Statistics and Application 统计学与应用, 2025, 14(11), 249-264
Published Online November 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.1411327

Hans汉斯

基于价值分级策略的商业银行潜在客户挖掘

袁国强[1,2]，陈真贺[3]，张梓萌[3]，刘　惠[3]

[1]河北金融学院河北省金融科技应用重点实验室，河北　保定
[2]河北金融学院统计与数据科学学院，河北　保定
[3]河北金融学院金融与投资学院，河北　保定

收稿日期：2025年10月19日；录用日期：2025年11月10日；发布日期：2025年11月20日

摘　要

潜在客户挖掘一直以来都是商业银行重点关注的领域之一，为解决单一模型在进行潜在客户挖掘可能存在的如精确度不足等问题，首先利用熵权法对客户信息进行赋权生成新的一级指标；其次，通过K-means，对客户进行价值分级，最后使用经过超参数调优的随机森林、KNN、BP-神经网络的混合模型对商业银行的潜在客户进行挖掘。最终的数据分析结果显示，客户挖掘的精确度高达99%，充分证明了本文建立的混合模型在异构数据与动态场景适应性的优势。该模型可以为商业银行提供端到端的客户分级解决方案，助力其合理分配资源，实现精准营销。局限性在于依赖仿真数据验证，未来需要引入真实业务数据进一步检验模型的鲁棒性。

关键词

潜在客户挖掘，K-Means，随机森林，BP-神经网络，KNN

Potential Customers Mining of Commercial Banks Based on Value Stratification Strategy

Guoqiang Yuan[1,2], Zhenhe Chen[3], Zimeng Zhang[3], Hui Liu[3]

[1]Hebei Key Laboratory of Financial Technology Application, Hebei Finance University, Baoding Hebei
[2]School of Statistics and Data Science, Hebei Finance University, Baoding Hebei
[3]School of Finance and Investment, Hebei Finance University, Baoding Hebei

Received: October 19, 2025; accepted: November 10, 2025; published: November 20, 2025

Abstract

Potential customer mining has always been one of the key areas that commercial banks focus on. In

文章引用: 袁国强, 陈真贺, 张梓萌, 刘惠. 基于价值分级策略的商业银行潜在客户挖掘[J]. 统计学与应用, 2025, 14(11): 249-264. DOI: 10.12677/sa.2025.1411327

order to solve the problems that may exist in the potential customer mining of a single model, such as insufficient accuracy, the entropy weight method is first used to empower customer information to generate new first-level indicators. Secondly, the value of customers is graded through K-means, and finally the potential customers of commercial banks are mined using a hybrid model of random forest, KNN and BP-neural network optimized by hyperparameters. The final data analysis results show that the accuracy of customer mining is as high as 99%, which fully proves the advantages of the hybrid model established in this paper in heterogeneous data and dynamic scene adaptability. This model can provide commercial banks with end-to-end customer classification solutions, helping them allocate resources rationally and achieve precision marketing. The limitation is that it relies on simulation data verification, and it is necessary to introduce real business data to further test the robustness of the model in the future.

Keywords

Potential Customer Mining, K-Means, Random Forest, BP-Neural Network, KNN

1. 背景介绍

《2024 年中国金融科技创新发展洞察报告》显示 2023 年末我国金融机构科技投入总额达 3558.15 亿元，并且自 2019 年起金融机构科技投入规模，年均增长 14.48% [1]。基于此，当前商业银行面临客户需求多元化与互联网金融平台抢占市场份额的双重挑战。传统客户管理模式依赖经验驱动和静态数据分析，难以精准识别潜在客户并评估其价值，会导致出现识别潜在客户的成本高、营销效率低、资源分配不均等问题。

与此同时，数据挖掘、机器学习、用户画像分析、风险控制等大数据技术与人工智能的兴起，已经从不同业务环节为商业银行提供了全新的解决方案，有效推动了金融科技的快速发展，艾瑞咨询数据显示，2023 年第三方支付在零售支付市场份额为 70.1%，同比增长 1.4% [2]。客户需求逐渐从单一金融产品转向个性化、场景化服务，银行亟需构建基于数据的动态客户价值评估体系，以实现精准营销、风险控制和资源优化配置。研究表明，多维数据融合分析可使客户价值预测准确率提升至 82%以上[3]。利用机器学习算法处理商业银行海量的客户数据，包括交易记录、行为轨迹、信用信息、社交网络等，如果利用机器学习算法处理，可显著提升客户分层精度[4]，基于实时数据的客户价值模型可使营销响应率提升至 28%，远超传统方法的 12% [5]。然而，现有研究多聚焦于单一数据维度(如交易数据)的价值挖掘，缺乏对社交网络、实时行为等异构数据的动态融合分析，导致客户价值评估存在一定的偏差[6]。因此，构建数据驱动的客户价值分级体系，既是银行数字化转型的核心任务，也是应对行业竞争、提升客户黏性的战略选择。

1.1. 相关文献研究成果

1.1.1. 商业银行关键业务科技方法与策略

随着大数据技术与人工智能的快速发展，银行业利用金融科技在客户挖掘领域的研究与实践不断深化。现有一部分文献主要围绕客户细分、风险识别、精准营销及客户关系管理等方面展开，结合数据挖掘算法与业务场景，为提升银行服务效率、降低风险及优化客户体验提供了理论依据与技术路径。

1) 客户细分与聚类算法的应用

客户细分是银行精准服务的基础，传统人工分类效率低且难以应对海量数据。刘玥(2016)针对这一问题，提出改进的 K-means 算法以优化银行客户聚类效果[7]。其研究通过定义客户分类理论体系，结合银行实际交易数据验证了算法在准确性、效率上的优势，为自动化客户分群提供了技术方案。其核心贡献在于平衡了算法效率与分类精度，解决了传统方法成本高昂的痛点。此外，尹鹏、张剑等(2018)虽聚焦电力行业，但其基于大数据构建客户识别模型、绑定需求与策略的闭环管理思路，亦可迁移至银行业，辅助实现从客户分群到服务落地的全流程优化[8]。

2) 风险识别的数据挖掘技术

信贷风险控制是银行业的核心挑战。江继龙、李宇希(2018)通过数据挖掘技术对个人贷款客户进行多维度分析，建立交易行为模型与违约指数聚类，成功识别高风险客户并提前干预，有效降低不良贷款率。该方法的关键在于挖掘客户行为轨迹的隐含规律，体现了大数据在风险预测中的前瞻性价值[9]。容志(2019)虽以公共服务为背景，但其强调通过结构化数据关联揭示群体行为一致性的观点，同样适用于银行风险识别场景，如整合客户交易、社交等多源数据以增强风险评估的全面性[10]。

3) 精准营销与服务策略优化

大数据驱动的精准营销是提升银行竞争力的重要手段。尹鹏、张剑等(2018)提出基于行业类别和用户需求的分层服务策略，通过数据挖掘绑定客户等级与营销方案，实现服务闭环管理。这一思路与刘玥(2016)的客户聚类形成互补：前者侧重需求分析与策略匹配，后者侧重分群技术，共同构建了“识别–需求分析–策略执行”的完整链条。李博雷(2014)进一步指出，传统银行过度依赖交易数据而忽视客户体验，需借鉴互联网企业的用户洞察方法，将营销策略从“以产品为中心”转向“以客户体验为中心”[11]。

4) 客户关系管理与体验升级

互联网时代下，客户体验成为银行差异化竞争的关键。李博雷(2014)阐述了传统银行在客户互动与服务设计上的不足，提出重塑“以客户关系为纽带”的商业模式。其强调通过数据挖掘分析客户行为偏好，设计个性化互动场景，超越单一交易关系。这一观点与文献[10]中“整合多方参与者提升服务体验”的理念相呼应，为银行构建客户生态圈提供了方向，例如引入社会组织或市场机构共同优化服务链条。

1.1.2. 商业银行潜在客户挖掘方法与策略

随着银行业竞争的不断加剧，为提高银行利润，潜在客户的挖掘就成为了银行业关注的重要领域。而随着大数据技术与人工智能的发展，机器学习算法在潜在客户的挖掘方面展现出其独特的优越性，如何应用机器学习算法完成潜在客户的挖掘成为商业银行的重要研究课题。

1) 聚类算法的应用

K-means 算法作为主流聚类方法，在多个领域展现了客户细分能力。郑星(2024)通过消费特征指标实现体育用品客户精准定位[12]，沈俞江(2022)在保险领域通过 5 类聚类获得最佳区分效果[13]。研究普遍反映该方法需解决三个核心问题：分类变量处理、聚类数确定(常通过肘部法则验证，如安康 2014 取 K = 5)、离群值敏感性。改进方向包括：沈子垚(2021)基于 Spark 平台的并行化改进[14]，杜慧铭(2021)结合先验知识的价值分层[15]，杨厚贤(2022)融合 SMOTE 处理样本不平衡问题[16]。

2) 集成学习模型的预测优势

采用随机森林在信用风险评估(刘帅祺，2023) [17]和客户识别(王文学 2021) [18]中表现突出。特别在处理非平衡数据时，燕紫君(2018)验证其优于神经网络[19]。常与 SMOTE 等采样技术结合，通过特征重要性分析增强可解释性[20]。通过利用 XGBoost 或 LightGBM，在 5G 客户识别中成为主流模型(周雅婷 2022，AUC 达 0.80) [21]，其优势体现在：二阶泰勒展开降低过拟合(陈娟，2024) [22]、处理非平衡数据

能力(何俊江 2023，加权 Focal Loss 改进) [23]。LightGBM 在运算效率上更优(周琦，2023) [24]，银行领域应用显示 CatBoost 融合可提升预测效果(孙希，2022) [25]。

3) 混合方法的融合趋势

在采样–模型组合当中，SMOTE+集成学习成为处理样本不平衡标准流程[20] (吴博民，2022)。在模型融合创新中，Stacking 融合[24] (周琦，2023)、PSO 优化神经网络(冯亚辉，2023)等混合方法显著提升效果[26]。在特征工程协同中，MIC 特征选择(谭广，2022) [27]与模型形成优化组合。

4) 文献整合分析

对现有文献模型的应用情况进行分析发现，在现有的文献中，应用混合模型进行潜在客户关系管理及营销策略优化的仅占所有研究的 28.1%，这说明对于混合模型的研究较少。而在总体模型的应用情况中，随机森林、XGboost、LightGBM、决策树模型的出现次数最多，分别占据 15.9%、14.5%、13%、13%，总共占据了所有模型的 56.4%。

在表 1 中，本文整理了现有参考文献的研究主题，以及在不同主题中，主要算法的应用情况。如在进行潜在客户挖掘中，主要使用了 K-means、随机森林模型、BP-神经网络模型等；在营销策略优化过程中，主要使用了随机森林、决策树模型等。

Table 1. Reference list
表 1. 参考文献一览表

文献作者	研究主题	运用的关键模型
杨厚贤、沈子垚、周雅婷、吴博民、周琦、冯亚辉、何俊江、谭广	潜在客户识别	K-means、SMOTE、Logistic 回归模型、决策树模型、随机森林模型、XGBoost、LightGBM 的混合输出模型、BP-神经网络模型、CNN 模型，使用 PSO 算法对 BP-神经网络模型进行优化的融合模型
沈俞江、王文学等、王梦蓉	营销策略优化	K-means 聚类、随机森林、决策树、XGBoost、LightGBM
白燕燕、杜慧铭、郑星	客户细分	RFM 模型与支持向量机模型和决策树模型结合的混合模型、BP-神经网络模型
安康、尹胜燕	客户关系管理	K-means 聚类、RFM、聚类分析
刘帅祺	信用风险评估	随机森林
燕紫君	非平衡数据分类	灰色 DGM (2, 1)模型与 BP 神经网络的组合预测模型
陈娟、陈子阳、吴礼旺、周念、冯睿、乌文波、曹淑鹏、肖晓亮、刘克非、樊晓唯	潜在客户挖掘、金融产品挖掘	使用 RFM 模型进行赋权，用 Logistic 回归模型与 XGBoost 的混合模型进行回归预测、决策树、随机森林、粗糙集、基于 Stacking 构建的融合模型、Apriori 算法、LightGBM
牛亚琴	客户提升	XGBoost 与 Logistic 的融合模型
孙希	潜在客户预测	SMOTE+ENN 采样策略的 LightGBM 与 catBoost 融合模型
李慧	客户忠诚度分析	Apriori 算法、DBSCAN 聚类算法
郑星、刘韵、杨翌	客户价值分析	K-means、RFM 模型、随机森林、BP-神经网络、LightGBM

1.2. 相关文献评析

1.2.1. 现有成果评析

对以上现有文献进行分析发现：① 当前，在潜在客户挖掘的这类问题中，应用单一模型的情况比较常见，在总体中占据了 71.9%。并且在现有研究中，对机器学习算法的应用也比较广泛，总共应用了十余种算法。但是单一模型在精度、泛化能力及场景适应性上的不足方面存在显著的局限性。现有研究证实

由多种模型集成的模型(AUC 普遍>0.75)显著优于单一模型，但需平衡计算复杂度与可解释性。② 通过相关模型的应用，对潜在客户进行挖掘，并且基于潜在客户的特征，可以针对其提出定制化的服务。③当前研究在数据层面，部分研究数据字段可能不够全面，缺乏客户交往圈、家庭信息圈等，可能会影响潜在客户的挖掘。

1.2.2. 现有研究的不足及对策

现有研究多使用单一模型，而单一模型在动态数据融合与非线性规律挖掘方面能力不足、对数据噪声、异常值或数据分布的微小变化可能比较敏感，导致预测不稳定等不足之处，难以适应复杂多变的客户需求场景。本文首先使用熵权法构建动态指标体系，通过客观赋权筛选关键特征，剔除冗余信息。其次，结合不同的基础模型，针对不同的模型偏差和建模假设，预测结果可以覆盖更广泛的数据模式和关系，更加接近真实的数据生成过程，有效提高模型对于复杂数据的适应能力。本文构建的混合模型可以有效弥补个别模型的偏差，对于扰动的敏感性会被其他模型“平均掉”或“纠正”，从而使得整体预测更加趋于稳定。

2. 价值分级策略下潜在客户挖掘的特征指标

价值分级策略是指企业根据一定的评估标准和维度，将其目标对象(通常是客户、用户、产品或市场)划分为不同的价值等级，并为每个等级制定差异化的策略、投入相应的资源，以实现整体效益最大化的管理方法。

基于客户价值分级策略构建客户融合指标体系的流程如图 1 所示：

1) 依据现实商业银行的客户数据结构，使用数据生成器等价生成相关的客户数据结构，以此有效保留数据中存在的各种有效信息。

2) 将客户指标划分为四个新维度并且对数据进行编码，其中客户特征的四个维度分别为基本特征、违约风险、忠诚度、财务贡献。

3) 使用熵权法计算指标体系的权重。

4) 使用构建好的数据指标体系用 K-means 模型进行客户价值分级。

5) 输出潜在客户价值分级结果。

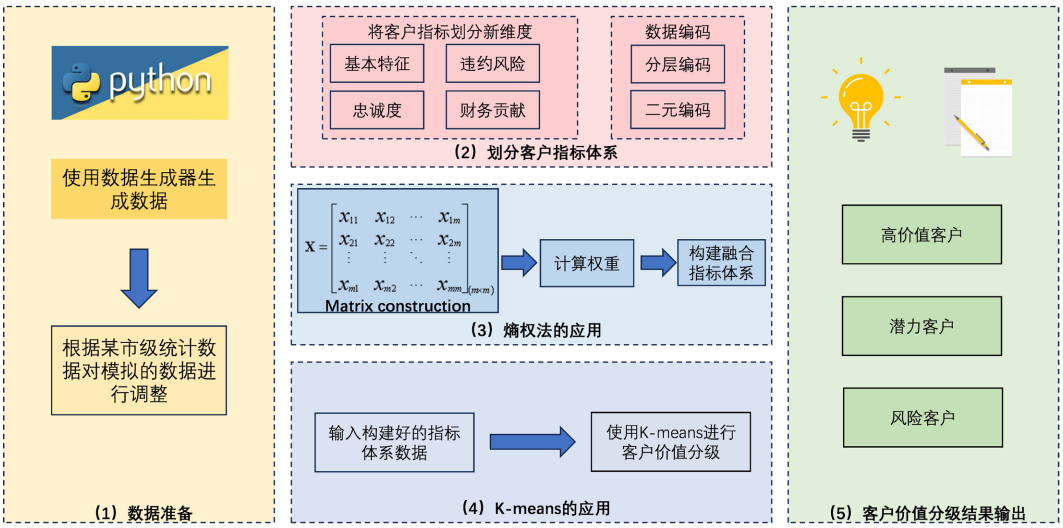


Figure 1. Construction of customer integration indicator system
图 1. 客户融合指标体系构建

2.1. 构建特征指标

为了构建一个相对全面且具有现实业务指导意义的商业银行客户价值体系评估指标，本文确定了表2 中的客户特征指标并将其划分为四个不同的维度。这些指标可以从不同的角度量化潜在客户对银行的综合贡献和潜在风险，进而实现对潜在客户的差异化管理与服务。其中，确定潜在客户的性别、年龄、文化水平和婚姻状况、收入水平等集中反映客户社会经济背景的特征归为基本特征指标；确定潜在客户的贷款违约记录、信用评分、负债率等直接关联到资产安全和风险成本的指标归为违约风险指标；确定潜在客户交易频率、距上次交易时间、持续使用年限等表现客户粘性及关系深度的指标归为忠诚度指标；确定潜在客户存款余额、持有理财的金额、净息差收入等最为直观体现客户为银行创造的收益归为财务贡献指标。

如表 2 所示，本文使用二元编码及分层编码进行数据处理。如对性别使用二元编码，1 代表男，0 代表女。对年龄使用分层编码，因为岁数较小的客户价值相对较低，而随着年龄的增长客户价值会升高，于是客户年龄编码结果为：1 = < 25 岁，2 = 大于 25 小于等于 35 岁，3 = 大于 35 小于等于 45 岁，4 = 大于 45 小于等于 55 岁，5 = 大于 55 岁。对交易频率进行分类编码：1 代表小于等于 1 次为休眠，2 代表 2~5 次为低频，3 代表 6~10 次为中频，4 代表大于等于 11 次为高频。

Table 2. Basic indicators of commercial bank customers
表 2. 商业银行客户基本指标

特征指标	数据处理规则
性别	二元编码(男 = 1，女 = 0)
年龄	1 = < 25 岁，2 = 大于 25 小于等于 35 岁，3 = 大于 35 小于等于 45 岁，4 = 大于 45 小于等于 55 岁，5 = 大于 55 岁
文化水平	1 = 高中及以下，2 = 本科，3 = 研究生及以上
婚姻状况	二元编码(已婚 = 1，未婚/其他 = 0)
收入水平	分段编码：1 = x ≤ 5 千，2 = 5 千 < x ≤ 1 万，3 = 1 万 < x ≤ 4 万，4 = 4 < x ≤ 6 万，5 = x ≥ 6 万
贷款违约记录	二元编码(有 = 1，无 = 0)
信用评分	1 = 小于等于 400，2 = 400 < x ≤ 500，3 = 500 < x ≤ 600，4 = 600 < x ≤ 700，5 = 大于 700
负债率	(贷款总额/总资产)，然后对其进行编码处理，1 = 小于等于 0.1，2 = 大于 0.1 小于等于 0.2，3 = 大于 0.2 小于等于 0.3，4 = 大于 0.3
交易频率	分类编码：1 = 休眠(≤1 次)，2 = 低频(2~5 次)，3 = 中频(6~10 次)，4 = 高频(≥11 次)
距上次交易相差的时间	逆向编码：4 = 活跃(≤30 天)，3 = 次活跃(31~90 天)，2 = 风险(91~180 天)，1 = 准流失(>180 天)
持续使用银行服务的年限	阶段编码：1 = 新客(≤1 年)，2 = 稳定(2~5 年)，3 = 长期(大于 5 年)
存款余额	分层编码：1 = 低(≤10 万)，2 = 中(10~100 万)，3 = 高(>100 万)
持有理财产品的金额	组合编码：1 = 保守型(低风险 ≤ 50 万)，2 = 平衡型(混合类大于 50 小于等于 150 万)，3 = 进取型(权益类 > 150 万)
净息差收入	贡献度分级：1 = 低(≤1 千元)，2 = 中(大于 1 千元小于等于 5 千元)，3 = 高(>5 千元)

2.2. 熵权法赋权计算

通过查找文献，发现众多学者使用熵权法解决了许多实际问题：李宏晨等(2024) [28]针对天然气客户分级问题，利用熵权法量化指标离散度，剔除冗余特征(如重复性客户数据)，构建三级评价体系，结合

TOPSIS 实现客户价值动态分级。谭锰锐(2023) [29]通过熵权法对吉林银行不良贷款率等指标客观赋权(权重占 15%)，动态调整财务风险权重，解决传统监管滞后问题。翟芸等(2022) [30]融合改进 AHP 与熵权法，筛选雷达装备“故障间隔时间”等关键指标，耦合主客观权重提升小样本评估精度。研究均凸显熵权法在动态赋权与数据降维中的优势，通过客观量化指标贡献，其核心优势在于动态筛选关键指标，剔除冗余信息(如高相关性或低差异指标)，提升指标体系的区分度。

而熵权法筛选动态关键指标，剔除冗余信息，科学的筛选关键特征，消除主观经验依赖，支持大规模客户分级的特点正好是本文研究所需要的，故本文利用熵权法构建指标，解决以上问题。

通过信息熵量化指标区分度，构建客观权重体系：

第一步，数据准备

构建初始决策矩阵，假设有 n 个要进行客户价值分级的客户，每个客户有 m 个评价指标，构建初始决策矩阵 X。

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}_{(n\times m)} \tag{2-1}$$

式中：x_{ij} 为第 i 个客户 x_i 的第 j 个指标的原始数值。

第二步，数据标准化

正变量：

$$X_{ij}^{*}=\frac{X_{ij}-\min\left(X_j\right)}{\max\left(X_j\right)-\min\left(X_j\right)} \tag{2-2}$$

X_{ij}^{*} 表示第 i 个客户的第 j 个标准化后的指标值；X_{ij} 表示第 i 个客户的第 j 个指标值；$\max(X_j)$表示第 j 个指标中的最大值，$\min(X_j)$表示第 j 个指标中的最小值。为避免结果结果出现零值，对处理后的数据进行平移处理，即

$$X_{ij}^{*}=X_{ij}^{*}+0.00001\text{。}$$

正变量是指那些对是否为潜在客户起支持作用的变量，正变量越大越好。

负变量：

$$X_{ij}^{*}=\frac{\max\left(X_j\right)-X_{ij}}{\max\left(X_j\right)-\min\left(X_j\right)} \tag{2-3}$$

为避免结果结果出现零值，对处理后的数据进行平移处理，即

$$X_{ij}^{*}=X_{ij}^{*}+0.00001\text{。}$$

负变量是指那些对是否为潜在客户起反对作用的变量，负变量越小越好。

第三步，概率矩阵的构建

计算每个指标的权重

$$P_{ij}=\frac{X_{ij}^{*}}{\sum_{i=1}^{n}X_{ij}^{*}} \tag{2-4}$$

n 表示样本个数，P_{ij} 表示每个客户 i 在对应的第 j 个指标在同类指标中所占的比重。

第四步，熵值计算

$$e_j = -k\sum_{n=1}^{n}\left(P_{ij} * \ln P_{ij}\right),\ \ k = 1/\ln(n) \tag{2-5}$$

e_j 表示第 j 个指标的熵值。

第五步，权重计算

$$d_j = 1 - e_j \tag{2-6}$$

$$w_j = \frac{d_j}{\sum_{j=1}^{n} d_j} \tag{2-7}$$

d_j 表示信息效用值，它是熵值的互补值，它直接反应了指标信息量大小。w_j 表示每个指标的权重。各个指标的权重计算结果，如表 3 所示：

Table 3. Weight calculation results table
表 3. 权重计算结果表

指标	权重(w)
性别	0.063
年龄	0.019
文化水平	0.098
婚姻状况	0.041
收入水平	0.006
贷款违约记录	0.016
信用评分	0.083
负债率	0.018
交易频率	0.010
距上次交易相差的时间	0.062
持续使用银行服务的年限	0.019
存款余额	0.375
持有理财产品的金额	0.153
净息差收入	0.037

2.3. 特征融合

将(表 3)以上十四个指标分为四个维度，利用上述熵权法得到的权重进行线性加权求和，得到(潜在客户的)新的四个指标，分别为基本特征、违约风险、忠诚度、财务贡献。

$$Y_i = \sum_{i=1}^{n} w_i * x_i' \tag{2-8}$$

Y_i 表示四个新的指标，w_i 表示用熵权法计算出的每个二级指标所占的比重，x_i' 表示经过标准化后的二级指标的值。

特征指标计算结果如表 4 所示：

Table 4. Weight calculation results table
表 4. 权重计算结果表

维度	混合权重
基本特征	22.7%
违约风险	11.7%
忠诚度	9.1%
财务贡献	56.5%

2.4. 实现客户价值分级

2.4.1. 数据预处理

对 2.3 中生成的四个维度的数据 Y_i 进行标准化，将第 2 列(违约风险)数值取反，使该指标方向与其他指标一致(数值越大代表风险越小)。

$$Z_i = (Y_i - u_i)/\sigma_i \tag{2-9}$$

Z_i 表示第 i 个指标标准化的数据，Y_i 表示四个新的指标，u_i 表示第 i 个指标的均值，σ_i 表示第 i 个特征的标准差。

2.4.2. 基于 K-means 聚类的客户价值分级方法

从客户价值分级的实际出发，为帮助银行的业务人员或者管理层在做出决策时提供更加直观、可靠的数据依据，并且对客户价值进行评级和完成客户等级划分，此处采用 K-means 的聚类方法，具体使用综合价值评分的方法衡量客户价值。

评价公式具体是：

价值评分 = 基本特征 + 忠诚度 + 财务贡献 − 违约风险。

模型的具体运行步骤为：

(1) 初始化 K-means 模型；设定聚类数为 3。(2) 使用标准化及方向调整后的指标数据训练模型。(3) 输出每个客户的初始聚类标签 0、1、2。(4) 计算各聚类综合价值评分：正向指标相加，负向指标相减。(5) 按客户价值分级得分排序，确定标签映射，输出计算结果，部分输出结果如表 5 所示。

2.4.3. 标签定义

风险客群(1)：信用风险显著偏高或价值贡献持续走低；潜力客群(2)：各维度表现均衡，存在价值提升空间；高价值客群(3)：价值贡献与行为粘性双高，风险评分低于均值。

部分客户价值分级结果如表 5 所示：

Table 5. Customer value segmentation results table
表 5. 客户价值分级结果表

客户 ID	Y1	Y2	Y3	Y4	客户价值
CUST00001	0.28	0.10	0.32	0.57	2
CUST00002	0.27	0.22	0.32	0.95	3
CUST00003	0.23	0.12	0.28	0.79	3
CUST00004	0.15	0.19	0.31	0.60	2
CUST00005	0.28	0.19	0.14	0.60	1

3. 商业银行潜在客户挖掘方法

3.1. 数据预处理

银行对为客户办理金融业务、银行间交易、保管账户等业务获取到的客户信息均负有保密义务，具体包括客户的基本信息、信用信息、交易信息和其他已获得但是未包含在以上类别中的信息。因此，本文将不会直接从银行获取具体的银行客户数据信息，从而使用数据生成器按真实客户数据结构来生成模拟客户数据，此方法具有高效便捷、高度的可控与定制化、可重现性以及稳定性等优点，于是本文使用python 中的数据生成器进行客户数据的生成。

使用 python 生成的银行客户模拟数据中，共有 100,000 组银行客户数据和 16 种客户关键特征，分别包括银行客户的 ID、性别、年龄、文化水平、婚姻状况、收入水平、总资产价值、贷款违约记录、信用评分、贷款金额、交易频率、距上次交易相差的时间、持续使用银行服务的年限、存款余额、持有理财产品的金额、为银行创造的净息差收入。

客户数据的生成逻辑为，先构建一些如基础收入、存款基数等基础数据，然后在这些基础数据的基础之上计算收入水平、违约概率等变量数据，完成客户数据的构建。其中基础数据的生成代码见表6：

Table 6. Code information for generating basic bank customer data
表 6. 银行客户基础数据的生成代码信息

代码	变量类别	客户特征
np.random.choice()	分类变量	性别、文化水平、婚姻状况
np.random.normal()	正态分布数据	年龄、基础违约概率、信用评分基础
np.random.lognormal()	对数正态分布数据	基础收入、独立资产部分、存款基础
np.random.exponential()	指数分布数据	贷款金额基数、交易时间
np.random.poisson()	泊松分布数据	交易频率基础
np.random.uniform()	均匀分布随机因子	生成均匀分布随机因子(各种调节系数)
np.random.binomial()	二项分布数据	贷款违约记录、贷款申请决策

对于变量数据的生成逻辑如下所述。最终生成的收入水平计算公式为：

$$\text{收入水平} = \text{基础收入} \times (1 + \text{年龄} - 30)/100 \times \text{随机因子} \times \text{教育水平}。$$

最终信用评分的计算公式为：信用评分 $= 0.6 \times$ 信用评分基础 $+ 0.1 \times$ 年龄因子 $+ 0.1 \times$ 收入因子 $+ 0.1 \times$ 服务因子 $- 0.1 \times$ 违约因子 $+$ 随机因子。

为了对模型的有效性进行验证分析，本文对代码进行了如下操作：1) 使用多种统计分布模拟真实世界变量的不同形态，使得生成的客户数据可以更加贴合现实情况；2) 引入大量的随机成分，如在计算收入水平时，引入了教育水平因子、年龄因子等等；3) 创建了一些非线性关系，从而避免简单的线性关系。4) 降低了特征之间的相关度，从而使得数据更加贴合现实情况。

3.2. 商业银行潜在客户挖掘模型的构建

3.2.1. 模型构建

本文通过以下流程确认最优模型，流程如图 2 所示，实施步骤如下：

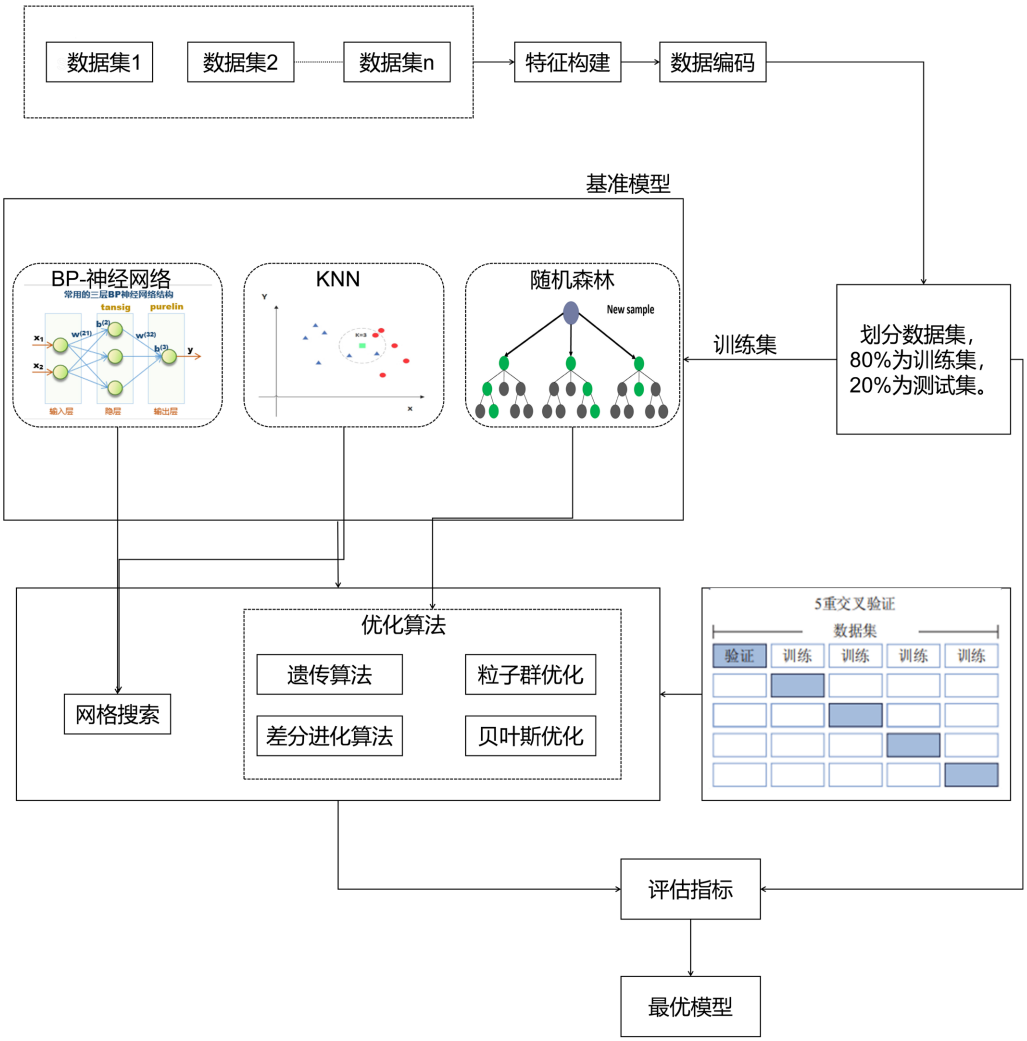


Figure 2. The process of establishing a customer discovery model
图 2. 客户挖掘模型的建立流程

1) 使用数据生成器生成相关数据，对数据进行编码。

2) 在进行融合指标构建时，首先利用客户价值分级将客户分为高价值客户、潜力客户、风险客户并将其作为目标变量。然后，对输入变量与目标变量的潜在关系进行探索。根据输入变量的不同，将全部样本划分为不同的数据集，在同一数据集下将 80%的样本作为训练集；剩余的 20%作为测试集，用于测试构建模型的精准度。

3) 首先，运用网格搜索法、智能优化算法优化机器学习算法的超参数。在训练过程中，采用 5 重交叉验证的方法，将训练集进行随机分割；以此寻找最优的超参数组合。

4) 计算并比较超参数优化后的模型在测试集上的性能指标，寻找最优的机器学习模型。

3.2.2. 超参数优化

如前所述，本文通过数据集的训练集对机器学习模型进行关键训练，并且为了使模型的性能根据所使用的数据进行进一步的优化，参考严格齐等(2024) [31]的优化方法，对 B-P 神经网络模型、KNN 模型使用参数网格法寻找最优的超参数，对随机森林模型使用遗传算法、差分进化算法、粒子群优化以及贝叶斯优化方法寻找最优的超参数。

遗传算法(GA)模仿自然选择和遗传过程，通过适应度评估和遗传操作(如交叉、变异)来不断优化路径[32]。该算法通过数学的方式，利用计算机仿真运算，将问题的求解过程转换成类似生物进化中的染色体基因的交叉、变异等过程。在求解较为复杂的组合优化问题时，相对一些常规的优化算法，该算法通常能够较快地获得较好的优化结果。遗传算法已被人们广泛地应用于组合优化、机器学习、信号处理、自适应控制和人工生命等领域。

差分进化(Differential Evolution, DE)是一种基于群体差异的启发式随机搜索算法，该算法是 R. Storn 和 K. Price 为求解切比雪夫多项式而提出。DE 算法优点显著：① 独立性强，与求解问题的文字描述相对独立；② 原理简单，容易实现；③ 群体搜索，针对个体具有自动识别最优解的能力；④ 协同搜索，能够综合个体局部信息和群体全局信息，指导算法进一步搜索；⑤ 通用性强，易与其他算法融合贯通，构造出更优算法[33]。

粒子群优化算法(PSO)是一种进化计算技术(evolutionary computation)，1995 年由 Eberhart 博士和 kennedy 博士提出，该算法起源于对鸟群捕食行为的研究。最初灵感来自于飞鸟集群活动的规律性，进而利用群体智能建立的一个简化模型。粒子群算法在对动物集群活动行为观察基础上，利用群体中的个体对信息的共享使整个群体的运动在问题求解空间中产生从无序到有序的演化过程，从而获得最优解。

贝叶斯优化(Bayesian Optimization)是一种十分高效的全局优化算法，主要用于机器学习调参，贝叶斯优化是一种不需要计算导数的系统化调优算法，采用高斯过程建立概率代理模型，考虑之前的参数信息，不断更新先验，使用采集函数来确定下一个评估点，可以在较短的时间内确定最佳参数[34]。

具体的模型优化参数如表 7 所示：

Table 7. Machine learning algorithms for customer mining and hyperparameter tuning of models
表 7. 客户挖掘的机器学习算法及模型调优的超参数

算法	超参数	范围
随机森林模型 Random forest (RF)	决策树棵树	(200, 1000)
	树的最大深度	(10, 100)
	划分内部节点所需要的最小样本	(2, 50)
	叶子结点最少样本数	(1, 20)
	分裂时考虑的最大特征数	(0.2, 0.9)
	每棵树使用的最大样本数	(0.6, 0.95)
BP-神经网络模型 BP-Neural network model	隐藏层结构	(50), (100), (50, 50), (100, 50), (100, 100)
	学习率	[0.001, 0.01, 0.1]
	优化器	“adam”, “sgd”
	批量大小	[32, 64, 128]
	激活函数	“relu”, “tanh”, “logistic”
	正则化率	[0.0001, 0.001, 0.01]
KNN 模型	邻居数	3, 5, 7, 10, 15, 20
	权重类型	“uniform”, “distance”
	距离度量方式	“euclidean”, “manhattan”, “minkowski”
	p 参数	10, 20, 30, 40, 50

4. 实证结果分析

4.1. 模型性能评估指标

本文使用精确度(precision)、召回率(recall)、F1 分数(F1-score)来对模型的性能进行有效评估，上述系数的计算公式如下：

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \tag{1}$$

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \tag{2}$$

$$\text{F1} = 2*(\text{Precision}*\text{Recall})/(\text{Precision}+\text{Recall}) \tag{3}$$

公式(1)、公式(2)和公式(3)适用于二分类问题，由于本文所预测的问题为多分类问题，故对于不同类别所预测的精确率、召回率、F1 分数要进行算数平均，进而得到最终的评估结果：

宏精确度(Macro Precision)：

$$\text{Macro Precision} = \frac{1}{C}\sum_{i=1}^{C}\text{Precision}_i \tag{4}$$

宏召回率(Macro Recall)：

$$\text{Macro Recall} = \frac{1}{C}\sum_{i=1}^{C}\text{Recall}_i \tag{5}$$

宏 F1 (Macro F1)：

$$\text{Macro F1} = \frac{1}{C}\sum_{i=1}^{C}\text{F1}_i \tag{6}$$

其中，TP (True Positive)表示真正例(实际为正类，预测为正类)；FP (False Positive)表示假正例(实际为负类，预测为正类)；FN (False Negative)表示假负例(实际为正类，预测为负类)；Macro Precision、Macro Recall、Macro Fl 加权平均后的精确率、召回率、F1 分数；C 表示预测结果的分类数量；i 表示所预测的分类结果；Precision_i、Recall_i、F1_i 分别表示第 i 中预测结果的精确度、召回率、F1 分数。

4.2. 实证结果分析

4.2.1. 客户挖掘结果分析

最终预测结果中有 20217 人被预测为高价值客户，有 55,101 人被预测为潜力客户群，剩余的 24,682 人被预测为风险客户群。该预测结果，清晰的展示了客户的结构特征，在银行客户整体质量较好的时候，可以为银行提供积极的信号；并且当银行的风险客户较多时，可以为银行提供风险提示，当风险客户达到一定的比例时，就需要制定针对性的风险管控策略。清晰的客户结构为银行提供了清晰的管理框架及资源配置依据。这一客户价值分层结果为企业的精细化运营、精准营销和风险管理提供了数据支撑和决策依据，有助于实现客户价值最大化和企业资源最优化。

4.2.2. 潜在客户挖掘模型性能评价

进行超参数优化时，BO (贝叶斯)算法的初始点数为 5，最大迭代次数为 15；GA 算法、PSO 算法、DE 算法的种群数分别为 30、20、15。

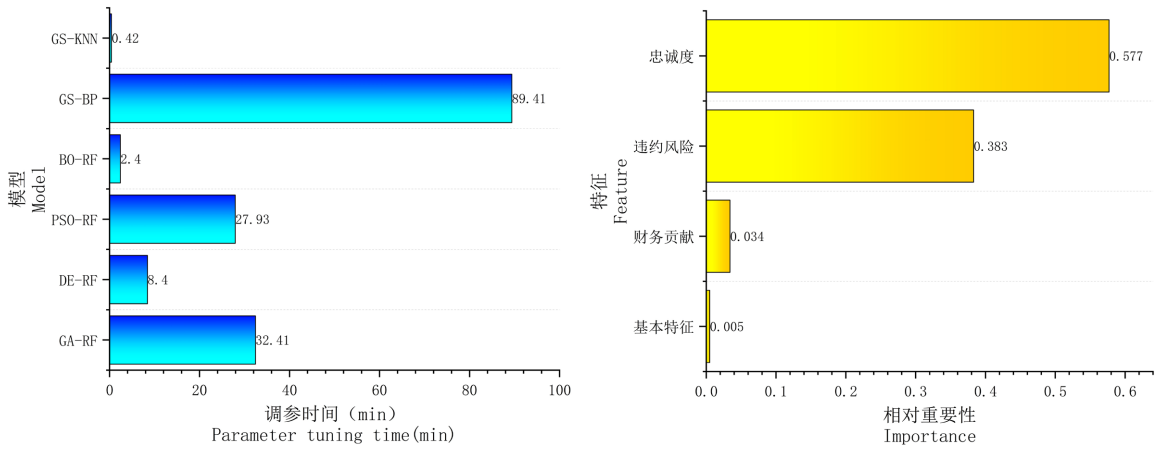
表 8 为模型在测试集上的性能评估指标结果，预测结果显示，这几种模型评估指标都在 0.99 左右，说明经过超参数优化的模型，可以通过他们强大的学习能力，学习和模仿特征维度与目标变量之间的关系，实现对潜在客户的精准挖掘。

模型调整参数的时间(图 3(a))显示，GS-KNN 模型的训练耗费时间最短，为 0.42 分钟，而 GS-BP 模型的训练耗费时间最长，为 89.41 分钟。综合考虑训练时间以及客户挖掘的精度，DE-RF 模型以较少的时间实现了较高的精度，综合表现最佳。而 GS-KNN 模型的训练时间较短的原因可能是由于所需调整的参数较少的原因导致的。因此 DE-RF 模型为进行商业银行客户挖掘的最优模型，而 DE(差分进化算法)在

调整参数时也展现出其独特的优越性。

Table 8. Performance evaluation metrics of the potential customer mining model on the test set
表 8. 潜在客户挖掘模型在测试集上的性能评估指标

模型 Medal	Precision	Recall	F1
GA-RF	0.9992	0.9993	0.9992
DE-RF	0.9992	0.9994	0.9993
PSO-RF	0.9992	0.9994	0.9993
BO-RF	0.9985	0.9988	0.9986
GS-BP	0.9991	0.9990	0.9991
GS-KNN	0.9975	0.9974	0.9974



(a) (b)

Figure 3. (a) Model parameter tuning is time-consuming; (b) Feature importance results
图 3. (a) 模型调参耗费时间；(b) 特征重要性结果

为了体现本文所提出的混合模型的优越性，本文将引入逻辑回归模型通过客户原始特征完成潜在客户的挖掘。经过模型训练后完成客户挖掘最终输出的结果显示，逻辑回归进行潜在客户挖掘的精确度为0.9343，召回率为0.9244，F1 分数为0.9292。与本文提出的 DE-RF 模型的 F1 分数 0.9993 相比，这显著的性能差距凸显了在客户数据中存在高度的非线性相关关系，证明了采取随机森林进行潜在客户挖掘的必要性。对于逻辑回归无法捕捉到的复杂交互关系，随机森林通过构建多棵决策树的方式可以有效解决这一问题。尽管 DE-RF 模型的训练耗费时间(8.4 分钟)高于逻辑回归模型的 1 分钟，但是考虑到随机森林模型在性能上的优越性，这种额外的投入是完全值得的。特别是在商业银行这种对准确度要求极高的应用场景，极小的性能提升都可能带来巨大的商业价值。因此本文所提出的混合模型，在性能上显著优于逻辑回归模型，且充分证明了本文混合模型的优越性与实用性。

4.2.3. 客户特征重要性分析

如图 3(b)所示，客户的忠诚度是目前最重要的特征，其重要性得分超过了其他得分的总和，这表明客户是否忠诚是预测其未来行为的最关键指标。因此，在进行客户关系管理时，首先应该考虑的是维护和提升客户的忠诚度。其次客户的违约风险的相对重要性也很高，这说明违约风险同样是一个非常关键

的考量因素，违约风险通常与客户的历史付款、还款、信用状况等行为相关联，客户的违约分线较高的话会给企业带来直接的经济损失，所以风险控制同样是客户管理中不可或缺的一环。

同时，为了将特征重要性分析转化为可操作的业务洞察，本研究将高层指标忠诚度、违约风险等落实到最初的、可以被业务人员理解和干预的子指标上。如将忠诚度指标落实在距上次交易相差的时间或者交易频率上面。通过距上次交易相差的时间来定义是否为沉睡客户，若是超过阈值(比如说 90 天)，则认为其为沉睡客户，存在流失风险。这时候银行就可以对其采取激活策略，通过向其推送个性化的激活方案，如限时的高息存款产品、专属理财顾问服务等，目标是为了在 30 天内促成其发生一次有效交易。对于交易频率小的客户，银行也可以采取习惯培养策略，如积分换礼品、支付返现，或者说引导客户将其日常的消费绑定至银行卡账户，培养其使用习惯。而违约风险我们可以落实在客户的贷款违约记录上，存在贷款违约记录的客户通常是风险客户群的组成部分，可能会造成银行资产的损失。对着这些客户银行可以将其纳入重点监控的名单，限制其业务权限，由风险部门对其风险进行定期的评估。

对这些指标进行分析，强调商业银行应该将重点的资源应用到对客户忠诚度的培养，并且对风险进行控制。具体操作就是通过对距上次交易相差的时间与交易频率，通过激活和培养策略，找寻价值增长的机会，同时通过违约风险控制客户风险。

5. 结论

为了在大数据时代，为商业银行客户管理以及精准营销提供新的路径，本文通过构建指标体系，对客户进行价值分级，最后使用遗传算法选用遗传算法(genetic algorithm, GA)、差分进化算法(differential evolution, DE)、粒子群优化(particle swarm optimization, PSO)算法、贝叶斯优化(Bayesian optimization, BO)算法结合随机森林(random forest, RF)以及使用网格搜索(grid search, GS)的 BP-神经网络模型和 KNN 模型进行了对比，主要结果如下：

1) 客户忠诚度对于潜在客户挖掘的结果的影响程度最高，相对重要性为 0.577，而客户的基本特征对客户挖掘的影响程度最低。

2) 四种优化算法下的随机森林模型，除 BO-RF 模型以外，挖掘性能都优于 GS-BP、GS-KNN 模型，并且 DE-RF 模型展现出较高的精确度(Precision = 0.9992, Recall = 0.9994, F1 = 0.9993)，并且优化耗时仅为 8.4 分钟，综合表现最佳。

3) 模型有效显著，在对使用 python 代码生成的复杂数据集进行潜在客户挖掘的过程中，优化后的模型精确度高达 0.99，这表明模型可以对客户价值分级的复杂映射关系进行有效的学习，验证了混合模型在进行潜在客户挖掘及客户价值分级的有效性。

综上所述，本文提出以下建议。

一是在进行模型构建时，尝试模型的堆叠和机器学习算法进行融合的多种优化算法，进而提高模型的预测能力。二是客户忠诚度和客户的违约风险在进行客户挖掘时，起到相对重要的作用，因此在构建模型时，充分考虑以上因素的作用对于提高预测的精确度有重要意义。

基金项目

河北省金融科技应用重点实验室项目“河北省城市商业银行潜在客户挖掘与系统开发研究”(项目编号：2024004)；河北省研究生创新资助项目“基于大数据分析的商业银行潜在客户挖掘与价值分级研究”(项目编号：CXZZSS2025133)。

参考文献

[1] 易观分析. 2024 年中国金融科技创新发展洞察报告[R]. 北京: 易观智库, 2024.

[2] 艾瑞咨询. 2023 年中国第三方支付行业研究报告[R]. 上海: 艾瑞研究院, 2023: 7-9.

[3] Chen, H., Chiang, R.H.L. and Storey, V.C. (2012) Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, **36**, 1165-1188. https://doi.org/10.2307/41703503

[4] Ranjan, J. and Agarwal, R. (2009) Application of Segmentation in Customer Relationship Management: A Data Mining Perspective. *International Journal of Electronic Customer Relationship Management*, **3**, 402-414. https://doi.org/10.1504/ijecrm.2009.029298

[5] Rust, R.T. and Huang, M. (2014) The Service Revolution and the Transformation of Marketing Science. *Marketing Science*, **33**, 206-221. https://doi.org/10.1287/mksc.2013.0836

[6] Kumar, V. and Reinartz, W. (2016) Creating Enduring Customer Value. *Journal of Marketing*, **80**, 36-68. https://doi.org/10.1509/jm.15.0414

[7] 刘玥. 基于改进的 K-Means 算法的银行客户聚类研究[D]: [硕士学位论文]. 长春: 吉林大学, 2016.

[8] 尹鹏, 张剑, 董兵, 等. 基于大数据的电力优质客户识别及市场营销服务策略分析[C]//国网天津市电力公司, 朗新科技股份有限公司. 2018 智能电网新技术发展与应用研讨会, 2018.

[9] 江继龙, 李宇希. 数据挖掘在个人贷款潜在风险客户识别中的应用[J]. 电子技术与软件工程, 2018(9): 189.

[10] 容志. 大数据背景下公共服务需求精准识别机制创新[J]. 上海行政学院学报, 2019, 20(4): 44-53.

[11] 李博雷. 超越交易重塑银行与客户关系——互联网时代银行业客户体验管理策略探索(上篇) [J]. 清华金融论, 2014(9): 77-80.

[12] 郑星. 基于消费数据挖掘的体育用品客户精细化定位方法[J]. 赤峰学院学报(自然科学版), 2024, 40(12): 53-57.

[13] 沈俞江. 基于聚类分析的 A 人寿保险公司苏州分公司营销策略研究[D]: [硕士学位论文]. 南京: 东南大学, 2022.

[14] 沈子垚, 袁晓玲. 基于并行化 K-Means 的综合能源服务客户识别[J]. 电力工程技术, 2021, 40(2): 107-113.

[15] 杜慧铭. 基于先验知识的新能源汽车潜在客户识别研究[D]: [硕士学位论文]. 武汉: 武汉理工大学, 2021.

[16] 杨厚贤. 5G 潜在客户评分模型研究[D]: [硕士学位论文]. 成都: 西南财经大学, 2022.

[17] 刘帅祺. 商业银行互联网贷款信用风险评估及客户特征发现[D]: [博士学位论文]. 北京: 北京科技大学, 2023.

[18] 王文学, 李公平, 许经伟, 等. 基于用户行为的智能组网营销挖掘模型[J]. 信息技术与信息化, 2021(5): 157-160.

[19] 燕紫君, 熊英, 吴明芬. 灰色 DGM(2,1)与 BP 神经网络的组合预测模型研究[J]. 信息技术与信息化, 2023(6): 72-75.

[20] 吴博民. 基于数据挖掘的 5G 潜客识别的研究[D]: [硕士学位论文]. 桂林: 广西师范大学, 2022.

[21] 周雅婷. 基于数据挖掘识别移动 5G 潜在客户[D]: [硕士学位论文]. 重庆: 西南大学, 2022.

[22] 陈娟. 基于电商数据的高价值客户挖掘[D]: [硕士学位论文]. 上海: 东华大学, 2024.

[23] 何俊江. 基于改进 XGBoost 算法的潜在 5G 客户分析[D]: [硕士学位论文]. 武汉: 华中科技大学, 2023.

[24] 周琦. 基于机器学习的 5G 套餐潜在用户识别[D]: [硕士学位论文]. 重庆: 重庆大学, 2023.

[25] 孙希. 基于机器学习算法的银行信用卡潜在客户预测[D]: [硕士学位论文]. 天津: 南开大学, 2022.

[26] 冯亚辉. 基于机器学习的 5G 套餐潜在客户识别研究[D]: [硕士学位论文]. 北京: 中国石油大学, 2023.

[27] 谭广. 基于 E-LightGBM 算法的 5G 套餐潜在客户识别研究[D]: [硕士学位论文]. 重庆: 重庆工商大学, 2022.

[28] 李宏晨, 于泽平, 刘妍, 等. 基于组合赋权-TOPSIS 方法的天然气销售企业客户评价[J]. 天然气技术与经济, 2024, 18(2): 60-69.

[29] 谭镒锐. 基于熵权-TOPSIS 法的吉林银行财务风险评价及控制研究[D]: [硕士学位论文]. 吉林: 吉林财经大学, 2023.

[30] 翟芸, 胡冰, 施端阳. 基于改进 AHP-熵权法的雷达装备可靠性评估指标赋权方法[J]. 现代防御技术, 2022, 50(4): 148-155.

[31] 严格齐, 赵婉莹, 于镇伟, 等. 基于超参数优化算法的随机森林模型预测奶牛呼吸频率[J]. 农业工程学报, 2024, 40(11): 195-203.

[32] 田雅琴, 胡梦辉, 刘文涛, 等. 基于跳点搜索-遗传算法的自主移动机器人路径规划[J]. 工程设计学报, 2023, 30(6): 697-706.

[33] 周伟, 谭振江, 朱冰. 基于差分进化算法的大数据智能搜索引擎研究[J]. 情报科学, 2018, 36(5): 85-89.

[34] 胡丹青, 赵为华. 基于遗传算法的多结构变点检测及其应用[J]. 统计与决策, 2022, 38(6): 21-25.