**摘要**

随着目前银行产品同质化现象的增多，客户选择产品和服务的途径也越来越

多，因此对产品的忠诚度越来越低。为了解决此类问题，提高客户对银行的忠诚度和银行营销量，商业银行需要转变经营理念。可以从“产品销售导向”业务模式向“以客户为中心”转变，实现共赢。因此需要从客户忠诚度这方面来考虑，客户忠诚度有需要从客户的行为和态度来进行考虑。短期客户忠诚度分析是通过产品的购买数据，分析不同指标客户对银行产品的购买依赖度从而提供更好的销售服务；长期客户忠诚度分析则是从客户资源信息数据中挖掘客户流失因素、预测可能流失的客户，尽可能留住高价值客户。

**针对任务一：**对短期数据和长期数据进行探索和预处理，再对短期数据的字符型数据进行特征编码，利用python相关的机器学习库。对短期数据的缺失值和重复值进行删除处理，对长期数据的异常值也进行删除处理。得到相关结果见附件“result1\_1”、“result1\_2”、“result1\_3”

**针对任务二：**对产品营销数据可视化分析，利用python和相应库我们可以得出相关的图片的可视化。得出结果见下文。

**针对任务三：**对客户流失因素可视化分析，任务三和任务二类似我们只需要利用python对相关问题进行可视化。得出结果见附件“result3.xlsx”相关图片的可视化见下文。

**针对任务四：**特征的构建，针对这一问题我们只需要利用python编程建立循环判断即可得出结果，得出相关结果见附件“result4.xlsx”

**针对任务五：**对银行客户长期忠诚度预测模型的建立，根据我们在生活中所得出的经验建立模型特征，再利用SMOTE过采样方法解决数据类型不平衡问题。利用python相关的机器学习库建立了11种机器学习模型逻辑回归、KNN、GaussianNB、决策树、Bagging、随机森林、ExTree、Adaboost、DBDR、lgb、XGBoost对模型进行训练。利用AUC作为模型的评价标准，最后选出XGBoost方法预测来进行预测。再利用SelectKBest方法进行特征筛选以提高模型的正确性，最后再利用网格搜索法进行调参。得出相关结果见附件“result5.xlxs”

**Abstract**

With the current increase in homogenization of bank products, customers have more and more ways to choose products and services，more and more, and therefore less and less loyalty to products. In order to solve such problems and increase customer loyalty and bank marketing volume, commercial banks need to change their business philosophy. This can be done by shifting from a "product sales-oriented" business model to a "customer-centric" one that achieves a win-win situation. Therefore, customer loyalty needs to be considered in terms of customer behavior and attitude. Short-term customer loyalty analysis is to analyze the product purchase data, analyze different indicators of customers' reliance on bank products and provide better sales services; long-term customer loyalty analysis is to dig out customer loss factors from customer resource information data, predict the potential loss of customers and retain high-value customers as much as possible.

**For task 1:** Exploration and pre-processing of short-term data and long-term data, and then feature coding of character-based data of short-term data, using python related machine learning library. The missing and duplicate values of the short-term data were removed, and the abnormal values of the long-term data were also removed. The related results are shown in the annexes "result1\_1", "result1\_2" and "result1\_3".

**For task 2:** visualization and analysis of product marketing data, using python and the corresponding library, we can visualize the relevant images. The results are shown below.

**For task 3:** visualization and analysis of customer churn factors, similar to task 2, we only need to visualize the related issues using python. The results are attached as "result3.xlsx" and the visualization of the related images is shown below.

**For task 4:** construction of features, for this problem we only need to use python programming to establish a circular judgment to get the results, and the related results can be found in the attachment "result4.xlsx".

**For task 5:** the establishment of a long-term customer loyalty prediction model for banks, we build a feature model based on our experience in life, and then use the SMOTE oversampling method to solve the data type imbalance problem. Eleven machine learning models logistic regression, KNN, GaussianNB, decision tree, Bagging, random forest, ExTree, Adaboost, DBDR, lgb, XGBoost were built to train the models using python related machine learning libraries. AUC is used as the evaluation criterion of the model, and finally XGBoost method prediction is selected for prediction. SelectKBest is then used to filter the features to improve the correctness of the model, and finally the grid search method is used for tuning. The related results are shown in the attachment "result5.xlxs".

**第1章 问题分析**

**1.1问题重述**

（1） 对客户数据进行预处理，并对字符型数据进行特征编码。

（2） 基于短期客户产品购买数据，分析不同指标客户对银行产品的购买依

赖度，并进行可视化呈现。

（3） 基于长期客户资源信息数据，分析客户流失因素，并进行可视化呈现。

（4） 依据长期客户资源信息数据的分析结果构建相关指标，对银行客户长

期忠诚度进行预测。

**第2章 数据探索与清洗**

**2.1任务1.1-1.2数据处理**

本问需要分别对短期客户产品购买数据和长期客户资源信息数据的训练集进行数据探索与清洗。针对任务1.1，需要探索短期数据各个指标数据的缺失值和“user\_id”列的重复值，并且要删除缺失值以及重复值。还需要删除长期数据中的客户年龄“Age”列存在的数值为1,0和“-”的异常值，并删除“Age”列存在空格和“岁”等异常字符但必须得保留年龄数值。针对任务1.2，要对短期数据中的字符型数据进行特征编码，将信用违约情况{‘否’，‘是’}编码变为{0,1}。

**2.1.1任务1.1：数据探索与预处理**

针对本问，我们需要对缺失值和重复列进行清理删除，并且对异常字符进行

删除，这就是数据预处理的过程。然后得到一个完整的，没有缺失值，重复项和异常字符的表格。

**1.短期数据缺失值的处理：**

从表格可以看出，缺失值可能为0，可能为空值，因此需要将这些缺失值进

行查找与删除。首先需要引入Python中的numpy库和pandas库，其次利用命令short\_data.isnull().sum()对数据进行缺失值的个数进行查看。可以得到一共有10698个缺失值。重要代码如下：

short\_data.shape

short\_data.isnull().sum() #缺失值查看

data=short\_data.dropna() #缺失值处理

data.isnull().sum()

data.shape

**表1.短期数据缺失项个数**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **缺失项** | **job** | **marital** | **education** | **default** | **housing** | **loan** |
| **个数** | 330 | 80 | 1730 | 8596 | 990 | 990 |

**2.短期数据重复项的处理：**

从表格可以看出，有重复值，我们以“user\_id”为基础对重复值进行查找，

可以得到一共有33个重复值，重要代码如下：

data.duplicated().sum() #对行

data.duplicated(['user\_id']).sum() #对userid

data=data.drop\_duplicates(['user\_id']) #删除重复值

data.shape

**3.长期数据异常值的处理**

从表格可以看出，有异常值，先将值分别为1,0，‘-’的异常值的个数求出

来，在将其删掉，重要代码如下：

long\_data[long\_data['Age']=='1'].shape #Age为1的个数

long\_data[long\_data['Age']=='-'].shape #Age为‘-’的个数

long\_data[long\_data['Age']=='0'].shape #Age为0的个数

long\_data\_1=long\_data.drop(long\_data[long\_data['Age']=='-'].index) #删除Age为1的异常值

long\_data\_1=long\_data\_1.drop(long\_data\_1[long\_data\_1['Age']=='1'].index) #删除Age为‘-’的异常值

long\_data\_1=long\_data\_1.drop(long\_data\_1[long\_data\_1['Age']=='0'].index) #删除Age为0的异常值

从表格可以看出，有；“Age”列存在空格和“岁”等异常字符，将其存在

列表中，使用for循环将其找出来，重要代码如下：

age\_list=[]

for i in age\_:

age=i.replace(" ", "")

age=age.replace("岁","")

age\_list.append(age)

**2.1.2任务1.2：对短期数据的字符型数据进行特征编码**

针对本问，需要将短期数据中的字符型数据进行特征编码，需要将信用违约情况{‘否’，‘是’}编码为{0,1}。首先对default housing loan contact y d独热编码，就是有效编码，使用N位状态寄存器来对N个状态进行编码，每个状态都有它独立的寄存器位，并且在任意时候，其中只有一位有效。还使用到了二进制。同时还需要将一个特征变量变位计算机能读懂的特征距离，因此使用了pandas.get\_dummies()函数将特征编码变为{0,1}。重要代码如下：

#对default housing loan contact y d独热编码

data['default']=pd.get\_dummies(data[['default']],drop\_first=True,prefix=[None])

data['housing']=pd.get\_dummies(data[['housing']],drop\_first=True,prefix=[None])

data['loan']=pd.get\_dummies(data[['loan']],drop\_first=True,prefix=[None])

data['contact']=pd.get\_dummies(data[['contact']],drop\_first=True,prefix=[None])

data['y']=pd.get\_dummies(data[['y']],drop\_first=True,prefix=[None])

**第3章 产品营销数据可视化分析**

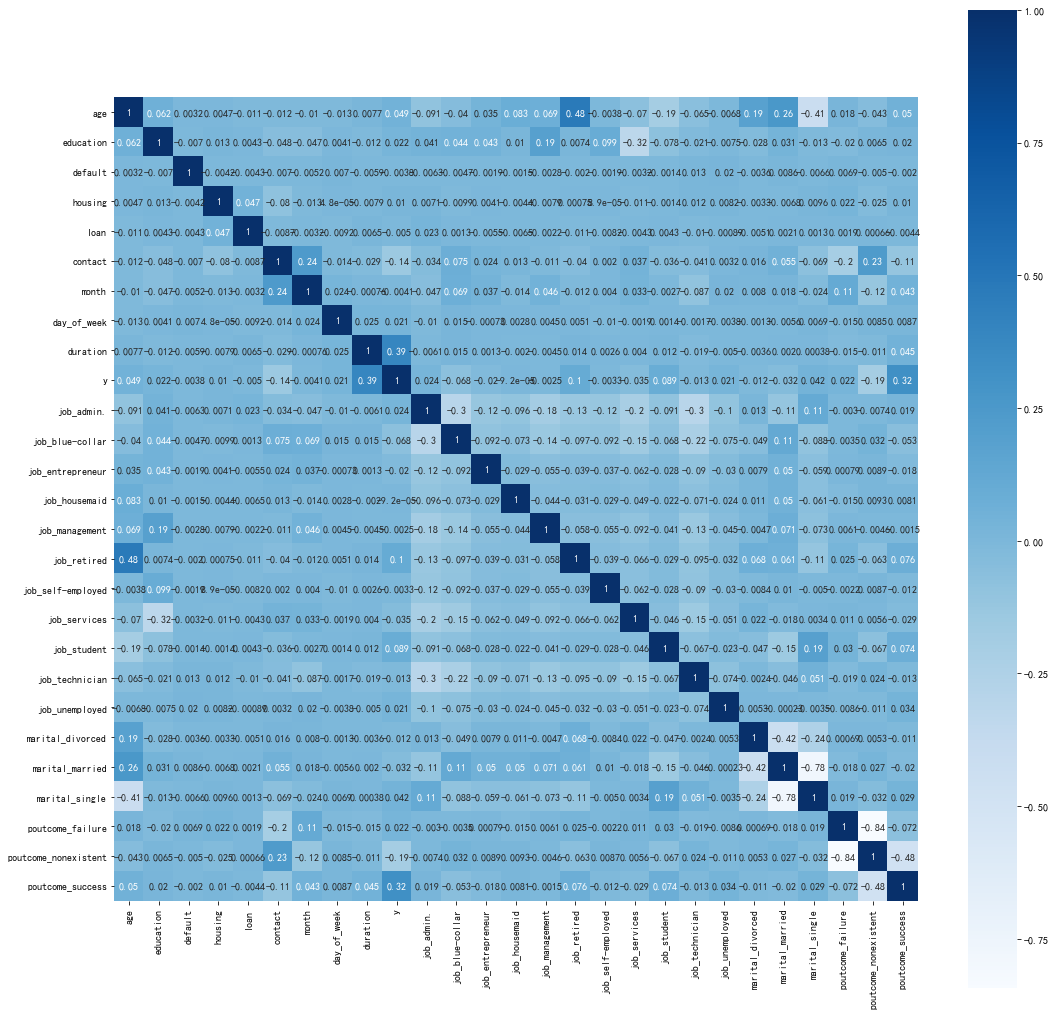
**3.1任务2.1-2.4可视化分析**

本问需要基于短期数据分析不同指标客户与购买银行产品行为的关联性，挖掘短期客户对银行的忠诚度。针对任务2.1，需要计算短期数据所有指标之间的相关性，绘制相关系数热力图。针对任务2.2，需要绘制反映两种产品购买结果下不同年龄客户量占比的分组柱状图并画在同一个图中，要求x为年龄，y为占比数据。针对任务2.3，需要绘制蓝领与学生的产品购买情况并画在同一个图中。针对任务2.4，以产品购买结果为x轴，拜访客户的通话时长为y轴，绘制拜访客户的通话时长箱线图。

**3.1.1任务2.1绘制相关系数热力图**

针对本问，是需要基于上一道题的数据处理过后的值来进行作图。使用corr

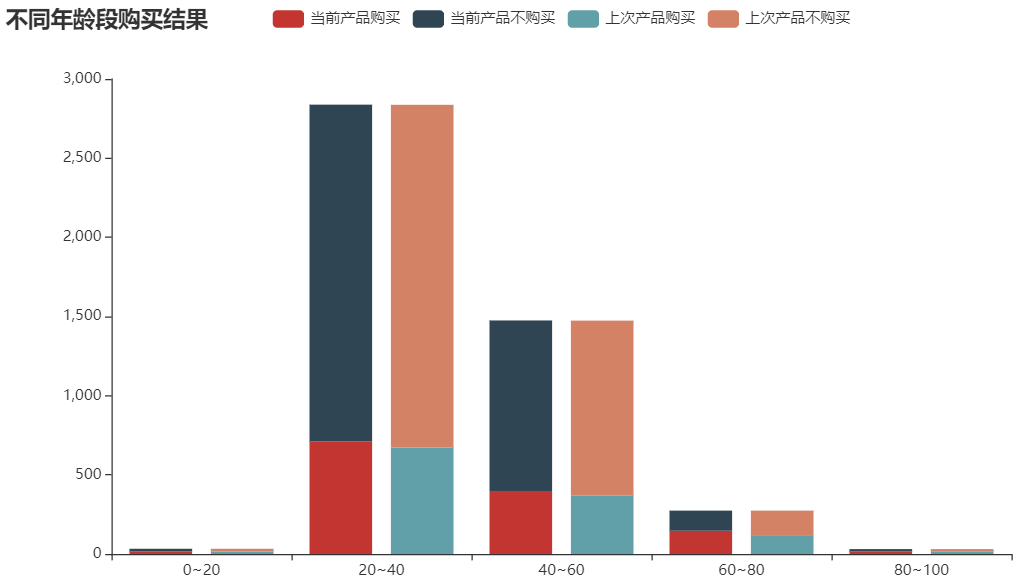
（）函数，给出相关系数矩阵，得到短期数据所有指标之间的相关系的热力图。具体如下图所示：

****

**图1.相关系数热力图**

**3.1.2任务2.2绘制反映两种产品购买结果下不同年龄客户量占比的分组柱状图**

通过分析可得，以x轴为年龄，y轴为占比数值，具体分组柱状图如下所示：



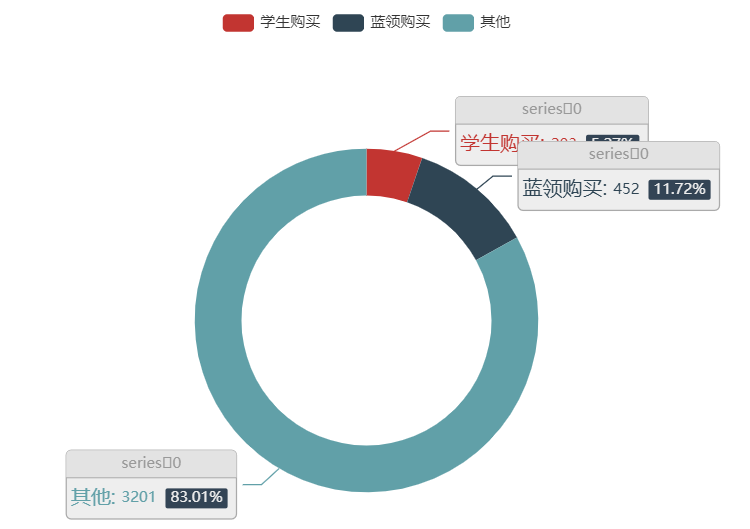
**图2.两种产品购买结果下不同年龄客户量占比的分组柱状图**

可以看出年龄在20-40的时候购占比数值最高，并且主要为当前产品不购买和上次产品不够买占比较大。其次为年龄在40-60岁的，同样还是当前产品不购买和上次产品不够买占比较大。第三个为年龄在60-80的。因为年龄过大和年龄过小，年龄在80-100和0-20的购买结果占比值较低。可以从图中看出该图类似于正态分布图。

**3.1.3任务2.3绘制蓝领与学生的产品购买情况饼图**

通过分析可计算出蓝领购买数量为占比量为452，学生购买占比量为203，

其他购买占比量为3201，因此可以画出相对应的产品购买情况饼图，具体产品购买情况饼图如下所示：

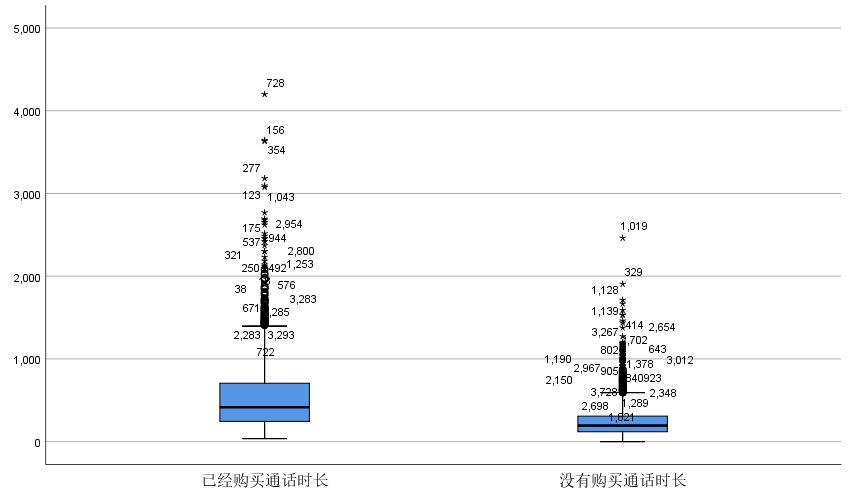


**图3.蓝领与学生的购买情况饼图**

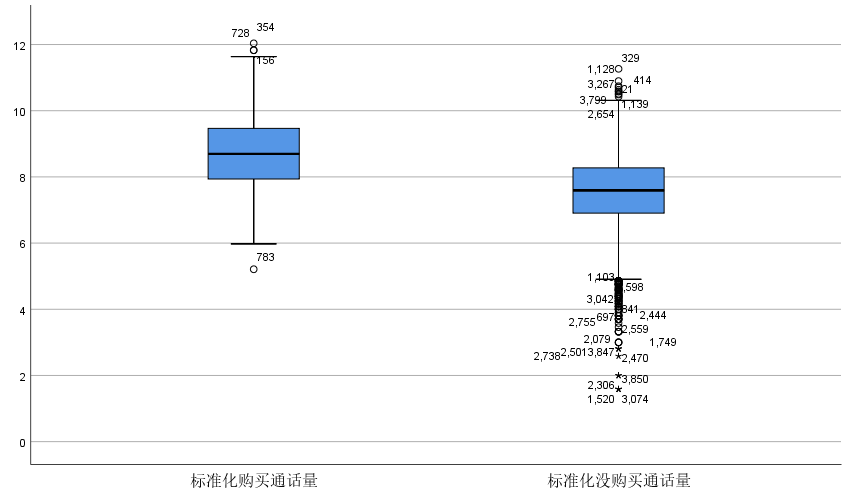
从图中可以看出，蓝领购买所占比例为11.72%，学生购买占比为5.27%，其他占比为为83.01%

**3.1.4任务2.4绘制拜访客户的通话时长箱线图**

因为考虑到该数值为离散型数据，因此考虑用箱线图来进行可视化分析。还可以看出哪里是异常值。我们将一组数据按照从小到大的顺序排列，把该数据四等分第一四分位数的数值为25%，第二四分位数也叫中位数为50%，第三四分位数为75%，而第三四分位数与第一四分位数的差距叫四分位距。



**图4.未标准化的拜访客户的通话时长箱线图**



**图5.标准化过后的拜访客户的通话时长箱线图**

图4表示的是未将数据标准化的时候拜访客户的通话时长箱线图，可以看出这时已购买通话时长和没有购买通话时长的异常值很多。因此对数据进行标准化处理，可以通过图5看出，数据标准化过后购买通话的量异常值几乎没有，而没购买通话量的异常值还是居多。在后续建立模型是要对相关数据集进行预处理。

**第4章 客户流失因素可视化分析**

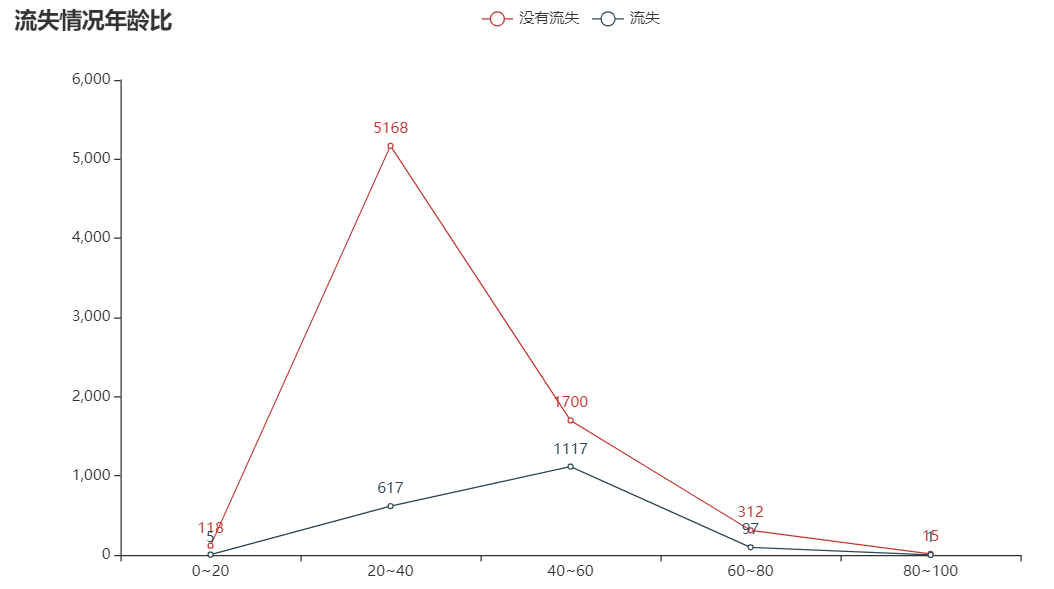
**4.1任务3.1-3.4可视化分析**

在本问中，需要基于长期数据分析导致银行客户流失的因素来进行可视化的分析，针对任务3.1，需要绘制反映两种流失情况下不同年龄客户量占比的折线图。针对任务3.2，需要绘制反映两种流失情况下客户信用资格与年龄分布散点图。针对任务3.3，需要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并且绘制反映两种流失情况的客户各账号户龄占比量的堆叠柱状图。针对任务3.4，需要对新老客户各资产阶段的客户流失情况分析。

**4.1.1任务3.1绘制两种流失情况下不同年龄客户占比折线图**

在本问中，通过处理好的数据，让x轴为年龄，y轴为信用资格来对两种流

失情况下不同年龄的客户占比进行画图，如下所示：

****

**图6.两种流失情况下不同年龄客户占比折线图**

通过折线图可以看出，年龄在20-40的时候没有流失的情况占比最高，其次

是年龄为40-60的，年龄为60-80的，年龄为0-20的，年龄为80-100的。年龄在40-60的流失的情况占比最高，其次是年龄在20-40的，年龄在60-80的，年龄在0-20的，年龄在80-100的

**4.1.2任务3.2绘制两种流失情况下客户信用资格与年龄分布散点图**

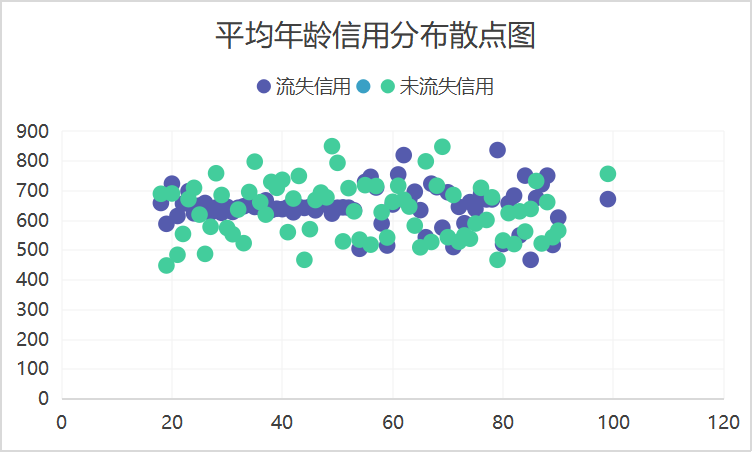
考虑到数据的冗杂，我们可以利用计算平均用户信用度来对两种流失情况下

客户信用资格与年龄进行分析绘制散点图。因此可以设表示每个年龄段，设为第个年龄段对应的用户，代表第个年龄段对应的第个用户的信用度，代表平均用户信用度，因此可以建立如下模型：



通过模型的建立与求解，可以得出相对应的数据并且画出平均年龄信用分布

散点图，具体如下图：

****

**图7.平均年龄信用分布散点图**

可以从图中看出，流失信用对应的年龄集中的规律的分布在20-60，0-20几

乎没有，60-100之间分布的离散。对于未流失信用的年龄集中分布在20-90之间。0-20和90-100几乎没有。

**4.1.3任务3.3构造包含账号户龄在不同流失情况下的客户量占比透视表以及绘制反映两种流失情况的客户各账号户龄占比量的堆叠柱状图**

针对本问，要构造占比透视表以及占比量的堆叠柱状图需要根据前面整理出

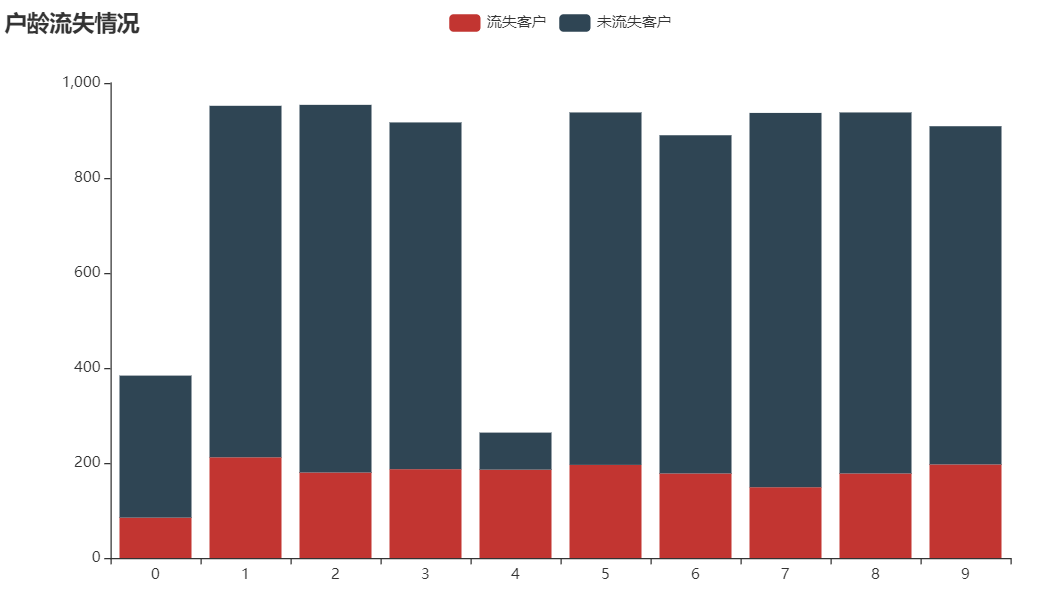
来的数据进行绘制。其中堆叠柱状图的x轴为客户的户龄，y轴为占比量。

**表1.账号户龄在不同流失情况下的客户量占比透视表**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Tenure**  **Exited** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| 0 | 85 | 212 | 180 | 187 | 186 | 196 | 178 | 149 | 178 | 197 |
| 1 | 299 | 740 | 774 | 730 | 728 | 742 | 712 | 788 | 760 | 712 |

通过表格可以看出未流失的都比流失的多，并且户龄为7的时候未流失的占

比最大，户龄为5的时候，流失的占比最大。

****

**图8.两种流失情况的客户各账号户龄占比量的堆叠柱状图**

通过图可以看出未流失的都比流失的多，并且户龄为7的时候未流失的占

比最大，户龄为5的时候，流失的占比最大。

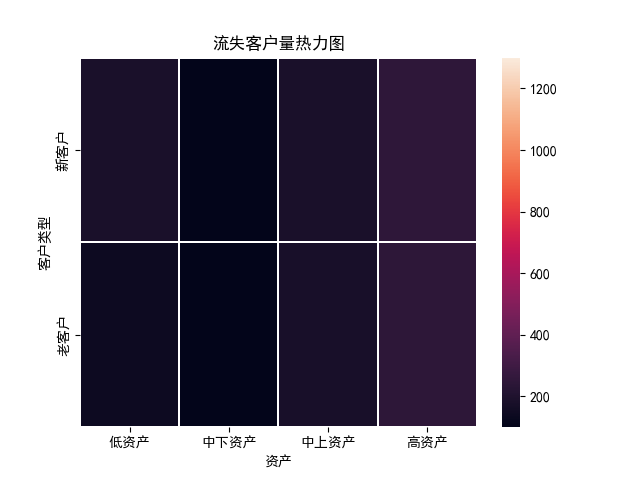
**4.1.4任务3.4新老客户各资产阶段的流失情况分析**

针对本问，需要按照题目中的表5表6对账号户龄和客户金融资产进行划分。

对于账号户龄来划分，在[0,3]之间为新客户，(3,6]之间为稳定客户，>6为老客户。对于客户紧张资产来划分，在[0,50000]为低资产，(50000,90000]之间为中下资产，(90000,120000]之间为中上资产，>120000为高资产。通过建立列表和使用for循环得到相对应的新老客户各资产阶段的流失情况：

**表12部分新老客户各资产阶段的流失情况**

|  |  |  |
| --- | --- | --- |
| **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 |
| …… | …… | …… |
| 15815645 | 老客户 | 高资产 |
| 15815656 | 老客户 | 中上资产 |
| 15815660 | 新客户 | 高资产 |
| 15815690 | 新客户 | 中上资产 |

****

**图9.新老客户在各资产阶段中流失的客户量的热力图**

**第5章 特征构建**

**5.1任务4.1特征构建**

在本问中，需要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。在这里，需要构建新老客户活跃程度表，不同存款客户活跃程度特征表，不同金融资产信用卡持有状态特征表。通过前面数据的整合，使用for循环，考虑到活跃程度、资产和信用卡持有状态来进行分析得到不同情况下的新老客户活跃程度值，不同金融资产客户活跃程度值，不同金融资产信用卡持有状态值，以下为部分值：

**表3.特征构建后**

|  |  |  |  |
| --- | --- | --- | --- |
| **CustomerId** | **IsActiveStatus** | **IsActiveAssetStage** | **CrCardAssetStage** |
| 15553251 | 3 | 9 | 9 |
| 15553256 | 5 | 6 | 6 |
| 15553283 | 2 | 2 | 5 |
| 15553308 | 0 | 0 | 6 |
| 15553387 | 0 | 0 | 0 |
| **……** | **……** | **……** | **……** |
| 15815628 | 2 | 2 | 5 |
| 15815645 | 5 | 9 | 9 |
| 15815656 | 5 | 8 | 9 |
| 15815660 | 3 | 9 | 9 |
| 15815690 | 3 | 8 | 9 |

**第6章 银行客户长期忠诚度预测建模**

**6.1任务5.1预测建模**

在本问中，由于长期数据存在特征分布不均衡，各项数值分布跨度大的现象，

有些未流失客户量是已流失客户量的3倍以上；有些客户信用资格最大数值达到25万，但客户活动状态为0和1等。所以要进行预测。

**一、特征的构建**

首先需要训练数据要使用异常处理后的数据，前面已经将异常值处理完成，

目的是为了提高模型的准确性。通过分析可得，一共建立了两个特征工程：①每户龄的平均资产。②每户龄的平均产品数量。有了特征工程的建立，就可以开始建立模型。

##每户平均资产的建立

zi=[]

for i,j in zip(all\_data['Balance'].values.tolist(),all\_data['Tenure'].values.tolist()):

if j==0:

zi.append(0)

else:

zi.append(i/j)

all\_data['Balance\_per']=zi

##每户龄平均产品数量

num =[]

for i,j in zip(all\_data['NumOfProducts'].values.tolist(),all\_data['Tenure'].values.tolist()):

if j==0:

num.append(0)

else:

num.append(i/j)

all\_data['NumOfProducts\_per']=num

**二、模型的构建**

**1.首先需要利用SMOTE综合采样人工合成数据算法解决数据类别不平衡的问题**，具体算法如下：

##平衡数据

smote = SMOTE(random\_state = 402)

X\_smote, Y\_smote = smote.fit\_resample(train\_data\_1,target)

sns.countplot(Y\_smote, edgecolor = 'black')

**2.利用相关机器学习库建立模型**

一共建立了11个机器学习方法，分别为逻辑斯蒂回归模型，KNN模型，GaussianNB模型，决策树模型，bagging模型，随机森林模型，ExTree模型，Adaboost模型，GBDT模型，lgb模型和XGboost模型。通过模型的建立并且对模型的测试值Train AUC Score和Test AUC Score分别进行比较，可以得到AUC越大，效果越好，正确率越高，具体结果见下表所示：

**表4.不同机器学习对应的AUC值**

|  |  |  |
| --- | --- | --- |
| 机器学习方法 | **Train AUC Score** | **Test AUC Score** |
| 逻辑斯蒂回归模型 | 0.770 | 0.780 |
| KNN模型 | 0.967 | 0.911 |
| GaussianNB模型 | 0.822 | 0.825 |
| 决策树模型 | 1.0 | 0.849 |
| bagging模型 | 0.920 | 0.826 |
| 随机森林模型 | 0.913 | 0.915 |
| ExTree模型 | 1.0 | 0.956 |
| Adaboost模型 | 0.929 | 0.931 |
| GBDT模型 | 0.928 | 0.929 |
| lgb模型 | 1 | 0.955 |
| XGboost模型 | 0.973 | 0.955 |

通过以上表格可以分析得到，决策树模型、ExTree模型和lgb模型的训练值

都为1，说明该方法过拟合，不采用。通过对其余模型的分析比较，最终选择使用XGboost模型。

**3.特征筛选**

需要对训练数据进行特征的筛选，提高正确性。利用SelectBest()函数进行

筛选，使用for循环相关数据及特征。重要代码如下所示：

def select(train,goal):

a\_score=[]

b\_score=[]

for i in range(2,train.shape[1]+1):

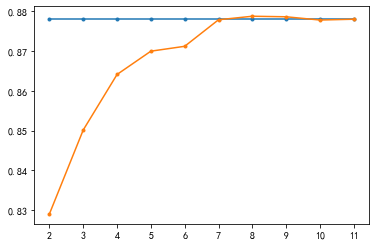
sel = SelectKBest(mutual\_info\_classif, k=i)

sel = sel.fit(train, goal)

train\_sel = sel.transform(train)

print('训练数据未特征筛选维度', train.shape)

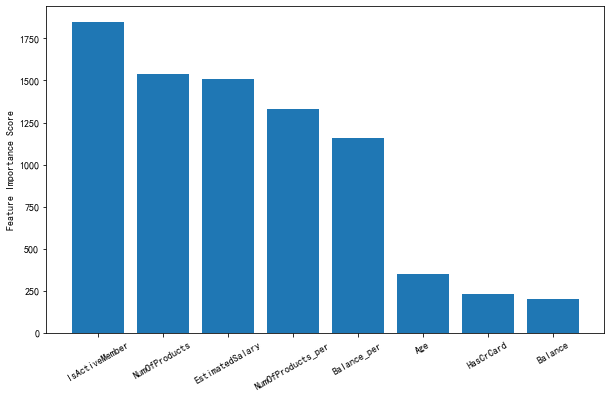
print('训练数据特征筛选维度后', train\_sel.shape)

****

**图10.选取特征数量个数**

图中横坐标代表了特征数量个数，纵坐标代表精确率，橙色的线代表对应最

优特征数量的精确率，蓝色的线代表11个特征的精确率。通过橙色的线可以看出，特征数量为2的时候，精确率只能达到83.95%左右。还可以看出，当特征数量个数为8的时候，精确率最高。

****

**图11.选取特征数量个数为8时的模型特征重要度**

**4.模型调参**

通过网格搜索法对模型进行调参，可以了解到，网格搜索法是制定参数值的

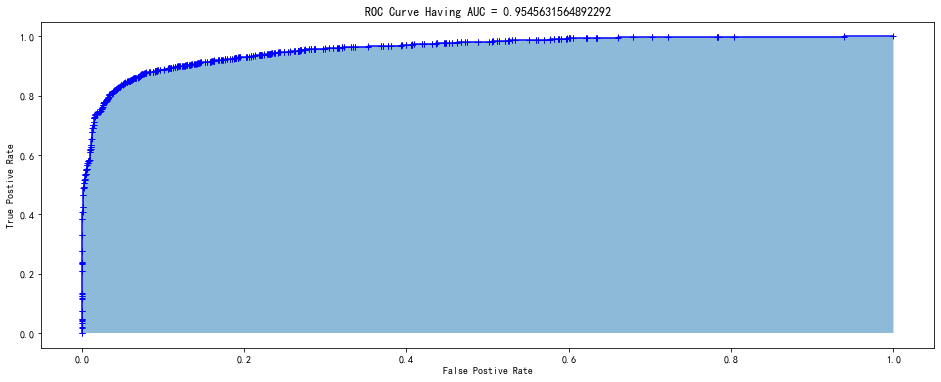
一种穷举搜索方法，通过将估计函数的参数通过交叉验证的方法进行优化来得到最优的学习算法。调完模型参数过后，再XGboost模型对训练值和测试值进行预测，可以得到如下值：

**表5.模型调参过后的训练值和测试值**

|  |  |  |
| --- | --- | --- |
| **机器学习方法** | **Train AUC Score** | **Test AUC Score** |
| XGboost | 0.9999913 | 0.9579750 |

可以从表格当中看出训练集的AUC值为0.9999913，测试集的AUC值为

0.9579750，

****

**图12.ROC图**

可以从ROC图看出其AUC值为0.955

**二、混淆矩阵**

**表6.最终模型的混淆矩阵**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **标签** | **precision** | **recall** | **F1-score** | **support** |
| 0 | 0.96 | 0.96 | 0.96 | 2192 |
| 1 | 0.96 | 0.96 | 0.96 | 2214 |

标签为1时也就是流失时，预测的正确性为0.96，召回率为0.96，F1的值

为0.96，支持他的数据为2192。标签为0时也就是无流失时，预测的正确性为0.96，召回率为0.96，F1的值为0.96，支持他的数据为2214。

**三、5个客户ID的预测结果**

**表7.5个客户ID预测结果**

|  |  |
| --- | --- |
| **CustomerID** | **Exited** |
| 15579131 | 0 |
| 15674442 | 0 |
| 15719508 | 1 |
| 15730076 | 1 |
| 15792228 | 1 |