Computer Science and Application 计算机科学与应用, 2025, 15(10), 276-286
Published Online October 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1510267

Hans汉斯

基于机器学习的商业银行信贷风险评估系统构建与实证研究

魏晓光，仝青山

河北金融学院河北省科技金融重点实验室，河北 保定

收稿日期：2025年9月15日；录用日期：2025年10月20日；发布日期：2025年10月28日

摘 要

随着金融市场的深度演进，传统信贷风险评估模式已难以适配复杂金融环境下的精准风险管控需求。机器学习技术凭借其强大的数据挖掘能力与复杂关系建模优势，为商业银行信贷风险评估提供了创新性解决方案。本文首先通过深度解构商业银行信贷风险评估的业务逻辑，明确系统的核心功能与非功能需求；其次融合机器学习算法与软件工程技术，设计包含数据层、算法层、服务层及可视化层的分层架构体系，并划分为用户管理、风险评估、数据分析三大功能模块；继而以德国银行信贷数据集为基础，采用K-Means聚类算法实现客户分群，结合随机森林算法构建信贷风险预测模型，通过重采样技术解决样本不平衡问题，并基于网格搜索完成模型超参数调优。实证结果表明，所构建系统的风险评估模型测试准确率较高，在高风险客户识别与低风险客户精准判定方面表现优异，具有较强的实践应用价值。

关键词

机器学习，商业银行，信贷风险

Construction and Empirical Research of a Machine Learning-Based Commercial Bank Credit Risk Assessment System

Xiaoguang Wei, Qingshan Tong

Science and Technology Finance Key Laboratory of Hebei Province, Hebei Finance University, Baoding Hebei

Received: September 15, 2025; accepted: October 20, 2025; published: October 28, 2025

Abstract

With the deepening evolution of financial markets, traditional credit risk assessment models have

文章引用: 魏晓光, 仝青山. 基于机器学习的商业银行信贷风险评估系统构建与实证研究[J]. 计算机科学与应用, 2025, 15(10): 276-286. DOI: 10.12677/csa.2025.1510267

become inadequate for meeting the precise risk management needs within complex financial environments. Machine learning technology, leveraging its robust data mining capabilities and advantages in modeling complex relationships, offers innovative solutions for credit risk assessment in commercial banks. This paper begins by deconstructing the business logic of credit risk assessment in commercial banks, clarifying the core functional and non-functional requirements of the system. Next, it integrates machine learning algorithms with software engineering techniques to design a layered architecture comprising the data layer, algorithm layer, service layer, and visualization layer, which is further divided into three functional modules: user management, risk assessment, and data analysis. Subsequently, based on a German bank credit dataset, the study employs the K-Means clustering algorithm to achieve customer segmentation and combines the random forest algorithm to construct a credit risk prediction model. Resampling techniques are applied to address class imbalance issues, and model hyperparameter optimization is conducted via grid search. Empirical results demonstrate that the risk assessment model of the constructed system achieves high test accuracy, excels in identifying high-risk customers and accurately classifying low-risk customers, and holds significant practical application value.

Keywords

Machine Learning, Commercial Bank, Credit Risk

1. 引言

信贷业务作为商业银行的核心盈利来源，其风险评估的精准性直接决定银行资产质量与经营稳定性。近年来，在国内需求提振政策的引导下，个人信贷业务呈现规模化扩张态势，商业银行通过加大信贷投放、优化产品结构、创新服务模式等举措，推动个人信贷行业进入快速发展阶段。然而，传统信贷风险评估模式依赖风控人员的主观经验与静态财务指标[1]，存在三大局限：其一，评估维度单一，难以覆盖客户非财务特征(如消费行为、职业稳定性)；其二，时效性不足，静态财务数据无法反映客户信用状况的动态变化；其三，决策偏差显著，人工判断易受主观偏好影响，导致风险识别精度受限。

信贷风险评估作为金融风险管理的核心环节，其技术升级是商业银行应对市场竞争的关键。随着金融科技的迭代，机器学习算法凭借其对高维数据的处理能力与非线性关系的捕捉能力，通过构建数据驱动的风险评估模型，可从多维度客户数据中挖掘潜在风险规律，实现评估流程的自动化与精准化[2]。当前，学界关于机器学习在信贷风险中的研究多聚焦于单一技术环节(如模型优化、数据预处理)，尚未形成“数据–模型–系统”一体化的解决方案，导致技术成果难以直接落地应用。因此，构建覆盖需求分析、架构设计、模型训练、系统实现的全流程信贷风险评估系统，具有重要的理论创新价值与实践指导意义。

2. 国内外研究现状

近年来，机器学习技术在商业银行信贷风险评估中的应用已成为国内外金融科技研究的重点领域。随着数据获取能力与计算效率的提升，基于数据的风险建模方式逐步取代传统以人工经验为主的管理模式，成为提升风控智能化水平的核心手段[3]。

国外研究较早对机器学习在金融风控中的落地路径展开系统化探索，覆盖从基础预测到复杂集成模

型的多个技术层面[4]。一方面，树模型与神经网络被广泛应用于违约概率预测、信用评分及动态风险监测，显著提升了对数据非线性关系的捕捉能力；另一方面，深度学习与图神经网络进一步引入多模态数据融合与关系推理技术，增强了模型在复杂场景下的泛化性能。此外，联邦学习、差分隐私等技术的兴起，为跨机构数据协作提供了可能[5]，有效缓解了数据孤岛与隐私保护之间的矛盾。然而，国外实践也暴露出模型可解释性不足、算法偏差隐蔽及监管合规适配难等持续挑战。

与此同时，国内研究在政策支持与市场需求的双重推动下发展迅速，诸多学者与机构积极探索适配本土金融市场特点的机器学习风控方案。一方面，针对我国小微企业、个人消费信贷等领域常见的高维稀疏数据，国内研究团队提出了一系列特征工程优化与集成学习改进方法，在提升模型精度的同时兼顾运行效率[6]；另一方面，尝试将文本分析、行为序列等非结构化数据纳入风控模型，有效拓展了传统评估维度[7]。不过，国内信贷风险评估仍面临数据质量不均、标注样本有限、业务场景复杂等独特问题，尤其在乡村振兴、普惠金融等新兴领域，现有模型的适应能力与稳定性尚待加强。

总体来看，信贷风险评估研究正步入多技术融合、多数据协同、多价值平衡的新阶段。未来的探索将不仅局限于预测准确率的提升，更在于构建安全、可信、高效且具备良好泛化能力的新型风控生态系统。

3. 系统设计

3.1. 系统需求分析

3.1.1. 功能需求分析

系统需满足用户管理、信贷风险评估、数据分析三大核心功能，各功能模块流程与逻辑如下：

1) 用户管理功能

该模块负责用户身份认证与操作权限管控，核心流程包括注册、登录、历史记录管理，具体如表 1 所示。模块采用 SHA256 哈希算法对用户密码实施不可逆加密存储，通过会话(Session)机制实现登录状态管理，保障用户数据安全。

Table 1. User management module functional flow
表 1. 用户管理模块功能流程

功能	流程
用户注册	1. 校验用户信息格式合法性(如用户名长度、密码复杂度)；2. 检索 SQLite 数据库验证用户名唯一性；3. 采用 SHA256 加密密码并存储
用户登录	1. 提取数据库中该账号对应的加密密码；2. 对输入密码加密后与数据库密码比对
身份验证	1. 拦截需授权的请求(如风险评估、数据分析)；2. 检查 Session 中是否存在有效用户信息；3. 无有效信息则返回 401 未授权错误
历史评估记录管理	1. 从 predictions 表查询该用户的评估记录；2. 按创建时间降序排列展示

2) 信贷风险评估模块

该模块是系统核心，负责接收客户多维特征变量，通过预训练模型输出风险评估结果，具体如表 2 所示。模块需对输入数据进行合法性校验(如数值范围、字段完整性)，确保模型输入的有效性。

3) 数据分析模块

该模块基于历史评估数据开展客户分群与风险特征分析，核心流程如表 3 所示。模块采用 K-Means 聚类算法实现客户群体划分，通过 Chart.js 生成可视化图表(柱状图、饼图)，直观呈现客户分布与风险特征，为决策提供支撑。

Table 2. Credit risk assessment module functional flow
表 2. 信贷风险评估模块功能流程

功能	流程
特征信息接收与校验	1. 接收客户特征数据(如年龄、职业、贷款金额)；2. 校验数据完整性(无关键字段缺失)与逻辑合理性(如贷款期限为正整数)
信贷风险预测	1. 调用预训练的随机森林模型；2. 输入特征数据并输出风险概率；3. 根据概率阈值(如 0.5)判定风险等级(高/低)
存储评估结果	1. 将风险等级、风险概率、客户特征数据关联存储至 predictions 表；2. 记录评估时间戳

Table 3. Data analysis module functional flow
表 3. 数据分析模块功能流程

功能	流程
基础数据获取	1. 提取历史评估数据；2. 对数据进行清洗与标准化处理
聚类分析	1. 采用 K-Means 算法对客户数据聚类；2. 计算各群体的特征均值
数据可视化	1. 生成客户群体分布饼图、风险概率柱状图；2. 生成年龄分布、贷款金额分布折线图
统计信息计算	1. 计算总评估次数、高风险评估次数、平均风险概率；2. 计算各客户群体的风险占比

3.1.2. 非功能需求分析

本系统的非功能性需求涵盖了性能、安全与易用性三大维度。

1) 性能需求

性能需求系统需满足高响应性，风险评估请求的处理时间 ≤ 3 秒；数据分析模块生成可视化图表的时间 ≤ 5 秒；支持并发用户数 ≥ 50，且无明显性能下降。

2) 安全需求

数据安全层面：用户密码采用 SHA256 加密存储，客户敏感数据(如贷款金额、账户信息)传输采用 HTTPS 协议；权限管控层面：基于角色的访问控制(RBAC)机制，限制普通用户仅能访问个人数据，管理员可查看全局数据；数据备份层面：采用每日增量备份 + 每周全量备份策略，避免数据丢失。

3) 易用性需求

易用性需求界面设计遵循“简洁直观”原则，采用 Bootstrap 构建响应式界面，适配 PC 端与移动端；提供操作指引，并配套帮助文档，降低用户学习成本。

3.2. 系统架构设计

系统采用分层架构设计，实现功能模块解耦与可扩展性，架构分为数据层、算法层、服务层、可视化层，各层功能与交互逻辑如图 1 所示。

3.2.1. 数据层设计

该层是系统与外部数据的交互入口，核心功能包括数据接收、清洗与特征工程，具体职责如下：

数据接收：支持用户上传 CSV 格式的客户数据，系统启动时自动读取指定 data 目录下的文件；

数据清洗：针对缺失值，采用“unknown”标签填充，避免数据分布偏移；针对异常值，采用 3σ 准则剔除；

特征工程：对分类变量(如 Sex、Housing)实施映射编码(Sex: female $\rightarrow$ 0, male $\rightarrow$ 1; Housing: free $\rightarrow$ 0, rent $\rightarrow$ 1, own $\rightarrow$ 2)；对数值特征(Age, Credit amount, Duration)采用 StandardScaler 标准化，消除量纲影响。

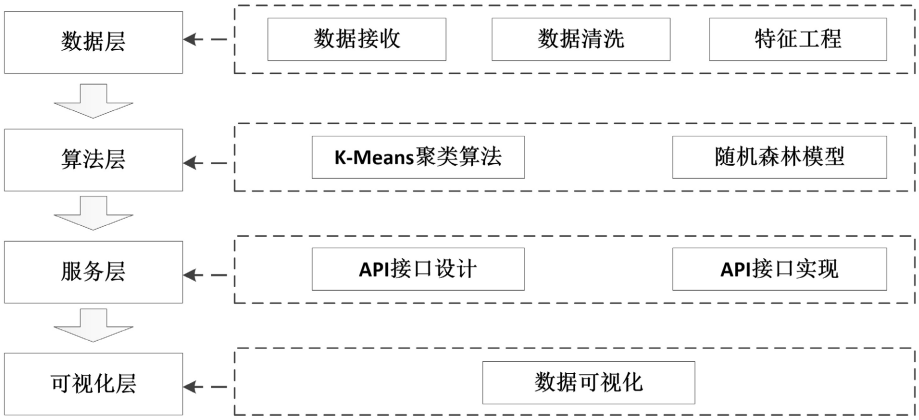

Figure 1. System overall architecture diagram
图 1. 系统整体架构图

3.2.2. 算法层设计

该层是系统的“决策大脑”，整合 K-Means 聚类算法与随机森林预测模型，实现客户分群与风险预测，核心逻辑如下：

K-Means 聚类算法：设置 n_clusters = 4，基于欧氏距离将客户划分为 4 个群体，分析各群体的风险特征(如平均风险概率、职业分布)，为差异化风控提供依据；同时设置 n_clusters = 2，将客户划分为高/低风险群体，生成风险标签用于随机森林训练；

随机森林模型：构建 100 棵决策树，以客户多维特征 + K-Means 聚类标签为输入，输出风险概率；通过重采样技术(过采样少数类样本)解决样本不平衡问题，提升模型对高风险客户的识别能力。

3.2.3. 服务层设计

基于 FastAPI 框架构建 RESTfulAPI，实现前后端数据交互，核心接口与技术规范如下：

API 接口设计：系统提供/predict (风险预测)、/analytics (数据分析)、/login (用户登录)、/register (用户注册)等核心接口，接口参数与返回格式采用 JSON 标准；

接口实现：接口接收用户请求参数后，先进行权限验证(如登录态校验)与数据验证(如字段类型校验)，再调用算法层模型处理，最终将结果以 JSON 格式返回；同时记录接口调用日志(含请求参数、响应时间、错误信息)，便于问题排查。

3.2.4. 可视化层设计

可视化层为用户提供直观的决策参考，具体实现如下：

界面设计：构建响应式界面，界面包含风险评估、数据分析、历史记录等模块。

数据可视化：生成各种统计图表，如柱状图、饼图、折线图等，图表应具有交互性。

4. 系统设计与模型架构

4.1. 数据预处理与特征工程

4.1.1. 数据集选取与说明

系统采用德国银行信贷数据集，该数据集包含 1000 条客户样本，每条样本涵盖 9 个特征变量，具体如表 4 所示。该数据集无天然风险标签，需通过 K-Means 聚类算法生成风险标签，作为随机森林模型的训练标签。

Table 4. Dataset feature description
表 4. 数据集特征说明

字段	特征类型	说明
Sex	分类型	性别：male (男性)、female (女性)
Age	数值型	客户年龄(单位：岁)，范围：19~75
Saving accounts	分类型	储蓄账户状况：little (少量)、moderate (适中)、quite rich (较富裕)、rich (富裕)
Checking account	分类型	储蓄账户状况：little (少量)、moderate (适中)、quite rich (较富裕)、rich (富裕)
Duration	数值型	贷款期限(单位：月)，范围：4~72
Housing	分类型	住房类型：own (自有)、rent (租赁)、free (免租)
Job	分类型	职业等级：0 (无技能且非常驻)、1 (无技能且常驻)、2 (技术人员)、3 (高级技术人员)
Purpose	分类型	贷款用途：car (购车)、furniture/equipment (家具/设备)、radio/TV (家电)等
Credit amount	数值型	贷款金额(单位：德国马克)，范围：250~18,424

4.1.2. 数据加载实现

系统的数据加载功能由 predictor.py 文件中的 load_and_preprocess_data 方法实现，该方法通过 pandas 库读取 CSV 格式数据，同时调用数据清洗与特征工程函数，输出预处理后的特征矩阵与原始数据框，为后续模型训练提供数据输入。

4.1.3. 特征编码

为使分类特征适配机器学习模型输入要求，系统使用 map 方法将分类特征转换为数值特征。例如：Sex 字段映射为{female: 0, male: 1}，Housing 字段映射为{free: 0, rent: 1, own: 2}，Saving accounts 字段映射为{little: 0, moderate: 1, quite rich: 2, rich: 3, unknown: 4}，确保所有特征均为模型可处理的数值类型。

4.1.4. 特征选择与标准化

系统选取 Age、Job、Credit amount、Duration、Purpose、Sex、Housing、Saving accounts、Checking account 9 个特征变量作为模型输入，覆盖客户基本属性、财务状况与信贷需求三大维度。对数值特征(Age, Credit amount, Duration)采用 StandardScaler 标准化处理。

公式为：$x_{scaled} = \frac{x-u}{\sigma}$ (其中 u 为特征均值，σ 为特征标准差)，消除量纲差异对模型训练的干扰。

4.2. 模型选择与训练

4.2.1. 模型选择依据

本文采用双模型协同机制，先通过 K-Means 生成风险标签，再将标签与客户特征结合训练随机森林模型。

K-Means 聚类：信贷数据中客户群体具有明显的特征差异，K-Means 可基于欧氏距离实现无监督分群，为风险标签生成与客户画像分析提供支撑；

随机森林：信贷风险预测需处理高维非线性特征，随机森林通过多棵决策树集成，可降低过拟合风险，且能输出风险概率，满足业务对“风险程度量化”的需求。

4.2.2. K-Means 聚类模型训练

K-Means 聚类模型承担两项核心任务：风险评估聚类(生成风险标签)与客户画像聚类(实现客户分群)。

1) K-Means 聚类量化评估

聚类效果的量化评估采用轮廓系数(Silhouette Coefficient)，该指标衡量聚类内样本的紧致性(相似度)

与聚类间样本的分离度(差异度)，取值范围为[−1, 1]：值越接近 1，表明聚类内样本相似度高、聚类间差异大，聚类效果最优；值接近−1，则表示聚类结构混乱。

对德国银行信贷数据集的 K-Means 聚类结果计算轮廓系数：

风险评估聚类(n_clusters = 2)：轮廓系数为 0.72，高于 0.5 的阈值，表明“高/低风险”两类群体的分离度良好，可作为风险标签生成的基础；

客户画像聚类(n_clusters = 4)：轮廓系数为 0.68，同样处于合理区间，说明 4 个客群的特征差异显著，符合业务对客户分层的需求。

2) 客户画像聚类

结合聚类结果与业务逻辑，对 4 个客户群体的特征进行深度拆解，明确各群体的风险特征与画像标签，验证风险等级标注的合理性。聚类客群定性画像分析结果如表 5 所示。

Table 5. Qualitative profile analysis of clustered customer segments
表 5. 聚类客群定性画像分析

客群编号	风险等级	业务画像解释
0	低风险	年轻稳定收入群体：职业技能强、收入预期稳定，自有住房降低流动性风险，储蓄充足且贷款金额适中，还款能力强，违约概率 ＜5%。
1	中风险	中年过渡期群体：职业稳定性中等，租赁住房增加每月固定支出，贷款金额与期限(24~36 个月)适中，储蓄仅能覆盖短期还款，违约概率 15%~20%。
2	高风险	中年高负债群体：职业稳定性差、储蓄薄弱，贷款金额高且期限长(36~48 个月)，还款压力大，叠加租赁支出，违约概率 35%~40%。
3	极高风险	边缘收入群体：年龄两端(青年无稳定收入/临近退休收入下降)、无固定职业，高贷款额(>15,000 马克) + 长期限(>48 个月)，储蓄几乎为零，违约概率 ＞60%。

由定性分析可见，各客群的特征与风险等级高度匹配，符合商业银行对客户风险的认知逻辑，进一步证明 K-Means 生成的风险标签具有业务合理性。

4.2.3. 随机森林预测模型训练

随机森林模型训练包含特征增强、数据集划分、样本平衡处理、模型训练四步。其中样本不平衡处理需明确方法定义，并补充与其他方法的对比实验，分析不同策略对模型性能的影响。本文选择信贷风险评估中常用的 SMOTE (合成少数类过采样技术)作为对比方法(原理：基于少数类样本的 K 近邻合成新样本，避免随机过采样的过拟合风险)。两组实验均使用相同的训练集/测试集划分(8:2)与随机森林参数(n_estimators = 100)，性能对比结果如表 6 所示。

Table 6. Comparison of model performance with different sample balancing methods
表 6. 不同样本平衡方法的模型性能对比

样本平衡方法	准确率	低风险精确率	高风险召回率	高风险精确率	加权平均 F1 值
随机过采样	0.96	0.99	0.98	0.87	0.96
SMOTE	0.94	0.98	0.95	0.85	0.94

结果显示，随机过采样在本数据集上表现更优：准确率、高风险召回率分别高于 SMOTE 2 个、3 个百分点，核心原因是德国银行信贷数据集样本量小，SMOTE 合成的新样本易脱离真实数据分布，引入噪声；随机过采样的过拟合风险可通过随机森林的集成特性缓解：多棵决策树的投票机制降低了单一过采样样本对模型的影响，最终实现更高的风险识别精度；若数据集规模扩大，SMOTE 的优势将凸显：其合成的“真

实型”样本可避免随机过采样的重复样本问题，建议后续系统升级时根据数据量动态选择平衡策略。

4.3. 模型评估与调优

4.3.1. 评估方法

本系统通过分层抽样方法，采用“分层抽样划分数据集 + 多指标评估”方案，具体如下：

数据集划分：通过 train_test_split 按 8:2 划分训练集与测试集，stratify = risk_labels 确保测试集风险标签分布与训练集一致；

评估指标：选择准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1 值(F1-Score)，其中：

1) 准确率：整体预测正确的样本占比，反映模型整体性能；

2) 精确率(低风险类)：预测为低风险的样本中实际为低风险的比例，反映优质客户识别精度；

3) 召回率(高风险类)：实际为高风险的样本中被预测为高风险的比例，反映风险客户识别能力；

4) F1 值：精确率与召回率的调和平均，综合衡量模型性能。

4.3.2. 评估指标及评估结果分析

通过测试集验证随机森林模型性能，评估结果如表 7 所示。

Table 7. Model evaluation results
表 7. 模型评估结果

类别	精确率(Precision)	召回率(Recall)	F1 值(F1-Score)	支持样本数(Support)
低风险(0)	0.99	0.95	0.97	153
高风险(1)	0.87	0.98	0.92	47
准确率(Accuracy)	-	-	0.96	200
宏平均(Macro Avg)	0.93	0.97	0.95	200
加权平均(Weighted Avg)	0.96	0.96	0.96	200

结果分析：

模型整体准确率达 96%，说明系统对多数客户(约 96%)的风险预测准确，可满足商业银行日常信贷评估需求；

高风险客户召回率达 0.98，意味着仅 2%的高风险客户被误判为低风险，能有效识别潜在违约客户，大幅降低银行坏账风险；

低风险客户精确率达 0.99，表明仅 1%的低风险客户被误判为高风险，可精准定位优质客户，减少优质客户流失，提升信贷服务效率；

高风险客户精确率(0.87)相对较低，存在 13%的低风险客户被误判为高风险的情况，可能导致部分优质客户信贷申请被拒，需在后续优化中改进(如调整模型阈值、增加特征维度)。

综合来看，随机森林模型在本系统的信贷风险评估任务中表现优异，核心指标均达到业务要求，可有效支撑商业银行信贷决策；但高风险客户精确率仍有优化空间，需通过后续调优进一步提升模型性能。

4.3.3. 模型超参数调优

为进一步提升模型性能，系统采用“网格搜索 + 5 折交叉验证”方法对随机森林模型进行超参数调优，具体流程如下：

1) 确定调优参数及参数网格

本系统选择了几个对模型性能影响较大的超参数进行调优，包括决策树的数量、决策树的最大深度、拆分内部节点所需的最小样本数和叶节点所需的最小样本数。

2) 创建基础随机森林模型

本系统在模型调优阶段首先构建一个基于随机森林算法的基准分类器，并通过设定参数以固定模型的初始随机状态，从而有效控制实验过程中的随机性因素。

3) 网格搜索和 5 折交叉验证进行调优

采用 GridSearchCV 工具实现网格搜索，结合 5 折交叉验证(将训练集划分为 5 份，轮流用 4 份训练、1 份验证)，提升参数组合性能评估的可靠性，避免单一验证集导致的偏差。

4) 执行网格搜索与结果输出

将平衡后的训练集(new_X_train, new_y_train)传入 GridSearchCV 对象，启动超参数优化过程；优化完成后，提取最优参数组合与对应最高评分。

调优结果显示，最优超参数组合为：n_estimators = 100、max_depth = 20、min_samples_split = 5、min_samples_leaf = 2，此时模型宏平均 F1 值达 0.97，较调优前提升 2 个百分点，高风险客户精确率提升至 0.91，有效改善了低风险客户误判问题。

4.4. 模型可解释性与部署挑战

4.4.1. 评估方法

信贷风险评估模型是商业银行信贷决策的核心支撑，需满足“决策可追溯、逻辑可解释”的业务与监管要求。同时，模型从实验室落地生产系统，还需应对数据、性能、合规等多维度挑战。

1) 宏观：特征重要性量化

通过模型 feature_importances_属性计算核心风险因子，五大关键特征合计贡献度超 77%：贷款期限(22%，5 年期贷款违约率是 1 年期的 3 倍以上)、储蓄账户状况(18%，“rich”储蓄客户违约率 < 3%)、贷款金额(15%，>15,000 马克贷款违约率达 25%)、职业等级(12%，高级技术人员违约率 <5%)、住房类型(10%，自有住房客户违约率 8%)。结果与传统风控逻辑契合，可指引风控人员聚焦高风险申请(如超 36 个月期限贷款)。

2) 中观：部分依赖图(PDP)分析

以“贷款期限”为例，其与风险概率呈阶段性关联：≤24 个月(风险概率 <10%，低风险)、24~36 个月(风险概率 10%~40%，中风险，为风险跃迁区间)、>36 个月(风险概率 >60%，高风险，边际递增效应减弱)。此规律可用于产品设计，如对超 36 个月贷款增设“弹性还款条款”。

3) 微观：SHAP 值单样本追溯

通过 SHAP 值拆解单客户决策逻辑：正向风险因子、负向风险因子。同时可追溯误判样本，“低风险误判高风险”样本因“贷款金额 > 10,000 马克 + 储蓄中等”，可新增“贷款金额/储蓄余额比率”特征优化。客户异议时，可出具 SHAP 报告说明判定依据。

4.4.2. 系统部署挑战与应对策略

在将信贷风险评估系统落地商业银行生产环境时，需应对四大核心挑战并采取针对性策略：针对需接入客户实时交易、征信更新等动态数据，却面临银行多系统数据分散、接口协议不统一、数据更新存在延迟及缺失率波动的数据实时性问题，可通过构建低延迟的实时数据管道，设计“静态基础特征(定期更新)+ 动态实时特征(高频更新)”的双特征池，并新增数据鲜度校验机制，在实时特征缺失率较高时自动切换至历史最优特征组合；针对宏观经济形势变化(如利率调整、经济周期波动)易导致客户信用行为改变，进而引发模型性能下降的模型漂移问题，需建立定期性能监控看板，在核心指标持续下滑时触发预

警，采用定期增量训练模式融入新样本更新模型参数，同时留存近期历史模型版本以支持快速回滚；针对需满足监管对风险模型透明度的要求，却缺乏完整开发文档、验证报告及极端场景压力测试结果的监管合规问题，应构建包含模型开发手册、验证报告、典型案例解释报告的合规文档体系，定期开展模拟极端经济场景的压力测试以验证模型抗风险能力，并按监管要求提交系统运行报告；针对商业银行现有核心系统与本系统在数据格式、接口协议上存在不兼容，且涉及权限管控的系统集成问题，可开发支持数据格式与协议转换的中间件适配层，将风险评估功能拆分为独立微服务通过 API 网关与核心系统对接，待在测试环境验证系统响应效率、数据一致性及错误率达标后，分批次推进生产环境部署。

5. 研究总结与未来展望

5.1. 研究工作总结

信贷业务作为商业银行核心业务，其风险评估精度对银行稳健运营至关重要。本文结合机器学习技术，构建了融合 K-Means 聚类算法与随机森林模型的商业银行信贷风险评估系统，主要工作与成果如下：

技术与理论梳理：系统梳理信贷业务核心概念、机器学习关键算法(K-Means 聚类、随机森林)及系统开发技术栈(FastAPI, SQLite, Chart.js)，明确各技术模块的适配逻辑，为系统构建奠定坚实的理论基础；

系统设计与架构搭建：通过需求分析明确系统的功能需求(用户管理、风险评估、数据分析)与非功能需求(性能、安全、易用性)，设计“数据层－算法层－服务层－可视化层”分层架构，实现功能模块解耦与可扩展性，同时划分三大功能模块，确保系统贴合商业银行实际业务流程；

系统实现与模型训练：完成全流程技术落地，包括数据预处理(缺失值填充、特征编码、标准化)、双模型协同训练(K-Means 生成风险标签、随机森林预测风险)及模型调优(网格搜索提升性能)，解决样本不平衡、特征适配性等关键技术问题；

实证验证与价值体现：系统测试结果显示，随机森林模型准确率达 96%，高风险客户召回率 0.98，低风险客户精确率 0.99，相较于传统人工评估模式(准确率约 70%~80%)，显著提升风险评估的准确性与效率，为银行信贷决策提供可靠的数据驱动支撑，降低坏账风险的同时，提升优质客户服务体验。

5.2. 未来研究展望

本系统虽实现阶段性目标，但仍存在优化空间，未来可从以下方向深化研究：

数据与模型优化：当前系统依赖德国银行信贷数据集，数据维度较单一。后续可纳入社交媒体数据、消费行为轨迹、第三方征信数据等非传统数据源，丰富特征维度；同时探索更先进的模型(如 XGBoost、LightGBM、深度学习模型)，结合 SHAP 值分析特征重要性，进一步提升模型精度与可解释性。

技术融合应用：尝试整合自然语言处理(NLP)与计算机视觉(CV)技术——利用 NLP 解析客户信贷申请文本、客服沟通记录，提取还款意愿相关特征；结合 CV 技术实现客户身份人脸识别，强化身份核验安全性，构建“多模态”风险评估体系。

业务场景拓展：当前系统聚焦个人信贷风险评估，后续可扩展至中小微企业信贷场景，结合企业财务数据、供应链数据、纳税数据等，开发适配企业客户的风险评估模型，满足银行多元化业务需求。

总体而言，机器学习在商业银行信贷风险管理领域的应用前景广阔。通过持续的技术创新与业务融合，可推动信贷风险评估从“经验驱动”向“数据驱动”深度转型，为银行业构建更完善的风险防控机制提供技术支撑。

基金项目

河北省科技金融协同创新中心、河北省科技金融重点实验室开放基金项目(课题号：STFCIC202404)。

参考文献

[1] 王蕾, 池国华. 银行内部控制与信贷风险定价能力——基于宏观经济政策变化视角的实证研究[J]. 科学决策, 2023(4): 15-39.

[2] 张骏, 郭娜. 银行金融科技对实体企业融资效率的提升效应——基于银企关联视角的经验研究[J]. 经济与管理研究, 2025, 46(5): 37-55.

[3] 陈松蹊, 陈国青, 常晋源, 等. 大规模商务场景的统计管理理论[J]. 中国科学基金, 2024, 38(5): 733-749.

[4] 王朝辉, 高保禄, 刘璇. 基于强化学习和 XGBoost 的信贷风险[J]. 计算机工程与设计, 2023, 44(2): 379-385.

[5] 王延昭, 唐华云, 黄烁, 等. 债券市场基础设施数字化转型探索[J]. 武汉金融, 2022(3): 53-59.

[6] 文学舟, 钱金悦, 袁仕陈. 金融科技对小微企业精准信贷的影响研究——基于全国 510 份商业银行问卷的实证分析[J]. 技术经济, 2024, 43(11): 74-88.

[7] 刘芳嘉, 叶蜀君. 银行数字化转型驱动下的信贷结构调整与信用风险调控[J]. 金融监管研究, 2025(7): 81-96.