

基于高维特征因子的券商客户流失预警模型研究

高天辰 曲浩 王菲菲 周静*

摘要: 随着证券行业的竞争日趋激烈,如何留住现有客户并预防潜在的客户流失已成为该行业管理者普遍关心的重要问题之一。本文针对证券行业客户流失进行研究。首先,结合证券行业的实际业务背景,探索了证券行业客户流失的定义;其次,提出了基于高维特征因子的独立性筛选方法;最后,基于筛选的因子,分别构建了客户流失预警日模型和周模型。研究表明,资产类因子和非资产类因子对预测客户流失具有显著效果,复合因子的预测效果不显著。客户流失预警日模型的外样本 AUC 值平均可以达到 0.95 以上,说明模型具有良好的预测精度。客户流失预警周模型的预测效果与日模型基本一致,并且具有计算成本低、预测效率高、模型更加稳定的特点。本文的研究结果可以为企业进行客户挽回提供策略分析,划分客户群体,针对不同流失风险的客户制定不同的挽回策略。另外,本文提出的流失预警模型在企业实际环境测试中也具有良好表现。

关键词: 客户流失; 证券行业; 高维特征因子; 独立性筛选; 逻辑回归

中图分类号: F490.6

JEL 分类号: G20

一、引言

证券行业被认为是数据密集度最高和业务场景最繁多的行业之一,由于该行业存在较为严格的监管和准入门槛,近年来新增证券公司数量增长缓慢,而行业竞争却日趋激烈。为了扩大市场份额,券商通过各种方法来获取新客户并提供差异化的营销服务,以提升自身的客户关系管理能力。然而,除了获取新客户,如何留住现有客户以预防潜在的客户流失也成为券商进行客户关系管理的一项重要工作。本文旨在深入研究证券公司的证券经纪类客户流失现象。客户流失 (Customer Churn) 是客户关系管理领域非常重要的研究问题之一 (Fornell and Wernerfelt, 1987; Ahn et al., 2006)。客户流失是指客户不再购买或使用企业的产品或服务,与企业脱离了存续关

* 高天辰,厦门大学经济学院, E-mail: gaotc_cufe@126.com; 曲浩,方正证券股份有限公司, E-mail: quhao@foundersc.com; 王菲菲,中国人民大学应用统计科学研究中心,统计学院, E-mail: feifei.wang@ruc.edu.cn; 周静 (通信作者),中国人民大学应用统计科学研究中心,统计学院, E-mail: zhoujing_89@126.com。本文为中国人民大学科学研究基金 (中央高校基本科研业务费专项资金资助) 项目 (21XNA027) 的研究成果。作者感谢匿名审稿人和编辑部的宝贵意见,当然文责自负。

系的一种状态。在过去的文献中,研究人员已对客户流失的影响因素进行了大量研究 (Ganesh et al., 2000; Reinartz and Kumar, 2003; Jamal and Bucklin, 2006; Nitzan and Libai, 2011; Kayaalp, 2017),但这些文献大多是基于电信行业的背景。例如,研究表明客户的人口统计学特征 (Azeem et al., 2017; Coussement et al., 2017; Mahmoud, 2019)、客户通话相关信息 (Wei and Chiu, 2002; Ahn et al., 2006; Sandhya et al., 2019)、客户满意度 (Bolton et al., 2000; Gerpott et al., 2001)、客户体验 (郭红丽, 2006)、客户社交互动关系 (Giudicati et al., 2013; Haenlein, 2013)、客户忠诚度 (Hwang et al., 2004; Lewis, 2004)、客户的自我网络特征 (周静等, 2017)、客户所使用的服务类型及其与客户核心需求的匹配程度 (Wei and Chiu, 2002; 李季等, 2020)等都是影响客户流失的重要因素。除了对影响因素进行研究,学者们对构建流失预警模型也做了大量研究 (Asamoah et al., 2016; Lee et al., 2017; Eria and Marikannan, 2018)。由于流失与否是一个典型的二分类问题,因此大量基于机器学习的分类器模型被广泛应用于流失预测 (Ahmad et al., 2019),例如逻辑回归、朴素贝叶斯、随机森林、决策树、神经网络、支持向量机等。

相比于电信行业大量的客户流失研究文献,对金融服务业,尤其是对证券行业的客户流失研究相对较少 (Ascarza et al., 2017)。例如,在金融服务业,学者对银行客户的研究表明客户的账户信息 (Nie et al., 2011)、人口统计学特征 (杜修平和王中, 2009; Nie et al., 2011)、交易活动信息 (Van den Poel and Larivière, 2004; Zhao and Dang, 2008)、对支付渠道的满意度 (Krishnan et al., 1999; Peppard, 2000)等是影响客户流失最重要的几方面因素。对证券行业,由于券商经纪业务的特点,一些学者利用 RFM 框架构建基于客户交易数据的各种变量来研究客户流失的影响因素,例如计算客户对各种产品的最近一次交易日期、交易频率、交易金额等 (Guo et al., 2006; Ahn et al., 2019)。除此之外,客户的人口统计信息和基金相关的财务指标也对客户流失有重要影响 (Maldonado et al., 2021)。尽管在证券行业有丰富的数据集,但是由于客户级别的数据可用性、指标繁杂、业务场景众多等问题,导致目前行业内缺乏一套系统的方法论以用于客户流失预警模型的建立。

通过对已有文献的研究以及现实券商企业的调研,我们发现阻碍客户流失研究的瓶颈主要有以下三方面因素。首先,客户流失的定义,由于券商的业务场景众多,导致对客户流失的定义缺乏统一的标准。总体来看,当券商提及客户流失时,主要包含两方面的含义,一是指客户从该证券公司转到其他证券公司;二是指客户的某种指标未达到公司设立的最低标准 (例如月平均交易量、月均资产等),从高价值客户转为低价值客户。对于第一种含义,这是最糟糕的情况,券商进行客户挽留的机会微乎其微,因此采用这种客户流失定义对企业开展客户流失预警的意义不大 (Gladly et al., 2009)。第二种含义虽然能在一定程度上提前识别潜在流失客户,但是对于选取何种指标、设立什么样的最低标准,以往文献并未给出一个统一的操作方

案 (Guo et al., 2006; Ahn et al., 2019)。其次,面对高维度的客户交易数据和行为数据,如何筛选出对预测客户流失最重要的因素,这也是企业面临的重要问题。过往文献对高维客户特征数据的研究均集中在电信行业 (Amin et al., 2017; Coussement et al., 2017; Tan et al., 2017),尚缺少可供证券行业借鉴的高维特征筛选方法。最后,客户交易行为瞬息万变,如何根据实际业务需求对客户流失状态进行动态预测,以制定不同的客户挽留策略,也是企业需要解决的一个重要现实问题。而现有文献的预测模型相对单一,大多数为静态模型且仅支持某一种维度的预测 (例如月到月),在实际应用中具有很大的局限性。

因此,基于目前证券行业在提前预警流失客户方面的迫切需求,本文拟提出一套适用于大数据背景下券商进行客户流失预警分析的方法论。具体而言,本文研究将回答以下三个问题:①定义客户流失需要遵循什么样的标准?②如何进行高维客户流失因子的筛选?③如何对客户流失状态进行动态预测?

余文结构安排如下:第二部分介绍本文研究的所有数据来源及变量基本情况;第三部分为本文重点部分,介绍客户流失预警模型的构建,具体将包括基于稳定性的客户流失定义方法、基于高维客户流失因子的独立性筛选准则以及客户流失预警模型的建立;第四部分为理论检验环节,将第三部分提出的方法论应用到企业实际中检验其具体效果;第五部分为结论与未来研究方向。

二、数据介绍

本文研究使用的用户级交易数据由国内某大型证券公司提供,并经过脱敏处理。数据时间跨度为2018年6月—2019年4月,涉及6个营业部共4万余名用户的历史交易数据和用户基本信息。上述数据储存在数据库中,共涉及16个数据表、831个字段。其中,资产表5个,总计407个字段;非资产表7个,总计407个字段;功能表4个,总计17个字段。表1详细介绍了数据字典中部分核心字段的情况,包括数据类型、取值范围等。其中,将原始数据表中资产表中的字段定义为资产类变量,非资产表中的字段定义为非资产类变量。功能表中的字段不属于上述两类,而是用于确定客户的流失状态。在原始数据中存在客户流失状态字段,这一字段并不是人工标记的流失状态,而是由业务人员根据一些规则进行的初步探索。每个客户每天都会有一个流失状态,研究发现该字段的流失状态极不稳定,流失变化率高达63.6%。这可能导致一个客户昨天是流失状态,今天又变成了未流失状态。这导致业务人员无法准确识别流失客户,其挽回客户的核心业务线也就无法开展。因此,一个稳定的客户流失状态与其核心业务息息相关,这促使我们在后文中提出具有稳定性的客户流失的定义。

表 1 数据字典中部分核心字段示例

类型	核心字段（部分）	数据类型	取值范围	缺失率（%）
资产类	普通 A 股交易量	定量变量	$0 \sim 9.86 \times 10^7$	0
	普通 A 股实收佣金	定量变量	$0 \sim 2.96 \times 10^4$	0
	股基权市值	定量变量	$0 \sim 2.15 \times 10^9$	0
	总保证金	定量变量	$0 \sim 2.00 \times 10^8$	0
	总负债	定量变量	$0 \sim 1.28 \times 10^8$	0
	净资产	定量变量	$-1.01 \times 10^5 \sim 2.09 \times 10^9$	26
	资金流入	定量变量	$0 \sim 2.01 \times 10^8$	0
	本日盈亏	定量变量	$-4.43 \times 10^8 \sim 2.00 \times 10^8$	0
非资产类	性别	定性变量 (2 水平)	男、女	0
	年龄	定量变量	18 – 108	0.29
	学历	定性变量 (6 水平)	1、2、3、4、5、6	95.36
	婚姻状况	定性变量 (2 水平)	未婚、已婚	95.11
	是否为小额客户	定性变量 (2 水平)	是、否	3.78
	电话回访	定性变量 (2 水平)	是、否	0
功能类	客户流失状态	定性变量 (2 水平)	是、否	—
	客户登录日期	文本变量	示例：2019 年 3 月 1 日	—
	客户是否为有效户	定性变量 (2 水平)	是、否	—

注：数据字典中部分核心字段的情况，包括数据类型、取值范围等。其中，将原始数据表中资产表中的字段定义为资产类变量，非资产表中的字段定义为非资产类变量，功能表中的字段用于确定客户的流失状态。学历的取值范围“1、2、3、4、5、6”分别代表“博士、硕士、本科、大专、中专、高中”。

由于原始数据十分庞杂且可能存在缺失值和异常值的情况，因此首先需要对数据质量进行仔细检查。在此过程中，结合实际业务需求以及数据字段类型，首要考虑资产类变量和非资产类变量。其中，资产类变量通常为定量变量（连续型），而非资产类变量则大多数为定性变量（离散型）。数据质量检查的主要内容包括以下几个方面：数据间的匹配度、字段缺失率、样本量、时间跨度、平均值、中位数、最大值、最小值、方差和标准差等。对于定量变量，首先筛选出所有定量字段，并仔细核对这些字段与客户 ID 之间的匹配情况。其次，需要检视每个定量字段的缺失率和分布情况，确保数据的完整性和合理性。其中，字段的缺失率计算如下：对于不随

时间变化的字段，例如性别、学历等，其缺失率的计算公式为：缺失率 = 与客户 ID 匹配后的非缺失数量/总客户 ID 数量；对于随时间变化的字段，例如净资产、资金流入等，其缺失率的计算公式为：缺失率 = 与客户 ID 匹配后的非缺失数量/总客户 ID 数量 × 时间跨度。最后，需要计算描述性统计指标并进行汇总。表 2 展示了部分核心原始字段的描述分析结果。

表 2 部分核心原始字段的描述分析结果

核心字段（部分）	最小值	平均值	中位数	最大值	标准差
普通 A 股交易量	0	7.24×10^4	1.73×10^4	9.86×10^7	3.63×10^5
普通 A 股实收佣金	0	3.98×10	1.43×10	2.96×10^4	1.29×10^2
股基权市值	0	1.77×10^5	3.36×10^4	2.03×10^9	5.15×10^6
总保证金	0	1.71×10^4	6.41×10^2	2.00×10^8	3.01×10^5
总负债	0	1.34×10^4	0	1.28×10^8	5.62×10^5
净资产	-1.01×10^5	2.54×10^5	5.01×10^4	2.09×10^9	6.13×10^6
资金流入	0	4.88×10^4	0	2.01×10^8	8.66×10^5
本日盈亏	-4.43×10^8	-1.31×10^2	0	2.00×10^8	2.19×10^5

注：作者基于研究结果自行整理。

在处理定性变量时，需要解决关键字段格式不统一以及变量水平繁多的问题。为了能够继续进行更深入的研究，必须在上述流程的基础上增加数据清洗的工作。最后，对于数据库中的所有 814 个资产与非资产类原始字段进行上述数据质量检查。经检查得出，资产类变量的平均缺失率约为 17%，数据质量较好。例如，净资产变量的缺失率为 26%，数据质量较好，能够作为一个可供考量的指标进行分析。而非资产类变量的平均缺失率约为 61.9%，数据质量较差。例如，学历变量的平均缺失率为 95.36%，数据质量较差。在后续建模过程中，对于缺失率较严重的变量采取了将变量剔除的处理原则。

三、客户流失预警模型

（一）基于稳定性的客户流失定义方法

在构建客户流失预警模型之前，首先需要明确定义良好的因变量 Y，即流失（Y=1）与未流失（Y=0）。根据企业的业务经验，客户流失与 APP 的登录次数以及连续若干交易日内的账户资产有关。然而，目前对于登录次数、交易日数和账户资产尚缺乏一定的理论指导数值。因此，本文研究首先基于稳定性来定义客户的流失状态。我们假设每个客户每天都有一个流失状态，并且一个良好的客户流失定义应该使得流失状态的变化尽量稳定。表 3 展示了 2 位客户在 10 个交易日内的流失状态

变化。以客户 1 为例，最开始他是一个正常的客户，在第 5 个交易日变成了流失客户，在这之后一直维持着流失状态，我们认为这是一个比较正常的情况。然而，有些变化是不合理的。我们假设流失代表客户不再使用公司的产品或服务，那么客户流失之后，其流失状态应该在一段时间内是稳定的，而不应该出现来回波动的情况。因此，客户 2 的状态变化是一个不合理的示例。在后续的流失预警模型构建中，准确定义客户的流失状态将是至关重要的一步。为此，需要进一步清洗和处理数据，以确保客户流失状态的稳定性和合理性。

表 3 客户流失状态变化

交易日	1	2	3	4	5	6	7	8	9	10
客户 1	0	0	0	0	1	1	1	1	1	1
客户 2	1	0	0	1	1	0	0	1	0	0

注：本表为 2 位客户在 10 个交易日内的流失状态变化。客户 1 的流失状态相对稳定，客户 2 的流失状态来回波动。

因此，基于以上假设，一个好的客户流失定义应该确保客户在流失后，其流失状态在一段时间内保持稳定。为了度量流失状态是否稳定，我们引入了流失状态变化率的概念。流失状态变化率衡量了在一个观察期 T （如 1 个月或 3 个月）内已流失客户的状态发生变化的人群比例。具体地，定义流失 - 未流失变化率 $R_{1-0}^{(t)}$ 为：

$$R_{1-0}^{(t)} = \frac{M^{(t)}}{N^{(t)}} \tag{1}$$

其中， $M^{(t)}$ 为 t 时刻到 T 时刻内状态由流失变为未流失的客户数量，在观察期内的客户出现一次流失到未流失的变化即被认为流失状态发生变化； $N^{(t)}$ 为 t 时刻到 T 时刻内已流失的客户数量，只要流失过一次即被认为是已流失客户。

接下来，我们将基于稳定性对客户流失的定义进行探索。首先，根据实际业务情况，我们将待探索的流失状态客户的范围限定在有效户，并要求其开户时间大于 90 天。具体来说，有效户是指一定时间内存在连续 20 个交易日资产不低于 1 万元的客户。每个客户每 6 个月进行一次判定。例如，如果一个客户在 2018 年 1 月 1 日被判定为有效户，则其有效户状态将持续到 2018 年 6 月 30 日。开户时间大于 90 天的条件是为了筛选出刚开户的新客户，这些客户不在客户流失预警的考虑范围内。接下来，我们将探索观察期（交易日数）、登录次数和账户资产三个因素的不同取值对流失状态的影响。具体地，在观察期方面，我们尝试了 20 个交易日、30 个交易日和 40 个交易日三个不同的取值；在登录次数方面，我们设置登录次数阈值为 m ，取值范围为 0 ~ 7；在账户资产阈值方面，我们尝试了账户资产阈值为 1 万元和 2 万元两个取值。最后，我们以 2019 年 2 月 ~ 4 月（3 个月）以及 2019 年 4 月（1 个月）为流失状态变化率的计算周期，对比分析不同数值组合下流失状态变化率的情况。控

制其他因素不变,账户资产阈值取1万元,登录次数阈值为0,流失状态变化率的计算周期为2019年2月~4月(3个月)。首先,探索不同观察期(20个交易日、30个交易日、40个交易日)对客户流失状态的影响,详细结果见表4。从表4中可以看出,观察期的变化对客户流失状态的变化影响不大。

表4 观察期对流失状态的影响

观察期	$R_{1-0}^{(t)}$
20个交易日	0.44
30个交易日	0.46
40个交易日	0.44

注:作者基于研究结果自行整理。

接下来,我们将对不同登录次数进行研究,其中账户资产阈值设置为1万元,观察期设置为20个交易日,流失状态变化率的计算周期为3个月。在控制上述因素不变的情况下,登录次数 m 的取值范围为0~7,然后分别计算流失状态变化率,详细结果见表5。结合前期对流失状态的探索以及表5的结果,我们发现在控制其他因素不变的情况下,登录次数的变化对客户流失状态变化的影响较小。由于前期的探索结果并非本文主要内容,因此省去了相关结果的汇报。

表5 登录次数阈值对流失状态的影响

m	$R_{1-0}^{(t)}$
0	0.44
1	0.49
2	0.45
3	0.43
4	0.42
5	0.42
6	0.39
7	0.37

注:作者基于研究结果自行整理。

观察期设置为20个交易日,登录次数阈值设置为3次,账户资产阈值分别设置为1万元和2万元,流失状态变化率的计算周期分别为1个月和3个月,结果见表6。相比于账户资产阈值为1万元时,2万元时的流失状态变化率更低,而且在3个月的计算周期上下降明显。基于以上对观察期、登录次数以及账户资产阈值的探索,我们得出以下结论:账户资产阈值是对流失状态稳定性影响较大的一个因素。因此,在接下来的研究中,我们将客户流失定义为:在有效户的范围内,资产小于2万元

且持续 20 个交易日，并且期间登录次数不超过 3 次的情况下，视为流失客户。在该定义下客户流失状态变化率仅为 0.31。同时，该客户流失的定义也得到了业务人员的认可，与业务专家的实际判断相吻合。

表 6 账户资产阈值与观察期长度对客户流失状态的影响

资产阈值	1 个月	3 个月
1 万元	0.11	0.43
2 万元	0.09	0.31

注：作者基于研究结果自行整理。

（二）基于高维流失因子的独立性筛选准则

1. 高维流失因子生成

因变量确定之后，接下来是选择自变量，即影响客户流失的因素。由于原始数据涉及上百个变量，如何从中挖掘出有意义的流失影响因子是本文的重要研究内容。本文提出了一种基于高维流失因子的独立性筛选准则的方法。根据数据情况，我们将流失因子分为三类，分别是资产类因子、非资产类因子和复合因子。首先，将资产类变量划分为 8 个维度，并结合实际业务场景，经过与业务专家的交流与讨论，在原始数据中选择了 65 个资产类字段。随后，对上述筛选出的每一项原始字段，在自变量观察期内计算极值（最大值、最小值）、集中趋势（均值）以及离散程度（变异系数、标准差）。需要注意的是，观察期的选择是根据考察时间变化的，在前面有详细的计算说明，这里不再赘述。表 7 展示了资产类数据的测算维度、原始字段以及相应的衍生指标。每个原始字段都可以衍生出 5 个变量，总计 325 个变量。例如，原始字段为净资产，衍生出净资产均值、净资产最大值、净资产最小值、净资产标准差、净资产变异系数 5 个变量。每一个衍生指标都有实际含义，例如净资产最大值代表客户一段时间内流通资产与总负债差值的最大值，净资产最大值越大，表示客户越不容易流失。

表 7 资产类数据的测算维度、原始字段以及相应的衍生指标

测算维度	原始字段	衍生指标
市值（Market）	普通沪 A 股票市值等 16 个原始字段	
交易量（Volume）	普通 A 股交易量等 18 个原始字段	
佣金净佣金（Fare）	普通 A 股实收佣金等 19 个原始字段	最大值
盈亏（Profit）	本日盈亏	最小值
保证金（Balance）	总保证金	均值
负债（Debit）	两融总负债	变异系数
资产（Assets）	总资产等 7 个原始字段	标准差
流入和流出（Fund in & out）	资金流入、资金流出	

注：作者基于研究结果自行整理。

其次，我们将非资产类变量划分为个人信息、回访行为、购买行为和登录行为四个维度，保留了 32 个非资产类原始字段，并删除了无效以及敏感字段（如手机号、邮箱、回访业务人员等）。表 8 展示了非资产类数据的测算维度、原始字段以及相应的衍生指标。在非资产类数据中，一共有 32 个原始字段，除个人标签外，每个字段可以衍生出 1~4 个变量，共生成了 73 个衍生变量。例如，以是否有过短信回访为例，该指标代表一段时间内客户是否被进行过短信回访。而针对这一字段，我们衍生出了 4 个变量，分别是：是否有过短信回访、短信回访次数、客户已读次数以及最近一次回访的时间间隔。一般而言，有过短信回访的客户往往是高价值客户，其流失概率较低。

表 8 非资产类数据的测算维度、原始字段以及相应的衍生指标

测算维度	原始字段	衍生指标
个人信息	年龄、性别、学历等 20 个原始字段	年龄分段（40 岁以下、40~50 岁、50~60 岁、60 岁以上）、是否期权潜在户等
回访行为	短信回访、电话回访、当面拜访	是否有过短信回访、是否有过当面拜访等
购买行为	申请、购买、持有等 8 个原始字段	是否持有金融产品、是否持有服务产品等
登录行为	登录	登录次数、是否有过登录行为等

注：作者基于研究结果自行整理。

进一步地，根据实际业务场景，我们对资产类与非资产类因子进行比值或乘积计算，得到了 10 个复合流失因子。表 9 展示了复合流失因子的字段构成及计算方式，并附有实际业务思考的说明。以“股基权市值 - 场外基金市值”为例，如果股基权市值减少，即场内交易量减少，一般来说该客户可能会流失；但是如果该客户的场外基金市值持仓增加，则很有可能不会流失。再者，考虑复合流失因子“是否有投顾类订单 - A 股 B 股交易量”，一般来说持有投顾类订单的客户不会流失，但如果其交易量变低，则有可能会流失。

表 9 复合流失因子的字段构成、计算和说明

字段 1	字段 2	计算	说明
股基权市值	场外基金市值	比值	场内交易量减少，但场外基金持仓增加，可能为非流失
登录次数	场外基金市值	比值	登录次数降低，但场外基金持仓增加，可能为非流失
流通资产	登录次数	比值	资产降低，登录次数增加，可能为非流失
是否有投顾类订单	A 股 B 股交易量	乘积	持有服务类产品，但交易量变低，可能会流失
投顾类产品是否退订	场外基金市值	乘积	服务类产品退订，但交易量均衡或有场外基金持仓，可能为非流失
投顾类产品是否大于 1	A 股 B 股交易量	乘积	签署多个服务产品，但交易量降低，可能会流失

(续)

字段 1	字段 2	计算	说明
投顾类产品是否大于 1	总保证金	乘积	签署多个服务产品，但资金量下降，可能会流失
本日盈亏	总保证金	比值	亏损率增高（本日盈亏减少），资产量变多，可能为非流失
总保证金	登录次数	比值	资金量下降，但是登录活跃，可能为非流失
总保证金	A 股 B 股交易量	比值	资金量下降，但是交易量增加，可能为非流失

注：作者基于研究结果自行整理。

2. 高维流失因子预筛选

接下来，需要对生成的资产类因子、非资产类因子和复合因子，共计 408 个因子进行筛选。首先，对于 325 个资产类因子，研究发现它们的自相关程度较高，缺失率较低。本文的资产类因子均为连续型变量。因此，我们采用单变量 AUC 的方法（Fan and Lv, 2008）进行筛选。针对资产类因子，首先根据变量含义将它们划分为 8 个组别，分别是交易量、佣金以及净佣金、盈亏、保证金、负债、市值、资产、流入和流出。然后，针对每个组别下的变量，计算与其所对应的流失状态建立逻辑回归模型后的 AUC 值。我们选择每个组别下 AUC 值最高的变量作为代表。此处，我们选取了 2019 年 3 月第 1 周内 5 个交易日的数据，对所有的资产类因子（共 325 个）进行了单变量 AUC 的计算。表 10 展示了全部变量的单变量 AUC 排名前 10 的变量。研究发现，各个交易日的结果高度一致，因此，用此方法筛选资产类因子是十分稳定的。最终，我们筛选出了 8 个资产类因子：股基权市值最大值、总保证金标准差、流动资产最大值、普通 A 股交易量均值、总负债标准差、普通 A 股实收佣金均值、本日盈亏标准差以及资金流入均值。

表 10 5 个交易日单变量 AUC 值 TOP10

变量名称	AUC_1	AUC_2	AUC_3	AUC_4	AUC_5
流动资产最大值	0.943 80	0.943 32	0.942 98	0.942 24	0.943 51
净资产最大值	0.943 70	0.943 27	0.942 92	0.942 17	0.943 45
总资产最大值	0.942 47	0.942 99	0.942 65	0.941 78	0.943 03
流动资产均值	0.942 84	0.942 27	0.941 86	0.941 26	0.941 60
净资产均值	0.942 73	0.942 18	0.941 77	0.941 17	0.941 50
总资产均值	0.942 57	0.942 00	0.941 59	0.940 85	0.941 23
流动资产最小值	0.933 47	0.933 46	0.932 95	0.932 58	0.933 01
总资产最小值	0.933 28	0.932 78	0.932 78	0.932 41	0.932 86
净资产最小值	0.933 22	0.932 69	0.932 69	0.932 20	0.932 61
总交易性总资产最大值	0.929 39	0.928 70	0.928 70	0.927 89	0.929 23

注：作者基于研究结果自行整理。

对于预筛选出的因子,如果其缺失率高于30%,我们将重新定义这些因子为两水平的离散型变量,即“缺失”与“非缺失”。举例来说,如果预筛选出的普通A股交易量均值的缺失率高于30%,那么将其重新定义为“普通A股交易量均值-缺失”与“普通A股交易量均值-非缺失”这两个水平。其实际含义可以理解是否为进行了普通A股交易,非缺失表示有进行交易,缺失则没有。在观察期内进行过普通A股交易的客户更不容易流失。对于缺失率低于30%的资产类因子,我们将使用均值进行插补,并利用盖帽法剔除异常值,即保留1%分位数与99%分位数之间的数据。然后,我们进行标准化处理,以便在建模过程中处理不同尺度的数据。

对于非资产类因子,我们不仅要从统计学的角度进行筛选,还要结合实际业务进行判断。由于非资产类因子的缺失率较高,我们需要针对其缺失情况和实际含义进行处理。首先,我们剔除了缺失比例超过85%的变量。举例来说,婚姻状况是一个非资产类因子,它有三个水平,即1、2、3,但缺失率高达94.7%。即使按照是否缺失处理为两水平的离散型变量也无法提取有效信息,因此予以剔除。根据以上筛选规则,我们最终预筛选出了21个非资产类因子,包括年龄、性别、学历、客户标签、职业、新客户价值度、风险承受能力、分级基金状态、是否高净值客户、是否小额、是否掘金用户、是否小方重点推广、是否期权潜在户、是否两融客户、是否有过当面拜访、是否有过电话回访、是否有过短信回访、是否购买、是否持有金融产品、是否持有服务产品以及是否有过登录行为。接下来,针对上述预筛选出的因子进行具体处理。举例来说,年龄这一变量存在少数缺失值,我们按照年龄的中位数“50”进行插补,并分段处理成四个区间:40岁以下、40~50岁、50~60岁、60岁以上。而对于新客户价值度这一变量,按照是否缺失处理为两水平的离散型变量。如果该变量缺失,代表该客户的价值较低,更容易流失。

复合因子的预筛选规则为剔除缺失比例超过80%的变量。其中,除了本日盈亏-总保证金(缺失率为0%)和总保证金-A股B股交易量(缺失率为67.4%)两个复合因子之外,其他复合因子都因为缺失率过高而无法使用。针对总保证金-A股B股交易量这个复合因子,我们使用了最大值进行插补,而不是转换为离散型变量进行处理。这个复合因子越大说明客户的交易占比越低,可能越容易流失。综上所述,经过对资产类因子、非资产类因子以及复合因子的预筛选,我们最终筛选出了8个资产类因子、21个非资产类因子和2个复合因子进行建模。详细的变量说明见表11。

表 11 变量说明表

变量类型		变量名称	详细说明	备注
因变量		是否流失	定性变量	流失占比 14%
自变量	资产类	股基权市值最大值	定量变量	市值维度
		总保证金标准差	定量变量	保证金维度
		流动资产最大值	定量变量	资产维度
		普通 A 股交易量均值	定量变量	交易量维度
		总负债标准差	定量变量	负债维度
		普通 A 股实收佣金均值	定量变量	佣金维度
		本日盈亏标准差	定量变量	盈亏维度
		资金流入均值	定量变量	资金流入维度
	非资产类	年龄、性别、学历、客户标签、职业、新客户价值度、风险承受能力、分级基金状态、是否高净值客户、是否小额、是否掘金用户、是否小方重点推广、是否期权潜在户、是否两融客户	定性变量	个人信息维度
		是否有过当面拜访、是否有过电话回访、是否有过短信回访	定性变量	回访行为维度
		是否购买、是否持有金融产品、是否持有服务产品	定性变量	购买行为维度
		是否有过登录行为	定性变量	登录行为维度
	复合类	本日盈亏 - 总保证金	定量变量	
		总保证金 - A 股 B 股交易量	定量变量	

注：作者基于研究结果自行整理。

（三）客户流失预警模型

1. 客户流失预警日模型

客户流失日模型是针对每日数据进行建模，并对客户实现天到天的流失状态进行预测。每日建模结果都有一套独立的参数估计结果。为了计算时刻 T 的因变量与自变量，我们选择 $T-19$ 到 T 这 20 个交易日作为因变量观察期，同时选择 $T-39$ 到 $T-20$ 这 20 个交易日作为自变量观察期。以 2019 年 3 月 1 日的日模型为例，图 1 展示了客户流失预警日模型的逻辑结构。首先，我们以 2019 年 1 月 28 日到 2019 年 3 月 1 日这 20 个交易日为因变量观察期，用于确定因变量 Y ，即客户的流失状态。其次，我们以 2018 年 12 月 27 日到 2019 年 1 月 25 日这 20 个交易日为自变量观察期，用于计算各种自变量的取值。值得一提的是，本文提出的模型根据给定日期确定因变量和自变量的观察期，以便选择相应的因变量和自变量。在这样的设计下，选取

数据集的最新数据并不能带来更多的信息。结合业务人员的建议，本文选择 2019 年 3 月 1 日作为示例日期来展示客户流失预警日模型的逻辑结构。

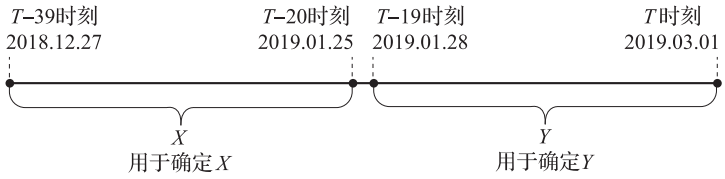


图 1 客户流失预警日模型逻辑结构示意
注：作者基于研究结果自行绘制。

为了对流失预警模型的效果进行评估，我们将样本划分为训练集与测试集，其中 70% 为训练集，30% 为测试集。训练人群与测试人群的划分方式不随时间改变。以客户流失状态为因变量，利用表 11 中展示的各类流失因子作为自变量，我们建立了逻辑回归模型，并采用 BIC 变量筛选准则（Schwarz, 1978）对变量进行筛选。BIC 变量筛选准则能够辅助从待选模型中筛选出“真模型”。表 12 展示了 2019 年 3 月 1 日的建模结果，模型选择出了 10 个重要的流失因子，其中 6 个为资产类因子，4 个为非资产类因子，复合因子对客户流失状态的影响不显著。

表 12 2019 年 3 月 1 日客户流失预警日模型建模结果

变量名称	变量系数	P 值	备注
截距项	-24.09	<0.001	
股基权市值最大值	-11.74	<0.001	
总保证金标准差	2.73	<0.001	
总负债标准差	-3.60	0.234	
流动资产最大值	-27.57	<0.001	
资金流入均值 - 缺失	0.61	<0.001	基准组：非缺失
普通 A 股交易量均值 - 缺失	1.38	<0.001	基准组：非缺失
年龄 - 40 ~ 50 岁	0.41	<0.001	
年龄 - 50 ~ 60 岁	0.75	<0.001	基准组：40 岁以下
年龄 - 60 岁以上	1.11	<0.001	
是否有过登录行为 - 是	-2.61	<0.001	基准组：否
是否持有金融产品 - 是	0.37	<0.001	基准组：否
是否持有服务产品 - 是	-0.54	<0.001	基准组：否

注：作者基于研究结果自行整理。

为了验证日模型筛选因子的稳健性，我们对 2019 年 3 月 1 日到 3 月 7 日内的 5 个交易日分别建立客户流失预警日模型，使用 BIC 变量筛选准则进行因子筛选，筛选结果基本一致。图 2 展示了 6 个资产类因子 2019 年 3 月 1 日到 3 月 7 日内的 5 个交易日的模型系数，从中可以看出，除了总负债标准差外，其余因子的系数基本一致可比。

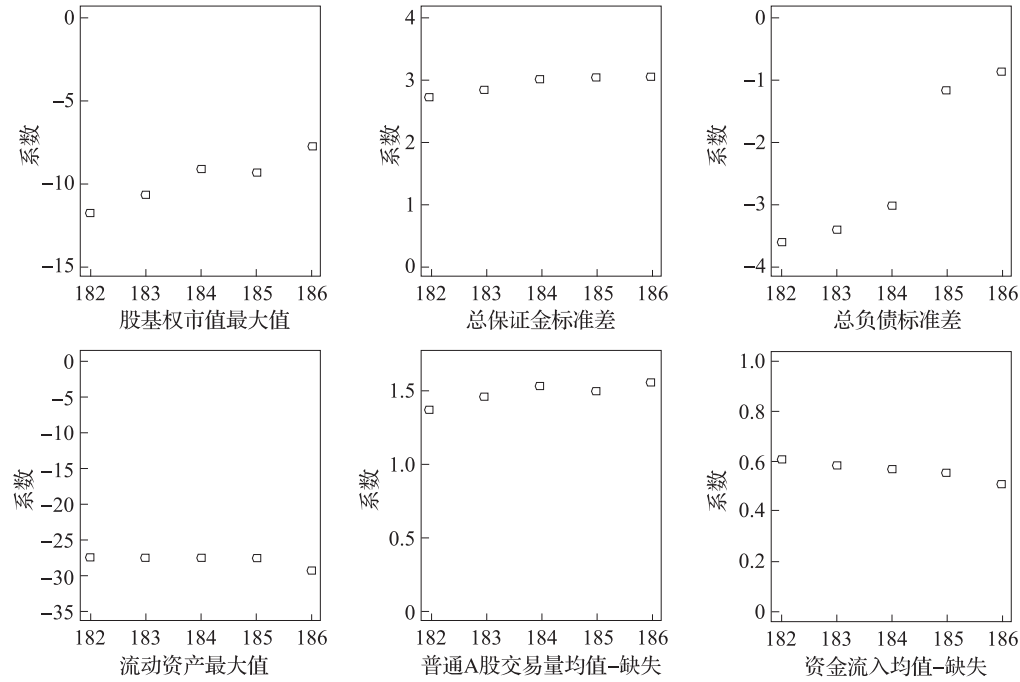


图 2 资产类因子系数

注：作者基于研究结果自行绘制。

在控制其他因素不变的前提下，本文对变量系数进行解读。其中，股基权市值最大值系数为负，这意味着股基权市值最大值与客户流失呈负相关，较低的股基权市值最大值可能暗示着客户的投资表现不佳或者客户对于当前市场形势不满意，从而增加了客户流失的可能性；总保证金标准差系数为正，这意味着较高的总保证金标准差可能导致客户面临更大的风险，特别是在波动性较大的市场环境下，这样的风险增加了客户流失的概率，因为客户可能会担心自己的投资面临较大的损失；总负债标准差为负，较低的总负债标准差可能反映了客户负债的稳定性较高，即客户在还款能力和债务管理方面表现较好，从而降低了客户流失的风险；流动资产最大值系数显著为负，较低的流动资产最大值可能暗示着客户的资金较少或者可能将资金用于其他投资或支出，从而增加了客户流失的可能性；普通 A 股交易量均值 - 缺失系数为正，交易量均值的缺失可能反映了客户的交易活动较少或者对于该证券公司的投资兴趣不高，从而增加了客户流失的可能性；资金流入均值 - 缺失系数为正，说明资金流入均值 - 缺失与客户流失呈正相关，资金流入均值的缺失可能表明客户的活跃程度较低或者对于该证券公司的交易活动不频繁，从而增加了客户流失的可能性。

图 3 展示了 4 个非资产类因子 2019 年 3 月 1 日到 3 月 7 日内的 5 个交易日的模型系数，因子系数基本一致可比。从中可以看出，控制其他因素不变的情况下：在年龄方面，客户的流失概率 60 岁以上 > 50 ~ 60 岁 > 40 ~ 50 岁 > 40 岁以下，说明客

户年龄越大, 越容易流失。特别是对于年龄在 50 ~ 60 岁之间的客户容易发生流失, 可能是因为他们接近退休年龄, 对于财务规划和退休安排有更多的需求和考虑。年龄在 60 岁以上这个年龄段的客户最容易发生流失, 可能是因为退休后财务规划和服务需求的变化, 或者其他因素导致他们更有可能转移投资或服务机构。是否有过登录行为 - 是系数为负, 表示有过登录行为的客户与流失之间存在负相关关系。这意味着有过登录行为的客户更不容易发生流失, 可能是因为他们对于证券公司的关注和参与较多, 显示了较好的客户活跃度。需要注意的是, 是否持有金融产品 - 是系数为正, 这可能表示持有金融产品的客户更容易发生流失, 可能是因为他们对于投资产品的满足程度较低。是否持有服务产品 - 是系数为负, 这可能表示持有服务产品的客户更不容易发生流失, 可能是因为他们对于证券公司提供的服务和支持较满意, 显示了较好的客户忠诚度。

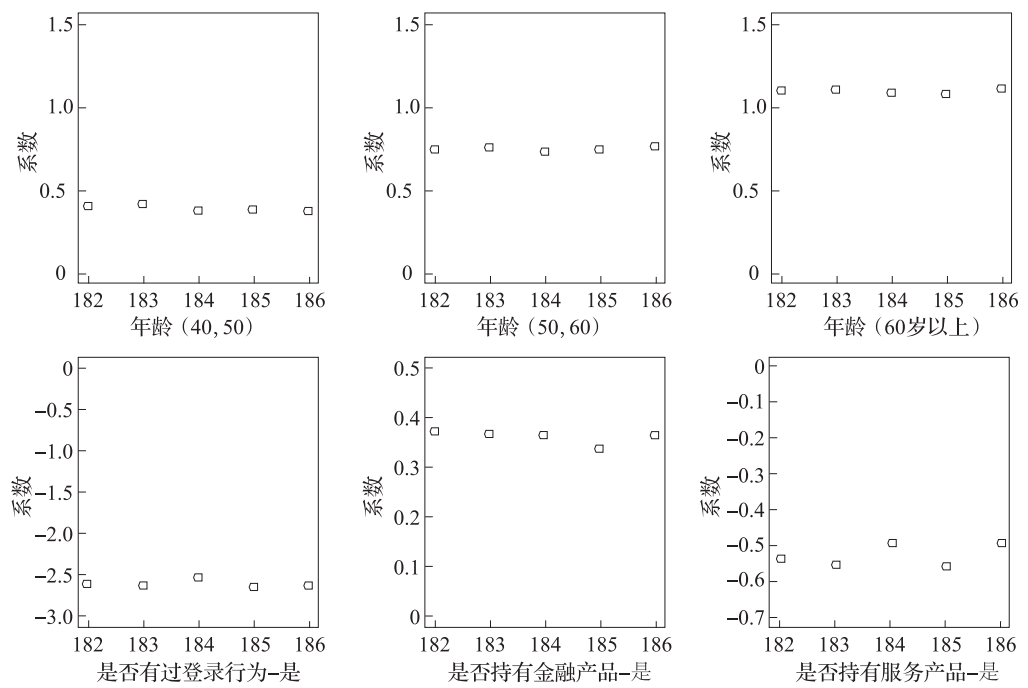


图 3 非资产类因子系数

注: 作者基于研究结果自行绘制。

为了进一步评价模型的预测效果, 我们对 2019 年 3 月 1 日客户流失预警日模型计算了混淆矩阵, 见表 13。测试集预测时选取的阈值是训练集中真实的客户流失比例。由混淆矩阵可得, 该模型的精确率为 0.963, 召回率为 0.469。图 4 展示了 2019 年 3 月 1 日客户流失预警日模型建模结果的 ROC 曲线。从图中可以看出, 模型的 AUC 值为 0.955, 模型的预测效果尚可。另外, 根据企业的实际业务需求, 可以适当调整预测的模型阈值, 在牺牲一定精确率的前提下, 保证较高的召回率。

表 13 2019 年 3 月 1 日客户流失预警日模型混淆矩阵

	预测值 -0	预测值 -1
真实值 -0	8 304	1 490
真实值 -1	51	1 317

注：作者基于研究结果自行整理。

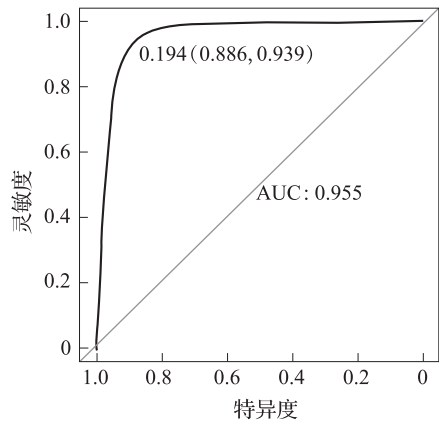


图 4 2019 年 3 月 1 日客户流失预警日模型 ROC 曲线

注：作者基于研究结果自行绘制。

为了进一步探究模型预测效果的稳定性，我们以 2019 年 3 月 1 日到 2019 年 3 月 7 日（5 个交易日）为例。图 5 展示了 5 个交易日建立客户流失预警日模型的 AUC 值、精确率和召回率。从图中可以看出，模型的效果是比较稳定的。模型的 AUC 值和精确率都处于较高水平，说明本文建立的模型具有一定的稳定性，这将更有利于将模型应用于实际业务数据。

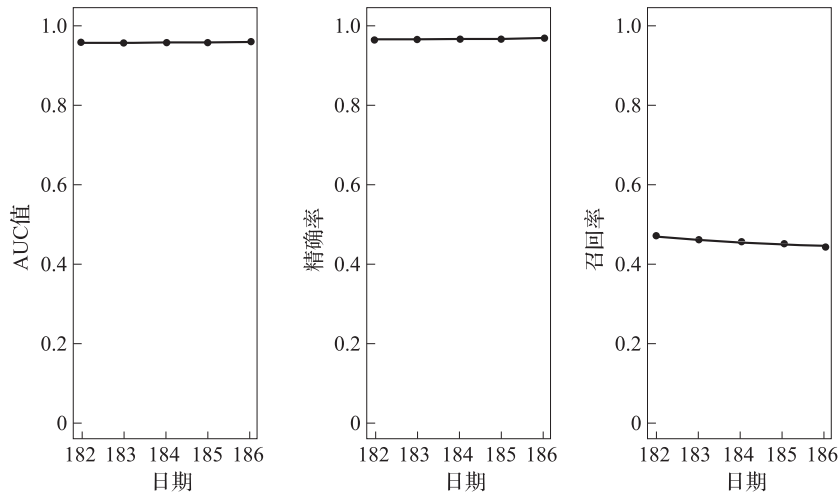


图 5 2019 年 3 月 1 日到 2019 年 3 月 7 日（5 个交易日）客户流失预警日模型预测效果

注：作者基于研究结果自行绘制。

2. 客户流失预警周模型

日模型的特点是时效性强，可以进行实时的流失状态更新，但是也存在模型计算成本高、系数稳定性较差等问题。基于此，本文采用模型平均（Model Averaging）方法（Raftery et al., 1997），将多个日模型的系数估计值进行平均，从而形成更加稳定的流失预测模型，用于更长周期的预测。这能够大大节约计算成本，提高计算效率，使模型更加稳定。具体地，将最近一周（5 个交易日）的客户流失预警日模型的系数求平均，生成周均模型，该周均模型的预测公式为：

$$\hat{Y} = \frac{1}{T} \sum_{t=1}^T \hat{\beta}_{0,t} + \sum_{i=1}^p \left(\frac{1}{T} \sum_{t=1}^T \hat{\beta}_{i,t} x_i \right)$$

(2)

其中， T 为交易日个数， p 为变量总个数。表 14 展示了 2019 年 3 月 1 日到 2019 年 3 月 7 日（5 个交易日）构成的周均模型的系数情况。

表 14 客户流失预警周均模型系数

变量名称	变量系数	备注
截距项	-22.19	
股基权市值最大值	-9.23	
总保证金标准差	2.98	
总负债标准差	-2.23	
流动资产最大值	-27.62	
资金流入均值 - 缺失	0.66	基准组：非缺失
普通 A 股交易量均值 - 缺失	1.48	基准组：非缺失
年龄 -40 ~ 50 岁	0.42	
年龄 -50 ~ 60 岁	0.75	基准组：40 岁以下
年龄 -60 岁以上	1.12	
是否有过登录行为 - 是	-2.60	基准组：否
是否持有金融产品 - 是	0.34	基准组：否
是否持有服务产品 - 是	-0.53	基准组：否

注：作者基于研究结果自行整理。

接下来对周均模型的预测效果进行测试。图 6 展示了 2019 年 3 月 1 日到 2019 年 3 月 7 日 5 个交易日内，周均模型与日模型的效果对比。其中，实线表示客户流失预警日模型的结果，虚线表示客户流失预警周均模型的结果。可以看出，周均模型的预测效果与日模型基本一致。周均模型相较于日模型具有计算成本低、预测效率高、模型更加稳定的特点。因此，在实际业务中，周均模型是一个很好的选择。

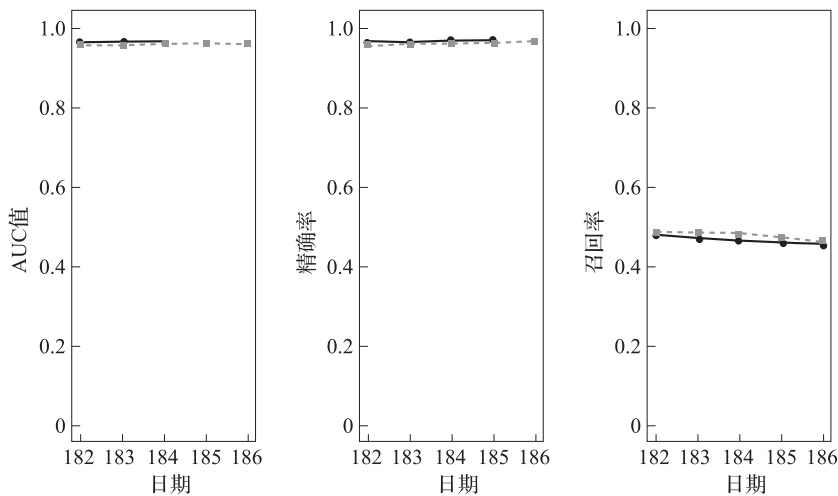


图 6 周均模型与日模型的预测效果对比

注：实线表示客户流失预警日模型的结果，虚线表示客户流失预警周均模型的结果。周均模型的预测效果与日模型基本一致。

四、企业实践检验

（一）基于模型的策略分析

根据提出的客户流失预警日模型，可以计算每个客户的预测流失概率值，并将样本按照预测值从高到低进行排序。本文提出了覆盖率－捕获率曲线，用于评估企业实践中的模型预测精度。该曲线的原理类似于 ROC 曲线，但更利于管理人员的理解与实际应用。具体地，覆盖率定义为在排序好的样本中需要覆盖（如营销触达）的人群比例，例如前 $q\%$ 的样本；捕获率定义为，当覆盖率为 $q\%$ 时，对应的真实流失样本数占所有流失样本数的比例。如果在较低的覆盖率下可以获得较高的捕获率，那么说明模型的精度比较高。图 7 展示了本文构建模型的覆盖率－捕获率曲线，从中可以看出，本文提出的模型可以利用 20% 的覆盖率达到 93% 的捕获率，说明覆盖预测概率值最高的前 20% 的人，可以抓住 93% 的流失客户。

进一步地，可以根据客户预测的流失概率值进行客群划分，从而明确高流失风险人群。本文根据建模结果将客户划分为 10 类，根据预测出的流失概率从高到低进行排序。前 10% 的客户被划分为高流失风险客户，接下来的 10% 的客户被划分为次高流失风险客户，依此类推，最后 10% 的客户则是流失风险最低的。图 8 展示了流失客户的划分情况，其中在的第一组高流失风险客户中，真实的流失比例为 69.9%，次高流失风险客户群的真实流失比例为 36.8%。在实际业务中，企业可以根据不同流失风险的客户制定不同的挽回策略。例如，企业可以重点将营销预算投入高流失风险客户群，采取当面拜访、电话回访等措施，使用较为直接的沟通方式，让业务

人员直接与客户对接,有针对性地了解客户的状态和意愿。并针对客户的需要,采取适当降低佣金、提供更加便捷的服务等措施,从而将此类客户的流失风险降到最低。对于次高流失风险客户群,企业可以适当采用短信回访、电话回访等成本相对较低的策略,在控制成本的前提下,起到挽回客户的目的。

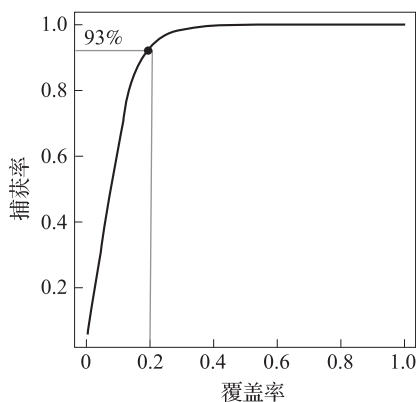


图7 覆盖率-捕获率曲线

注:说明覆盖预测概率值最高的前20%的人,可以抓住93%的流失客户。

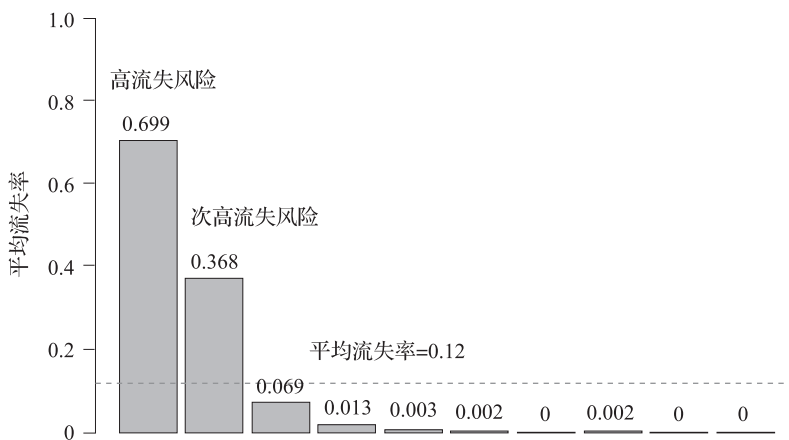


图8 流失客户划分

注:将客户划分为10类,其中高流失风险客户群的真实流失比例为69.9%,次高流失风险客户群的真实流失比例为36.8%。

不同流失因子对客户流失状态的影响程度也有所不同。图9展示了各个影响因子对流失状态的影响情况。从中可以看出,流动资产最大值的系数最大并且为负值,表示流动资产最大值与客户流失呈负相关关系。换句话说,客户的流动资产最大值越低,发生流失的可能性越高。股基权市值最大值的系数其次,也是负值,较低的股基权市值最大值可能暗示着客户的投资表现不佳或者对于当前市场形势不满意,从而增加了流失的可能性。在年龄方面,随着年龄的增长,客户流失的可能性呈现逐渐增加的趋势。其中,年龄在60岁以上的客户系数(1.11)最大,说明该年龄段

的客户更容易发生流失；而年龄在40~50岁之间的客户系数（0.40）和年龄在50~60岁之间的客户系数（0.76）也都为正值，表示这两个年龄段的客户相对于较年轻的客户，更容易发生流失。相对而言，是否持有金融产品与是否持有服务产品对客户流失状态的影响程度较为有限。因此，把握核心流失因子，能够更加高效地对客户的流失状态进行判断，更加有针对性地制定相应的挽留措施。

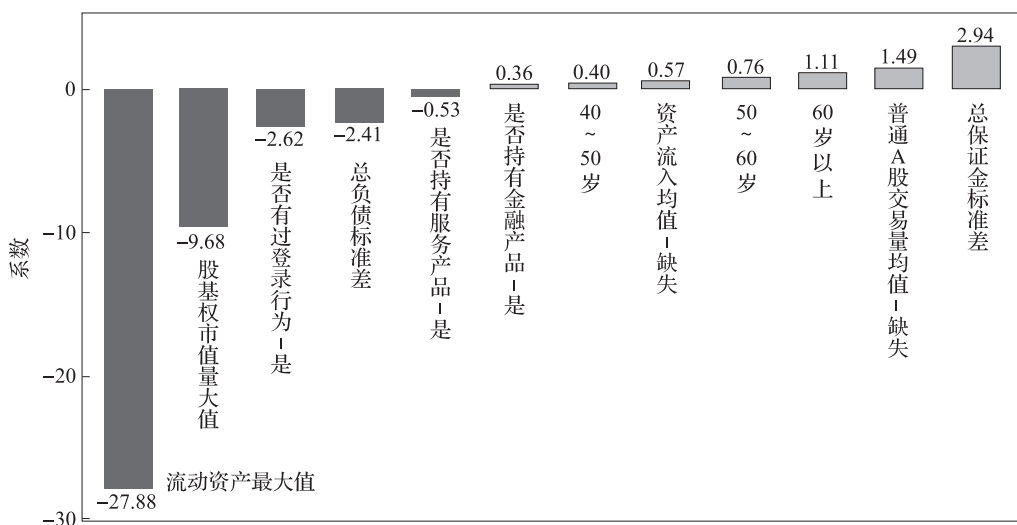


图9 流失因子的影响程度分析

注：各个影响因子对流失状态的影响程度不同，其中流动资产最大值的系数的影响程度最大，股基权市值最大值次之。

（二）企业实际环境测试

最后，我们将本文提出的基于高维特征因子的独立性筛选的客户流失预警方法应用于企业的线上生产环境，可以实时地给出客户的流失状态。为此，我们与合作企业对2020年9月到12月的全部400余万有效客户进行了测试。为验证线上模型的精度，我们从这400余万有效客户中随机抽取了5%的客户作为样本。在线上测试中，模型效果良好，预测精确率为39.2%，召回率为93.6%。对企业而言，拥有较高的召回率是相当重要的，因为召回率越高，说明在原始样本中我们抓住真正流失的客户的概率就越高。图10a为线上测试模型的ROC曲线，从中可以看出AUC值为0.914，这说明我们的模型能够较为准确地对客户流失状态进行预测。图10b为覆盖率-捕获率曲线，在线上测试中25%的覆盖率可以达到87%的捕获率，说明覆盖预测概率值最高的前25%的人，可以抓住87%的流失客户。图10c展示了线上测试中流失客户的划分情况，其中高流失风险客户群的真实流失比例为62.3%，次高流失风险客户群的真实流失比例为36.9%。线上测试结果符合预期，说明本文提出的模型能够准确识别出客户的流失情况，对流失客户做出预警。通过将模型应用于企业

的线上生产环境,我们实现了实时监控客户流失状态的目标,并且预测效果良好,有助于企业更好地进行客户关系管理和制定挽回策略。

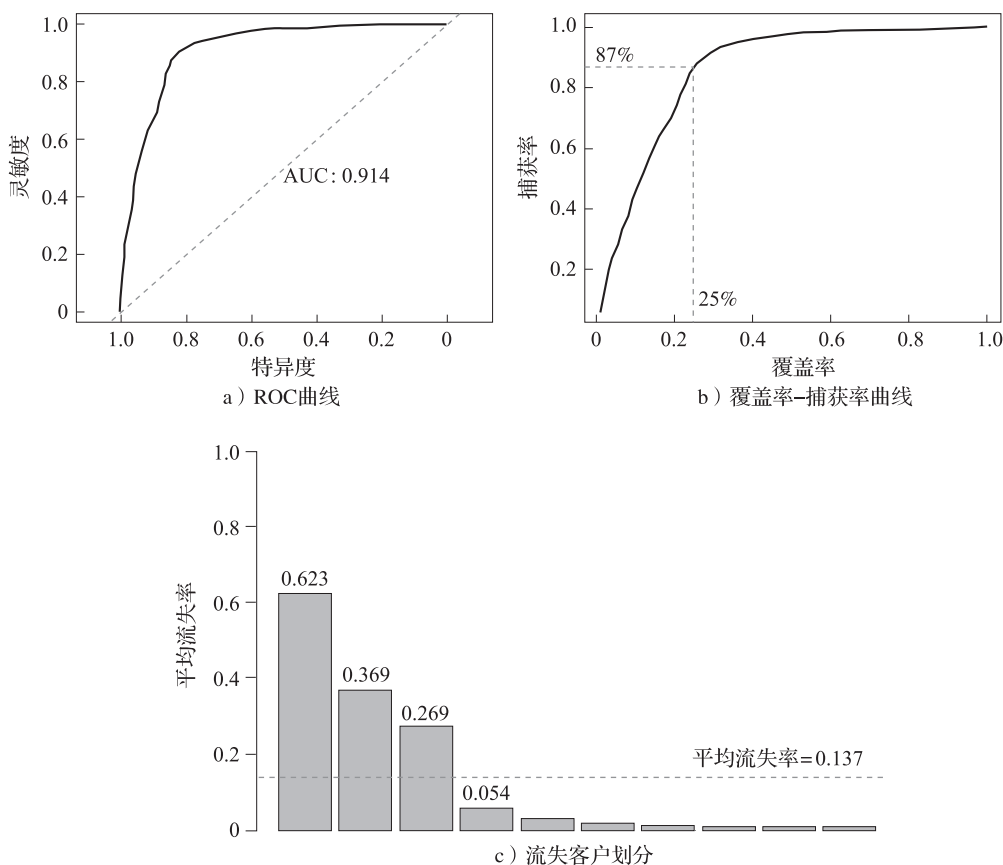


图10 线上测试效果

五、结论与未来研究方向

本文旨在研究证券公司流失客户的行为,并建立了一套在大数据背景下完整的客户流失预警模型的方法论体系,该方法论框架包括客户流失定义、关键因子挖掘、流失预警模型的建立以及实际线上测试。这套方法论可以拓展到其他证券公司。具体地,本文的方法论体系总结如下:首先,对客户流失状态进行定义,并确保在该定义下客户流失状态的稳定性,从而为本文的研究奠定了基础;其次,本文通过从市值、交易、持有服务情况、基本信息、行为信息等资产类和非资产类的多个维度挖掘高维流失因子,从而全方面地刻画流失客户的行为;再次,提出了基于高维流失因子的独立性筛选流程,用于探究对流失状态具有显著影响的因子;最后,建立了高捕获率、高准确率为客户流失预警模型,并通过对客户进行实时监控,帮助公

司尽早发现潜在流失客户群体。基于流失预警模型进行客户群体细分和策略研究,可以帮助公司更好地进行客户关系管理,增强证券公司抵抗客户流失风险的能力。

虽然本文建立的模型能够较为准确地对流失客户进行预警,并能够在实际生产环境中对客户实时流失状态进行跟踪预测,但仍有以下不足:一是由于时间跨度与数据管理的问题,本文无法提供更多关于原始字段的可视化结果;二是本文仅展示了部分示例日期下的模型表现,没有展示更长范围内的预测结果以及更多模型的对比结果。但是,本文提出的模型框架具有良好的泛化性能。例如,企业利用本文模型在2019年10月份的小范围测试结果的精确率为0.986、召回率0.769,模型效果良好。

对于证券行业的客户流失问题,今后仍可在以下方面进行拓展:首先,可以探索客户行为数据库,从客户行为的角度构建更多流失因子,以进一步丰富模型的特征;其次,可以基于证券公司的营销实践构建更多流失因子,并尝试追踪营销触达效果,为客户挽留提供可借鉴的思路和方法;最后,可以探索不同客户群体的流失特点,从而为不同客户群体提出更加有针对性的挽回策略,提高挽回效率。这些拓展将进一步提升模型的准确性和实际应用价值,有助于更好地解决证券行业面临的客户流失问题。

参考文献

- [1] 杜修平,王中,2009. 基于决策树的证券客户流失模型[J]. 计算机应用与软件, 26(9): 230-233.
- [2] 郭红丽,2006. 客户体验维度识别的实证研究——以电信行业为例[J]. 管理科学, 19(1): 59-65.
- [3] 李季,张帅,周静,2020. 服务与需求的匹配度对客户流失的影响研究:基于电信行业的客户数据实验[J]. 管理评论, 32(5): 192-204.
- [4] 周静,周小宇,王汉生,2017. 自我网络特征对电信客户流失的影响[J]. 管理科学, 30(5): 28-37.
- [5] AHMAD A K, JAFAR A, ALJOUAA K, 2019. Customer churn prediction in telecom using machine learning in big data platform[J]. Journal of Big Data. DOI: 10.1186/s40537-019-0191-6.
- [6] AHN J H, HAN S P, LEE Y S, 2006. Customer churn analysis: churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry[J]. Telecommunications Policy, 30(10/11): 552-568.
- [7] AHN Y, KIM D, LEE D J, 2019. Customer attrition analysis in the securities industry: a large-scale field study in Korea[J]. International Journal of Bank Marketing. DOI: 10.1108/IJBM-04-2019-0151.
- [8] AMIN A, ANWAR S, ADNAN A, NAWAZ M, ALAWFI K, HUSSAIN A, HUANG K, 2017. Customer churn prediction in the telecommunication sector using a rough set approach[J]. Neurocomputing, 237: 242-254.
- [9] ASAMOAH D A, SHARDA R, KALGOTRA P, OTT M, 2016. Teaching case who renews?who leaves?identifying customer churn in a telecom company using big data techniques[J]. Journal of Information Systems Education, 27(4): 223-232.
- [10] ASCARZA E, NESLIN S A, NETZER O, ANDERSON Z, FADER P S, GUPTA S, HARDIE B G S, LEMMENS A, LIBAI B, NEAL D, PROVOST F, SCHRIFT R, 2017. In pursuit of enhanced customer retention management: review, key issues, and future directions[J]. Customer Needs and Solutions, DOI: 10.1007/s40547-017-0080-0.

- [11] AZEEM M, USMAN M, FONG A C M, 2017. A churn prediction model for prepaid customers in telecom using fuzzy classifiers[J]. *Telecommunication Systems*, 66: 603–614.
- [12] BOLTON R N, KANNAN P K, BRAMLETT M D, 2000. Implications of loyalty program membership and service experiences for customer retention and value[J]. *Journal of the Academy of Marketing Science*, 28(1): 95–108.
- [13] COUSSEMENT K, LESSMANN S, VERSTRAETEN G, 2017. A comparative analysis of data preparation algorithms for customer churn prediction: a case study in the telecommunication industry[J]. *Decision Support Systems*, 95(3): 27–36.
- [14] ERIA K, MARIKANNAN B P, 2018. Systematic review of customer churn prediction in the telecom sector[J]. *Journal of Applied Technology and Innovation*, 2(1): 7–14.
- [15] FAN J, LV J, 2008. Sure independence screening for ultra-high dimensional feature space[J]. *Journal of the Royal Statistical Society*, 70(5): 849–911.
- [16] FORNELL C, WERNERFELT B, 1987. Defensive marketing strategy by customer complaint management: a theoretical analysis[J]. *Journal of Marketing Research*, 24(4): 337–346.
- [17] GANESH J, ARNOLD M J, REYNOLDS K E, 2000. Understanding the customer base of service providers: an examination of the differences between switchers and stayers[J]. *Journal of Marketing*, 64(3): 65–87.
- [18] GERPOTT T J, RAMS W, SCHINDLER A, 2001. Customer retention, loyalty, and satisfaction in the German mobile cellular telecommunications market[J]. *Telecommunications Policy*, 25(4): 249–269.
- [19] GIUDICATI G, RICCABONI M, ROMITI A, 2013. Experience, socialization and customer retention: lessons from the dance floor[J]. *Marketing Letters*, 24(4): 409–422.
- [20] GLADY N, BAESENS B, CROUX C, 2009. Modeling churn using customer lifetime value[J]. *European Journal of Operational Research*, 197(1): 402–411.
- [21] GUO L, ZHANG M, SUN L, WANG Z, 2006. Churn analysis model of securities business based on the decision tree [C] // 2006 6th World Congress on Intelligent Control and Automation. DOI: 10.1109/WCICA.2006.1714241.
- [22] HAENLEIN M, 2013. Social interactions in customer churn decisions: the impact of relationship directionality[J]. *International Journal of Research in Marketing*, 30(3): 236–248.
- [23] HWANG H, JUNG T, SUH E, 2004. An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry[J]. *Expert Systems with Applications*, 26(2): 181–188.
- [24] JAMAL Z, BUCKLIN R E, 2006. Improving the diagnosis and prediction of customer churn: a heterogeneous hazard modeling approach[J]. *Journal of Interactive Marketing*, 20(3–4): 16–29.
- [25] KAYAALP F, 2017. Review of customer churn analysis studies in telecommunications industry[J]. *Karaelmas Fen ve Mühendislik Dergisi*, 7(2): 696–705.
- [26] KRISHNAN M S, RAMASWAMY V, MEYER M C, DAMIEN P, 1999. Customer satisfaction for financial services: the role of products, services, and information technology[J]. *Management Science*, 45(9): 1177–1288.
- [27] LEE E B, KIM J, LEE S G, 2017. Predicting customer churn in mobile industry using data mining technology[J]. *Industrial Management & Data Systems*, 117(1): 90–109.
- [28] LEWIS M, 2004. The influence of loyalty programs and short-term promotions on customer retention[J]. *Journal of Marketing Research*, 41(3): 281–292.
- [29] MAHMOUD M A, 2019. Gender, e-banking, and customer retention[J]. *Journal of Global Marketing*, (5): 1–19.
- [30] MALDONADO S, DOMÍNGUEZ G, OLAYA D, VERBEKE W, 2021. Profit-driven churn prediction for the mutual fund industry: a multisegment approach[J]. *Omega*. DOI: 10.1016/j.omega.2020.102380.
- [31] NIE G, ROWE W, ZHANG L, TIAN Y, SHI Y, 2011. Credit card churn forecasting by logistic regression and decision tree[J]. *Expert Systems with Applications*, 38(12): 15273–15285.
- [32] NITZAN I, LIBAI B, 2011. Social effects on customer retention[J]. *Journal of Marketing*, 75(6): 24–38.
- [33] PEPPARD J, 2000. Customer relationship management (CRM) in financial services[J]. *European Management Journal*, 18(3): 312–327.

- [34] RAFTERY A E, MADIGAN D, HOETING J A, 1997. Bayesian model averaging for linear regression models[J]. Journal of the American Statistical Association, 92(437): 179–191.
- [35] REINARTZ W J, KUMAR V, 2003. The impact of customer relationship characteristics on profitable lifetime duration[J]. Journal of Marketing, 67(1): 77–99.
- [36] SANDHYA N, SAMUEL P, CHACKO M, 2019. Feature intersection for agent-based customer churn prediction[J]. Data Technologies and Applications, 53(3): 318–332.
- [37] SCHWARZ G, 1978. Estimating the dimension of a model[J]. Annals of Statistics, 6(2): 461–464.
- [38] TAN Y F, SHUAN L H, YAN L J, GUO X, SOO W K, 2017. Prediction on customer churn in the telecommunications sector using discretization and Naïve Bayes classifier[J]. International Journal of Advances in Soft Computing and Its Applications, 9(3): 23–35.
- [39] VAN DEN POEL D, LARIVIÈRE B, 2004. Customer attrition analysis for financial services using proportional hazard models[J]. European Journal of Operational Research, 157(1): 196–217.
- [40] WEI C P, CHIU I T, 2002. Turning telecommunications call details to churn prediction: a data mining approach[J]. Expert Systems with Applications, 23(2): 103–112.
- [41] ZHAO J, DANG X H, 2008. Bank customer churn prediction based on support vector machine: taking a commercial bank's VIP customer churn as the example[C]//4th International Conference on Wireless Communications, Networking and Mobile Computing. IEEE, 1–4. DOI: 10. 1109/WiCom. 2008. 2509.

Customer Churn Model Based on High-Dimensional Factors —A Study on Securities Companies

Tianchen Gao

(School of Economics, Xiamen University)

Hao Qu

(Founder Securities Co., Ltd.)

Feifei Wang Jing Zhou*

(Center for Applied Statistics, School of Statistics, Renmin University of China)

Summary: The securities industry is widely recognized as one of the most data-intensive sectors, characterized by diverse business scenarios. However, due to stringent regulations and high entry barriers, the growth of new securities companies is sluggish. Nevertheless, competition among the industry is intensifying. In order to expand their market share, securities companies have employed various strategies to attract new customers. They have invested considerable effort in providing personalized marketing services to enhance their customer relationship management capabilities. However, a top priority for securities companies in their customer relationship management efforts is retaining existing customers and preventing potential customer churn.

This research focuses on customer churn in the securities sector. The data used in this study consists of user-level transaction data provided by a prominent domestic securities company. Initially, we conduct data inspection and categorize the raw data into asset variables and non-asset variables. After cleaning the original data and filtering out relevant variables for later analysis, we define customer churn based on the stability of the churn status, drawing from practical experience within the securities industry. We then investigate how the number of trading days, logins, and assets impact the state of customer churn, ultimately arriving at a viable and effective definition of customer churn. Having established the response variable, the subsequent crucial task is to identify meaningful impact factors from hundreds of raw variables. To address this, we propose an independent screening method based on high-dimensional features. We first divide the churn factors into asset variables and non-asset variables, each represented by 8 and 4 aspects, respectively. From these aspects, we derive a series of indicators. We also calculate the ratios or products of asset and non-asset factors to obtain 10 composite churn factors. Subsequently, we employ the univariate AUC method, considering data quality and actual business context, to screen the churn factors. As a result, 8 asset-type factors, 21 non-asset-type factors, and 2 compound factors are selected for modeling.

* Corresponding Author: Jing Zhou, Center for Applied Statistics, School of Statistics, Renmin University of China, E-mail: zhoujing_89@126.com.

Building upon the logistic regression model framework, we propose separate daily and weekly customer churn models, both based on the screening features. The results demonstrate that both asset-type and non-asset-type variables significantly influence customer churn prediction. The daily customer churn prediction model achieves an AUC above 0.95, indicating strong prediction accuracy. Moreover, compared to the daily model, the weekly average model exhibits lower computational cost, higher prediction efficiency, and greater stability. The model also reveals underlying mechanisms behind customer churn. For example, the coefficient of the maximum value of stock fund option market capitalization is negative, indicating an inverse correlation between the maximum value of stock fund option market capitalization and customer churn. A lower maximum value of stock fund option market capitalization might suggest poor investment performance or dissatisfaction with the current market situation, thereby increasing the likelihood of customer churn.

Regarding age, the probability of customer churn follows the pattern: 60 years and above > 50 ~ 60 years > 40 ~ 50 years > below 40 years. This suggests that older customers are more prone to churn. Particularly, customers aged between 50 and 60 years are more likely to churn, possibly due to their proximity to retirement age, leading to higher financial planning and retirement considerations. Among the age groups, customers aged 60 years and above are most susceptible to churn. This could be attributed to changes in financial planning and service requirements after retirement or other factors that make them more likely to transfer their investments or switch service providers.

Our research findings can provide strategic analysis for companies to recover customers. Based on the proposed daily customer churn model, it is possible to calculate the predicted churn probability value for each customer and sort the samples in descending order according to the prediction values. We introduced the coverage-capture rate curve to evaluate the model's prediction accuracy in practical business scenarios. Based on the modeling results, we classified customers into ten categories. In real-world business operations, companies can develop specific recovery strategies for customers with different churn risks. For instance, they can focus marketing budgets on high-risk churn customers, engaging in face-to-face visits, phone follow-ups, and direct communication with customers. By gaining targeted insights into customer status and preferences, they can implement appropriate measures such as reducing commissions or offering more convenient services, effectively minimizing the risk of customer churn.

Finally, we applied the customer churn model based on high-dimensional features proposed in this paper to the online production environment of the enterprise, which enables real-time identification of customer churn status. To achieve this, we conducted tests in collaboration with a partner company on over four million valid customers from September to December 2020. To validate the accuracy of the online model, we randomly sampled 5% of these four million valid customers as the testing sample. During the online testing, the performance was excellent, with a prediction accuracy of 39.2% and a recall rate of 93.6%. For the enterprise, having a higher recall rate is particularly crucial because a higher recall rate indicates a higher probability of capturing truly churned customers from the original sample.

Keywords: Customer Churn; Securities Industry; High-Dimensional Factors; Independent Screening; Logistic Regression

JEL Classification: G20