Toc27193)

[2.1 任务1.1缺失、异常、重复值处理 6](#_Toc379)

[2.1.1 缺失、重复值的处理 6](#_Toc8390)

[2.1.2 异常值的处理 6](#_Toc25445)

[2.2 任务1.2特征编码 7](#_Toc2809)

[第三章 任务2产品营销可视化分析 8](#_Toc4888)

[3.1 任务2.1相关性分析 8](#_Toc22167)

[3.2 任务2.2分组柱状图 9](#_Toc12081)

[3.3 任务2.3产品购买情况分析 10](#_Toc22546)

[3.4 任务2.4拜访客户时长通话分析 11](#_Toc12694)

[第四章 任务3客户流失因素可视化分析 12](#_Toc9217)

[4.1 任务3.1不同年龄客户流失分析 12](#_Toc7561)

[4.2 任务3.2客户信誉资格与年龄分布分析 13](#_Toc5512)

[第五章 任务4特征构建 16](#_Toc15098)

[5.1 任务4 构建客户长期忠诚度相关特征 16](#_Toc12481)

[第六章 任务5客户长期忠诚度预测 17](#_Toc12740)

[6.1 任务5.1建立模型 17](#_Toc2468)

[6.2 任务5.2模型评估 18](#_Toc11131)

[6.3 任务5.3模型预测 20](#_Toc23299)

[参考文献 22](#_Toc6382)

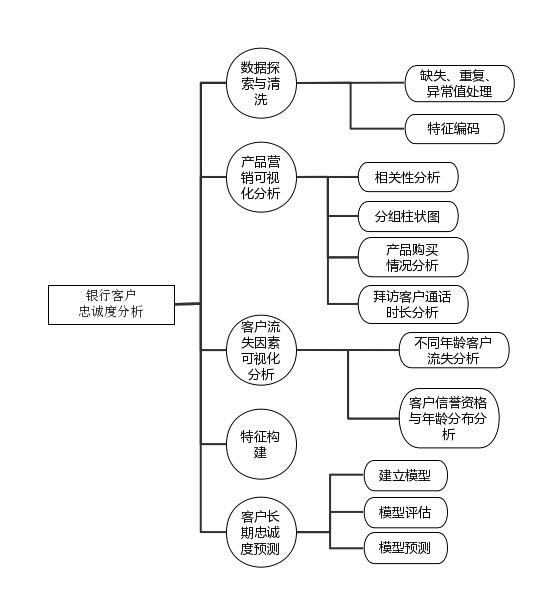
# **第一章 问题分析**

## 1.1 问题重述

1. 对客户数据进行数据预处理，基于题目中的客户数据，分析不同指标客户对银行产品的购买依赖度和客户流失原因，并进行可视化呈现。
2. 依据长期客户资源信息数据的分析结果，对银行客户长期忠诚度进行预测。

## 1.2 思路分析

根据题中所给的五个任务，本文概括出的解题思路与大体步骤如下：



基于上述思维导图，可以大体分为五个步骤，首先对数据进行预处理，删除缺失、重复值并对异常值进行处理，在保证了数据的可靠性之后，再对特征进行编码，针对任务一，运用Kappa相关性分析，画出热力图分析相关性。然后分别对产品营销数据以及客户流失因素进行可视化分析，即针对任务二、任务三进行可视化，并分析其原因。最后进行特征的构建，建立机器学习模型，分别进行模型的评估以及预测客户长期的忠诚度。

# **第二章 任务1数据探索与清洗**

## 2.1 任务1.1缺失、异常、重复值处理

### 2.1.1 缺失、重复值的处理

针对任务1.1，根据题目要求，探索短期数据各指标数据的缺失值和“user\_id”列的重复值，删除缺失值和重复值的所在行数据。

针对该要求，运用Python处理步骤如下：

1. 使用Pandas库读入“short-customer-data.csv”表，对全表数据使用dropna()方法去除缺失值所在行。
2. 使用drop\_duplicates()方法去除重复值，subset选择“user\_id”列，keep参数选择first表示保留第一个出现的重复数据。
3. 最后将上述处理过的表读出到“result1\_1.xlsx”中保存。

针对编程思想，部分重要代码示例如下：

|  |
| --- |
| Data\_short = pd.read\_csv("short-customer-data.csv")  #丢弃缺失值 dropna()删除缺失值所在行(axis=0)或列(axis=1) 默认为 axis=0  Data\_short = Data\_short.dropna()  #去重  Data\_short.drop\_duplicates(subset=['user\_id'],keep='first',inplace=True)  Data\_short.to\_excel("result1\_1.xlsx",index=None) |

最终去除行数据共12716条，删除重复的行数据33条，最后共剩下30445条数据。

### 2.1.2 异常值的处理

针对任务1.2，根据题目要求，探索长期数据中的客户年龄“Age”列的存在数值为-1、0 和“-”的异常值，删除存在该情况的行数据“Age”列存在空格和“岁”等异常字符，删除这些异常字符但须保留年龄数值。

针对该要求，运用Python处理步骤如下：

1. 使用Pandas库读入“long-customer-train.csv”表，对全表数据使用Data\_long = Data\_long[~Data\_long['Age'].isin(['-1'])]方法去除存在数值为-1异常值所在行，同理对0 和“-”的异常值所在行进行删除
2. 使用replace()方法对空格和“岁”异常字符进行替换，同时可以保留元数据；同时还对该列的数据类型进行修改，使用astype()方法进行修改。
3. 最后将上述处理过的表读出到“result1\_2.xlsx”中保存。

针对编程思想，部分重要代码示例如下：

|  |
| --- |
| Data\_long = Data\_long[~Data\_long['Age'].isin(['-1'])]  Data\_long = Data\_long[~Data\_long['Age'].isin(['0'])]  Data\_long = Data\_long[~Data\_long['Age'].isin(['-'])]  Data\_long['Age'].replace('\s+','',regex=True,inplace=True)  Data\_long['Age'].replace(regex=True,inplace=True,to\_replace=r'岁',value=r'')  Data\_long['Age'] = Data\_long['Age'].astype(int)  Data\_long.to\_excel("result1\_2.xlsx",index=None) |

最终去除存在数值为-1、0 和“0”的异常值行数据共33条，将特殊字符和空格进行去除保留数据最后共剩下9198条数据。

## 2.2 任务1.2特征编码

根据题目要求，对短期数据中的字符型数据进行特征编码。

针对该要求，运用Python处理步骤如下：

1. 使用Pandas库读入“short-customer-data.csv”表，根据题目所给示例对各项指标值进行编码，给出具体编码要求如下：由下到上（部分示例）

|  |  |
| --- | --- |
| 指标 | 编码 |
| Unemoloyed  Technician  Admin  Single  Married  Divorced  Illiterate  Junior college  High school  Undergraduate | 0  1  10  0  1  2  0  1  2  3 |

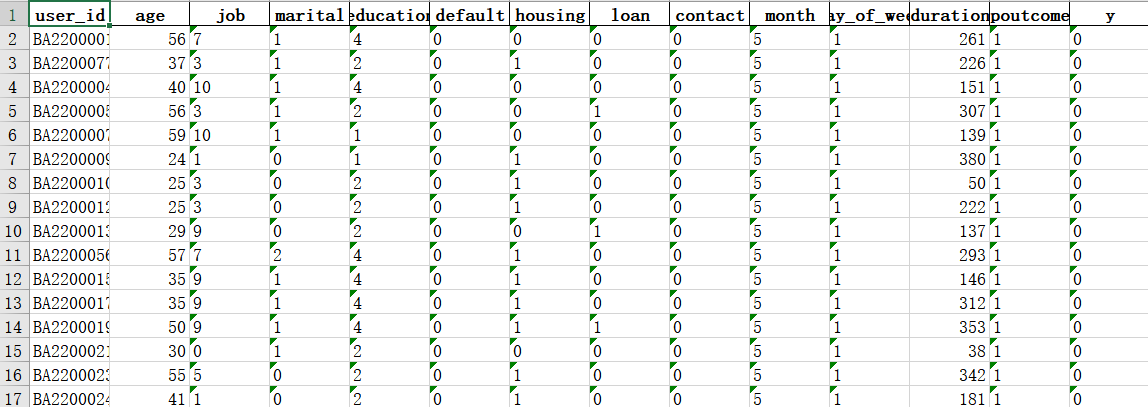
1. 使用repalce()方法将原数据中的字符数据用以上编码进行替换。
2. 最后将上述处理过的表读出到“result1\_2.xlsx”中保存。

针对编程思想，部分重要代码示例如下：

|  |
| --- |
| Data\_short = pd.read\_excel("result1\_1.xlsx")  Data\_short['job'].replace(regex=True,inplace=True,to\_replace=r'unemployed',value=r'0')  Data\_short['marital'].replace(regex=True,inplace=True,to\_replace=r'single',value=r'0')  Data\_short['default'].replace(regex=True,inplace=True,to\_replace=r'no',value=r'0') |

得到表的数据部分展示：

表 1 特征编码表



# **第三章 任务2产品营销可视化分析**

## 3.1 任务2.1相关性分析

运用SPSSPRO[1]对result1\_1中的表进行Kappa一致检验性分析[2]所给数据中的定类和定量之间的相关性，计算出短期数据里所有指标之间的相关性系数，绘制出热力图如下所示：



图 3-1 相关性热力图

从相关性热力图中不难看出，在所有指标中，poutcome（上一次银行活动，客户购买产品的结果）与本次银行活动，客户是否购买产品有较大的相关性。

## 3.2 任务2.2分组柱状图

基于两种产品购买结果下（本次银行活动客户购买产品的结果，分别为no、yes ）不同年龄客户量占比的分组柱状图，如下所示：

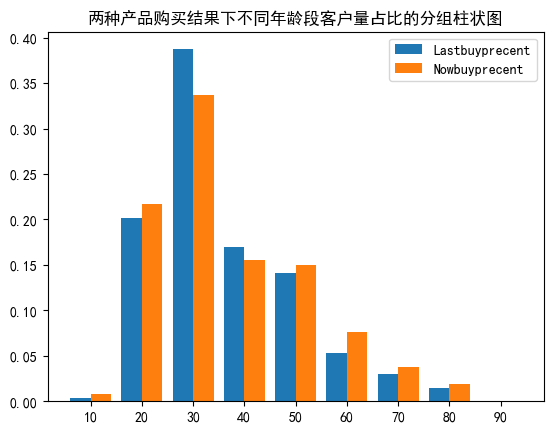


图 3-2 两种产品购买结果下不同年龄段客户量占比分组柱状图

根据上述的图可知，从两次产品的购买结果来看，在10-20岁、50-60岁、60-70岁、70-80岁、80-90岁的人群较上次购买产品人数占比有小幅度上升，而90岁以上的人群在两次产品的购买中都未进行购买，作为购买主力的30-40岁的人群本次购买人数占比有较大下滑。

## 3.3 任务2.3产品购买情况分析

针对任务2.3，基于短期数据分析蓝领（bule-collar）与学生的产品购买情况，利用Pyhton中的numpy和matplotlib库绘制出饼图，如下图所示：

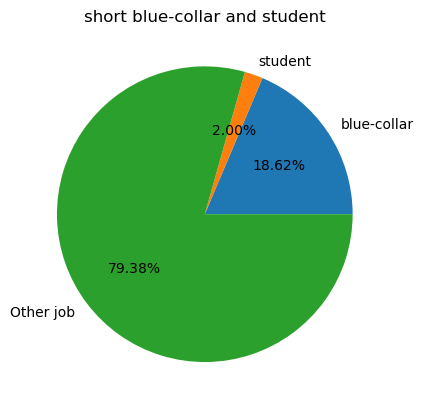


图 3-3 blue-collar与student的产品购买情况饼图

从该饼图可以看出，学生（student）群体在银行产品购买中仍占少数仅占2%，可能与自身经济实力与需求有关，而蓝领（blue-collar）在银行产品购买中占18.52%，是银行产品购买的主力军，是不可忽视的客户群体之一。

## 3.4 任务2.4拜访客户时长通话分析

根据从任务2.1的相关性分析得出的poutcome（上一次银行活动，客户购买产品的结果）与本次银行活动客户是否购买产品有较大关联，针对任务2.4，绘制拜访客户的通话时长箱线图，如下图所示：

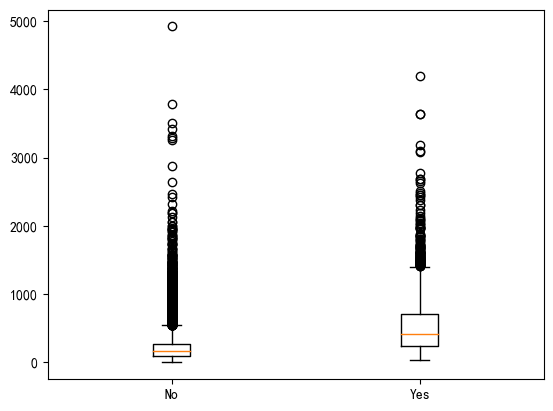


图 3-4 拜访客户的通话时长箱线图

箱线图是一个能够通过5个数字来描述数据的分布的标准方式，这5个数字包括：最小值、第一分位数、中位数、第三分位数、最大值，箱线图能够明确的展示离群点的信息，同时能够让我们了解数据是否对称，数据如何分组、数据的峰度。

从图中的数据分析得出没有购买银行产品的拜访客户群体通话时长普遍偏短，而购买银行产品的拜访客户群体通话时长则普遍偏长，由此可以通过拜访客户的通话时长来基本确定客户是否有购买的可能性，为后续预测提供良好的标准。

# **第四章 任务3客户流失因素可视化分析**

## 4.1 任务3.1不同年龄客户流失分析

先对表格进行处理，筛选出已流失的客户群和未流失的客户群并分别建表。引用python的matplotlib库，使用jupyter对两表进行可视化处理，绘制双折线图，如下图所示：

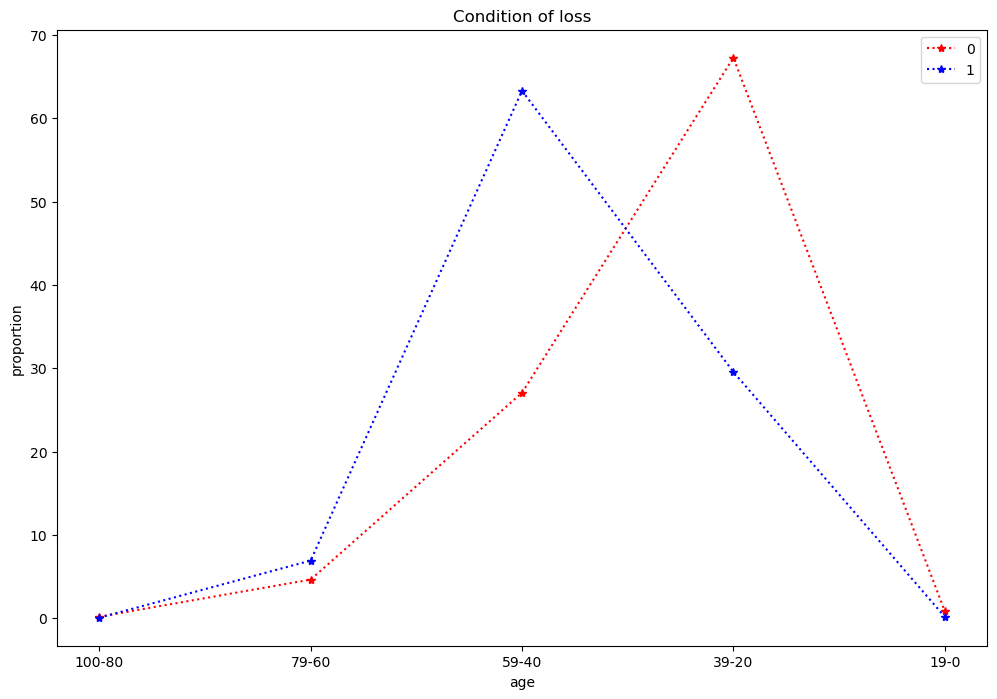
****

图 4-1 不同年龄客户占比流失图

根据该折线图可以得出以下结论：未流失客户主要集中在20-39岁这个年龄段，并且随着年龄的增长，客户的流失占比快速下降；而流失的客户则主要集中在40-59这个年龄段。因此可得出：未流失的客户为青壮年人群；而流失的主要客户为中年人群。

## 4.2 任务3.2客户信誉资格与年龄分布分析

针对任务3.2，运用python的matplotlib库，对已流失的客户群和未流失的客户群两表分别取出其年龄分布和信用资格两属性，使用jupyter对两属性进行关联性分析，并进行可视化处理，绘制散点图，如下图所示：

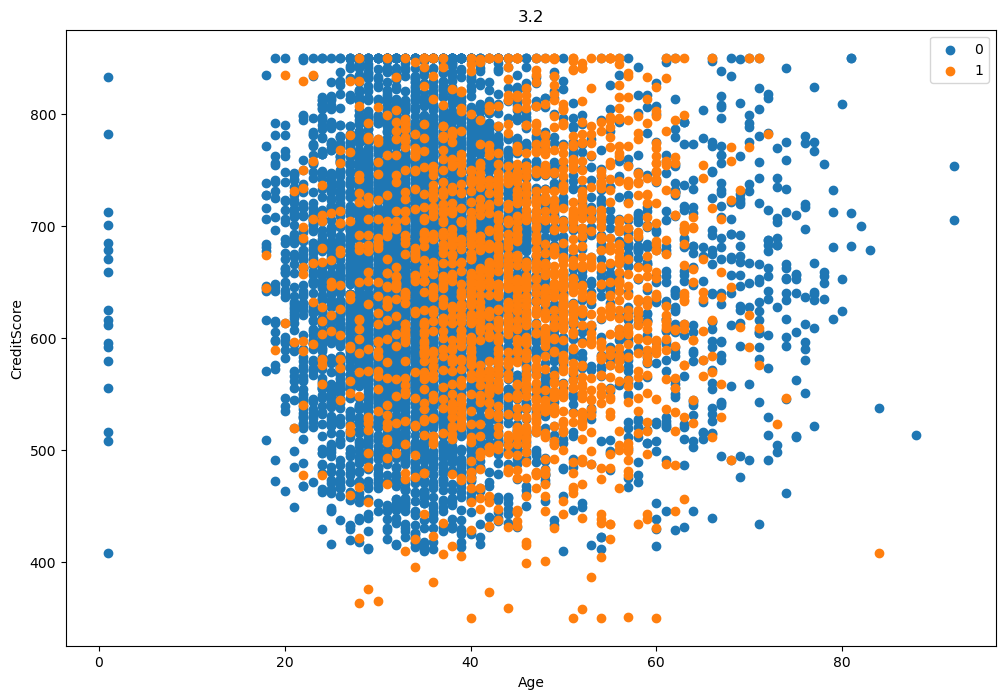


图 4-2 客户信用资格与年龄分布散点图

该散点图显示出：流失客户主要年龄在40-60岁左右，信用分数集中在500-700之间，高信用分数人群较少，低信用分数人群较多；而未流失客户则主要集中在20-40岁左右，高信用分数人群较多，低信用分数人群较少。同时，年龄较大于60岁的人群和年龄小于40岁的人群信用分数总体来说不会很低，而40-60岁人群信用分数低的人数较多。

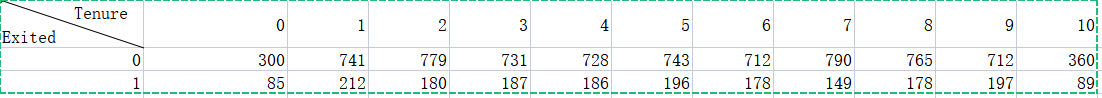
由此可得出结论：一般来说，未流失客户总体来说信用度较高，而流失客户信用度较低。同时中年客户存在信用度较低客户的概率比老年和青年客户概率更高。

部分重要代码示例如下：

|  |
| --- |
| plt.figure(figsize=(12,8))  detail1=pd.read\_excel(r'E:\program\0.xls')  product\_D=detail1['CreditScore']  product\_B=detail1['Age']  detail=pd.read\_excel(r'E:\program\1.xlsx')  product\_C=detail['CreditScore']  product\_A=detail['Age']  plt.scatter(product\_B,product\_D,marker="o")  plt.scatter(product\_A,product\_C,marker="o")  plt.ylabel('CreditScore')  plt.xlabel('Age')  plt.legend(["0","1"])  plt.title('3.2') |

针对任务3.3，先整理出不同户龄下的用户流失情况，并绘制出透视表。表格如下所示：

表 2 各账号户龄在不同流失情况下的客户量占比透视表



根据透视表数据，对用户的户龄和流失情况进行关联性分析，并绘制堆叠柱状图，更好地展示各户龄下用户的流失情况，如下图所示：

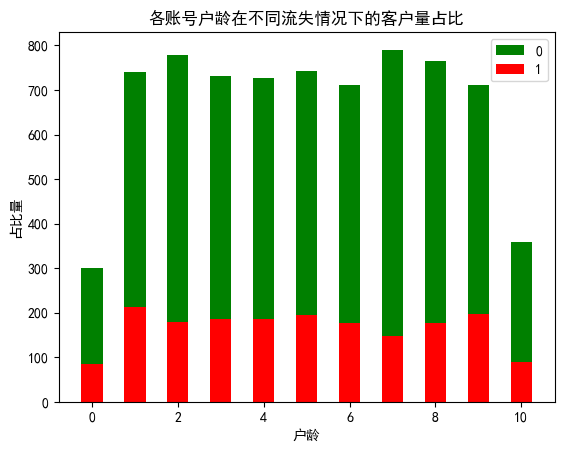


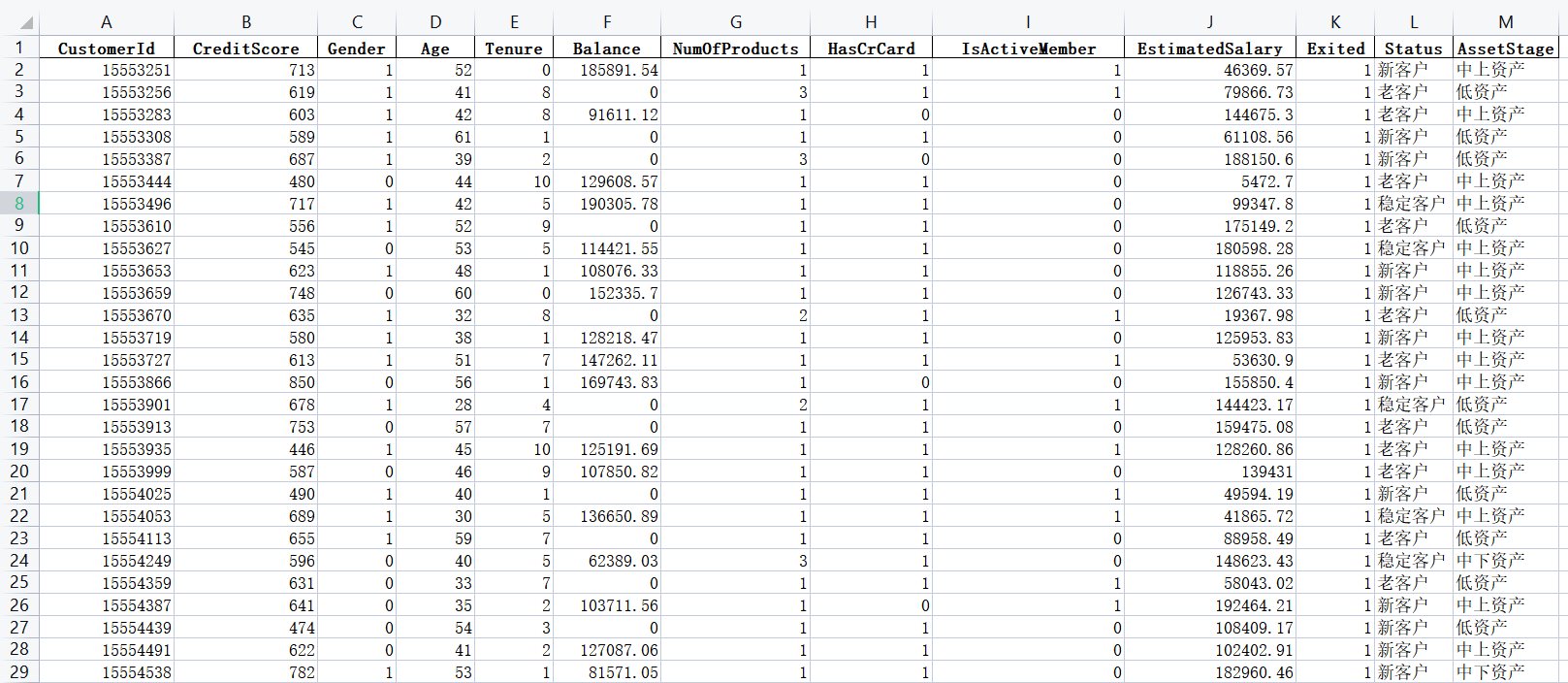
图 4-3 各账号户龄在不同流失情况下的客户量占比图

该透视表和堆叠柱状图显示出：户龄在1-5年的客户流失占比较大，而从第6年到第8年流失占比则比较小，到了第9年又有上升，第10年流失占比则是最低。

由此可得出结论：客户在刚开户的前五年较容易流失，从第六年开始则趋于稳定，第九年又迎来上涨，到了第十年则完全稳定下来，不易流失。

针对任务3.4，按照所给材料的表 5 和表 6 对账号户龄和客户金融资产进行划分，并分别进行特征编码作为新的客户特征，将编码结果建表,如下表所示：

表 3 特征编码部分结果表



同时统计新、老客户在各资产阶段中流失的客户量，绘制热力图，进行关联性分析，如下图所示：

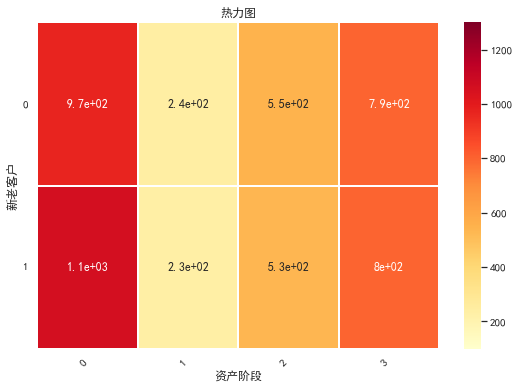


图 4-4 新、老客户在各资产阶段中流失的客户量热力图

图片说明（x与y轴）：00：970，10：241，20：546，30：794，01：1063

11：232，21：534，31：798

该图显示：对比不同资产人群，低资产人群为主要流失人群，其次为高资产人群，接下来是中上资产人群。而中下资产人群则较为稳定，不易流失。对比新老客户：在低资产人群中老客户更易流失，而在其他资产阶级下则相差不大。

由此可得出结论：高资产和低资产人群更易流失；中下和中上资产人群则较为稳定。因此需要提高对低资产人群和高资产人群的服务品质，尽最大努力争取这两类客户，避免银行客户人群的流失。

# 任务4特征构建

## 5.1 任务4 构建客户长期忠诚度相关特征

根据题目要求，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的特征，将结果保存到文件“result4.xlsx”中

针对该要求，运用Python处理步骤如下：

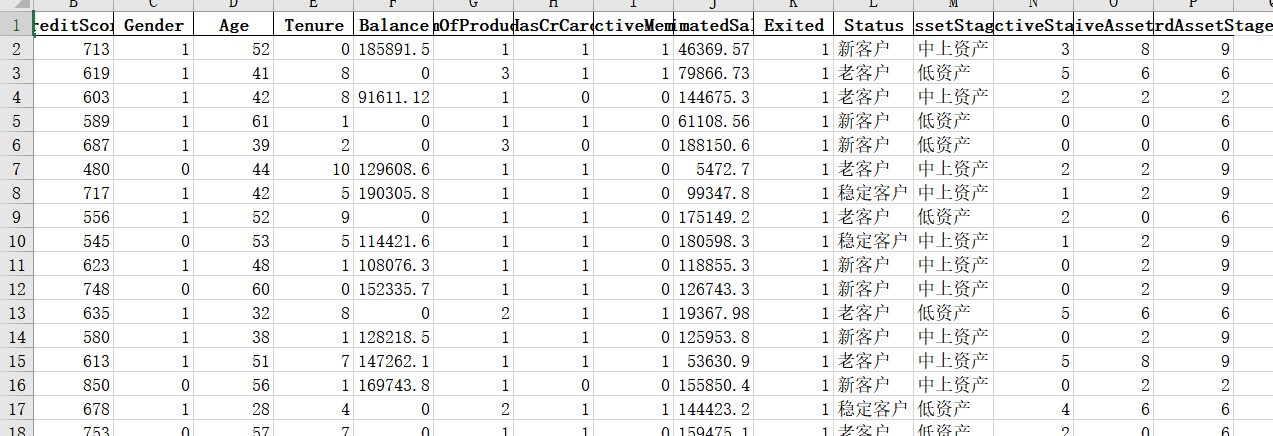
1. 使用Pandas库读入“result3.xlsx”表，根据题目所给表对各项指标值进行编码。
2. 通过insert()方法进行添加“IsActiveStatus”列、“IsActiveAssetStage”列、“CrCardAssetStage”列，使用np.where()方法查找原数据进行替换。
3. 最后将上述处理过的表读出到“result4.xlsx”中保存。

针对编程思想，部分重要代码示例如下：

|  |
| --- |
| Data\_1 = pd.read\_excel("result3.xlsx")  Data\_1.insert(loc=13, column='IsActiveStatus', value='0')  Data\_1['IsActiveStatus'] = np.where((Data\_1['Status'] == "新客户")&(Data\_1['IsActiveMember'] == 0),0,Data\_1['IsActiveStatus'])  Data\_1['IsActiveAssetStage'] = np.where((Data\_1['AssetStage'] == "低资产")&(Data\_1['IsActiveMember'] == 0),0,Data\_1['IsActiveAssetStage'])  Data\_1['CrCardAssetStage'] = np.where((Data\_1['AssetStage'] == "低资产")&(Data\_1['HasCrCard'] == 0),0,Data\_1['CrCardAssetStage'])  Data\_1.to\_excel("result4.xlsx",index=None) |

最终得到表的数据部分展示：

表 4 特征构建结果表



# 任务5客户长期忠诚度预测

## 6.1 任务5.1建立模型

利用SPSSPRO，根据任务4的特征建立的启发，选取了年龄（age）和不同金融资产信用卡持有状态（CrCardAssetStage）、不同金融资产客户活跃程度（IsActiveAssetStage）的特征、新老客户活跃程度（IsActiveStatus）的特征作为四条特征依据来进行预测，建立**随机森林分类模型**来对数据进行预测。

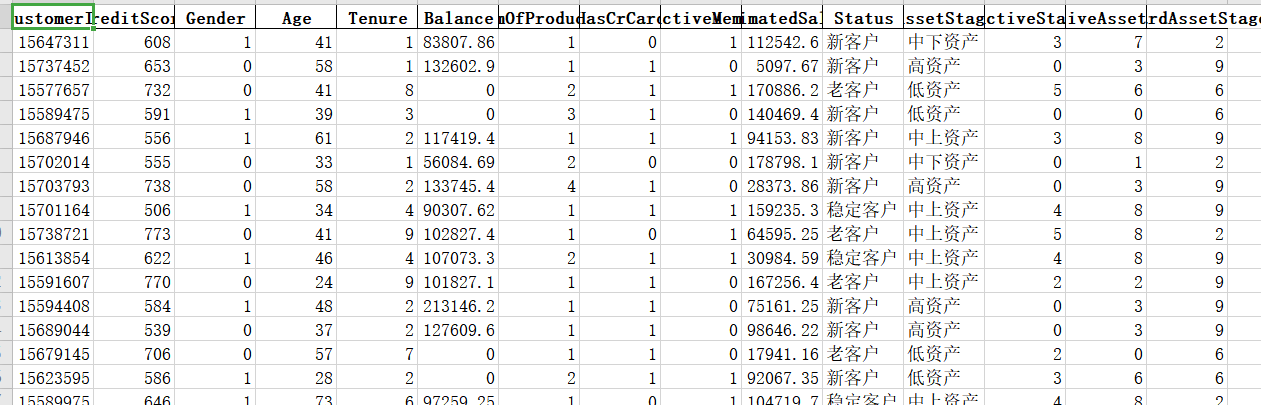
选取依据：年龄大小可作为用户忠诚度依据，年龄低和年龄高的受外界影响导致的流失率低，对忠诚度的影响小，中等年龄因为社会各方面原因导致流失影响大；新老客户活跃程度的特征能够反应用户对该银行忠诚的程度，使用频率等；不同金融资产客户活跃程度的特征能反应不同资产的人对银行使用活跃程度来体现出其各个资产的忠诚程度；不同金融资产信用卡持有状态的特征，反应了用户对银行存款的多少可以体现用户对银行的忠诚程度，由此选取以上四项作为特征依据。

随机森林分类模型建立步骤：

对测试集进行按照任务3.4的模式对其进行对账号户龄（Tenure）和客户金融资产（Balance）进行划分，并分别进行特征编码作为新的客户特征，使用Python进行编码将其保存到“result\_test.xlsx”中。

1. 再按照任务4的模式使用Python进行特征构建，得到新老客户活跃程度的特征、不同金融资产客户活跃程度的特征、不同金融资产信用卡持有状态的特征，将其存储到“result\_test处理.xlsx”中
2. 模型建立原理：一个样本容量为的样本，有放回的抽取次，每次抽取1个，最终形成了个样本。这选择好了的个样本用来训练一个决策树，作为决策树根节点处的样本。当每个样本有个属性时，在决策树的每个节点需要分裂时，随机从这个属性中选取出个属性，满足条件。然后从这个属性中采用某种策略（比如说信息增益）来选择1个属性作为该节点的分裂属性。决策树形成过程中每个节点都要按照步骤2来分裂。
3. 部分处理结果：

表 5 随机森林预测结果表



## 6.2 任务5.2模型评估

Precision（精确值）和recall（召回率）多用于二分类问题，可以结合混淆矩阵介绍，如表1所示。

表 6 混淆矩阵

|  |  |  |  |
| --- | --- | --- | --- |
| 混淆矩阵 | | 预测类型 | |
| Positive | Negative |
| 实际类别 | Positive | TP | FN |
| Negative | FP | TN |

这张表包含了三个方面，真实值、预测值以及预测值和真实值之间的关系，如果知道任意两者都可以确定第三者，显然，TP+FP+FN+TN = 样本总数。

通常取预测值和真实值之间的关系、预测值对矩阵进行划分：

1. TP是被预测为正的正样本
2. TN是被预测为负的负样本
3. FP是被预测为正的负样本
4. FN是被预测为负的正样本

所以对于二维混淆矩阵来说，TP和TN的样本数量尤为重要，这是反映一个模型预测好坏的标准之一。

精确率（p）和召回率（r）的定义为：

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |

在条件允许的情况下，精确率和召回率的值都越高越好。但是事实上，当精确率高时，召回率低；当精确率低时，召回率高，所以就需要F1-score值来进一步说明。

F1-scroe值是精确率和召回率的调和平均值：

|  |  |
| --- | --- |
|  | (3) |

f值可泛化为对精确率和召回率基于不同权重进行的加权调和：

|  |  |
| --- | --- |
|  | (4) |

精确率和准确率是两个不同的的评估指标，精确率是一个二分类指标，而准确率能应用于多分类，其计算公式为：

|  |  |
| --- | --- |
|  | (5) |

针对任务5，给出的随机森林分类模型的精确率和F1-socre以及召回率如下所示：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 准确率 | 召回率 | 精确率 | F1 |
| 训练集 | 0.839 | 0.839 | 0.825 | 0.81 |
| 测试集 | 0.83 | 0.83 | 0.99 | 0.903 |

混淆矩阵热力图如下所示：

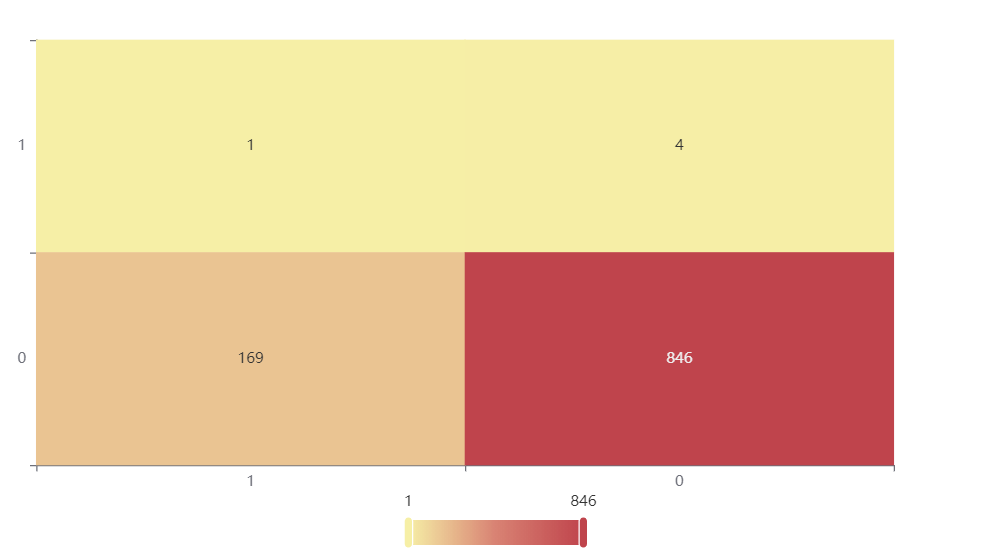


图 6-1 混淆矩阵热力图

图表说明：上表以热力图的形式展示了混淆矩阵。

## 6.3 任务5.3模型预测

随机森林分类模型预测步骤：

1. 对训练集“result4.xlsx”和测试集“result\_test处理.xlsx”进行拼接保存得到“YC.xls”文件。
2. 运用SPSSPRO的随机森林分类进行预测得到预测结果。
3. 将预测结果进行整合，与“result\_test处理.xlsx”联系保存放入“result5.xlsx”中。

测试数据预测评估部分结果如下所示：

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 预测结果Y | Exited | 预测结果概率\_0.0 | 预测结果概率\_1.0 | IsActiveStatus | IsActiveAssetStage | CrCardAssetStage | Age |
| 0 | 0 | 0.9106697025634223 | 0.0893302974365775 | 1 | 0 | 0 | 34 |
| 0 | 0 | 0.7328094813950821 | 0.2671905186049182 | 3 | 6 | 6 | 56 |
| 0 | 1 | 0.6243078246600751 | 0.37569217533992494 | 1 | 2 | 9 | 43 |
| 0 | 0 | 0.7850127338142185 | 0.21498726618578134 | 0 | 2 | 9 | 38 |
| 1 | 1 | 0.4029040475830763 | 0.5970959524169238 | 2 | 1 | 7 | 50 |
| 0 | 0 | 0.5590212174768847 | 0.44097878252311523 | 5 | 8 | 2 | 47 |
| 0 | 0 | 0.8321741558755249 | 0.16782584412447504 | 0 | 2 | 9 | 34 |
| 1 | 0 | 0.44325271844804326 | 0.5567472815519569 | 2 | 2 | 9 | 46 |
| 0 | 1 | 0.7430377360199014 | 0.2569622639800987 | 0 | 2 | 2 | 40 |
| 0 | 0 | 0.7149957939256096 | 0.28500420607439053 | 2 | 1 | 2 | 42 |
| 0 | 0 | 0.8699273534296157 | 0.1300726465703847 | 3 | 8 | 9 | 37 |
| 0 | 0 | 0.7808444433292575 | 0.21915555667074205 | 4 | 8 | 9 | 42 |
| 0 | 0 | 0.8967023885177757 | 0.10329761148222437 | 5 | 8 | 9 | 74 |
| 0 | 0 | 0.8910305935283059 | 0.10896940647169417 | 4 | 8 | 9 | 36 |
| 0 | 1 | 0.8787306609708248 | 0.12126933902917511 | 3 | 7 | 7 | 63 |

部分预测结果如下所示：

表 7 result5.xlsx 部分预测结果

|  |  |
| --- | --- |
| **CustomerId** | **Exited** |
| 15647311 | 0 |
| 15737452 | 1 |
| 15577657 | 0 |
| 15589475 | 0 |
| 15687946 | 0 |
| 15702014 | 0 |
| 15703793 | 1 |
| 15701164 | 0 |
| 15738721 | 0 |
| …… | …… |

表 8 指定的 5 个客户 ID 的预测结果表

|  |  |
| --- | --- |
| CustomerId | Exited |
| 15579131 | 0 |
| 15674442 | 0 |
| 15719508 | 0 |
| 15730076 | 0 |
| 15792228 | 1 |

# 参考文献

[1] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.

[2] 夏邦世, 吴金华. Kappa一致性检验在检验医学研究中的应用[J]. 中华检验医学杂志, 2006, 29(1):83-84.

[3] 周志华. 机器学习[M]. 清华大学出版社, 2016.