

**本科毕业论文（设计）**

**开题报告书**

**题 目 基于数据挖掘的借款人贷款决策影响因素及风**

**险评估**

**——以Lending Club网络贷款数据为例**

**学　　院 信息工程学院**

**班 级 大数据管理2001**

**姓 名 应敏婕 学 号 202021130214**

**指导教师 祝启虎 职 称 讲师**

2024年1月13日

|  |
| --- |
| 1. 本选题研究的目的及意义   1.1研究背景和意义  根据中国人民银行的统计数据显示2023年三季度末，金融机构人民币各项贷款余额234.59万亿元，同比增长10.9%，前三季度人民币贷款增加19.75万亿元，同比多增1.67万亿元，可以看出我国对于贷款需求的活跃程度显著增加。随着金融科技的快速发展和互联网金融的普及，信用贷款业务也越来越受到追捧，信用贷款业务因其具有客观的收益使得银行和金融机构的放款意愿增加。  中央有关部委也提出要深入研究金融健康问题,推进普惠金融高质量发展。人民银行2023年2月10日召开金融市场工作会议指出要及时分析研判经济金融运行苗头性趋势性变化，有力支持实体经济高质量发展，积极防控金融市场重点领域风险，统筹推进金融市场运行机制建设和改革发展稳定，保障金融基础设施高效稳健运行，为全面建设社会主义现代化国家开好局起好步提供有力有效的金融支持。[1]随着网贷新规等政策的出台,消费金融市场的规范性也逐步增强,消费金融公司也开始向利用大数据等金融科技技术,为消费者提供更好的金融产品与服务的公司进行转变。  但贷款风险的存在令金融机构和借贷双方都面临着潜在的不确定性和损失风险。因此如何提高银行和金融机构抵御信用风险的能力成为了目前研究的热点。传统的统计分析方法和理论分析方法已经无法满足日益增长的信贷风险预测业务需求,也无法保证预测质量。基于数据挖掘的借款人贷款决策影响因素及风险评估是以借款消费者的借款登记和历史行为信息作为判断依据，对借贷人能否履行财务承诺或还款承诺的提前预测工作。了解借款人贷款决策的影响因素以及如何评估风险，精准识别出潜在违约用户并降低商业银行的不良贷款率，通过数据挖掘技术从海量个人信贷数据中智能化识别违约客户，能够高效地为金融机构提供预警信息，对提高贷款市场的稳和借贷效率至关重要。  1.2研究目的  降低信息不对称造成的逆向选择和道德风险。对借款人进行预筛选、信用评分等是贷款必须要做的前期工作。随着用户贷款信息的海量增加,如何有效且快速地预测用户违约风险和评估用户信用等级,是当前网络贷款领域面临的一个重要而紧迫的问题。故对借款人进行风险评估，是互联网到金融科技时代发展，以及市场借贷稳健运营需要关注的重点。  帮助机构或平台贷款准确决策。传统的贷款决策主要依赖于借款人的信用报告和少量的财务信息，但是这些信息无法全面反映借款人的真实情况。本研究意在基于借款人的个人历史信息：个人职位名称，请求类别，信用额度，额度开通时间，按揭账户数目，破产记录数量等等对贷款人进行风险评估。有助于银行和其他金融机构能做出更明智的贷款决策，降低坏账风险；定量表示借款人的特征与偿还能力之间的关系，为贷款机构提供更科学的风险评估模型，降低不必要的损失和风险。降低违约风险，提高盈利能力。  帮助个性化贷款产品开发。数据挖掘一类的基于大样本的评估和预测方法可以揭示借款人的个性化需求和偏好，帮助银行或金融机构构建借款人群的用户特征和画像，推出更加贴合借款人需求的个性化贷款产品。通过了解借款人的个人信息、消费行为等信息，可以针对性为借款人提供更加灵活、定制化的贷款服务，提高客户满意度和忠诚度。  优化贷款进度流程。传统的贷款流程通常需要借款人提供个人信息以及历史数据，流程繁琐且耗时。基于数据挖掘的贷款决策可以更加智能化的方式筛选出更有效的信息，排除相关度低的其他干扰信息，减少后台的借款人风险评估计算量，实现贷款流程速度的优化，提高贷款审批的效率和速度。  预警还款计划和稳定金融市场。借款人根据自己的风险等级和画像特征可以预测自己是否有能力如期偿还贷款，有助于还款人提前做好还款规划，避免逾期或违约情况的发生，保护个人信用记录。另外本研究还可以为金融行业提供参考，推动贷款风险管理和信用评估的发展，促进金融市场的稳定和健康发展。  本研究旨在利用开源数据集或某个借贷交易平台的用户贷款申请数据，运用数据挖掘技术，探索借款人贷款决策的影响因素，并构建一个风险评估模型。通过建立基于该数据集的风险预测模型，分析借款人的特征与其偿还能力之间的关系。这样的模型可以帮助贷款金融机构根据借款人的历史数据和贷款注册信息预测其是否会偿还贷款，为贷款机构制定更准确的风险评估模型提供依据。此外，该研究还旨在帮助借款人了解贷款决策的关键因素，从而更好地规划自身的贷款行为，提高借款成功率和还款能力。 |
| 1. 本选题国内外研究状况综述   2.1国内研究现状  网络借贷行为研究方面。苏亚，成春林（2017）认为在对网贷借款人违约行为的影响分析中应多维综合考虑借款人的“软信息”和“硬信息”。于晓虹等（2016）研究发现网贷借款人历史违约期数、单位性质、住房条件、婚姻状况以及年龄等指标对贷后违约依次存在递减效应。沈永健（2022）在客户数据的统计分析和业务材料收集的基础上,发现研究对象在客户准入、贷款资金用途跟踪和贷款违约催收等环节存在问题,大数据应用存在明显不足。赵志红等（2021）将采集的信贷欺诈数据作为风险预测模型的数据支持，定义信贷欺诈风险产生的信号频谱建立风险信息函数。李羿霖（2022）研究了首次违约事件给投资者学习行为带来的净影响。戴杨文意（2021）采用抽样调查的方式通过自填问卷法考察大学生网贷行为的现状及其影响因素。  机器学习风险评估和预测方法选择上，沈玉溪，徐浩（2018）运用Lending Club2017年第一季度的交易数据通过决策树模型对借款人的违约风险特征的重要性进行分析。谭中明，谢坤，彭耀鹏（2018）利用Logistic回归方程式的方法筛选变量建立梯度提升决策树模型来对借款人的信用风险进行评测。霍江林（2016）等运用神经网络模型测度P2P借款人信用风险, 吴利娟，孟泽远（2023）的研究表明该模型具备较优的评测效果。结果表明，BP神经网络可以很好地对银行机构绿色信贷风险进行预测。张宇幸（2023）通过主成分分析法构建指标体系,分别使用Logistic模型、决策树模型、KNN模型来进行对比分析,结果表明KNN模型分类效果最佳。  2.2国外研究现状  网络借贷行为研究方面。Herzenstein（2011）等指出信用状况以及背景特征等指标反映的是网贷主体偿付资金的效力, 因而会影响到其还贷行为。Lim M（2009）等研究了社交资本对网贷借款人贷后行为的影响。Collier（2010）指出借款金额、财务状况以及竞拍方式等指标对网贷违约影响显著。  机器学习风险评估和预测方法选择上，由Xue等（2020）于2020年提出的麻雀算法（sparrow search algorithm,SSA）是一种新型群智能优化算法，表现出了较高的寻找最优解能力。PRISCILLA C V提出的极端梯度提升（extreme gradient boosting,XGBoost）算法被广泛应用于该领域并取得了较好的效果。常规的XGBoost集成模型具有参数过多，计算复杂等特点，因此优化模型参数具有重要的意义。  2.3研究现状述评  学者们对网络贷款风险研究多基于单向和消费者个人信息影响因素探究，往往易忽略现实中多方面因素的影响。数据的选取上，多用于传统的问卷与业务材料收集信息进行分析，基于互联网背景的贷款人大样本分析较少。在网贷信用风险评测模型层面, 定性与定量相结合以及契合风险评估的多类型复杂系统建模方法被广泛应用到违约风险评估领域，多种机器学习的方法被用于风险预测是目前的发展趋势，但在金融科技发展背景下基于大数据挖掘的消费者信贷风险预测研究仍存在明显不足。根据“贷款风险预测”关键词检索可知大部分的学者对这类风险的识别与预测研究集中在P2P市场，且经调研了解后发现P2P网贷市场的准入门槛较低，样本庞大易获取，有助于进行贷款产品的战略研究。因此本文综合借款人的基础历史信息与借款信息，结合文本类数据与时间序列数据进行风险建模，讨论借款人的用户特征与分类。 |

|  |
| --- |
| 1. 本选题研究的主要内容及写作提纲   主要内容  1) 研究背景和意义：金融科技的发展，道德风险下机构对借款人风险评估的价值和意义；数据挖掘在贷款决策和风险评估中的应用背景和意义；  2) 相关理论和方法：综述数据挖掘在贷款决策和风险评估中常用的理论和方法，包括特征选择、分类算法、聚类分析等。  3) 数据收集和预处理：介绍数据来源，全球 P2P 市场是一个巨量市场，主要覆盖个人消费贷款和中小企业贷款。lending club是全球最大的个人对个人(p2p)借贷平台，是第一家向美国证券交易委员会(SEC)注册其证券产品并在二级市场提供贷款交易的p2p贷款机构，它的贷款用户数据对于研究互联网借贷贷款人的风险评估有重要参考价值。并对数据进行预处理，包括数据清洗、数据集成、数据变换等步骤。  4) 特征选择和构建：使用特征选择方法，从借款人的各项数据中选择出对贷款决策和风险评估有影响的特征，并构建特征向量。  5) 模型建立和评估：选择合适的分类算法或聚类分析方法，建立贷款决策模型或风险评估模型，并对模型进行评估。基于所得出的影响因素，建立贷款风险评估模型，例如使用机器学习算法如决策树、随机森林等，预测借款人的还款能力和偿还风险。  6) 结果分析和讨论：分析模型的结果，探讨借款人贷款决策的影响因素和风险评估的准确性。根据选取的模型和误差分析得到的准确率、召回率、精确率等指标，验证所建立的模型在真实数据集上的有效性，根据数据结果和现有资料参考，并探讨研究结果对金融机构风险管理和信贷政策的指导意义。  7) 研究局限和展望：总结研究的局限性，并提出未来进一步研究的展望，包括改进模型、扩大样本规模等。  一、 引言  1) 研究背景和意义  2) 国内外综述  3) 研究方法和数据来源  二、 相关理论和方法  1) 数据挖掘在贷款决策和风险评估中的应用概述  2) 特征选择方法的介绍  3) 分类算法和聚类分析方法的介绍  三、 数据处理与分析  (一) 数据收集和预处理  1) 数据准备（来源获取方式、预处理填充）  2) 数据的集成和变换处理  (二) 特征选择和构建  1) 特征选择方法的选择和介绍  2) 从借款人数据中选择对贷款决策和风险评估有影响的特征  3) 构建特征向量  四、 模型建立和评估  1) 选择合适的分类算法或聚类分析方法  2) 建立贷款决策模型或风险评估模型  3) 对模型进行评估和验证  五、 结果分析和讨论  1) 分析模型的结果  2) 讨论模型的有效性和可行性  3) 探讨借款人贷款决策的影响因素和风险评估的准确性  六、 研究局限和展望  1) 总结研究的局限性（数据指标局限、样本大小局限、时间局限等等）  2) 提出未来进一步研究的展望  3) 改进模型、扩大样本规模等建议 |
| 1. 主要参考文献   [1]. 人民银行召开2023年金融市场工作会议  [2]. 沈永健. C银行苏州分行个人信贷风险管理中的大数据应用研究[D].苏州大学,2022.DOI:10.27351/d.cnki.gszhu.2022.000610.  [3]. 陈盛邦. 基于深度学习的个贷信用风险预测研究[D].黑龙江大学,2023.DOI:10.27123/d.cnki.ghlju.2023.000391.  [4]. 吴利娟,孟泽远.基于BP神经网络的银行机构绿色信贷风险评价研究[J].时代经贸,2023,20(11):72-76.DOI:10.19463/j.cnki.sdjm.2023.11.013.  [5]. 张宇幸. L商业银行个人信贷风险预警与应对策略研究[D].山东财经大学,2023.DOI:10.27274/d.cnki.gsdjc.2023.001070.  [6]. 王雨萌. 基于多特征信息融合的网络借贷行为影响因素及预测研究[D].北京邮电大学,2022.DOI:10.26969/d.cnki.gbydu.2022.000986.  [7]. 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].北京科技大学,2023.DOI:10.26945/d.cnki.gbjku.2023.000344.  [8]. 王萧萧,王亭雯,马玉玲等.基于深度森林的P2P网贷借款人信用风险评估方法[J].计算机科学,2021,48(S2):429-434.  [9]. 徐紫秋. 网贷个人信用风险评估集成研究[D].安徽工业大学,2022.DOI:10.27790/d.cnki.gahgy.2022.000285.  [10]. 钟方强. P2P借贷平台风险预警研究[D].南昌大学,2022.DOI:10.27232/d.cnki.gnchu.2021.002232.  [11]. 赵竞达. 基于机器学习算法的网贷信用风险模型研究[D].对外经济贸易大学,2022.DOI:10.27015/d.cnki.gdwju.2021.000234.  [12]. 胡修远. P2P网贷借款人违约风险研究[D].厦门大学,2021.DOI:10.27424/d.cnki.gxmdu.2021.001992.  [13]. 周文攀. P2P投资者的风险识别与投资行为研究[D].吉林大学,2022.DOI:10.27162/d.cnki.gjlin.2021.003730.  [14]. 苏亚,成春林.P2P网贷借款人违约行为影响因素的实证研究[J].金融发展研究,2017(01):70-76.DOI:10.19647/j.cnki.37-1462/f.2017.01.010.  [15]. 谭中明,谢坤,彭耀鹏.基于梯度提升决策树模型的P2P网贷借款人信用风险评测研究[J].软科学,2018,32(12):136-140.DOI:10.13956/j.ss.1001-8409.2018.12.28  [16]. 霍江林,刘素荣.P2P网贷平台借款人信用风险评估研究[J].金融发展研究,2016,(12):43-47.DOI:10.19647/j.cnki.37-1462/f.2016.12.008  [17]. 于晓虹,楼文高.基于随机森林的P2P网贷信用风险评价、预警与实证研究[J].金融理论与实践,2016,(02):53-58.  [18]. Herzenstein M S. Sonenshein U M Dhola-kia. Tell me a Good Story and I May Lend You Money:The Role of Narratives in Peer-to-peer Lending Decisions[J]. Journal of Marketing Research, 2011, 48 (SPL) :138-149.  [19]. Lin M N R Prabhala, S Viswanathan. Judging Borrowers by the Company They Keep:Social Networks and Adverse Selection in Online Peer-to-peer Lending[J]. SSRN e Library, 2009.  [20]. Collier B, R Hampshire.Sending Mixed Signals:Multilevel Reputation Effects in Peer-to-Peer Lending Markets[M]. CSCW, Savannah, USA.2010. |

|  |
| --- |
| 1. 完成措施及进度安排   2024年1月10日前：数据收集和准备  1． 收集相关文献，了解借款人贷款决策的影响因素和风险评估的研究现状。  1月10日-1月14日：数据收集和处理  2． 收集借款人的相关数据，包括个人信息、财务状况、信用记录等。  3． 对收集到的数据进行预处理，包括数据清洗、缺失值处理等。  4． 进行数据探索性分析，包括降维、主成分分析、交叉验证等。  5． 提取借款人的贷款申请文本信息特征，初步探索文本类数据对贷款人风险的影响。  1月15日-1月21日：  1． 根据研究目标，确定数据集适合的统计或机器学习模型。  2． 完成代码逻辑梳理，进行模型训练和评估。  3． 划分训练集和测试集，并进行特征工程。  1月22日-1月28日：  1． 进行机器学习模型的训练和优化。  2． 根据模型的训练和优化结果，进行贷款人风险模型的误差分析，呈现数据分析和可视化结果。  3． 进行模型和指标的解释和解读，继续查找文献，分析影响因素的重要性。  1月29日-2月4日：  1． 进行风险评估模型的构建，结合借款人的特征和模型预测结果，对风险评估模型进行评估和优化。  2． 进行模型的解释和解读，分析风险因素的重要性。  3． 撰写论文的引言、研究方法和数据分析部分。  2月5日-2月11日：  1． 根据准确率、召回率、精确率等指标，进行实验结果的分析和讨论，验证研究目标的实现程度。  2． 撰写论文的引言、研究方法和数据分析部分。  3． 编写论文的结果和讨论部分。  2月12日-2月18日：  1． 编写论文的结论和展望部分。  2． 进行论文的初稿修改和完善。  3． 检查论文的格式和语法，进行最后的润色。  2月19日-2月25日：  1. 自行进行论文的初次提交终稿的修改和校对。  2. 准备论文的附录和参考文献。  3. 完成论文的最终提交。 |
| 指导教师意见  论文拟基于数据挖掘技术，探索贷款审批决策的影响因素，并建立风险评估模型，研究借款人的特征与其偿还能力之间的关系，为贷款机构制定更准确的风险评估模型提供依据，展，促进金融市场的稳定和健康发展。同时，研究还有助于还款人提前做好还款规划，避免逾期或违约情况的发生，保护个人信用记录。国内外研究现状综述较为充分，研究内容可行，论文结构较为合理，并针对研究内容安排了论文进度。  同意开题！  签名： 祝启虎 2024年1月13日 |
| 院（系）审核意见  **通过**  院（系）公章  2024年1月13日 |