聚知刊出版社
JZK publishing

银行客户流失预测

程烁文

青海民族大学，青海西宁，810000；

摘要：本文主要研究了银行客户流失预测问题。为了探究客户流失的原因，本研究选取了年龄、性别、所处国家、姓氏作为核心解释变量，并控制了 CreditScore、Tenure、Balance、NumOfProducts、HasCrCard、IsActiveMember、EstimatedSalary 等冗余变量。使用 Logistic 回归和随机森林模型对客户流失进行因果推断，并在训练集上启用 5 折交叉验证。经过分析，年龄对银行客户流失的影响存在门槛效应，在 50-60 岁之前，年龄越大的客户越容易流失；而在 50-60 岁之后，年龄越大的客户粘性越大。本文还发现客户的性别以及所处国家（或者说所说的语言）对客户流失也有显著的异质性影响：女性客户比男性客户更容易流失，说德语的客户比说法语以及西班牙语的客户更容易流失。最后，本文研究发现，客户的姓氏对客户流失并无显著影响。本文利用 logistic 回归、随机森林模型，从多个角度验证了结论，并且还使用替换代理变量的方法使得 logistic 回归的结论通过了稳健性检验。综上所述，本文从多种角度对银行客户流失进行了因果推断，结果表明年龄、所处国家和性别等因素对客户流失具有重要影响。本研究可为银行客户流失预测提供参考，并对相关领域的研究有所贡献。

关键词：银行客户流失；logistic 回归；机器学习；随机森林

DOI：10.69979/3029-2700.25.03.010

引言

随着技术的不断进步和市场的不断成熟，银行产品和服务的差异越来越小，以生产为中心、以销售产品为目的的管理理念逐渐被以客户为中心、以服务为目的的新思维所取代。银行通过准确掌握客户的行为趋势，有效发掘和管理客户资源，就能获得市场竞争优势，在激烈的市场竞争中立于不败之地。银行客户流失特别是优质客户的流失，一直是银行在发展中面临的重大挑战。优质客户的流失及联带效应减少了银行在优质客户市场的份额，造成了银行收益的损失。因此，进行银行客户流失分析，并探究客户流失的原因，从而积极主动地进行有针对性的维系挽留工作，正成为各家银行的重点需求。

目前国内外研究文献中，对银行客户流失的研究主要集中在流失预警模型和算法的研究上，因此本文以影响银行客户流失的因素为基础，探究客户流失的原因。本文假设年龄、性别、所处国家、姓氏四个因素会对客户流失造成影响，使用 Logistic 回归和随机森林模型对客户流失进行因果推断，从而找出真正的影响因素以及这些变量对客户流失的影响程度。

1 文献综述

1.1 国外研究现状

近年来，机器学习逐渐被应用于高风险领域，若模型预测出现错误，造成影响也是重大的，这意味着本文不仅要知道模型预测结果，还要分析模型做出这个预测的原因。而数据挖掘可以看作是一门应用型学科，用来发现事物存在的更多价值。在实际应用中，更多的是体现在商业价值中。在国外，许多银行为预测客户流失情况，都使用了数据挖掘工具，其中有很多案例都取得了不错的效果。在机器学习和数据挖掘算法用于预测银行客户流失情况较少的时候，DoD[1]等在对客户流失情况进行预测时，主要是解决了客户特征数据不平衡问题，采用 SMOTE 算法，该方法主要是增加少数类样本的数目来使得不同分类数据量平衡，通过使用该方法，在客户流失预测上获得了一定的效果；在 2012 年，RezaAllahyariSoeini[2]使用了分类算法和聚类算法，将两种方法相融合，通过这种模型，对银行客户流失情况预测，并且分析了银行客户流失的原因，提出了挽留客户的建议和方案；VafeiadisT[3]等人通过对支持决策树、向量机、朴素贝叶斯[4]、Logistic[5]和人工神经网络这五种比较典型的机器学习分类方法进行实验和比较，最终分析得出这几种模型在预测分类时的效果；Lu[6]等人对客户预测模型准确度的提高探索出利用 Boosting 算法，基于 L

ogistic 的 Boosting 算法将客户划分为两个群体，对每个客户群体建立流失预警模型，能够较好的对流失数据进行分割，更好的确定具有较高流失概率的客户；VafeiadisT[7]等人通过对包括人工神经网络、支持向量机、决策树、朴素贝叶斯和 Logistic 回归在内的典型的机器学习分类方法进行了对比分析研究，得到了这五种分类模型的预测效果。

1.2 国内研究现状

数据挖掘技术在国内银行也有很广泛的应用，并且国内学者对数据挖掘在客户流失预测方面的应用也有了一定的研究，取得了显著的成果。方匡南[8]等人在构建个人信用风险评估模型时，将 Lasso[9]和 Logistic 模型进行结合，基础模型的精度得到了明显提升；李会[10]等人使用决策树构建流失预测模型，通过二项式回归算法对客户流失因素进行分析；柳婷[11]通过采用 Logistic 模型查看客户流失预警风险，得出了客户流失的概率，从而减少客户流失；郑宇晨[12]等人使用 K-均值算法来得出客户的流失状态，依据客户的交易指标情况和业务特点等对特征进行筛选，以筛选出来的特征为基础，构建出 Logistic 模型，再对流失客户进行预测，最后根据预测结果提出有效地预防客户流失建议，取得了不错的效果。从以上研究可以看到有很多不足之处：算法的选择和模型构建比较简单，且分析客户的数据比较局限。高海燕[13]通过以数据挖掘和客户关系管理为基础，以银行业为背景，基于 SAS9.1 数据挖掘平台，提出了运用 Logistic 回归模型进行客户流失预警，得到了客户的流失概率，为银行的经营提供量化支持；李长山[14]，罗晓光[15]等人运用 Logistic 回归法对企业和商业银行财务风险构建了预警模型，通过因子分析法对影响银行风险状况的财务指标赋权，进行量化评估，为企业提前应对财务风险提供了保障；兰军[16]提出了基于客户综合特征进行客户分群的新研究方法，将银行客群细分为七个客群，运用客户流失分析方法，在实际的银行客户数据集上进行实验并进行流失观察，为系统性分析商业银行客户行为提供了新的思路和手段。

2 实证模型

2.1 Logistic 回归

本文首先采用 logistic 回归来对年龄、性别、所处国家、姓氏和银行客户流失进行因果推断. Logistic 回归是一种广泛应用于二分类问题的统计模型。它可以被用于分析哪些因素对于某二元变量(如银行客户是否流失)的发生有影响，并提供预测新数据的能力。Logistic 回归的模型形式如下：

$$\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k$$

其中，p 表示二元变量发生的概率，x1, x2, • • •, xk 表示解释变量， β 0, β 1, β 2, • • •, β k 表示未知参数。左边的式子也可以写成：

$$p=\frac{1}{1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k\right)\right)}$$

本文旨在使用 Logistic 回归模型预测银行客户是否会流失，并探究年龄、性别、所处国家和姓氏系数等解释变量对客户流失的影响。

本文使用了一份银行客户流失数据集，其中包括了客户的年龄、性别、所处国家和姓氏系数等解释变量、CreditScore、Tenure、Balance、NumOfProducts、HasCrCard、IsActiveMember、EstimatedSalary 等冗余变量, 以及客户是否流失的二元变量。本文使用 Python 中的 scikit-learn 库将数据集按照 80:20 的比例随机拆分为训练集和测试集，并在训练集上拟合 Logistic 回归模型。

本文使用了 scikit-learn 库中的 LogisticRegression 类，在训练集上拟合了一个 Logistic 回归模型。模型的形式如下：

$$P(y=1|x)=\frac{e^{\beta_0x_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_{control}}}{\left[1+e^{\beta_0x_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_{control}}\right]} \quad (1)$$

其中，y 为二元变量表示客户是否流失，x1 至 x4 分别表示客户的年龄、性别、所处国家和姓氏, 为核心解释变量 xcontrol 为控制变量; β 1 至 β 5 是待估计的系数。

为了评估模型的拟合效果和预测能力，本文采用了 5 折交叉验证的方法。具体地，本文将训练集按照 5 个子样本的方式分割，并在每个子样本上训练模型，在剩余的子样本上进行预测。

2.2 随机森林模型

本文使用 Python 中的 scikit-learn 库中的 RandomForestClassifier 类来构建随机森林模型。随机森林模型的超参数包括了树的个数、每棵树的最大深度、每

个节点最少样本数等。本文使用 GridSearchCV 函数在训练集上搜索最优的超参数组合，并在验证集上进行模型评估。具体地，本文设置了树的个数 n_estimators 的范围为[50,100,150,200,300,400,500,600]，每棵树的最大深度 max_depth 的范围为[5,10,15]，每个节点最少样本数 min_samples_leaf 的范围为[3,5,7,9,12]。

同样的,本文使用了 5 折交叉验证的方法来评估随机森林模型的预测能力,将训练集按照 5 个子样本的方式分割，并在每个子样本上训练模型，在剩余的子样本上进行预测。

3 数据描述与处理

在此处撰写正文。

按 Ctrl+=插入公式，如公式 1 所示（自动编号）。

$$\begin{pmatrix} x_{1t} \\ x_{2t} \\ \cdots \\ x_{pt} \end{pmatrix} = f\begin{pmatrix} x_{1,t-1} & x_{1,t-2} & \cdots & x_{1,t-k} \\ x_{2,t-1} & x_{2,t-2} & \cdots & x_{2,t-k} \\ \cdots & & & \\ x_{p,t-1} & x_{p,t-2} & \cdots & x_{p,t-k} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdots \\ \epsilon_p \end{pmatrix} \#(1)$$

where$\epsilon_1, \epsilon_2, \ldots, \epsilon_p$arerandomvaluesfollowsadistributionsatisfying$E(\epsilon) = 0$.

自动编号的方法：在公式末尾输入#()然后在该英文括号中按 Ctrl+F9（如果你的 F1-F12 是媒体按键，请根据电脑型号锁定媒体案件，例如 Fn+Shift），输入 seqeq 然后再次按 Ctrl+F9.

3.1 数据来源

数据来源于一份银行客户流失数据集。

3.2 数据来源

1.对于客户姓氏、所处国家、性别这三个字符串数据,本文采用的方法是将它们转变为虚拟变量进行转化为数值型数据。

2.对于剩下的自变量,本文采取了归一化的手法。这是一种常见的平滑数据的手法,首先从 sklearn.preprocessing 模块中导入 MinMaxScaler 类,然后创建了一个名为 mms 的 MinMaxScaler 对象。最后，使用 fit_transform()方法对这些自变量进行了处理，将其转换为在[0,1]范围内的浮点数。也即以下公式：

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

其中，X 为原始数据，Xnorm 为归一化后的数据，Xmin 和 Xmax 分别为数据的最小值和最大值。

表 1:本表格利用 stata 得出

Variable	Obs	Mean	Std. dev.	Min	Max	Missing values
Age	10,000	38.9218	10.48781	18	92	No
Geography	10,000	1.7463	0.8275293	1	3	No
Gender	10,000	1.5457	0.497932	1	2	No
Surname	10,000	1508.774	846.2043	1	2932	No
Tenure	10,000	5.0128	2.892174	0	10	No
IsActiveMe r	10,000	0.5151	0.4997969	0	1	No
Balance	10,000	76485.89	62397.41	0	250898.1	No
NumOfProdu s	10,000	1.5302	0.5816544	1	4	No
HasCrCard	10,000	0.7055	0.4558405	0	1	No
CreditScore	10,000	650.5288	96.6533	350	850	No
EstimatedS y	10,000	100090.2	57510.49	11.58	199992.5	No

3.3 描述性统计

上述是本文的描述性统计表,需要注意的就是 Gender、Surname、Geography 这三个虚拟变量。

$$Gender = \begin{cases} 1, & Gende = Female \\ 2, & Gende \neq male \end{cases} \quad (3)$$

$$Geography = \begin{cases} 1, & Geography = France \\ 2, & Geography = Germany \\ 3, & Geography = Spain \end{cases} \quad (4)$$

至于 Surname 就是将相同姓氏的人分为一组,一共分了 2932 组。

4 实证结果

4.1 多重共线性检验

由于存在大量的冗余变量,所以可能会导致模型中出现多重共线性的问题,如果存在多重共线性,那么本文就需要利用因子分析法提取自变量中的公因子进而解决多重共线性。所以为了严谨,本文有必要去进行多重共线性的检验。(表 2)

本文利用经典的 VIF 检验法进行检验,如果 VIF<5 就说明模型比较稳定,不存在明显的多重共线性;如果 VIF>10,则说明存在严重的多重共线性。本文的检验的 VIF 显著小于 5,故数据不存在明显的多重共线性问题。

表 2:本表格利用 stata 得出

Variable	VIF	1/VIF
Balance	1.11	0.901432
NumOfProdu s	1.10	0.905064
Age	1.01	0.989671
IsActiveMe r	1.01	0.990109
Geography	1.01	0.993328
Gender	1.00	0.997672
Tenure	1.00	0.997789
Surname	1.00	0.998596
HasCrCard	1.00	0.998758
CreditScore	1.00	0.998894
EstimatedS y	1.00	0.998931
Mean VIF	1.02	

表 3:本表格利用 python 得出

variable	coef	std err	z	P>|z|	[0.025	0.975]
const	-3.6630	0.195	-18.778	0.000	-4.045	-3.281
Age	0.0718	0.003	25.380	0.000	0.066	0.077
Gender	-0.5428	0.060	-9.025	0.000	-0.661	-0.425
Geography	0.1068	0.037	2.911	0.004	0.035	0.179
Surname	-4.914e-05	3.54e-05	-1.386	0.166	-0.000	2.03e-05
CreditScore	-0.3451	0.155	-2.222	0.026	-0.649	-0.0412
Tenure	-0.1433	0.104	-1.382	0.167	-0.347	0.060
Balance	1.2179	0.128	9.478	0.000	0.966	1.470
NumOfProducts	-0.1451	0.156	-0.933	0.351	-0.450	0.160
HasCrCard	-0.0084	0.065	-0.128	0.898	-0.137	0.120
IsActiveMember	-1.0723	0.064	-16.812	0.000	-1.197	-0.947
EstimatedSalary	0.0691	0.105	0.659	0.510	-0.136	0.275

这里需要注意的是:大样本情况下 Z 值和 T 值差别很小,所以此时一般会用 Z 值来代替 T 值进行显著性检验

4.2 Logistic 回归

本文使用 logistic 回归发现：姓名对客户流失不显著，客户年龄对客户流失显著为正、客户所处国家对客户流失显著为正、客户性别对客户流失显著为负。(表格 3)并且对于整体的 F 检验结果为：F(11, 9988)=143.70，Prob>F=0.0000 从目前的实证结果本文可以粗略的得出以下结论：

1 由于客户年龄对客户流失在百分之一的条件下显著为正，所以当年客户龄增长 1%，客户流失率就会提高 e^7.18%。

2 由于客户的性别对客户流失在百分之十的条件下显著为负，所以客户性别对客户流失是有显著影响的。关于具体的男性还是女性客户更容易流失这个问题，本文在下面的异质性分析部分进行回答。

3 由于客户所处国家对客户流失在百分之五的条件下显著为正，所以客户所处国家对客户流失是有显著影响的，关于具体的法国、西班牙、德国哪个国家的客户更容易流失这个问题，本文同样在下面的异质性分析部分进行回答。

4 由于客户的姓氏对客户流失不显著，所以本文可以认为客户的姓氏对银行客户流失无关。

4.3 稳健性检验-替换代理变量

本文采用最为经典的替换代理变量的方法对上述的 logistic 回归进行稳健性检验。本文之前是将每位客户的姓氏都作为一个虚拟变量进行回归；而现在本文进行如下处理：

1. 将每位客户的姓氏的首字母提取出来生成新的字符串变量 first_letter；

2. 本文将改字符串变量转换为虚拟变量 Surname2

这样操作，本文就把之前的 10000 位客户产生的 2932 个分组构成的虚拟变量 Surname，转换为了 26 个分组构成的虚拟变量 Surname2。此时，本文再次进行 logistic 回归，结果如下(表格 4)

表 4:本表格利用 stata 得出

variable	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0332847	0.0369498	-0.90	0.368	-0.1057137	0.0391444
Age	0.0112916	0.0003589	31.46	0.000	0.0105881	0.0119952
Gender	-0.0774357	0.0075292	-10.28	0.000	-0.0921944	-0.0626769
Geography	0.0114415	0.0045403	2.52	0.012	0.0025416	0.0203414
Surname2	-5.93e-06	4.43e-06	-1.34	0.180	-0.0000146	2.75e-06
CreditScore	-0.0000929	0.0000388	-2.40	0.017	-0.0001689	-0.0000169
Tenure	-0.0018938	0.0012962	-1.46	0.144	-0.0044346	0.000647
Balance	6.82e-07	6.32e-08	10.79	0.000	5.58e-07	8.06e-07
NumOfProducts	-0.0048697	0.0067672	-0.72	0.472	-0.0181348	0.0083954
HasCrCard	-0.0028992	0.00822	-0.35	0.724	-0.019012	0.0132137
IsActiveMember	-0.1433166	0.0075297	-19.03	0.000	-0.1580763	-0.1285568
EstimatedSalary	7.27e-08	6.51e-08	1.12	0.265	-5.50e-08	2.00e-07

4.4 异质性分析

为了进一步探究客户的性别、客户所处的国家对客户流失的具体影响,本文对性别、国家进行了分组处理:

$$France = \begin{cases} 1, & Geography = France \\ 0, & Geography \neq France \end{cases} \quad (5)$$

$$Spain = \begin{cases} 1, & Geography = Spain \\ 0, & Geography \neq Spain \end{cases} \quad (6)$$

$$Germany = \begin{cases} 1, & Geography = Germany \\ 0, & Geography \neq Germany \end{cases} \quad (7)$$

$$Gender2 = \begin{cases} 1, & Gender = Male \\ 0, & Gender = Female \end{cases} \quad (8)$$

本文将三个国家分别处理为三个虚拟变量:France:法国为 1,不是法国为 0;Spain:西班牙为 1,不是西班牙为 0;Germany:德国为 1,不是德国为 0。而后分别进行逻辑回归,最后发现 France、Spain 显著为负,而 Germany 显著为正,也即如果客户是法国或西班牙客户,那么银行客户流失率就会降低,如果客户是德国客户,那么银行客户流失率就会升高;这就说明德国的客户更容易流失。同样本文对性别一样这样处理:Gender2:男性为 1,女性为 0。最后发现结果显著为负,这就说明客户如果是男性客户,那么银行客户流失率就会降低,也即男性客户更不容易流失,而女性客户更容易流失。(表格 5)

表 5: 异质性分析

variable	法国	西班牙	德国	男女
Age	0.0726 ***	0.0730***	0.0727***	0.4359**
	(28.36)	(28.57)	(28.24)	(2.16)
France	-0.4090***			
	(-7.32)			
Spain		-0.2360***		
		(-3.58)		
Germany			0.7629***	
			(12.04)	
Gender	-0.0292***	-0.5405***	-0.5286***	-0.5430**
	(-3.60)	(-10.01)	(-9.70)	(-10.06)
Surname	-0.0000337	-0.0000423	-.0000388	-0.0000393
	(-1.06)	(-1.33)	(-1.21)	(-1.24)
CreditScore	-.00067**	-0.00063**	-0.0006597**	-0.0006452***
	(-2.40)	(-2.28)	(-2.35)	(-2.32)
Tenure	-0.0150	-0.0152	-0.0162*	-0.0149
	(-1.61)	(-1.64)	(-1.73)	(-1.61)
Balance	4.27e-06***	4.83e-06***	2.64e-06***	5.04e-06***
	(9.01)	(10.40)	(5.13)	(10.96)
NumOfProducts	-0.0624	-0.0423	-0.1025**	-0.0375
	(-1.33)	(-0.91)	(-2.17)	(-0.81)
HasCrCard	-0.0314	-0.0334	-0.0456	-0.0297
	(-0.53)	(-0.57)	(-0.77)	(-0.51)
IsActiveMember	-1.0798***	-1.0761***	-1.0746***	-1.0788***
	(-18.82)	(-18.79)	(-18.63)	(-18.85)
EstimatedSalary	4.93e-07	5.05e-07	4.86e-07	5.03e-07
	(1.05)	(1.07)	(1.03)	(1.07)
_cons	-3.1087***	-3.3451***	-3.3264***	-3.4190***
	(-12.44)	(-13.57)	(-13.40)	(-13.91)
R^2	0.1443	0.1403	0.1533	0.1390
Control	Yes	Yes	Yes	Yes
N	10,000	10,000	10,000	10,000

* 本表格基于 stata 计算结果整理得到.

4.5 随机森林模型

本文同样采用五折交叉验证，并在交叉验证的验证集上使用网格搜索法搜索最优的超参数，本文发现随机森林决策树的最优最大深度为12，最优决策树数量为200。本文在最大深度为 12，数量为 200 的决策树基础上对数据集进行训练最终得到核心解释变量：年龄、性别、所处国家、姓氏的特征重要性(表格 6)

Variable	特征重要性	logistic 回归 Z 值	是否显著
Age	0.407511	25.380	是
Surname	0.349907	-1.386	否
Gender	0.142187	-9.025	是
Geography	0.100395	2.911	是

表 6: 本表格利用 python、stata 得出

本文发现特征重要性的排名为：年龄、姓名、性别、所属国家；如果排除姓名这一不显著的变量，这时排名为：年龄、性别、所属国家，而这个和本文之前 logistic 回归结果的 Z 值排名一致；也即：如果解释变量对被解释变量显著，那么解释变量的特征重要性也就越强，那么对被解释变量的影响也就越强，显著度也就越高，t 值越大。随后对随机森林模型进行了 F 检验，发现核心解释变量的特征重要性均显著，这说明本文的结论是稳健的。(表格 7)

最后，本文对随机森林进行了可视化，最后可视化的结果：德国的特征重要性最高，女性的特征重要性高于男性。(图 1)这和本文之前的结论一致：德国的客户更容易流失，女性客户更容易流失。而本文在进行年龄的特征重要性可视化的时候发现了新的结论：年龄对被解释变量的影响存在门槛效应(图 2)：在 50-60 岁之前，年龄越高，客户越容易流失，在50-60 岁之后年龄越高客户越不容易流失。

Variable	F 值	p 值
Surname	9.31259	0.002282
Age	3142.003457	0.0
Geography	217.966699	0.0
Gender	2125.408399	0.0

表 7: 本表格利用 python 得出

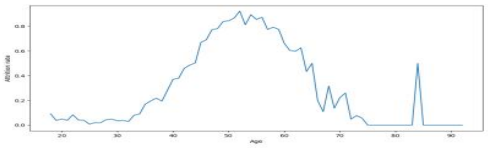
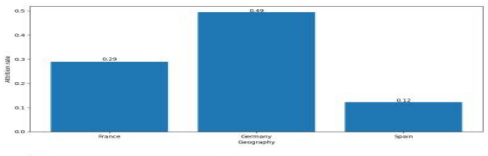


图 1: 年龄异质性



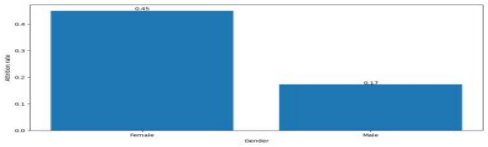


图 2: 所处国家、性别异质性

5 结论

本文分别使用 logistic 回归以及随机森林模型对客户的年龄、姓氏、所处国家、性别与银行客户流失之间进行了因果推断。最终本文得到如下的结论：（1）年龄对银行客户流失的影响存在门槛效应，在 50-60 岁之前，年龄越大的客户越容易流失；而在 50-60 岁之后，年龄越大的客户粘性越大。（2）客户的性别以及所处国家(或者说所说的语言)对客户流失也有显著的异质性影响：女性客户比男性客户更容易流失，说德语的客户比说法语以及西班牙语的客户更容易流失。（3）客户的姓氏对客户流失并无显著影响。

参考文献

[1]DoD,HuynhP,VoP,etal.Customerchurnprediction inaninternetserviceprovider[C]//2017IEEEIntern ationalConferenceonBigData(BigData).IEEE,2017.

[2]SoeiniRA,TashakorL,BafghiJT,etal.Supplierselectionbasedonmultiplecriteria[J].InternationalJournalofNewComputerArchitectures&ThrApplications,2012,2(1):259-274.
[3]VafeiadisT,DiamantarasKI,SarigiannidisG,etal.Acomparisonofmachinelearningtechniquesforcustomerchurnprediction[J].SimulationModellingPractice&Theory,2015,55:1-9.
[4]蒋良孝.朴素贝叶斯分类器及其改进算法研究[D].中国地质大学,2009.
[5]林怡婷,蔡涛,邓喜珊,等.MCP-Logistic 模型在银行客户流失数据的应用[J].宁波工程学院学报,2021,33(4):6.
[6]LuN,LinH,LuJ,etal.Acustomerchurnpredictionmodelintelecomindustryusingboosting[J].IEEETransactionsonIndustrialInformatics,2014,10(2):1659-1665.
[7]VafeiadisT,DiamantarasKI,SarigiannidisG,etal.Acomparisonofmachinelearningtechniquesforcustomerchurnprediction[J].SimulationModellingPracticeandTheory.2015,55:1-9.
[8]方匡南,章贵军,张惠颖.基于Lasso-logistic模型的个人信用风险预警方法[J].数量经济技术经济研究,2014,31(02):125-136.
[9]XiuliangWu,MaoyongCao,KaiSun,FengyingMa.Development of a novel variable selection algorithm for LSTM with LASSO[C]//.第 40 届中国控制会议论文集(14),2021:543-548.DOI:10.26914/c.cnkihy.2021.025127.
[10]李会，吴小兰，李侠.电信客户流失预测模型的构建及客户流失因素分析[J].内蒙古农业大学学报（社会科学版）,2017,19(03):23-27.
[11]柳婷.基于数据挖掘的银行客户流失模型分析研究[D].重庆大学,2008.
[12]郑宇晨,吕王勇.基于logistic模型的证券公司客户流失预警分析[J].郑州航空工业管理学院学报，2016,34(05):80-88.
[13]高海燕.基于数据挖掘的银行客户流失预测研究[D].西安理工大学,2007.
[14]李长山.基于 Logistic 回归法的企业财务风险预警模型构建[J].统计与决策,2018,34(06):185-188.
[15]罗晓光,刘飞虎.基于Logistic回归法的商业银行财务风险预警模型研究[J].金融发展研究,2011(11):55-59.
[16]兰军,严广乐.基于客户特征分群的银行客户流失分析[J].技术经济与管理研究,2014(05):105-108.

作者简介：程烁文（1999-），男，汉族，辽宁沈阳人，青海民族大学研究生在读，研究方向：金融投资
项目名称：青海民族大学研究生创新项目　项目编号：65M2024097