基于数据挖掘的贷款决策影响因素及风险评估

——以Lending Club 网络贷款数据为例

**应敏婕**

*Loan decision influencing factors and risk assessment based on data mining*

*-- Take Lending Club online lending data as an example*

***Aria.ying***

**摘 要**

本文通过数据挖掘技术，对贷款决策的影响因素进行深入研究，并结合风险评估模型，为金融机构提供更准确、有效的贷款决策支持。通过对大量贷款数据的分析和挖掘，对从2007年到2020年实际发放的贷款信息，总计有2925493行数据，142个字段。

总结重要的可用借款人信息，数据的选择和处理上，根据官方网站对客户的的二级评级进行等级分类赋分，同时对借款人数据进行包括空值填充，异常值去除，共线性检测等对观测数据初步处理。结合处理后数据集的数据不平衡特点，分别采用重采样和过采样对样本观测值进行采样评估，最终对向量化的特征进行分类，得到n类风险级别的客户，预测能够全部归还贷款或亏损冲销的客户，并建立了相应的风险评估模型。

实验结果表明，过采样与重采样相比准确率0.88385>0.4814，更适用于该借款数据集，能够有效地处理样本数据集。基于ANN算法和贷款系统性风险概率可以高效地为金融机构提供预警信息。准确率为 ，。。更具预测性的信贷决策模型可以在机构指定的风险承受能力范围内识别更多的客户，为平台带来更高收入。

**关键字：数据挖掘;风险评估；信用评分；人工神经网络；**

**Abstract**

Through data mining technology, the influential factors of loan decision are deeply studied, and the risk assessment model is combined to provide more accurate and effective loan decision support for financial institutions. Through the analysis and mining of a large number of loan data, there are 2,925,493 lines of data and 142 fields for the information of loans actually issued from 2007 to 2020. Summary of important information available to the borrower, the selection and processing of data, according to the official website of secondary ratings of the customer classification assigned points, including a null value for the borrower data populated, outlier removal, collinearity detection to preliminary observation data processing, etc. Combined with the data imbalance characteristics of the processed data set, the sample observations were sampled and evaluated by resamping and oversampling respectively, and finally the vectorized features were classified to obtain customers with N-class risk levels, predict customers who can fully repay loans or write off losses, and establish the corresponding risk assessment model. Compared to the experimental results show that the sampling and resampling, more applicable to the loan data sets, can effectively predict the loan default risk. Based on ANN algorithm and loan systemic risk probability, early warning information can be provided to financial institutions efficiently. A more predictive credit decision model can identify more customers within an institution's specified risk tolerance with an accuracy of... Bring higher revenue to the platform.

**key words:** **Data mining; Risk assessment; Credit score; Artificial neural network;**

目录

[基于数据挖掘的贷款决策影响因素及风险评估 - 1 -](#_Toc160036752)

[——以Lending Club 网络贷款数据为例 - 1 -](#_Toc160036753)

[*Loan decision influencing factors and risk assessment based on data mining* - 1 -](#_Toc160036754)

[*-- Take Lending Club online lending data as an example* - 1 -](#_Toc160036755)

[一、 绪论 - 5 -](#_Toc160036756)

[(一) 引言 - 5 -](#_Toc160036757)

[(二) 国内外综述 - 9 -](#_Toc160036758)

[(三) 研究方法和数据来源 - 11 -](#_Toc160036759)

[二、 相关理论和方法 - 12 -](#_Toc160036760)

[(一) 数据挖掘在贷款决策和风险评估中的应用概述 - 12 -](#_Toc160036761)

[(二) 特征选择方法的介绍 - 12 -](#_Toc160036762)

[(三) 分类算法和聚类分析方法的介绍 - 14 -](#_Toc160036763)

[三、 数据处理与分析 - 14 -](#_Toc160036764)

[(一) 数据收集和预处理 - 14 -](#_Toc160036765)

[(二) 特征选择和构建 - 15 -](#_Toc160036766)

[四、 模型建立和评估 - 17 -](#_Toc160036767)

[(一) 选择合适的分类算法或聚类分析方法 - 17 -](#_Toc160036768)

[(二) 贷款决策模型或风险评估模型构建 - 17 -](#_Toc160036769)

[(三) 模型评估和验证，优化 - 17 -](#_Toc160036770)

[五、 结果分析和讨论 - 17 -](#_Toc160036771)

[(一) 分析模型的结果 - 17 -](#_Toc160036772)

[(二) 讨论模型的有效性和可行性 - 17 -](#_Toc160036773)

[(三) 探讨借款人贷款决策的影响因素和风险评估的准确性 - 17 -](#_Toc160036774)

[六、 研究局限和展望 - 18 -](#_Toc160036775)

[(一) 研究的局限性 （数据指标局限、样本大小局限、时间局限等等） - 18 -](#_Toc160036776)

[(二) 提出未来进一步研究的展望 - 18 -](#_Toc160036777)

[(三) 改进模型、扩大样本规模等建议 - 18 -](#_Toc160036778)

[七、 参考文献 - 19 -](#_Toc160036779)

# 绪论

## 引言

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，可以看出贷款需求和信贷业务需求的活跃程度显著增加。

图表 1 来源：中国人民银行统计数据

大数据热潮下，随着金融科技的快速发展和互联网金融的普及，以网络为媒介的信用贷款业务也越来越受到追捧，点对点网络借贷的贷款方式迅速发展。作为交易便捷，手续简单且审核程序较为宽松的P2P贷款在此环境下快速发展，P2P借贷双方在平台上协商价格，交易完成后平台或中介收取一定的手续费和中介费，借贷人迅速拿到贷款，而投资人拿到高收益。

但与之对应的是这种贷款方式会面临极大的信用风险，由于网贷对个人的信用评分非常依赖，一旦出现诈骗或者失信行为，会给相关行业带来许多不良影响，引发一系列社会问题。以国内主流网络借贷中介为例，团贷网、玖富、PP、点融网等监管真空，非法集资、诈骗等乱象层出不穷，无论是贷款平台的持续运营还是投资者的信誉、流动资金都受到损害。为防止系统性风险的产生，国家已经在通过全额或打折现金退出、商城兑换、属地化债权置换、属地律所法催、全额大礼包等强制力的方式快速推进P2P平台的清偿事项。

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。其中核心的问题在于官方的征信系统中，具有大量贷款需求的借贷方信息未能够完全体现，贷款人在向银行借款时需要繁杂的担保手续。而P2P机构以盈利和业务扩张为导向，作为网络借贷信息中介，对于贷款信息资质的审核不够完善。同时相关法律法规和金融监管仍有空档，疫情结束后我国的经济发展并未完全回到正轨，大量的便捷信用贷款仍然被许多负债的企业和个人所需要。

### 研究目的

中央有关部委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。

在贷款需求和业务不断扩张的背景下，传统个人信用风险评估面临巨大的挑战：交易数量多，金额差异大，贷款信息复杂化多样化等。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。P2P作为当下紧缺的理财工具，如何高效准确识别单笔信贷业务的风险等级，评估贷款人的还款概率显得尤为重要。因此为防止P2P贷款发展为庞氏骗局，保证民间信贷业务的正常运行和持续发展。本文旨在构建一种有效的贷款违约预测模型，通过对新增贷款申请人进行预测是否会违约，帮助金融机构提高贷款决策的准确性和效率。

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，分析借款人的个人特征、借款目的、历史信用记录等因素对贷款决策的影响，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，构建一个可预测借款人贷款风险的评估模型，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。更具预测性的信贷决策模型可以在机构指定的风险承受能力范围内识别更多的客户，为平台带来更高收入。

### 研究意义

风控系统的核心方法要通过深入研究不同特征个人的违约概率,形成大数据风控模块和评分卡系统,将其固化在风控审核的业务过程和决策引擎中,了解不同个人的违约概率程度,从而引导风险审批业务进行。风险评级是平台进行利率分配的基础，充分的风险评估应有助于投资者做出更符合其风险偏好的投资决策，并降低平台客户的流失率。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。

因此个人信用风险研究的意义主要在于通过ANN算法和贷款系统性风险概率研究来丰富现有的个人信用风险研究模型，从以下几个方面为信用贷款各方的风险控制提供参考。

1. **减少道德风险**

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。通过预测借款人是否会违约，金融机构可以及时采取措施，避免放贷给高风险借款人，降低信息不对称造成的逆向选择和道德风险，保护金融机构的资金安全。

1. **完善欺诈检测**

传统的银行贷款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。针对贷款人和贷款申请信息的数据分析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，降低违约风险，提高盈利能力。

1. **控制市场细分**

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。

1. **优化进度流程**

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，获取到更合适的贷款产品，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。

1. **管理客户关系**

根据风险等级和画像特征，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。还可以为金融行业提供参考，减少不良贷款的发生，提高贷款的回收率，增强金融机构的稳健性和可持续性。

此外，丰富和发展数据挖掘在贷款决策领域的应用理论，还可以为金融机构提高贷款审批效率、降低信贷风险提供实践指导，为经济的健康增长提供有力支持。

## 国内外综述

### 国内研究现状

网络借贷行为研究方面。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。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”。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。

样本选择方面，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。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。张远（2018）构建WOE模型和logistic回归模型研究借款人的信用风险，评估中国P2P网络平台鸿菱创业，发现影响P2P借贷平台贷款成功率的主要因素包括贷款金额、年利率、竞价成交金额、按时还款比例等。

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。刘帅祺（2023）以违约风险和误分程度为视角的信贷客户细分，结合误分损失、误分程度等因素构建客户细分类的样本关注度矩阵,提出融合样本关注度矩阵的随机森林提升算法(FOMBRF)，验证了其在信贷风险预测中的有效性。

### 国外研究现状

网络借贷行为研究方面。Mohammad Fahmi Arkanuddin，Bernardus Yuliarto Nugroho等（2022）以分发问卷的形式，用定量方法和SEM-Amos分析工具，研究证明法律监管在大数据分析、自动化和机器人技术等方面的空缺。Shuqi Liang等人（2022）从客户基本属性、还款能力、信贷交易、财产状况、贷款属性、其他因素和时间窗口等方面获取数据，对个人信用进行评分和建模。Martin Španko（2024）分析逻辑回归，判别分析，分类和回归树，随机森林，朴素贝叶斯，K-最近邻居，AdaBoost和XGBoost算法，评估时使用混淆矩阵，ROC/AUC，基尼系数，Kolmogorov-Smirnovova和Brier分数。最终得到XGBoost和AdaBoost最有效，弹性网络逻辑回归排名第三，强调支付行为作为P2P贷款评估中信用风险预测的关键因素。

机器学习风险评估和预测方法选择上，Yuta Tanoue（2019）基于 Boosting 模型和向量回归模型，以日本的五家银行为样本，同时引入非线性因素来建立一个两阶段模型估计违约损失率，比较得出Boosting 模型结果更好；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。Jasmna Nalic等（2020）利用银行客户的数据集，分别建立 NB 朴素贝叶斯、GLM 泛化线性模型、DT 决策树和 SVM 向量自回归四种算法模型来分析银行客户的信用评分，最终证明数据预处理的重要性以及 GLM 性能最优。

### 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，均以欺诈和风险研究为导向，但现实中多方面因素的影响因素难以全面。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，且Mohammad Fahmi Arkanuddin，Bernardus Yuliarto Nugroho等人的研究进一步证明了金融科技发展环境下大数据在预测中的重要性。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”关键词检索结果可知大部分的学者对这类风险的识别与预测研究种P2P市场是目前的研究热点，且经调研了解后发现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。基于大样本的风险预测和研究中，大部分学者重视对模型的精确和评估，对理论特征模型和风险控制模型的构建仍有完善空间。

因此本文综合数据集的已知借款人的基础历史信息与借款信息，针对信贷数据繁杂且包含信息混杂等问题,基于业务理解进行数据理解,提出基于贷款申请信息的属性划分方法,并结合客户细分形成信贷数据对象。基于划分后的数据实现对不同客户细分类、以不同维度信息多角度挖掘客户特征。结合文本类数据与时间序列数据进行风险建模，讨论贷款的风险类别。根据预测结果对客户进行风险评级并以违约风险为视角将信贷客户分为两类,包括:最高违约风险类客户、目标客户、最易误分的非履约类客户、最高潜在风险类客户等,并挖掘了客户细分类间重要风险特征,为信贷客户细分和客户特征的研究提供了新的视角和方法。此外,在数据严重不均衡的信贷数据机器学习任务中,对于易被错分的少数类风险样本的准确识别对进一步模型优化及降低机构损失具有重要意义。(

## 研究方法和数据来源

### 研究方法

本研究采用文献研究法，实证分析法和数据挖掘技术等方法，对贷款决策影响因素及贷款客户风险评估进行深入研究。基于数据量比较大，本研究使用anaconda数据科学环境，jupyter notebook进行编译运行。数据预处理调用pandas库进行数据预处理，特征选择方法使用，，，风险预测算法采用scikit-learn、matplotlib、keras\_tuner。Matplotlib、seaborn进行可视化处理，

pandas库处理后的数据方便进行后续的分析和建模，特征选择方法的选择有助于筛选出对贷款决策和客户风险评估最具影响力的特征，提高模型的预测准确性。scikit-learn提供丰富的机器学习算法和工具，可以灵活建立和评估不同的风险预测模型，并进行参数修改。keras\_tuner可以自动化地搜索最佳的超参数组合，从而提高神经网络模型的性能和泛化能力。Matplotlib和seaborn是常用的数据可视化处理库，可以直观地展现数据的分布和关联关系，更好解释理解数据和模型的结果。

### 数据来源

Lending Club成立于2007年，是曾是美国最早、业务量最大、最为著名的P2P网贷机构，其主要业务是评估借款人的违约风险并设置不同的借款利率，借款人可以通过提交申请快速获得借款。投资人通过浏览借款人的信用记录和借款目的等信息，决定是否借款给不同利率的借款人，以此赚取利息收益。2014年上市时，其市值超90亿美元。

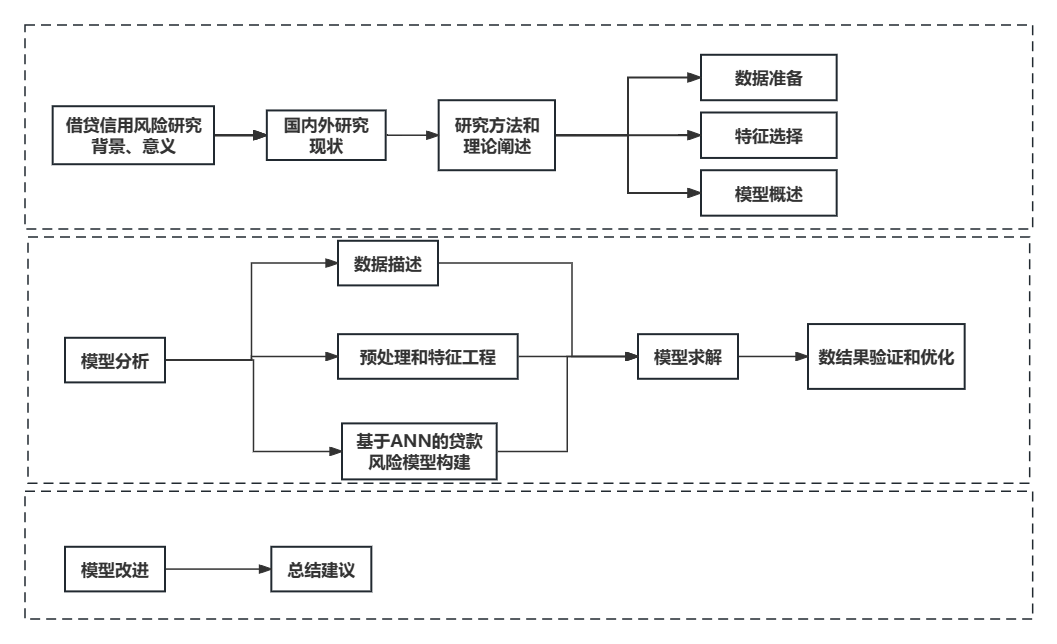
其官方发布的292万美国人（1/10人口）的网络贷款数据集跨时13年，包含年龄，收入，职位，地理位置，贷款目的，信用评级等海量信息。数据量大，变量丰富，是理想的机器学习建模各种算法实验的数据集，非常适用于数据清洗，变量筛选，调参，多算法比较，非平衡数据处理等测试。

### 论文预期结构

Lending Club有基于FICO信用数据的风控模型。将借款人归入A至G共7个等级，每个等级又包含1至5五个子级，共有35个贷款等级。Lending Club会根据借款人的信用报告对每笔借款申请制定不同的借款利率，实行差别定价，等级越高，利率越低。

首先对数据挖掘相关理论和技术进行梳理，分析贷款决策的影响因素，接着构建风险评估模型，通过实证数据验证模型的有效性和准确性，最终得出结论并提出建议。

将从以下几个方面进行详细展开：

1. 不同特征对借款人违约概率的影响；
2. 建立基于ANN机器学习算法的贷款违约预测模型；
3. 评估模型的准确性和稳定性；
4. 为金融机构提供科学、准确的决策支持。

# 相关理论和方法

解释为何选择这些方法，以及它们如何帮助回答研究问题。

## 数据挖掘在贷款决策和风险评估中的应用概述

数据挖掘在贷款决策和风险评估中发挥着至关重要的作用。在传统的贷款决策过程中，金融机构主要依赖借款人的征信、职业、收入等基础信息来评估其还款能力和意愿。然而，这些信息往往无法全面反映借款人的真实风险状况。数据挖掘技术可以从多个维度获取借款人的信息，对借款人的历史交易、行为轨迹等数据进行深入分析信用状况和潜在风险，预测其未来还款能力等信息，在风险出现前做出预警。

在贷款决策方面，数据挖掘技术可以帮助金融机构更准确地识别潜在的优质客户和不良客户。基于大样本的海量数据的分析和挖掘，我们可以发现隐藏在数据中的规律和模式，更准确地预测借款人的还款能力和意愿。这有助于金融机构制定更合理的贷款策略，提高贷款审批的准确性和效率。

在风险评估方面，数据挖掘技术可以应用于信用风险评估和诈骗检测等方面。通过对客户的信用历史、财务状况和交易记录等数据的深入分析，数据挖掘技术可以帮助金融机构更准确地评估客户的信用风险，降低坏账率。同时，数据挖掘技术还可以帮助金融机构检测异常交易和不寻常的模式，及时发现潜在的诈骗行为，保障金融机构的资金安全。

在算法模型建立的选择上，目前的研究主要选择XGBoost机器学习算法，BP神经网络，随机森林算法等，。

## 特征选择方法的介绍

特征选择是机器学习中的一个重要步骤，它涉及到从原始特征集中选择出最相关、最有代表性的特征子集，同时去除冗余和不相关的特征。以简化模型、提高模型的泛化能力，并减少计算开销，并提高模型的可解释性。

特征选择方法有很多种，常用的方法有：

1. 过滤器方法

过滤器方法是基于统计度量（如信息增益或卡方检验）评估特征的重要性

信息增益：衡量特征与目标变量之间的相关性，信息增益越高的特征越相关。

卡方检验：衡量特征值分布与目标变量值分布之间的差异，卡方检验值越高的特征越相关。

互信息：衡量两个特征之间的相关性，互信息越高的特征越相关。

相关性分析：计算特征与目标变量之间的皮尔逊相关系数或斯皮尔曼相关系数，相关系数越高的特征越相关。

1. 包装器方法

包装器方法：使用机器学习模型评估特征子集，并选择性能最佳的子集。

向前选择：从空特征子集开始，逐步添加最相关的特征，直到达到停止标准。

向后选择：从包含所有特征的子集开始，逐步删除最不相关的特征，直到达到停止标准。

递归特征消除（RFE）：使用线性模型（如逻辑回归）对特征进行排序，并逐步删除排名最低的特征。

1. 嵌入式方法

嵌入式方法：将特征选择过程集成到机器学习模型的训练过程中。

L1 正则化（LASSO）：对模型权重施加 L1 范数惩罚，导致不相关的特征的权重为零。

L2 正则化（岭回归）：对模型权重施加 L2 范数惩罚，导致不相关的特征的权重较小。

决策树：决策树算法在构建树的过程中自动执行特征选择。

1. 基于模型的方法

使用特定的机器学习模型进行特征选择，如随机森林、支持向量机等。

这些模型可以评估特征的重要性，并选择最重要的特征。

通常能得到较好的结果，但可能受到模型选择的影响。

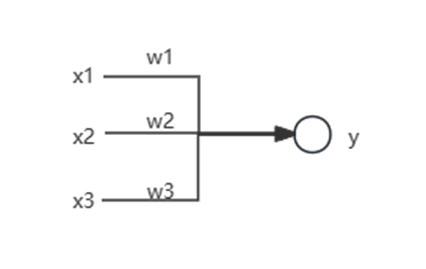
1. 基于相关性的方法

通过计算特征与目标变量之间的相关性来选择特征。

常用的相关性度量有皮尔逊相关系数、斯皮尔曼秩相关系数等。

这种方法简单直观，但可能忽略了特征之间的非线性关系。

## 分类预测算法介绍

本文用到的算法是人工神经网路（ANN算法），通过模拟大脑生物神经网络的突触节点，实现对数据之间的复杂关系进行建模。神经元是大脑的基本处理单元，人工神经网络通过模拟生物神经系统的组织结构实现数据的处理和转换。神经元也称“激活单元”，具体结构如下，一般具有多个输入值和其权重和一个恒为1的截距项，偏执单元，一个输出值y，如下图所示：

，i=1，2，3，…为该层神经元的输出，为权重。

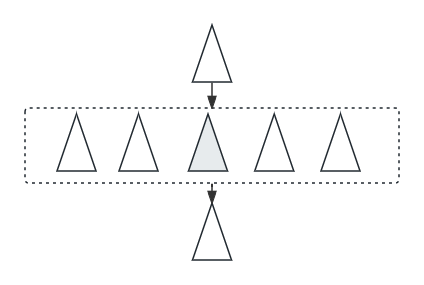
一般来说神经元可能有多层，第n层神经元接收上层传入的神经冲动刺激：

该激活函数g作用后，产生激活向量,即第n层第i个神经元获得的激活值：

即输出值可以表示为：

人工神经网络的特点在于极强的学习能力，得到训练集之后，能够通过学习提取观测到的各个部分特征，以网络节点连接，通过权重改变链接的强度，只到顶层的输出得到答案。在本研究中，ANN神经网络算法能够更真实直观地反映信用指标与信用的映射关系评级，通过合并邻近的区间统计测试集，重新进行划分间隔与每个属性值，并创建更大的间隔。该模型还有极强的自我学习能力，更好地拟合因变量和自变量之间的非线性关系，解释能力更贴近真实情况，比其他方法有更强的针对性。但其中所使用的参数非常难以确定，非常受初始值的影响，输入层初始值、隐藏层节点数、输出层层数的初始值确定后可能出现最终局部极小值、数据过拟合的现象。

随机森林是一种拓展的并行式集成学习方法，在训练基学习器时，也采用有放回采样的方式添加样本扰动，同时它还引入了一种属性扰动：在基决策树的训练过程中，在选择划分属性时，Random Forest先从候选属性集中随机挑选出一个包含K个属性的子集，再从这个子集中选择最优划分属性。

特点是每个子决策树都随机选择部分样本和特征，具有较强的抗噪声能力，性能稳定。创建多颗决策树，用它们分别进行预测，再根据少数服从多数的原则从多个预测结果中选择最终预测结果。如图所示为随机森林的工作原理。

# 数据处理与分析

## 数据收集和预处理

### 数据准备

* 1. 数据来源

本文采用的数据集包含了从2007年到2020年实际发放的贷款信息，总计有2925493行数据，142个字段。这些字段包括了借款人的基本信息（如年龄、性别、工作、工资收入等），借款详情（如借款金额、借款期限、借款利率、借款用途等），以及借款的还款状态等。

文本的研究目标是贷前信用风险评估，为了使数据更具有代表性，只考虑全额结清和未按时还款的的情况，常识性剔除贷款申请和发放无关特征，并剔除部分贷款发放后特征，最终得到个样本观测值。研究中所使用的数据集来源于Lending Club官方网站。此外，为了克服极端值的影响，对所有连续变量进行了去除离群值的处理。

* 1. 数据预处理

从该数据集获得的可用信息与研究出发，首先基于常识删除与贷款因素无关的列，并删除缺失值大于50%的特征，与贷款违约风险有关的数据列loan\_status。

* Fully Paid: The loan has been fully paid off by the borrower.
* Charged Off: The loan has not been fully repaid and Lending Club has charged off the remaining balance as a loss.
* Current: The loan is currently being repaid on schedule.
* Default: The borrower has failed to make payments on the loan and the loan is in default.
* Late (31-120 days): The borrower has missed payments and is between 31 and 120 days late on their payment schedule.
* In Grace Period: The borrower is in a grace period and has missed a payment.
* Late (16-30 days): The borrower has missed a payment and is between 16 and 30 days late on their payment schedule.

根据官方网站的解释，full Paid解释为借款人已全部付清贷款。Charge off解释为贷款未全部偿还，剩余余额作为亏损冲销。其他类型在所有数据列的占比不足1%，因此舍弃其它类型的贷款进行分析。最终获得2925493条，93列数据。对初始数据集本研究将数据集的可用数据分为分类特征和数值特征。

除此外 本文对主要变量进行相关度分析，结果显示解释变量、控制变量的相关性系数均小于0.5，且方差膨胀银子（VIF）均小于2，表明不存在较严重的多重共线性问题，变量选取合理。

变量选择与测度方法

|  |  |  |  |
| --- | --- | --- | --- |
| 变量 | 变量名 | 含义 | 数据类型 |
| 因变量 | loan\_status | 借款当前状态 | String |
| 贷款信息 | loan\_amnt | 贷款金额 | Float |
| term | 贷款期限(36个月或60个月) |  |
| int\_rate | 贷款利率 |  |
| installment | 按月支付贷款 |  |
| 借款人资料 | grade | 根据借款人的信誉给出的等级 |  |
| sub\_grade | 根据借款人的信誉给出的更详细的等级 |  |
| emp\_length | 借款人的工作年限 |  |
| home\_ownership | 借款人的房屋所有权类型 |  |
| annual\_inc | 借款人的年收入 |  |
| verification\_status | 收入是否经过LendingClub的核实 |  |
| 贷款详情¶ | purpose | 贷款的目的 |  |
| dti | 借款人的债务收入比 |  |
| delinq\_2yrs | 过去两年的拖欠次数 |  |
| inq\_last\_6mths | 过去6个月的查询次数 |  |
| open\_acc | 借款人信用档案中开放信用额度的数量 |  |
| pub\_rec | 贬损的公共记录的数量 |  |
| revol\_bal | 总信贷循环余额 |  |
| revol\_util | 借款人使用的信贷金额相对于其总信贷限额 |  |
| total\_acc | 借款人信用档案中的信用额度总数 |  |
| last\_pymnt\_amnt | 上次收到的付款金额 |  |

### 数据的集成和变换处理

1. 等级赋分

Lending club官方将人群特征画像贷款分为A、B、C、D、E、F、G 7 个等级，每个等级又包含了1、2、3、4、5 五个子级。为更好区分等级分数，这里将官方评定的登记进行赋分处理：

Grade:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| A | B | C | D | E | F | G |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 |

Subgrade:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A1 | A2 | A3 | A4 | A5 |
| 0 | 1 | 2 | 3 | 4 |
| B1 | B2 | B3 | B4 | B5 |
| 5 | 6 | 7 | 8 | 9 |
| … | … | … | … | … |
| G1 | G2 | G3 | G4 | G5 |
| 30 | 31 | 32 | 33 | 34 |
|  |  |  |  |  |

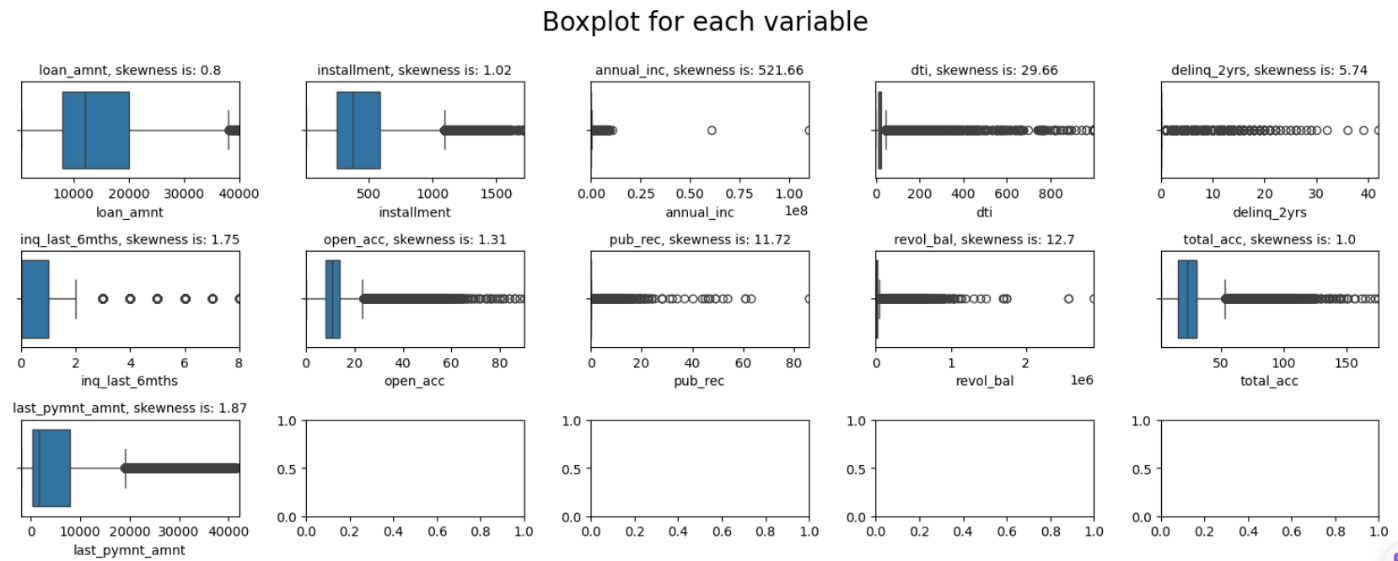
1. 独热编码与采样

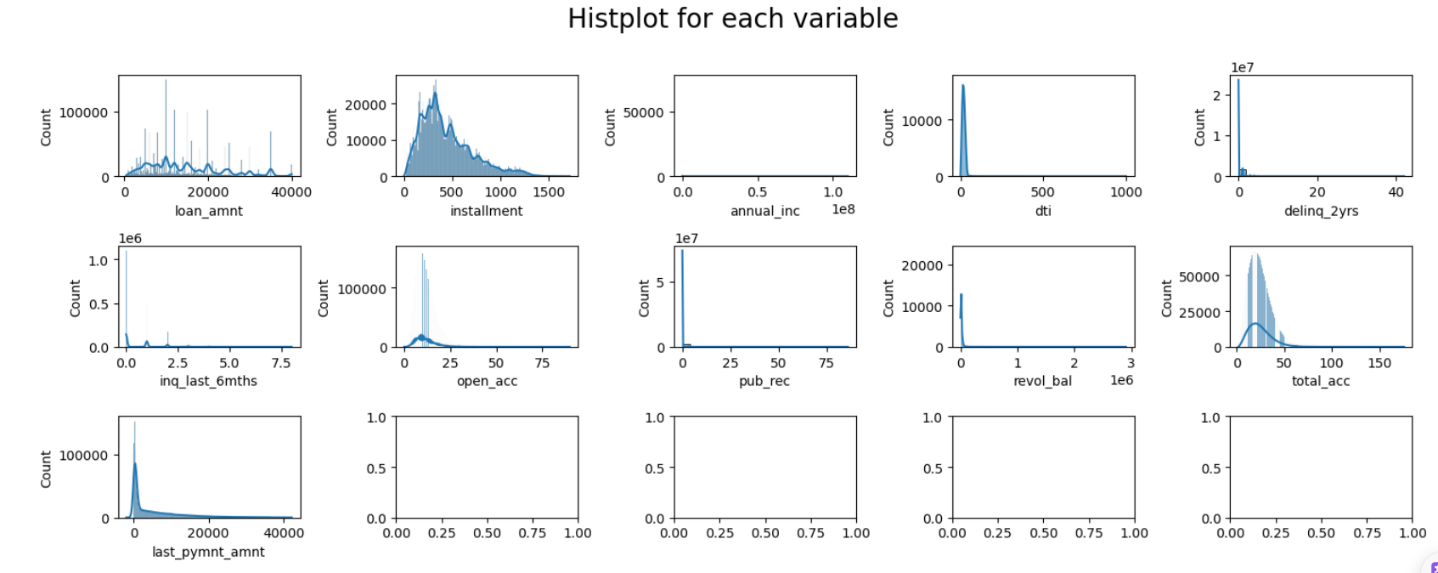
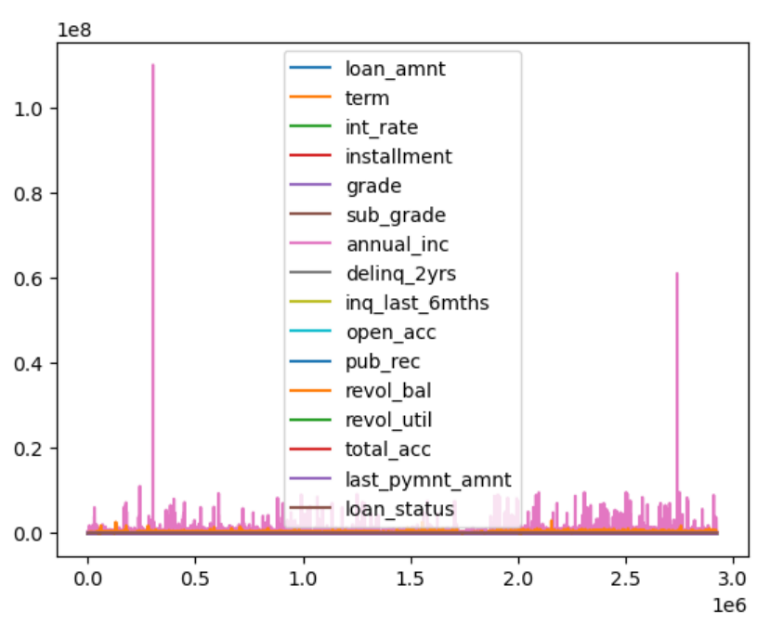
'home\_ownership','verification\_status','purpose'列进行独热编码处理，转化为二进制变量。

根据转换后得到的直方图和箱线图可知存在不平衡数据，根据数据列的分布情况对部分列进行均值填充。并通过尝试重采样和过采样方法平衡数据集的类别分布。采用最近邻策略对每个少数类样本，找到一定数量的与这些样本距离最近的样本，从多数类中随机选择一些样本进行删除，减少类别之间的不平衡。

|  |  |
| --- | --- |
| 方法 | 数据 |
| 重采样 | [[155705 218539]  [ 17997 72842]]  Accuracy Score:  0.4914112104721093  Classification Report:  precision recall f1-score support  0 0.90 0.42 0.57 374244  1 0.25 0.80 0.38 90839  accuracy 0.49 465083  macro avg 0.57 0.61 0.47 465083  weighted avg 0.77 0.49 0.53 465083 |
| 过采样 | [[345449 28795]  [ 25223 65616]]  Accuracy Score:  0.8838529896814117  Classification Report:  precision recall f1-score support  0 0.93 0.92 0.93 374244  1 0.70 0.72 0.71 90839  accuracy 0.88 465083  macro avg 0.81 0.82 0.82 465083  weighted avg 0.89 0.88 0.88 465083 |

总的来看过采样准确率0.88385>0.4814,且精确率，召回率F1分数综合表现由于重采样，因此我们选择重采样后的数据进行进一步分析。

1. 归一化处理



图表 2 变量示意图

如图所示数据集的变量差异较大，需要进一步的处理消除量纲，帮助后续算法更快收敛，提高泛化能力，

Xstd​=X.max(axis=0)−X.min(axis=0)X−X.min(axis=0)

​

Xscaled​=Xstd​∗(max−min)+min

Xstd为标准化结果，Xscaled为归一化结果。

X.min(axis=0)：每列中的最小值组成的行向量，X.max(axis=0)：每列中的最大值组成的行向量，max: 要映射到的区间最大值，默认是1，min：要映射到的区间最小值，默认是0。划分标准测试集为25%，便于后续模型训练。

## 特征选择和构建

### 特征选择方法的选择和介绍

针对初步处理后的数据列，将它们分为类别列和数值列：

|  |  |
| --- | --- |
|  |  |
| 类别列 | 'term', 'int\_rate', 'grade', 'sub\_grade', 'emp\_length', 'home\_ownership', 'verification\_status', 'purpose', 'revol\_util', 'loan\_status' |
| 数值列 | 'loan\_amnt', 'installment', 'annual\_inc', 'dti', 'delinq\_2yrs', 'inq\_last\_6mths', 'open\_acc', 'pub\_rec', 'revol\_bal', 'total\_acc', 'last\_pymnt\_amnt' |

因变量不同类型的占比（保留至小数点后两位）：

|  |  |
| --- | --- |
| 类别 | 比例 |
| Fully Paid | 80.51% |
| Charge Off | 19.49% |

对于本文来说有这些方法要用到：，同时，特征选择也需要与其他机器学习步骤（如模型选择、参数调整等）相结合，以得到最佳的模型性能。

优点

提高模型性能

减少训练时间

提高模型可解释性

避免过拟合

### 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

特征选择和构建是使用特征选择方法，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

根据类别列和数值列的归类，又可将选择的特征总结为三个方面的因素：

个人信用评估因素：这包括借款人的信用历史、债务水平、还款能力和财务状况等。文献可以探讨不同的信用评估模型和算法，并分析它们对借款人信用风险评估的影响。

社会经济因素：这包括借款人的职业、教育水平、收入水平和工作经验等。文献可以研究这些因素在借款人决策和风险评估中的作用，并分析它们与借款人还款能力之间的关系。

贷款本身因素：包括借款目的，借款前六个月的查询次数，总信贷循环余额，借款人使用的信贷金额相对于其总信贷限额，借款人信用档案中的信用额度总数，上次收到的付款金额等。

### 构建特征向量

# 模型建立和评估

## 贷款决策模型或风险评估模型构建

使用具有自定义超参数的Keras创建多层感知机模型。

//画图和理论数学推导

定义模型：

model = keras.Sequential() 创建模型， Keras构建神经网络。

添加隐藏层：

for i in range(hp.Int('num\_layers', 2, 20)):遍历一个指定范围内的层数，从 2 到 20 层。

model.add(layers.Dense(...)) 添加了一个全连接层（Dense层），其中包含 units\_ + str(i) 个神经元，这个数量在 32 到 512 之间，步长为 32。

activation使用 ReLU 激活函数，这是深度学习中常用的激活函数，因为它有助于缓解梯度消失问题。

3.

添加输出层：

model.add(layers.Dense(1, activation='sigmoid')) 添加了一个输出层，只有一个神经元，使用 sigmoid 激活函数。这通常用于二分类问题。

4.

编译模型：

编译模型，设置优化器、损失函数和评估指标。

设置优化器为 Adam，常用的优化算法，结合 RMSprop 和 Momentum 的优点。

learning\_rate 设置为 1e-2, 1e-3, 或 1e-4 中的一个，优化器调整权重的速率。

loss='binary\_crossentropy' 设置损失函数为二元交叉熵，这是二分类问题的常用损失函数。

metrics=['accuracy'] 设置评估指标为准确率。

## 模型评估和验证，优化

# 结果分析和讨论

## 分析模型的结果

val\_accuracy: 0.8219091892242432

Best val\_accuracy So Far: 0.8219091892242432

## 讨论模型的有效性和可行性

为了验证模型的有效性，我们使用了大量的真实贷款数据进行实验。实验结果表明，该模型在预测贷款违约风险方面具有较高的准确性和稳定性。与传统的风险评估方法相比，该模型能够更好地捕捉到隐藏在数据中的复杂关系，为金融机构提供更全面、准确的风险评估结果。

## 探讨借款人贷款决策的影响因素和风险评估的准确性

根据网站贷款记录提取的特征数据，可以概括出人群的特征画像：

人口特征：教育程度、收入水平、职业等。

财务特征：贷款金额、贷款期限、还款记录、信用评分等。

行为特征：贷款申请频率、贷款用途、还款方式等。

交互特征：浏览时间、查询行为等。

风险级别

基于特征画像，可以评估人群的风险级别。风险级别通常分为几个等级：

# 研究局限和展望

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。

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。

## 研究的局限性 （数据指标局限、样本大小局限、时间局限等等）

## 提出未来进一步研究的展望

创新点 包括视角创新，方法创新以及实践应用的创新。

## 改进模型、扩大样本规模等建议

未来,第三方消费平台付款的方向、额度、购物品牌及金额等都可以作为个人信用评估的重要数据参考,生活类信息服务网站中有关个人用户的日常生活消费信息如水费、电费、煤气费、物业费、电话费、上网费等都有可能反映用户的生活基础信息,为个人信用评估提供重要的依据,只有掌握所有的大数据分析信息才可以实现大数据分析的有效加工。

然而，由于数据的限制和模型的局限性，仍然存在一些问题需要进一步研究和改进。未来的研究可以进一步扩展数据集，引入更多的影响因素，提高模型的准确性和稳定性。

# 参考文献

<https://sc.panda321.com/> 文献规范化引用查询

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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
15. 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
16. 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.
17. HerzensteinMS.SonensheinUMDhola-kia.TellmeaGoodStoryandIMayLendYouMoney:TheRoleofNarrativesinPeer-to-peerLendingDecisions[J].JournalofMarketingResearch,2011,48(SPL):138-149.
18. LinMNRPrabhala,SViswanathan.JudgingBorrowersbytheCompanyTheyKeep:SocialNetworksandAdverseSelectioninOnlinePeer-to-peerLending[J].SSRNeLibrary,2009.
19. CollierB,RHampshire.SendingMixedSignals:MultilevelReputationEffectsinPeer-to-PeerLendingMarkets[M].CSCW,Savannah,USA.2010.