Computer Science and Application 计算机科学与应用, 2025, 15(11), 305-319
Published Online November 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1511306

Hans汉斯

基于SMOTE增强与多模型择优的银行客户忠诚度预测研究

刘政永，孙　娜

河北金融学院河北省金融科技应用重点实验室，河北　保定

收稿日期：2025年10月22日；录用日期：2025年11月19日；发布日期：2025年11月27日

摘　要

本研究针对银行客户流失预测问题，通过系统性数据处理、可视化分析与特征工程，构建了多种机器学习模型(包括逻辑回归、随机森林、AdaBoost和支持向量机)，并基于ROC曲线、F1分数等指标评估模型性能。核心发现表明，随机森林模型在应对数据不平衡和捕捉复杂特征关系方面表现最优(测试集F1分数达0.8546)，显著优于其他模型；方法贡献在于提出了一套结合可视化探索与特征优化的建模框架，强调了数据质量与衍生特征对预测性能的关键作用；研究局限包括数据来源单一性及模型对特定业务场景的泛化能力有待进一步验证。本研究为银行客户忠诚度管理提供了数据驱动的决策支持。

关键词

银行客户忠诚度，SMOTE增强，逻辑回归，随机森林，AdaBoost，支持向量机

Research on the Prediction of Bank Customer Loyalty Based on SMOTE Enhancement and Multi-Model Optimization

Zhengyong Liu, Na Sun

Hebei Key Laboratory of Financial Technology Application, Hebei Finance University, Baoding Hebei

Received: October 22, 2025; accepted: November 19, 2025; published: November 27, 2025

Abstract

This study addresses the problem of bank customer churn prediction. By systematic data processing,

文章引用: 刘政永, 孙娜. 基于 SMOTE 增强与多模型择优的银行客户忠诚度预测研究[J]. 计算机科学与应用, 2025, 15(11): 305-319. DOI: 10.12677/csa.2025.1511306

visualization analysis, and feature engineering, a variety of machine learning (including logistic regression, random forest, AdaBoost, and support vector machine) are constructed and evaluated based on ROC curves, F1 scores, and other metrics. Core findings show that the random forest model performs the best in dealing with data imbalance and capturing complex feature relationships (achieving a F1 score of 0.8546 the test set), significantly outperforming the other models. The methodological contribution lies in proposing a modeling framework that combines visualization exploration and feature optimization, emphasizing the critical roles of quality and derived features in prediction performance. Research limitations include the singularity of data sources and the need for further validation of the model's generalization capability in specific business scenarios. This study provides data-driven decision support for bank customer loyalty management.

Keywords

Bank Customer Loyalty, SMOTE Enhancement, Logistic Regression, Random Forest, AdaBoost, Support Vector Machine

1. 引言

目前银行产品存在同质化现象，客户选择产品和服务的途径越来越多，对产品的忠诚度越来越低[1]。为了提高客户对银行的忠诚度和银行营销量，商业银行迫切需要转变经营理念，从“产品销售导向”业务模式向“以客户为中心”转变，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[2]。

客户忠诚度主要体现为客户的行为和态度[3]。客户行为主要表现为产品重复购买的频率，而客户态度主要表现为情感的倾向[4]。为了有效挖掘客户忠诚度，需要从短期客户产品购买数据和长期客户资源信息中分析客户需求指标。其中，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。

本研究首先通过数据探索与清洗，产品营销数据可视化和客户流失因素可视化分析，构建特征，利用相关系数矩阵和热力图进行变量选择，构建逻辑回归模型进行预测；其次基于数据不平衡问题，采用SMOTE 方法进行不平衡数据处理，选择逻辑回归、随机森林、AdaBoost 和 SVM 模型四种模型进行建模分析，通过混淆矩阵，F1 sorce 等指标最终选择随机森林模型对预测数据进行预测，构建银行客户忠诚度模型。

2. 数据来源及说明

数据来源于第五届“泰迪杯”数据分析技能赛 B 题竞赛项目。银行客户忠诚度分析数据包括短期客户产品购买数据和长期客户资源信息数据，附件数据详细情况如表 1 所示。

长期客户资源信息数据记录了往期该银行客户流动状态及客户信息，包括客户基本信息、客户户龄与金融资产、客户活跃状态与流失情况等，具体的数据指标说明如表 2 所示短期客户产品购买数据记录了往期银行营销活动中客户购买产品的信息，包含客户的基本信息、上次活动后拜访客户信息和上次活动产品购买结果等，具体的数据指标说明如表 3 所示。

Table 1. Details of the attachment data
表 1. 附件数据详细情况

数据集	数据名称	备注
短期客户产品购买数据	short-customer-data.csv	客户购买产品的记录数据
长期客户资源信息数据	long-customer-train.csv	训练集数据
	long-customer-test.csv	测试集数据

Table 2. Explanation of data indicators for long-term customer resource information
表 2. 长期客户资源信息数据指标说明

序号	字段	说明
1	CustomerId	客户 ID
2	CreditScore	表示信用资格，数值越大表明信用越高
3	Gender	客户性别，0 表示男性，1 表示女性
4	Age	客户年龄
5	Tenure	账号户龄，客户在这家银行存款的时长，以年为单位
6	Balance	AUM，客户的金融资产
7	NumOfProducts	客户购买产品数量
8	HasCrCard	客户持有信用卡状态，客户有信用卡为 1，否则为 0
9	IsActiveMember	客户活动状态，客户处于活跃状态为 1，否则为 0
10	EstimatedSalary	客户个人年收入
11	Exited	客户流失情况，已流失为 1，否则为 0

Table 3. Explanation of short-term customer product purchase data indicators
表 3. 短期客户产品购买数据指标说明

	字段	说明
基本数据	user_id	客户 id，例 BA2200001
	age	年龄(数字)
	job	工作类型包含 11 种，分别为行政人员(admin.)，蓝领(blue-collar)、企业家(entrepreneur)、家政(housemaid)、企业管理层(management)、退休(retired)、个体经营者(self-employed)、服务行业人员(services)、学生(student)、技术员(technician)、失业(unemployed)
	marital	婚姻状况包含 3 种，分别为离婚(divorced)、已婚(married)、单身(single)，注：离婚指离婚或丧偶)
	education	教育情况包含 5 种，分别为研究生以上(postgraduate)、高中(high school)、文盲(illiterate)、专科(junior college)、大学学位(undergraduate)
	default	信用违约情况包含 2 种，分别为否(no)、是(yes)
	housing	住房贷款情况包含 2 种，分别为否(no)、是(yes)
	loan	个人贷款情况包含 2 种，分别为否(no)、是(yes)

续表

上次活动后拜访客户信息	contact	联系人通信类型包含 2 种，分别为蜂窝(cellular)、电话(telephone)
	month	最近一次拜访客户的月份，分别为一月(jan)、二月(feb)、三月(mar)……十一月(nov)、十二月(dec)
	day_of_week	最近一次拜访客户的星期，分别为星期一(mon)、星期二(tue)、星期三(wed)、星期四(thu)、星期五(fri)
	duration	最近一次拜访客户的通话时长，以秒为单位(数字)，如果通话时长 = 0，表示没有成功联系上客户
其他属性	poutcome	上一次银行活动，客户购买产品的结果包括 3 种，分别为失败(failure)、不存在(nonexistent)、成功(success)
产品购买结果	y	本次银行活动客户购买产品的结果，分别为否(no)、是(yes)

3. 建模过程

3.1. 数据探索与清洗

本研究分别对短期客户产品购买数据(简称“短期数据”)和长期客户资源信息数据(简称“长期数据”)进行探索性分析与清洗。数据清洗旨在处理缺失值、异常值及重复数据，以确保数据质量。缺失值可能导致模型偏差，异常值可能反映记录错误或特殊个案，重复数据则可能造成信息冗余。通过系统性清洗，为后续建模奠定基础。

3.1.1. 短期数据预处理

首先，对短期数据进行读取与初步检查，以了解数据规模、字段类型及分布情况。随后，识别并删除存在缺失值的行，确保样本完整性。进一步地，针对“user_id”字段检测重复记录，并剔除重复项以避免样本偏差。最终，将清洗后的数据保存至文件“result1_1.xlsx”中，用于后续分析。

3.1.2. 长期数据预处理

长期数据通过读取后保存为“result1_2.xlsx”文件。在年龄字段(“Age”)中，发现存在数值异常(如−1、0 及符号“−”)，这些异常值可能源于数据录入错误，因此将其所在行删除。此外，年龄字段中混杂空格与“岁”等非标准字符，通过文本处理保留纯数值信息，并将修正后的结果更新至原字段，确保数据格式统一。

3.1.3. 短期数据特征编码

为将字符型数据转化为数值型以适配机器学习算法，对短期数据中的分类变量进行编码。例如，将信用违约情况(“否”、“是”)分别映射为$\{0,1\}$。此过程通过构建映射规则实现类别变量的数字化转换，最终将编码结果保存至“result1_3.xlsx”。

3.2. 产品营销数据可视化分析

基于短期数据，通过可视化手段分析客户特征与产品购买行为间的关联性，挖掘客户忠诚度潜在规律。

3.2.1. 相关性热力图

通过计算各变量间的 Pearson 相关系数，构建相关性矩阵，并利用热力图可视化呈现，见图 1。

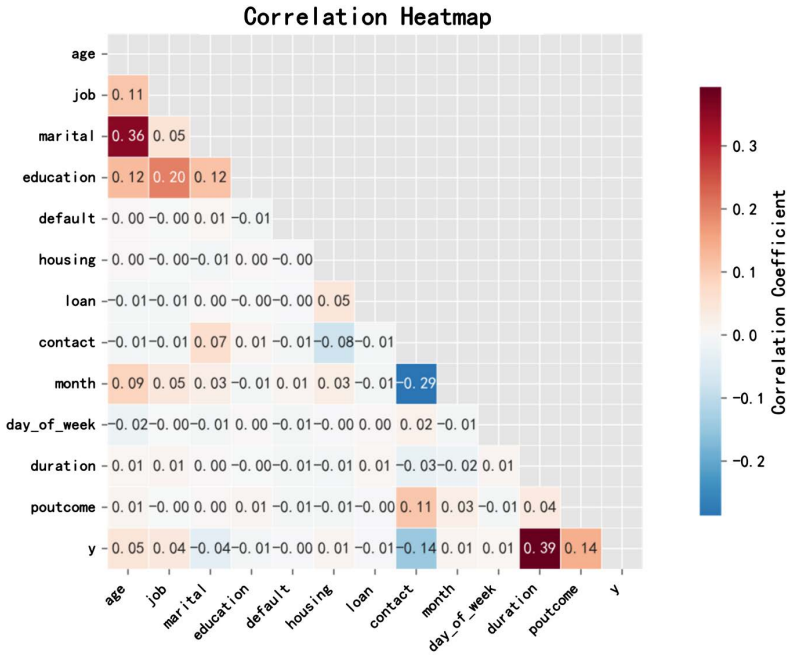

Figure 1. Heat map of correlation coefficients of variables
图 1. 变量相关系数热力图

根据图 1 显示，“duration”等变量与目标变量“y”相关性较高，为后续特征选择提供依据。

3.2.2. 年龄－购买行为分组柱状图

绘制分组柱状图，横轴为年龄分段，纵轴为客户占比，对比不同产品购买结果下各年龄段的分布差异，见图 2。

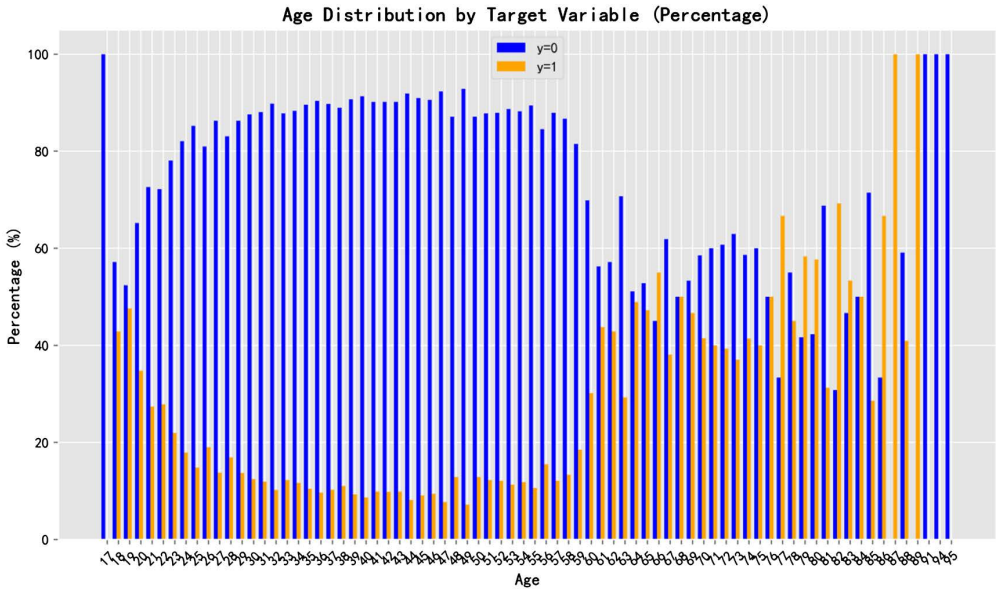

Figure 2. Age-purchasing behavior group column chart
图 2. 年龄－购买行为分组柱状图

根据图 2 显示，年轻客户在特定产品中的购买比例较高，而中年客户群体则表现出不同的偏好。

3.2.3. 职业与购买行为饼图

针对蓝领与学生群体，绘制产品购买情况的饼图，展示各职业群体中购买与未购买的占比，见图 3。

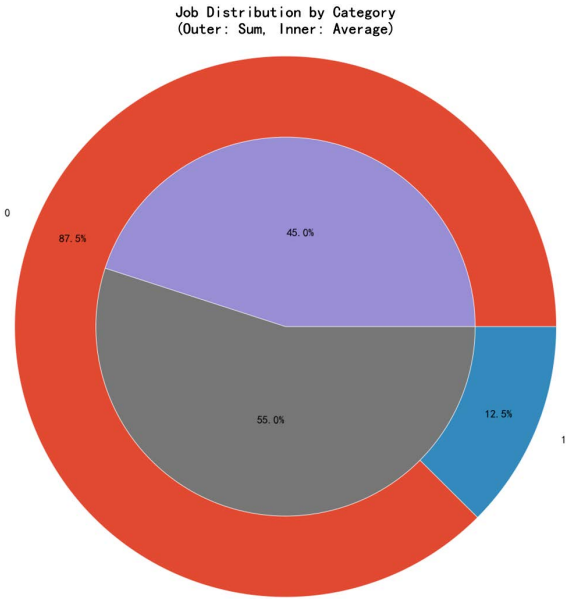

Figure 3. Pie chart of occupation and purchasing behavior
图 3. 职业与购买行为饼图

根据图 3 表明，学生群体的购买意愿显著高于蓝领群体。

3.2.4. 通话时长箱线图

以产品购买结果为分组变量，绘制通话时长的箱线图，见图 4。

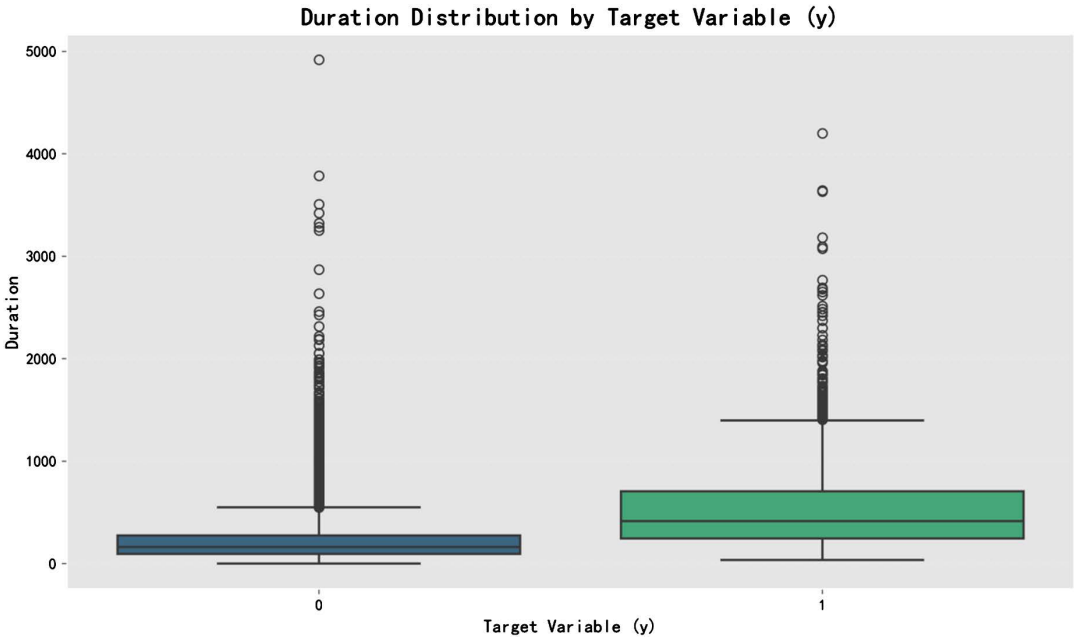

Figure 4. Box plot of call duration by group
图 4. 通话时长分组箱线图

根据图 4 显示，购买产品的客户通话时长中位数显著高于未购买客户，表明通话时长可能与购买意愿存在正相关。

3.3. 客户流失因素可视化分析

基于长期数据，通过多维度可视化探究客户流失的影响因素。

3.3.1. 年龄－流失占比折线图

按流失状态分组计算各年龄段的客户占比，绘制双折线图，见图 5。

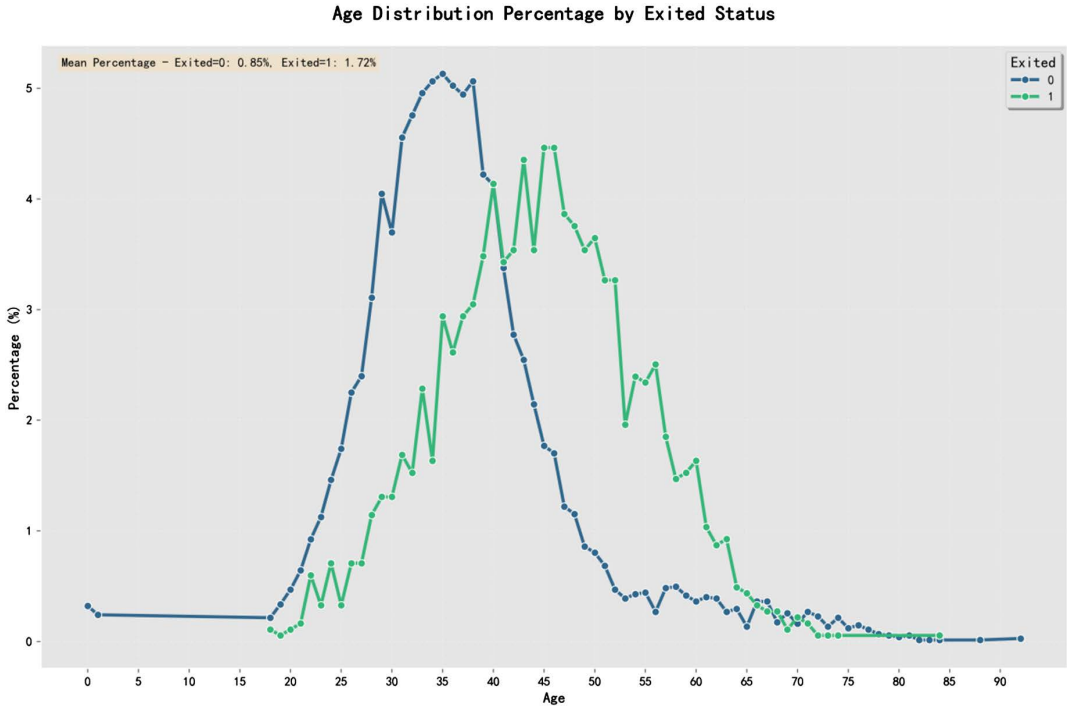

Figure 5. Line chart of age-churn ratio
图 5. 年龄－流失占比折线图

根据图 5 显示，随着年龄增长，流失客户占比上升速度高于未流失客户，表明高龄客户流失风险较高。

3.3.2. 信用资格－年龄散点图

通过散点图展示客户信用资格与年龄的分布，并按流失状态着色，见图 6。

根据图 6 分析发现，信用资格与年龄之间无明显相关性，说明二者在流失行为中可能独立发挥作用。

3.3.3. 户龄－流失堆叠柱状图

构建户龄与流失情况的交叉表，计算各户龄段中流失与未流失客户的占比，并绘制堆叠柱状图，见图 7。

根据图 7 显示，户龄较短的客户流失比例较高，而长期客户稳定性更强。

3.3.4. 资产阶段与流失热力图

根据表 4 和表 5 将客户按账号户龄和金融资产划分为新客户、稳定客户、老客户及低、中下、中上、高资产阶段。

统计各组合下的流失客户量并通过热力图可视化，见图 8。

根据图 8 热力图显示，中低资产阶段的新客户流失量较大，而高资产老客户的流失率较低。

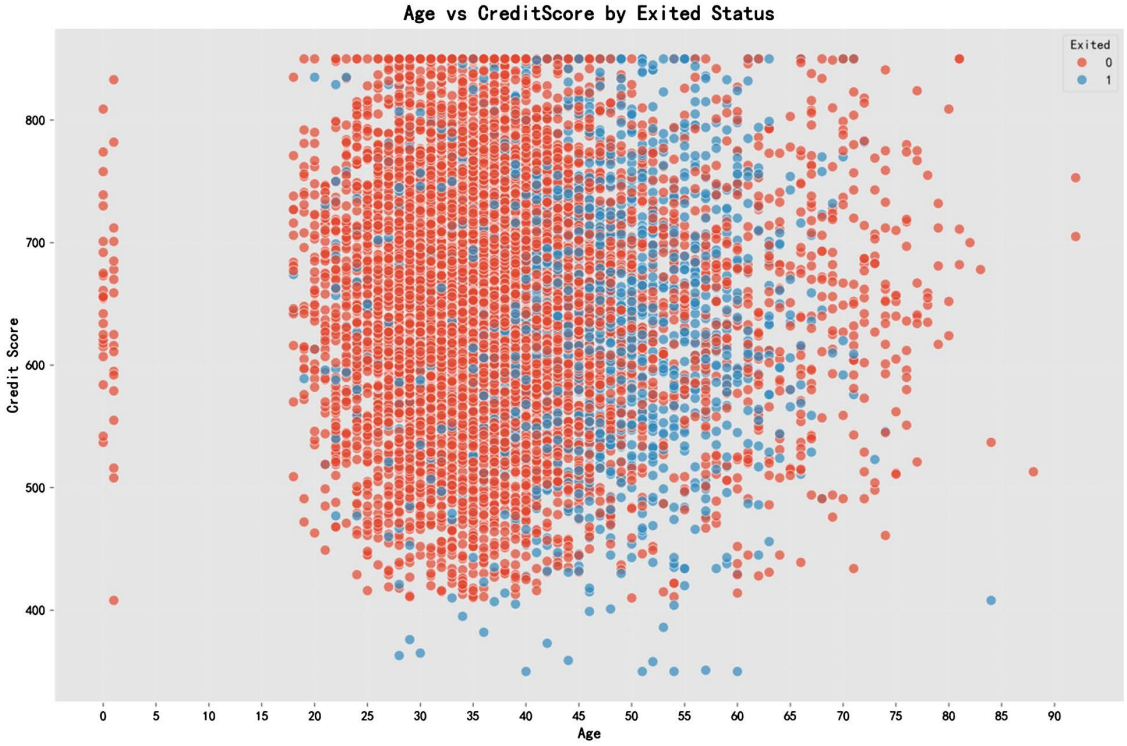

Figure 6. Scatter plot of credit qualification-age
图 6. 信用资格－年龄散点图

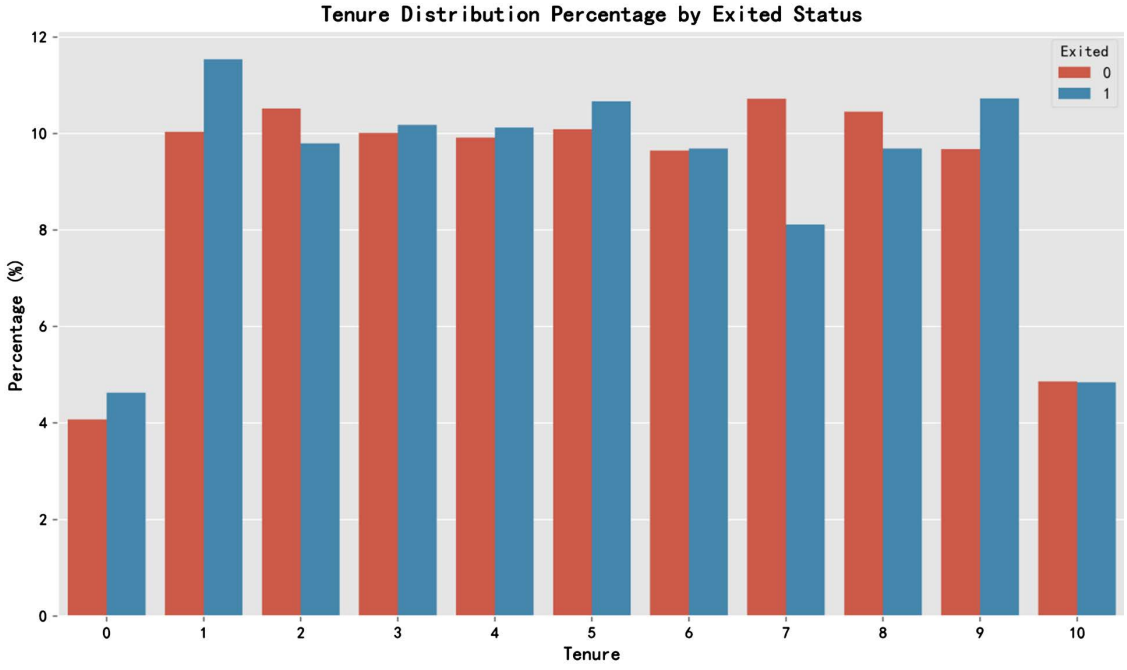

Figure 7. Stacked bar chart of churn by tenure
图 7. 户龄－流失堆叠柱状图

Table 4. Distribution of account age
表 4. 账号户龄划分情况

账号户龄区间	客户状态
[0, 3]	新客户
(3, 6]	稳定客户
>6	老客户

Table 5. Distribution of customer financial assets
表 5. 客户金融资产划分情况

客户金融资产区间	资产阶段
[0, 50,000]	低资产
(50,000, 90,000]	中下资产
(90,000, 120,000]	中上资产
>120,000	高资产

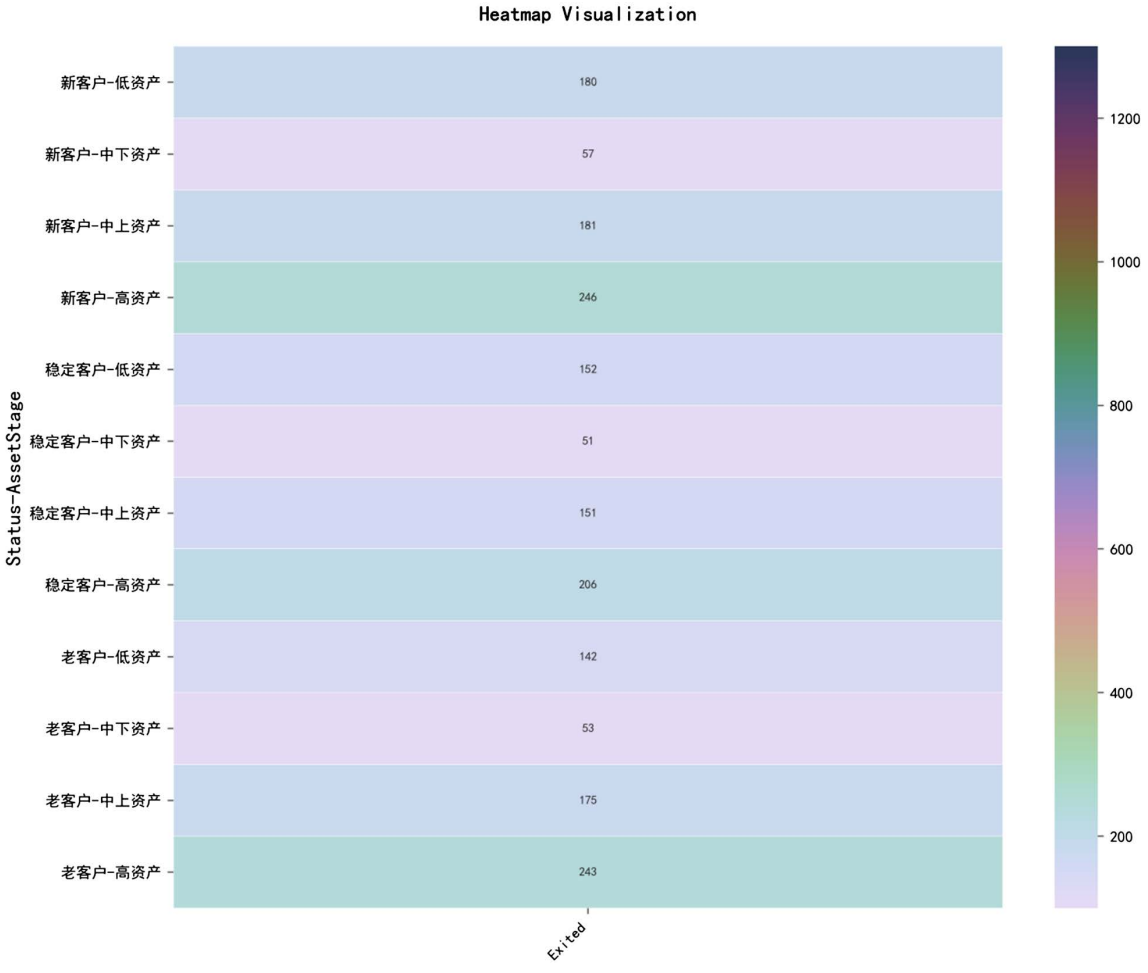

Figure 8. Heatmap of asset phases and leakage
图 8. 资产阶段与流失热力图

3.4. 特征构建

为增强模型表达能力，基于长期数据构建衍生特征。

3.4.1. 客户活跃程度特征

根据表 6 账号户龄划分客户状态(新客户、稳定客户、老客户)，结合活跃状态生成“IsActiveStatus”特征，数值化表征客户活跃度。

Table 6. Construction rules for new and old customer activity feature
表 6. 新老客户活跃程度特征构建规则

新老客户活跃程度		活跃状态	
		0	**1**
账号户龄	新客户	0	3
	稳定客户	1	4
	老客户	2	5

3.4.2. 资产活跃程度特征

依据表 7 金融资产分段(低、中下、中上、高资产)与活跃状态，构建“IsActiveAssetStage”特征，反映不同资产水平客户的活跃差异。

Table 7. Construction rules for customer activity feature with different deposit amount
表 7. 不同存款额客户活跃程度特征构建规则

不同金融资产客户活跃程度		活跃状态	
		0	**1**
资产阶段	低资产	0	6
	中下资产	1	7
	中上资产	2	8
	高资产	3	9

3.4.3. 资产－信用卡持有特征

结合表 8 资产阶段与信用卡持有状态，生成“CrCardAssetStage”特征，量化资产与信用工具的交互作用。

Table 8. Construction rules for the characteristics of credit card holding status of different financial assets
表 8. 不同金融资产信用卡持有状态特征构建规则

不同金融资产信用卡持有状态		信用卡持有状态	
		0	**1**
资产阶段	低资产	0	6
	中下资产	2	7
	中上资产	5	9
	高资产	5	9

3.5. 银行客户长期忠诚度预测建模

针对长期数据中存在的类别不平衡(未流失客户数远高于流失客户)及特征量纲差异问题，采用多种机器学习方法构建客户流失预测模型。

3.5.1. 特征选择

通过计算特征与目标变量“Exited”的相关系数，筛选绝对值大于 0.05 的变量(包括客户信用资格、年龄、户龄、金融资产等 8 项特征)，以降低维度并提升模型效率。

3.5.2. 逻辑回归模型

基于选定特征，将数据按 7:3 划分为训练集与验证集，建立逻辑回归模型。使用 ROC 曲线(见图 9)对模型进行评估。

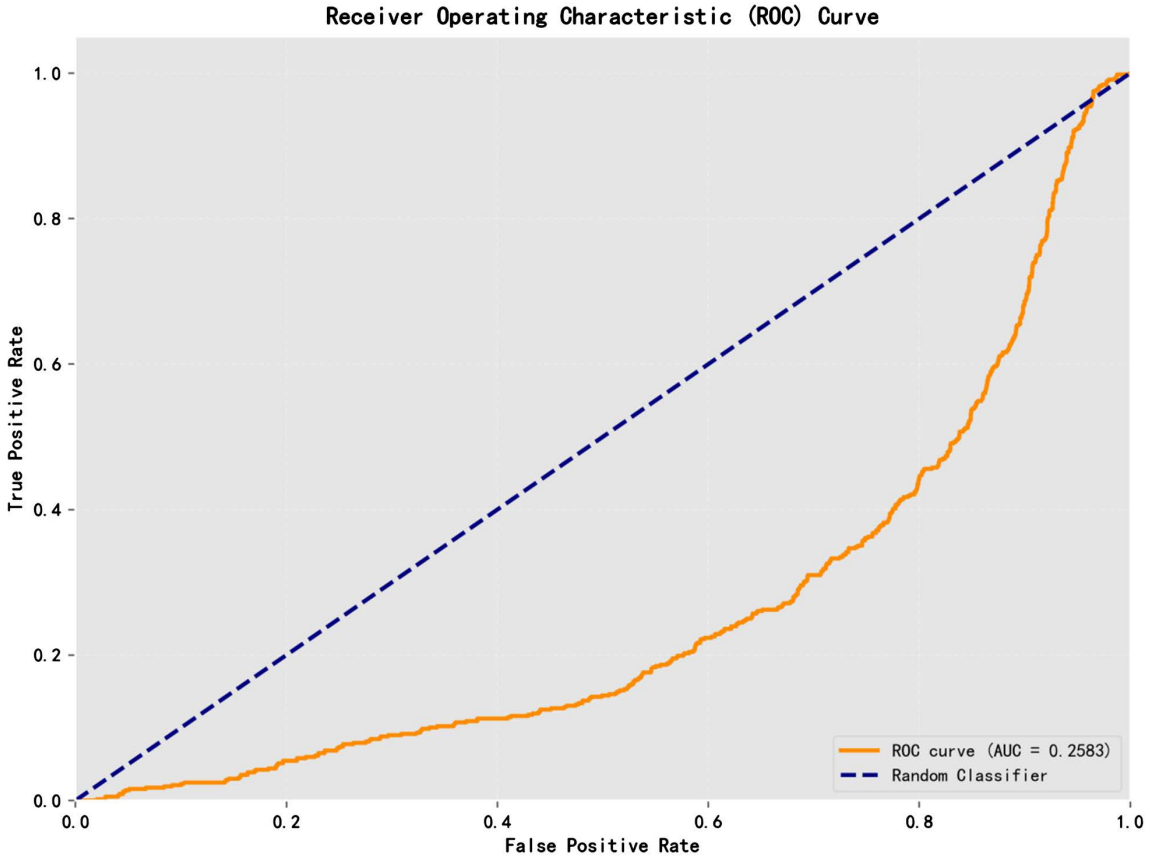

Figure 9. ROC curve of the logistic regression model
图 9. 逻辑回归模型 ROC 曲线

由图 9 可知，该模型表现极不理想，AUC 值仅为 0.2583，远低于随机分类器的 0.5 基准线，表明模型判别能力严重不足；曲线始终位于对角线下方，说明模型预测结果与真实类别完全相反，存在严重的系统性误判问题，有可能是数据平衡造成的。

3.5.3. SMOTE 过采样与逻辑回归

客户流失预测研究显示，复杂的机器学习技术能够有效解决数据不平衡问题，并提供具有高预测准确性的可解释模型。Ke Peng 等人(2023 年)特别强调了 SMOTEEN 比其他采样技术更有效，可用于平衡

银行数据[5]。Luong Thanh Tam 等人(2025 年)发现，集成模型优于单一模型，其中模型在准确率、精确度、召回率和 F1 分数方面均超过了 0.9 [6]。研究人员强调模型的可解释性。Jitendra Maan 等人(2023 年)提出 Shapley 值来解释特征重要性[7]，而 Ying Li 等人(2025 年)使用 SHAP 和因果推断来展示交易次数和金额等变量影响流失预测[8]。

为缓解数据不平衡，采用 SMOTE 方法对训练数据进行过采样，重新训练逻辑回归模型。继续利用 ROC 曲线(见图 10)对模型进行评估。

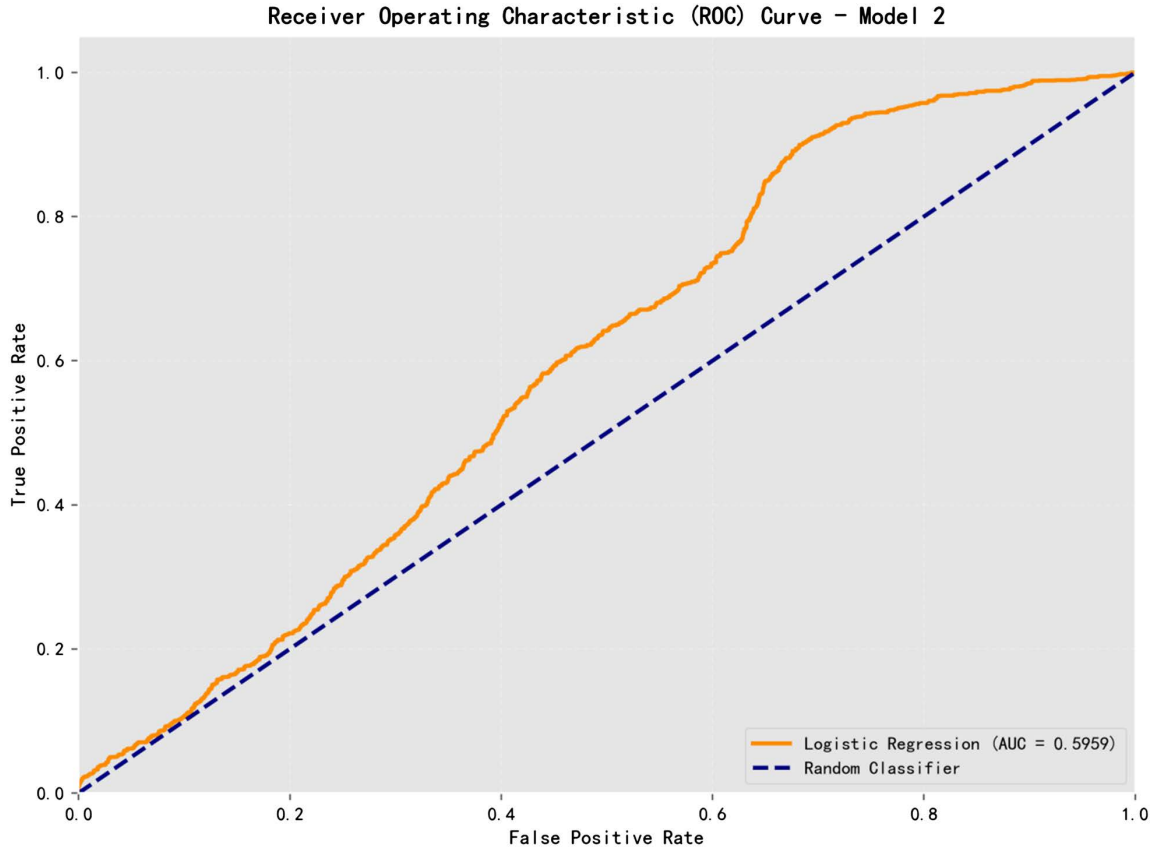

Figure 10. ROC curve of SMOTE oversampling logistic regression model
图 10. SMOTE 过采样逻辑回归模型 ROC 曲线

由图 10 可知，SMOTE 过采样后的逻辑回归模型 AUC 值为 0.5959，仅略高于随机分类器，表明模型判别能力有限，虽能一定程度上缓解类别不平衡问题，但整体性能提升不明显，需进选择其他模型以改善分类效果。

3.5.4. 多模型比较分析

我们引入随机森林、AdaBoost 和支持向量机(SVM)模型，并进行 5 折交叉验证。三类模型训练数据评估结果见图 11。

由图 11 可知，基于交叉验证 F1 分数分析，随机森林模型表现最佳，平均 F1 分数达 0.8443 且标准差最小(±0.0041)，显示出优异的预测性能和稳定性；AdaBoost 模型表现相对较弱(平均 F1 = 0.8022)，而 SVM 模型虽平均 F1 分数略高(0.8140)但标准差最大，表明其预测稳定性有待提升，综合来看随机森林是三者中最可靠的分类模型。

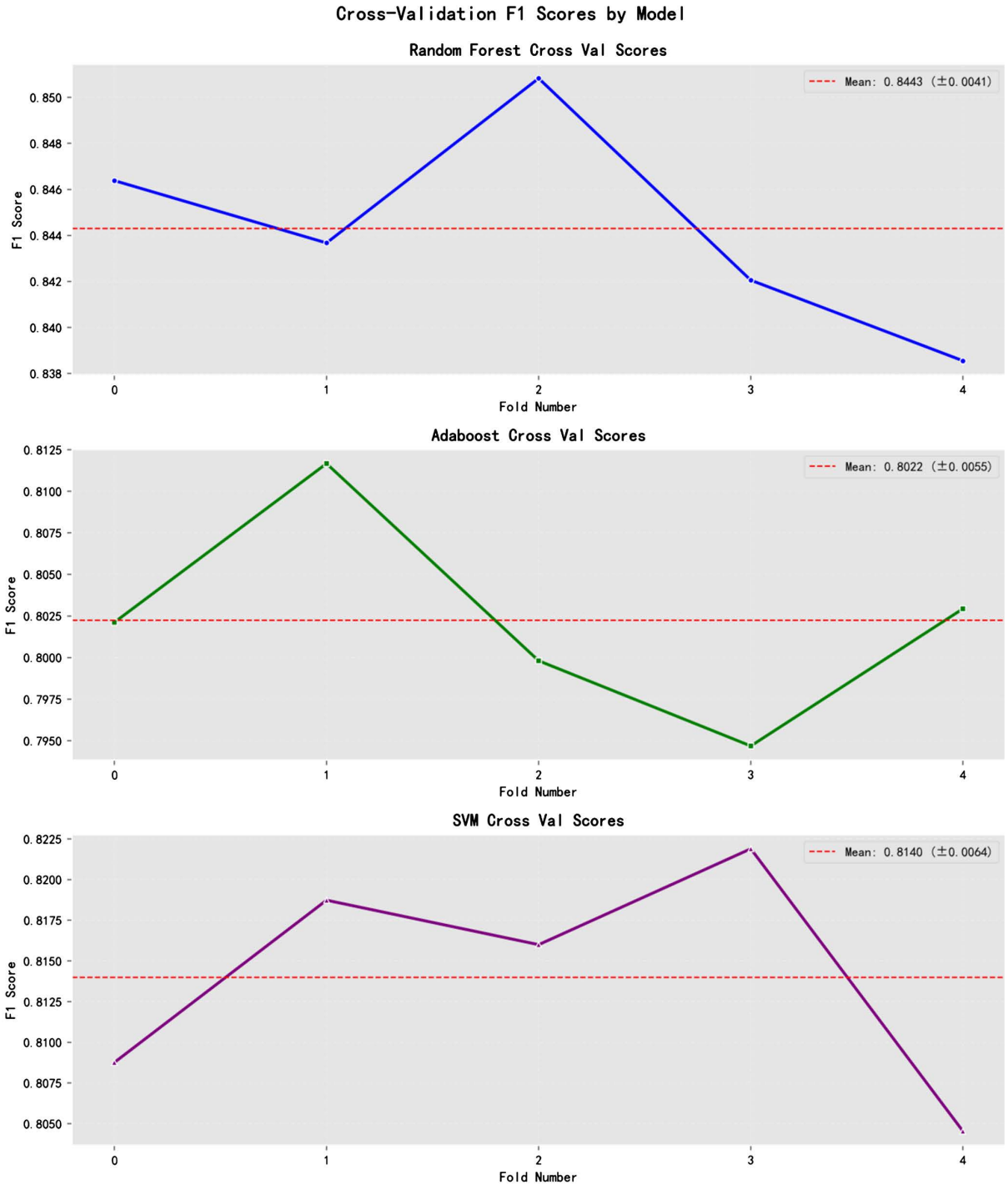

Figure 11. Cross-validation F1 score chart of training data of the three types of models
图 11. 三类模型训练数据交叉验证 F1 分数图

3.5.5. 预测的关键特征识别与业务启示

我们对随机森林模型特征重要性进行分析，见图 12。

根据随机森林模型的特征重要性分析结果，客户年龄(Age)以 40.69%的贡献度成为预测银行客户忠诚度的最关键特征，其重要性显著高于其他变量。账户余额(Balance, 20.40%)和持有产品数量(Num of

Products, 12.66%)分别位居第二和第三，这三个核心特征共同构成了近 75%的预测能力，形成了客户流失评估的基础框架。

年龄特征反映了客户生命周期的稳定性，年长客户通常表现出更强的忠诚度；账户余额直接体现了客户价值贡献和金融依赖程度；而持有产品数量则代表了客户与银行的关系广度。这些特征通过协同作用影响客户忠诚度。值得注意的是，客户活跃状态(IsActiveMember, 8.10%)和资产活跃阶段(IsActiveStatus, 7.09%)作为行为指标提供了重要的补充信息，表明即使具备良好财务特征的客户，如果活跃度不足，仍存在流失风险。

这些特征重要性排名为银行维护客户，增强客户忠诚度提供了业务启示。银行应当优先关注中年客户群体中那些余额呈现下降趋势且仅持有单一产品的用户，同时针对财务状况良好但活跃度较低的客户设计专门的唤醒和维系策略。相对而言，性别特征(Gender, 1.57%)的微弱影响表明客户忠诚度管理应更侧重于财务行为模式和人口统计特征，而非基于性别的差异化策略，从而实现客户维系资源的最优配置。

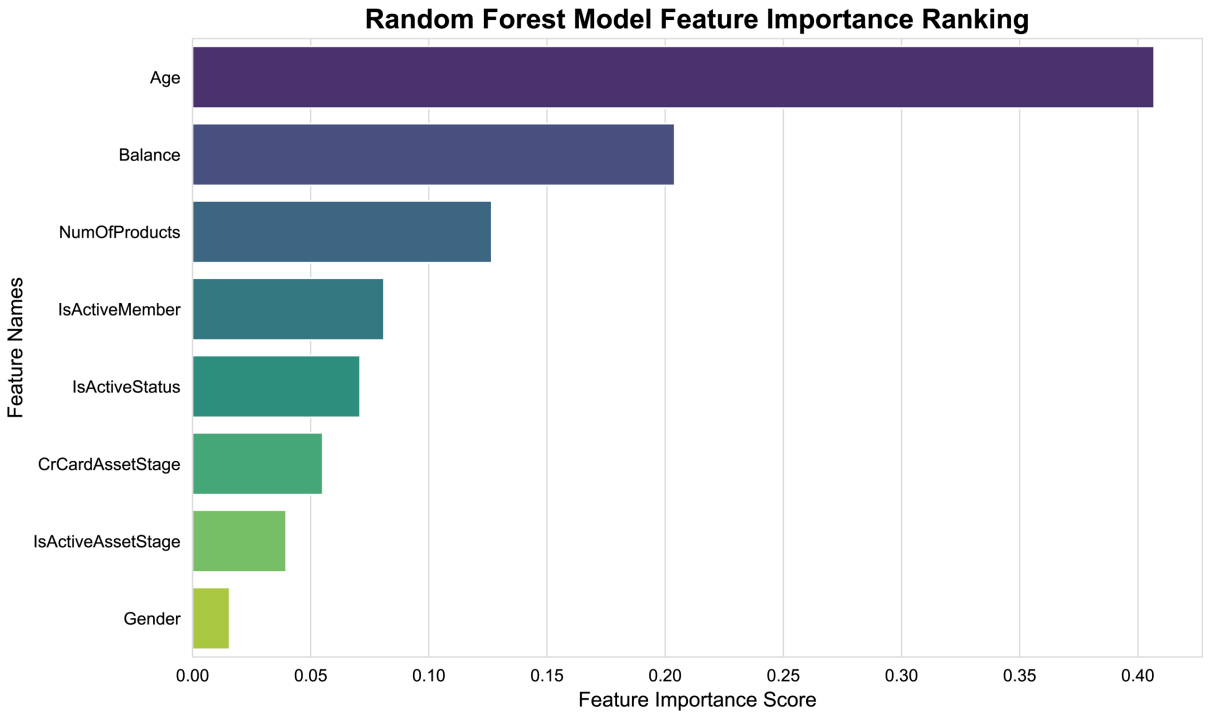

Figure 12. Bar plot of the ranking of feature importance in the random forest model
图 12. 随机森林模型特征重要性排名条形图

3.5.6. 测试集预测

对独立测试集进行预处理与特征构建后，应用训练好的随机森林模型进行预测。输出包括客户 ID 与流失预测结果，部分样本显示高龄或中低资产客户更可能流失，与前期分析一致。最终结果保存至“result5.xlsx”。

4. 研究结论与展望

4.1. 研究结论

本研究通过实证分析得出以下结论：核心发现显示，随机森林模型在客户流失预测中表现卓越(F1 分数 0.8546)，其优势源于对数据不平衡的鲁棒性和复杂特征关系的有效捕捉，同时可视化分析揭示了年龄、

资产状况与流失风险间的显著关联(如高龄低资产客户流失倾向更高)；方法贡献在于开发了集成数据处理、可视化探索与特征工程的系统性建模流程，提升了预测模型的可靠性和可解释性；研究局限包括依赖内部数据导致特征覆盖面不足，以及模型在动态业务环境中的适应性仍需长期验证，未来需扩展数据源并优化泛化能力。

4.2. 研究展望

未来研究可从以下几个方面深化。首先，应正视本研究所用竞赛数据在时间跨度、样本代表性及业务噪声模拟等方面的局限性，尤其是在时间序列维度上的稀疏性与静态性。在此基础上，未来可重点探索如何将本研究所验证的分析框架应用于具有强时间依赖特性的真实银行流水数据，引入如 LSTM、Transformer 等时序建模方法，以捕捉客户行为的动态演变规律。其次，在引入更丰富内外部数据的同时，需进一步增强模型的可解释性，借助 SHAP、时序注意力机制等技术解析动态预测决策的逻辑，保障业务应用的透明度与可信度。最后，应着力构建实时的动态风险监控与客户管理系统，将预测输出嵌入客户关系流程，形成“监测–预测–干预–反馈”的闭环机制，并通过持续学习优化实现数据驱动的智能客户管理。此外，可进一步探索该框架在反欺诈、信贷评估等多场景下的跨领域迁移潜力。

基金项目

2024 年度河北省金融科技应用重点实验室课题(2024003)。

参考文献

[1] 贾薇. 徽商银行 S 支行个人理财客户忠诚度影响因素研究[D]: [硕士学位论文]. 北京: 中国矿业大学, 2022.

[2] 余敏. Z 银行高端客户忠诚度维护研究[D]: [硕士学位论文]. 昆明: 云南财经大学, 2022.

[3] 蒋月. 基于客户感知价值理论的建行 H 支行个人客户忠诚度提升策略研究[D]: [硕士学位论文]. 大连: 东北财经大学, 2022.

[4] 冉丽. G 银行 C 分行信用卡客户忠诚度评价及提升策略研究[D]: [硕士学位论文]. 重庆: 重庆工商大学, 2022.

[5] Peng, K., Peng, Y. and Li, W. (2023) Research on Customer Churn Prediction and Model Interpretability Analysis. *PLOS ONE*, **18**, e0289724. https://doi.org/10.1371/journal.pone.0289724

[6] Tam, L.T., Vi, L.G. and Tuan, N.M. (2025) Comparison of Methods for Handling Imbalanced Data in Customer Churn Prediction with Feature Selection Using SHAP and mRMR Frameworks. *Cybernetics and Information Technologies*, **25**, 68-87. https://doi.org/10.2478/cait-2025-0023

[7] Maan, J. and Maan, H. (2023) Customer Churn Prediction Model Using Explainable Machine Learning. *International Journal of Computer Science Trends and Technology*, **11**, 33-38.

[8] Li, Y. and Yan, K. (2025) Prediction of Bank Credit Customers Churn Based on Machine Learning and Interpretability Analysis. *Data Science in Finance and Economics*, **5**, 19-34. https://doi.org/10.3934/dsfe.2025002