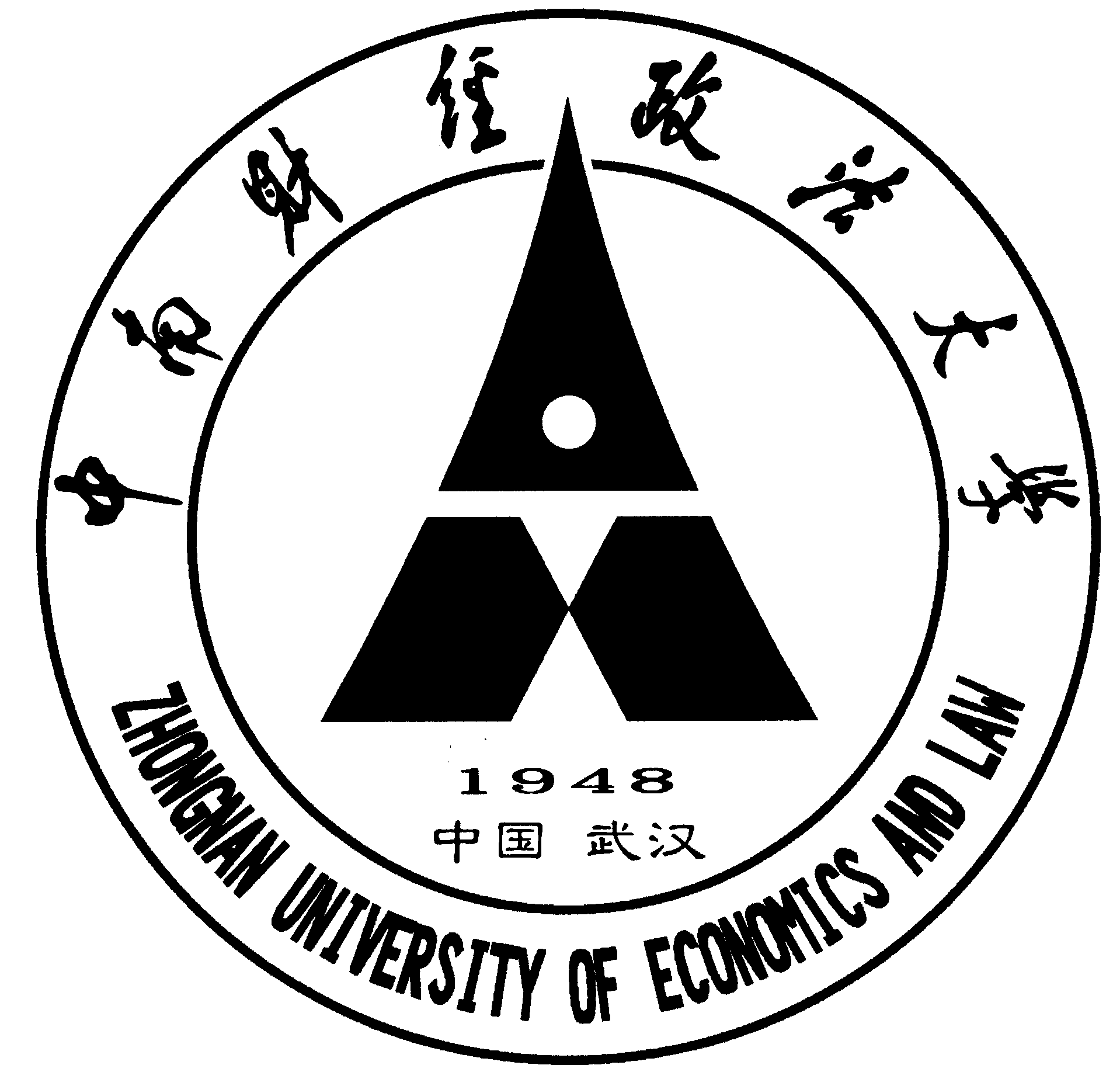


**本科毕业论文（设计）**



**论文题目：基于数据挖掘的贷款决策风险评估**

**——以Lending Club 网络贷款数据为例**

**姓 名： 应敏婕**

**学 号： 202021130214**

**班 级： 大数据管理2001班**

**专 业： 大数据管理与应用**

**学 院： 信息工程学院**

**指导教师： 祝启虎**

**完成时间： 2023年4月29日**

**作者声明**

本毕业论文（设计）是在导师的指导下由本人独立撰写完成的，没有剽窃、抄袭、造假等违反道德、学术规范和其他侵权行为。对本论文（设计）的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。因本毕业论文（设计）引起的法律结果完全由本人承担。

特此声明。

|  |  |  |
| --- | --- | --- |
| 作者专业 | ： | **大数据管理与应用** |
| 作者学号 | ： | **202021130214** |
| 作者签名 | ： |  |
|  | **年 月 日** | |

基于数据挖掘的贷款决策风险评估

——以Lending Club 网络贷款数据为例

**应敏婕**

*Risk Assessment of Loan decision based on Data Mining*

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

***Aria, ying***

**摘 要**

本文针对P2P网贷市场中的信贷风险预测问题进行了深入研究。目前P2P网贷研究主要集中在运营模式、交易行为、借贷风险和社会网络等方面，但现实中多方面因素的影响难以全面考虑。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，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。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，通过数据预处理技术，包括空值和重复行的删除、离群异常值的去除、单位和符号的去除，以及数据归一化处理，来确保数据集的质量和一致性。在剔除无关变量后，通过特征工程，包括特征衍生、特征抽象和特征缩放，对数据进行了进一步的优化，以提高模型的预测能力，初步处理得到18个有效列。

在特征衍生阶段，研究者利用主成分降维技术对六个高共线性特征进行降维处理，将六个特征简化为三个特征，以减少模型复杂度并提高预测精度。在特征抽象阶段，研究者采用了分类和独热编码处理对文本数据进行量化，并对客户的等级数据进行了数值映射，以便更好地适应机器学习模型。

结合处理后数据集的数据不平衡特点，分别采用随机欠采样和随机过采样对样本观测值进行采样评估，并通过实验验证了过采样技术在提高模型性能方面的有效性。使用Keras调谐器进行超参数调优得到最佳超参数，最终建立ANN网络模型对向量化的特征进行分类，并建立了相应的风险评估和预测模型，预测能够全部归还贷款，延期归还，欠款的客户，准确率达到84.52%。与逻辑回归模型和xgboost模型相比，ANN模型在预测效果上具有明显优势。

实验结果表明，随机欠采样和随机过采样相比准确率88%>13%，增加少部分样本的随机过采样在该不均衡大规模数据集更适用，能够有效地处理样本数据集。使用Keras调谐器进行超参数调优，构建了一个具有4到20个隐藏层和32到512个神经元的深度学习模型，以预测三分类问题。模型的输出层使用Softmax激活函数，并采用多类交叉熵作为损失函数。经过多次实验，模型的准确率达到84.52%，显示出较高的分类准确性。与传统的风险评估方法相比，该模型基于大数据集，能够捕捉数据中的复杂关系，为金融机构提供风险评估参考。模型的超参数调优灵活性、多分类处理能力、交叉熵损失函数的应用以及多性能指标评估，使其在实际应用中具有良好的可调性和适用价值。特征分析表明，贷款期限、信用账户数量、贷款利率等特征对于区分不同贷款客户类别至关重要。对于低风险的Paid类型客户，建议降低贷款门槛和简化申请流程；对于高风险的Late和Default类型客户，则需要加强监测、调整还款计划和贷后管理。为金融机构提供风险评估结果参考，为平台减小一定风险。

**关键字：**数据挖掘；贷款；风险评估；ANN人工神经网络；

**Abstract**

This paper focuses on the credit risk prediction in P2P lending market. At present, the research of P2P lending mainly focuses on the operation mode, transaction behavior, lending risk and social network. However, it is difficult to fully consider the influence of many factors in reality. Through the analysis and mining of a large number of loan data, there are a total of 2,925,493 lines of data for the actual loan information issued from 2007 to 2020. 142 fields summarize important available borrower information, and the basic historical information and loan information of known borrowers are integrated in the data set. Based on business understanding, this paper proposed an attribute division method based on loan application information, and combined with customer segmentation to form credit data objects. In the selection and processing of data, according to the official website, the secondary rating of customers is classified and assigned, and the quality and consistency of the data set are ensured by data preprocessing techniques, including the deletion of null values and duplicate rows, the removal of outliers, the removal of units and symbols, and data normalization. After eliminating irrelevant variables, the data were further optimized by feature engineering, including feature derivation, feature abstraction, and feature scaling, to improve the predictive power of the model, and 18 valid columns were obtained by preliminary processing.

In the feature derivation stage, the principal component dimension reduction technology is used to reduce the dimension of the six high collinearity features, and the six features are simplified into three features to reduce the complexity of the model and improve the prediction accuracy. In the feature abstraction stage, the researchers used classification and one-hot encoding processing to quantify the text data, and mapped the customer's rating data numerically, so as to better adapt to the machine learning model.

Combined with the data imbalance characteristics of the processed data set, random undersampling and random oversampling are used to sample and evaluate the sample observations respectively, and the effectiveness of oversampling technology in improving the performance of the model is verified by experiments. The Keras tuner was used to optimize the hyperparameters to obtain the best hyperparameters. Finally, an ANN network model was established to classify the vectorized features, and the corresponding risk assessment and prediction model was established. Compared with logistic regression model and xgboost model, ANN model has obvious advantages in prediction effect.

The experimental results show that the accuracy of random under-sampling and random over-sampling is 88%>13%, and the random over-sampling with a small number of samples is more suitable for the imbalanced large-scale data set, which can effectively process the sample data set. Hyperparameter tuning using Keras tuner, a deep learning model with 4 to 20 hidden layers and 32 to 512 neurons is constructed to predict three-class classification problems. The output layer of the model uses Softmax activation function and adopts multi-class cross-entropy as the loss function. After many experiments, the accuracy of the model reaches 84.52%, which shows high classification accuracy. Compared with traditional risk assessment methods, the model is based on large data sets, which can capture complex relationships in the data and provide risk assessment reference for financial institutions. The flexibility of hyperparameter tuning, multi-classification processing ability, the application of cross-entropy loss function and multiple performance index evaluation of the model make it have good adjustability and applicable value in practical applications. Feature analysis shows that features such as loan duration, number of credit accounts, and loan interest rate are crucial for distinguishing different loan customer categories. For low-risk Paid customers, it is recommended to lower the threshold of loan and simplify the application process. For high risk Late and Default type customers, it is necessary to strengthen the monitoring, adjustment of repayment plan and post-loan management. It provides a reference for the risk assessment results of financial institutions and reduces a certain risk for the platform.

**key words:** Data mining; Loan; Risk assessment; ANN artificial neural network;

**目 录**

[一、 绪 论 1](#_Toc163808143)

[(一) 研究背景及意义 1](#_Toc163808144)

[(二) 国内外综述 4](#_Toc163808145)

[(三) 研究方法和思路 7](#_Toc163808146)

[二、 相关理论和方法 10](#_Toc163808147)

[(一) 数据挖掘在贷款决策和风险评估中的应用概述 10](#_Toc163808148)

[(二) 分类预测算法介绍 10](#_Toc163808149)

[三、 数据收集与处理 15](#_Toc163808150)

[(一) 数据收集和预处理 15](#_Toc163808151)

[(二) 变量选取和特征工程 19](#_Toc163808152)

[四、 模型建立和评估 24](#_Toc163808153)

[(一) 预测模型构建 24](#_Toc163808154)

[(二) 结果分析和讨论 24](#_Toc163808155)

[五、 结 语 28](#_Toc163808156)

[(一) 研究结论 28](#_Toc163808157)

[(二) 研究展望 28](#_Toc163808158)

[主要参考文献 1](#_Toc163808159)

[附 录 3](#_Toc163808160)

[附录一 3](#_Toc163808161)

[后 记 4](#_Toc163808162)

# 绪 论

## 研究背景及意义

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，图表1金融机构人民币信贷收支表（部分）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 1 金融机构人民币信贷收支表（部分）**

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[[1]](#footnote-1)，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。

### 研究背景

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[[2]](#footnote-2)，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[[3]](#footnote-3)。

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贷款[[4]](#footnote-4)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，为平台获得更高收入。

### 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[[5]](#footnote-5)。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科学的机器学习或概率统计方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，以现有的历史数据作为参考对参加借款业务的客户进行标签化处理和风险性类别预测。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。

丰富和发展数据挖掘在贷款决策领域的应用理论，在已有贷款客户数据研究背景下挖掘机器学习方法对客户信息的预测，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和方法指导，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。翟金阳基于多元犯罪机会理论框架,分别从平台的高管特征,监管环境和运营业绩层面对网络贷款欺诈行为进行分析研究[[6]](#footnote-6)。明依东考虑投资者投资偏好和贷款的及时性，提出了基于神经网络的潜在因子交互作用模型[[7]](#footnote-7)。陈健巧基于技术接受模型和感知风险理论,构建出影响P2P投资者借贷行为的模型，并设计五级量表的问卷进行网络调查[[8]](#footnote-8)。刘雨杭引入投资者情绪，通过股票市场投资者情绪和网络借贷平台交易数据的关联性，证明股市投资者情绪变量仍然会对网络借贷市场交易具有统计上显著的解释力[[9]](#footnote-9)。

样本选择方面，戴杨文意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[[10]](#footnote-10)。沈永健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[[11]](#footnote-11)。张丽改选取“拍拍贷”2015年至2017年交易数据作为原始样本，使用Logistic回归分析对借款人的借款信息进行量化研究,稳健标准误的OLS模型对信用等级对投资者决策进行实证研究[[12]](#footnote-12)。

机器学习风险评估和预测方法选择上，姜晨利用随机森林模型确定特征重要性,发现历史成功借款金额的重要程度最高，总待还本金,借款金额与历史正常还款期数、借款次数等历史交易数据可以显著影响其违约风险，遗传算法优化后的BP神经网络模型预测误差仅20.88%[[13]](#footnote-13)。雷佳使用了k-近邻方法、决策树方法、神经网络方法、Logistic回归方法、支持向量机方法对个人信贷风险等级进行了综合分类预测,其中使用Logistic回归和支持向量机的实验结果(测试集准确率) 实现较高准确率的预测判别，分别达到97%和98%[[14]](#footnote-14)。吴利娟，孟泽远通过因子分析法构建风险评价指标体系和评价模型，研究表明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[[15]](#footnote-15)。张宇幸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[[16]](#footnote-16)。刘帅祺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[[17]](#footnote-17)。

### 国外研究现状

网络借贷行为研究方面。Mohammad Fahmi Arkanuddin，Bernardus Yuliarto Nugroho等以分发问卷的形式，分析画布商业模式（CBM）对风险的影响，同时引入COVID-19大流行风险，研究证明不良贷款风险可以通过实施CBM组件和客户细分来减轻[[18]](#footnote-18)。Garrett T. Senney & Jonathan R. Lhost利用在线P2P贷款拍卖数据，发现最优保留价在借款人质量（信用评分和债务收入比）上呈非线性变化，导致借款人支付更高的利率[[19]](#footnote-19)。Štefan Lyócsa，Petra Vašaničová对美国P2P借贷平台Lending Club的 472,106 笔贷款进行估计的分位数建立回归盈利评分模型，研究表明较高的贷款金额、贷款期限、利率和较低的收入与信用较差的借款人的较低回报有关[[20]](#footnote-20)。

样本选择方面。Ruqoyah Ruqoyah，Yenti Afriani等通过问卷调查获得的一手数据以及其他来源的二手数据，采用五个值的李克特量表进行分析数据，结果显示高利率和数据安全问题是社区用户的主要关注点[[21]](#footnote-21)。Martin Španko基于Zonky平台数据，分析了逻辑回归、判别分析、分类和回归树、随机森林、朴素贝叶斯、k近邻、AdaBoost和XGBoost等各类算法的有效性[[22]](#footnote-22)。Garrett T. Senney & Jonathan R. Lhost[[23]](#footnote-23)利用在线P2P贷款拍卖数据，Štefan Lyócsa，Petra Vašaničová[[24]](#footnote-24)对美国P2P借贷平台Lending Club进行大规模数据集的计算和预测。

机器学习风险评估和预测方法选择上，Bequé等人设计了一个用于贷款信用风险评估的深度神经网络，对传统神经网络的收敛方法进行调整。通过仿真实验，表明该网络所具有的参数少并且精确度高等优点[[25]](#footnote-25)。Martin Španko基于Zonky平台数据，使用混淆矩阵、ROC/AUC、基尼系数、Kolmogorov-Smirnov统计量和Brier Score等指标进行评价，分析了逻辑回归、判别分析、分类和回归树、随机森林、朴素贝叶斯、k近邻、AdaBoost和XGBoost。结果表明，XGBoost和AdaBoost是最有效的，并强调支付行为作为P2P贷款评估中信用风险预测的关键因素[[26]](#footnote-26)。

### 研究现状述评

根据研究[[27]](#footnote-27)，目前的P2P网贷研究集中在运营模式、交易行为、借贷风险和社会网络等四个方面。结合上述参考文献，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，有基于互联网背景的贷款人大样本分析的趋势，其中回归模型和神经网络是常用的计算模型，但基于不同视角进行的大数据挖掘类客户风险类别预测较少，且Mohammad Fahmi Arkanuddin，Bernardus Yuliarto Nugroho等人[[28]](#footnote-28)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建立层面,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空白。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中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，样本庞大易获取的特点更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三类,包括: 全额结清,拖欠、超时。并挖掘了客户细分类间的重要风险特征,为信贷客户细分和客户特征的研究提供了视角和方法的参考。另外在数据严重不均衡的信贷数据集机器学习任务中,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。

## 研究方法和思路

### 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，numpy库进行数据预处理，预测算法采用scikit-learn、keras\_tuner库进行超参数调优选择和模型建立求解。Matplotlib、seaborn进行可视化处理分析。

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，便于直观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。

### 论文思路

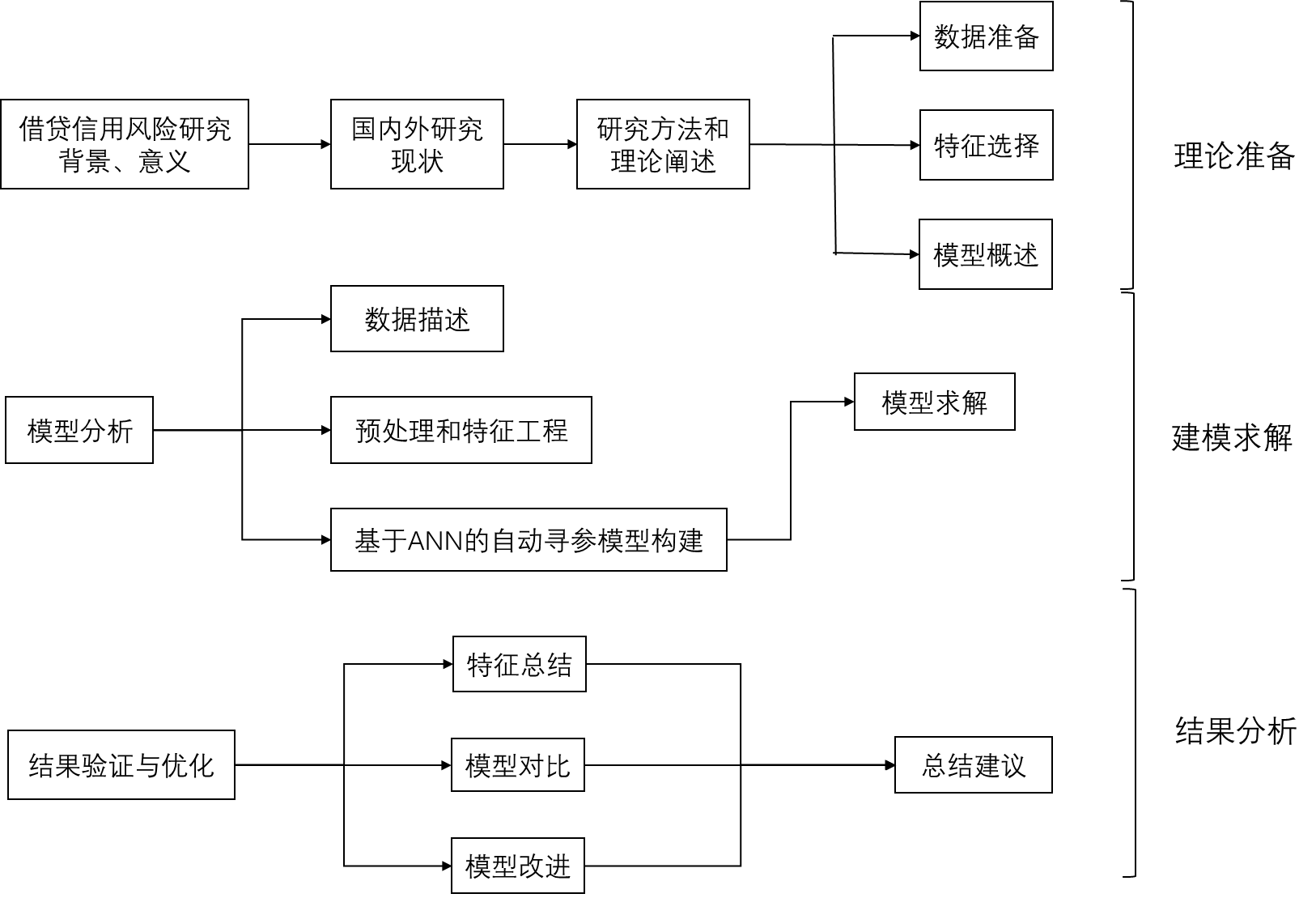
Lending Club有基于FICO信用数据的风控模型。将借款人归入A至G共7个等级，每个等级又包含1至5五个子级，共有35个贷款等级。平台会根据借款人的信用报告对每笔借款申请制定不同的借款利率，实行差别定价，等级越高，利率越低。

本研究将首先对数据挖掘相关理论和技术进行梳理针对P2P网贷市场中的信贷风险预测问题进行深入研究。目前P2P网贷研究主要集中在运营模式、交易行为、借贷风险和社会网络等方面，但现实中多方面因素的影响难以全面考虑。在数据选取上，传统问卷与业务材料收集信息进行分析的趋势逐渐转向基于互联网背景的贷款人大样本分析，回归模型和神经网络是常用的计算模型，因此本文将通过对不同特征的分析，讨论哪些特征是贷款客户的重要预测特征，并基于这些特征总结出贷款风险客户画像。

在构建风险分类模型方面，本文将采用先进的人工神经网络（ANN）机器学习算法，考虑超参数调优、多分类处理、损失函数的应用、指标的评估、特征和样本平衡的调整性。且该模型将基于对大量贷款数据的分析和对关键特征的识别，对贷款违约进行预测。模型将考虑包括借款人的信用历史、收入水平、债务负担等多个维度的数据，以确保预测的全面性和准确性。

最后基于构建的风险评估模型，为金融机构提供贷款申请的贷前、贷中、贷后角度的对应建议，帮助他们更好地管理贷款风险。通过参考本研究的数值结果与决策建议，金融机构将能够更加精准地预测贷款违约风险类别，迅速识别客户标签，从而做出针对性的更加明智的贷款决策。

本论文的结果预期将为金融机构提供参考思路，帮助他们改进贷款贷前审批，贷中运营以及贷后管理流程，降低不良贷款率，提高贷款业务的盈利性。本文预期结构见图片2所示：



**图 2 文章结构**

# 相关理论和方法

## 数据挖掘在贷款决策和风险评估中的应用概述

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。

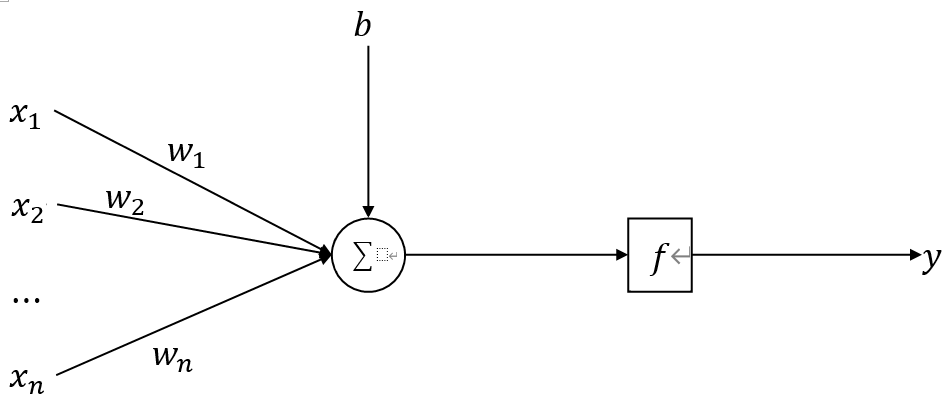
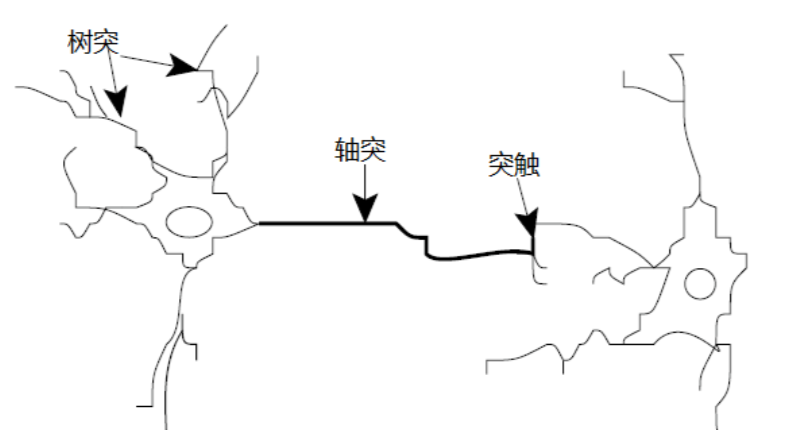
在进行贷款决策评估时，数据挖掘技术可以帮助金融机构更准确地识别潜在的优质客户和不良客户。基于大样本的海量数据的分析和挖掘，使用大数据挖掘技术可以根据科学的算法发现隐藏在数据中的规律和模式，更准确地预测借款人的还款能力和意愿。这有助于金融机构制定更合理的贷款策略，提高贷款审批的准确性和效率。

在客户风险评估时，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。同时还可以帮助金融机构提前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。

## 分类预测算法介绍

常用的多分类方法有支持向量机（SVM），决策树，随机森林，朴素贝叶斯等等，基于不同数据规模、维度、和噪声的考虑，需要选择合适的预测算法进行研究。针对本研究用到的长达13年的所有客户大数据集，首要的考虑因素是计算资源和模型复杂度的要求，且需要同时兼顾特征重要性对客户重要特征的解释，因此选用较为灵活的ANN人工神经网络模型进行分类预测，特别需要注意的是，ANN的超参数组合的选择多样，且大量神经元能够更好地处理复杂的数据模式，多个隐藏层能够学习到复杂的特征表示，激活函数高效处理非线性关系。最后将简单调用常用的逻辑回归LR模型，XGboost模型进行对比，证明ANN对该数据集的适用价值。

1. ANN神经网络

本文用到的算法是人工神经网络（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。它是一种模拟人脑神经网络结构和功能的计算模型，由大量的节点相互连接组成，节点可以类比为神经元。神经元（见图片3）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。

**图 3 神经元示意图**

**图 4 激活单元结构图**

神经元也称“激活单元”，具体结构见图片4激活单元结构图，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，如式1所示：

，i=1，2，3，…为该层神经元的输出，为权重，b为偏置。

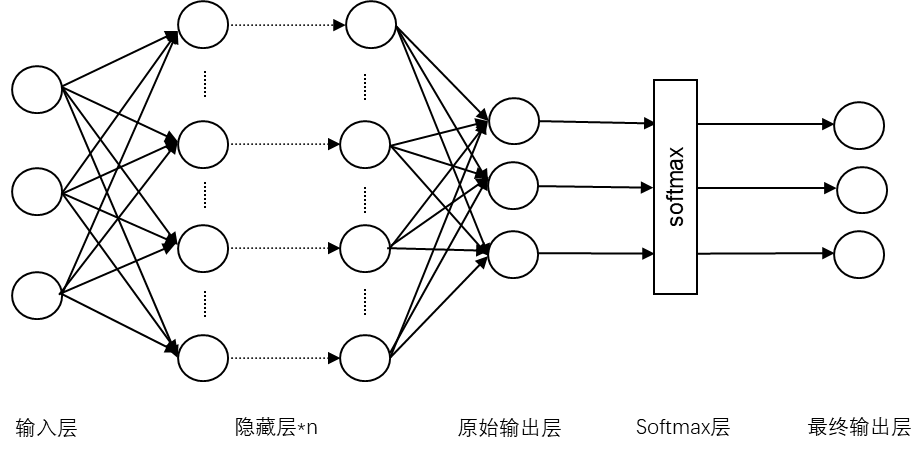
一般来说神经元可能有多层，第n层神经元接收上层传入的神经冲动刺激，如式2所示：

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，如式3所示：

即输出值可以表示为式4：

人工神经网络的基本结构包括输入层、隐藏层和输出层。输入层接收原始数据，隐藏层对数据进行加工处理，输出层则产生最终的预测或分类结果。极强的学习能力可以使模型得到训练集之后，拟合因变量和自变量之间的非线性关系，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。需要注意到本研究中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，但使用的参数非常难以确定，非常受初始值的影响。

1. Softmax激活函数

Softmax激活函数通常用于多分类问题的输出层。它可以将一个数值向量归一化为一个概率分布向量，各个概率之和为1。Softmax可以用来作为神经网络的最后一层（见图片5），用于多分类问题的输出。

**图 5 Softmax函数在神经网络中的位置**

以第i个节点输出为例，预测每个类别的概率，softmax函数详见式5:

为第i个节点的输出值，C为输出节点的个数(即分类的类别个数)。在多类分类问题中，超过两个类标签则需要类成员关系。对于长度为K的任意实向量，Softmax函数可以将其压缩为长度为K，值在 [0,1]范围内且向量中元素的总和为1的实向量。如图片6运算过程Softmax将一个[2.0,1.0,0.1]的向量转化为[0.7,0.2,0.2]，而且各项之和为1。

Softmax函数可以将上一层的原始数据进行归一化，转化为一个(0,1)之间的数值，这些数值可以作为概率分布，用来作为多分类的目标预测值。它一般作为神经网络的最后一层，接收来自上一层网络的输入值，然后将其转化为概率,常常和交叉熵损失函数一起结合使用。

|  |  |  |
| --- | --- | --- |
| Z | Softmax函数 | 概率P |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2.0 |  |  |  | P=0.7 |
| 1.0 |  |  | P=0.2 |
| 0.1 |  |  | P=0.1 |

**图 6 softmax函数处理过程示意图**

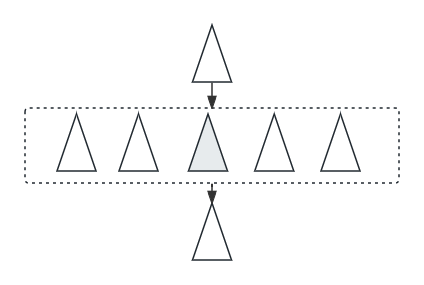
1. 交叉熵损失函数

交叉熵损失函数是一种在机器学习和深度学习中常用的损失函数，主要用于评估分类问题中模型的预测概率分布与真实概率分布之间的差异程度。交叉熵损失函数的值越小，表示模型的预测结果越接近真实结果。具体见式6：

M为类别的数量，为符号函数（0或1）如果样本的真实类别等于c取 1，否则取 0，为观测样本i属于类别c的预测概率。

1. 随机森林

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。

随机森林的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。

**图 7 随机森林工作原理示意图**

随机森林的工作原理如图片7所示，大部分预测结果为浅色，构建决策树时随机选择部分特征进行分割，每一个子树独立训练，根据少数服从多数的原则最终预测结果为浅色三角形。以此原理的多个决策树集成能够降低数据过拟合的风险。

# 数据收集与处理

## 数据收集和预处理

### 数据准备

* 1. 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。

研究中所使用的数据集来源于Lending Club官方网站。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实验。

* 1. 数据预处理
     1. 自变量处理

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，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18个样本观测值。此外，为了克服空值和极端值的影响，对部分连续变量进行了去除空值和离群值的处理。

针对初步处理后的数据列，参考文本特征和内容，结合现有文献和数据分析研究中对文本特征和借款信息的研究，将控制变量分为类别列和数值列，见表格1：

**表格 1 控制变量表**

|  |  |
| --- | --- |
| 类别列 | 'home\_ownership', 'verification\_status', 'purpose', 'loan\_status' |
| 数值列 | 'installment','loan\_amnt','int\_rate','grade','sub\_grade','term','annual\_inc','dti','emp\_length','collections\_12\_mths\_ex\_med','delinq\_2yrs','open\_acc','pub\_rec','revol\_bal','revol\_util' |

* + 1. 因变量处理

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80%的特征，将目标变量数据列loan\_status按照表格2的形式重新映射归类。

根据因变量不同类型的占比（保留至小数点后一位），其他类型在所有数据列的占比远小于1%，为了使数据更具有代表性，只考虑全额结清、拖欠、超时的情况。

**表格 2 目标变量重新映射表**

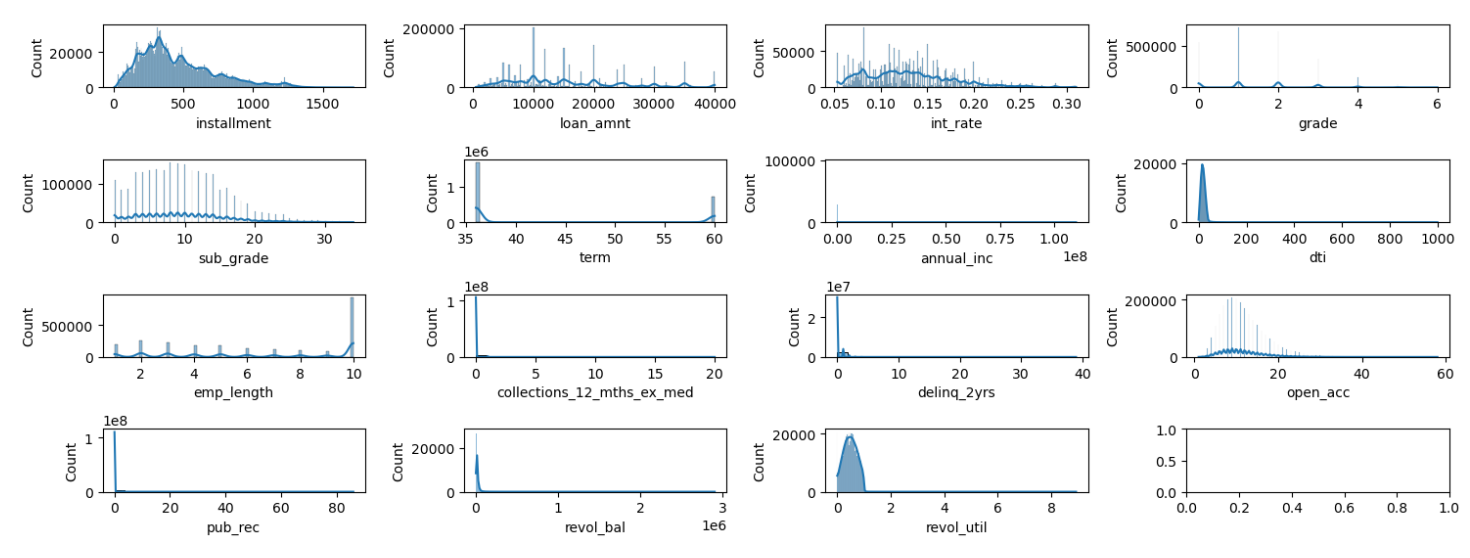
|  |  |
| --- | --- |
| Status | Loan\_status |
| Paid | Fully Paid |
| Current |
| Does not meet the credit policy. Status:Fully Paid |
| Default | Charged Off |
| Default |
| Does not meet the credit policy. Status:Charged Off |
| Late | Late (31-120 days) |
| Late (16-30 days) |
| In Grace Period |

### 描述性统计

表格3为描述性统计特征，涵盖贷款信息，贷款人信息，贷款人历史信息等多中内容。其中贷款金额标准差9458.46表明贷款金额有较大的个体差异。贷款期限中位数为36个月，标准差11.04个月；利率中位数12.49%，标准差4.86%，分布较为集中。过去12个月医疗除外的首款次数中位数0，标准差0.15表明大多数借款人没有医疗除外的收款记录。

**表格 3 描述性统计表**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | mean | std | min | 25% | 50% | 75% | max |
| loan\_amnt | 15568.22 | 9458.462 | 500 | 8000 | 14000 | 21000 | 40000 |
| term | 43.29054 | 11.03726 | 36 | 36 | 36 | 60 | 60 |
| int\_rate | 0.130423 | 0.048562 | 0.0531 | 0.0917 | 0.1249 | 0.1599 | 0.3099 |
| installment | 457.5659 | 272.2163 | 4.93 | 259.33 | 389.03 | 610.75 | 1719.83 |
| grade | 1.561019 | 1.236414 | 0 | 1 | 1 | 2 | 6 |
| sub\_grade | 9.774847 | 6.185629 | 0 | 5 | 9 | 14 | 34 |
| emp\_length | 6.473576 | 3.38378 | 1 | 3 | 7 | 10 | 10 |
| annual\_inc | 82296 | 116966.3 | 14 | 50000 | 70000 | 98000 | 1.10E+08 |
| dti | 18.96598 | 11.58053 | 0.01 | 12.14 | 18.05 | 24.7 | 999 |
| delinq\_2yrs | 0.296205 | 0.847697 | 0 | 0 | 0 | 0 | 39 |
| inq\_last\_6mths | 0.556026 | 0.861462 | 0 | 0 | 0 | 1 | 31 |
| open\_acc | 11.74239 | 5.527473 | 1 | 8 | 11 | 15 | 58 |
| pub\_rec | 0.175503 | 0.531385 | 0 | 0 | 0 | 0 | 86 |
| revol\_bal | 17290.92 | 23147.49 | 0 | 6209 | 11794 | 21047 | 2904836 |
| revol\_util | 0.495118 | 0.247741 | 0 | 0.305 | 0.492 | 0.686 | 8.923 |
| total\_acc | 24.03032 | 11.37382 | 1 | 16 | 22 | 31 | 63 |
| last\_pymnt\_amnt | 3618.596 | 6238.695 | 21033.4 | 275.59 | 611.84 | 4281.81 | 42163.62 |
| collections\_12\_mths\_ex\_med | 0.017803 | 0.148328 | 0 | 0 | 0 | 0 | 20 |

针对数值变量，由下列的直方图（图片8）可知15个数值型特征均为离散型连续变量，大部份客户月付金额在0-500之间，贷款金额峰值在10000左右，且与贷款利率有相似的分布，债务收入比非常低，均在50%以内。值得注意的是贷款期限出现了极值明显的情况，大部分客户均为36月与60月的贷款期限。

**图 8 数值型变量分布直方图**

### 数据检验

3.1多重共线性检验-VIF

VIF值可以表征自变量之间的共线性程度，其计算公式表达如下（见式7）：

其中计算公式表达如下（见式8）：

为观测因变量i的值，为观测值i的回归模型的近似值，为所有观测值中因变量的平均值。

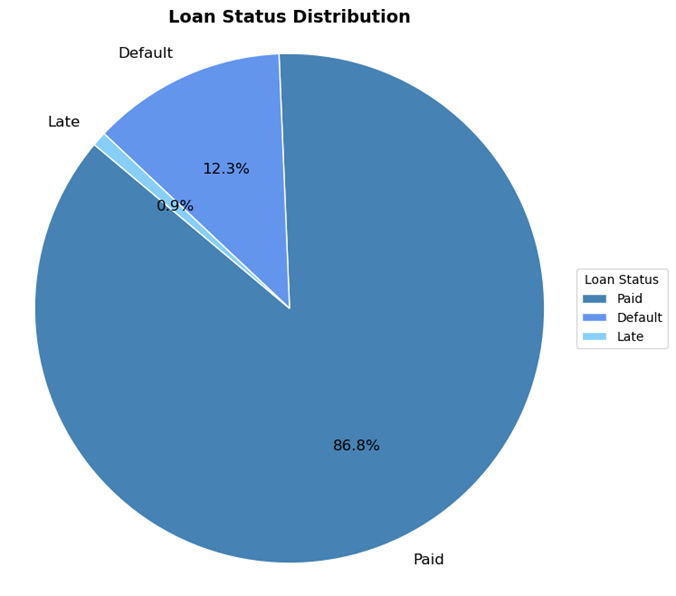
它的大小可以反映出自变量的观察值之间是否存在复共线性以程度，如果VIF值较高，表明该自变量与其他自变量高度相关，可能存在多重共线性问题。通常VIF值大于10时，可以认为该自变量存在较高的多重共线性问题。根据计算得到的结果，下列这些变量（见表格4）存在严重的多重共线性。

**表格 4 VIF大于10的变量表**

|  |  |
| --- | --- |
| feature | VIF |
| installment | 94.811915 |
| loan\_amnt | 107.505549 |
| int\_rate | 106.449467 |
| grade | 48.119343 |
| sub\_grade | 106.07003 |
| term | 46.435059 |

3.2交叉验证

交叉验证是一种统计方法，用于评估和比较机器学习模型的泛化能力。核心思想是将数据集分成多个小的、不重叠的子集，然后轮流使用其中一个子集作为验证集，其余子集作为训练集。这个过程重复进行多次，每次选择不同的子集作为验证集，从而确保每个子集都有机会被用作验证集。

根据描述性统计数据饼图（见图片9），数据集的类别非常不均衡，需要进行交叉验证减少模型评估的方差。多次使用少类别的数据，进行训练和验证，提高数据的利用率。

**图 9 类别数据占比图**

在交叉验证过程中，模型的超参数会被调整，以找到最佳的模型配置。每次调整后，模型都会在训练集上进行训练，并在相应的验证集上进行评估。通过比较不同超参数配置下的模型性能，研究者可以选择出在验证集上表现最佳的模型配置。

### 不均衡数据集处理

原数据类别不均衡会影响模型正常输出，本研究中三种类型的样本存在显著差异。Paid类别贷款占比最大为86.8%，其次是default类型贷款，值得注意的是late类型贷款在经过数据预处理和归类后仅为0.9%，需要后续的采样处理。基于数据集较大，为了保证预测模型的准确性，这里分别使用欠采样和过采样进行比较分析。

4.1欠采样

欠采样是指减少多数类（即数据集中占多数的类别）的样本数量，以平衡数据集中的类别分布。这里选择常用的欧氏距离进行计算（见式9）：

d(x,y)为两个点的欧氏距离，N维空间中，n维欧氏空间是一个点集，每个点表示为(, , …, )，其中 (i=1,2…n)是实数称为x的第i个坐标。通过选择距离最近的少数类样本来减少多数类的样本数量。通过计算每个多数类样本与最近的三个少数类样本之间的距离，并选择那些距离最近的多数类样本进行删除。保留那些对分类任务最有信息量的样本。

4.2过采样

过采样通过随机复制少数类样本来平衡数据集。复制的样本可能包含重复的信息，可能会导致模型过拟合。表格5为采样数据结果描述。

**表格 5 不均衡数据集采样结果描述表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 采样方法 | 类别 | 准确率 | 精确率 | 召回率 | F1分数 |
| 欠采样 | Default | 0.13 | 0.20 | 0.21 | 0.21 |
| Late | 0.01 | 0.84 | 0.02 |
| Paid | 0.96 | 0.11 | 0.19 |
| 过采样 | Default | 0.88 | 0.90 | 0.97 | 0.93 |
| Late | 0.56 | 0.39 | 0.39 |
| Paid | 0.00 | 0.00 | 0.00 |

根据计算得到的数值（见表11，12），过采样在Default类别上表现最好，准确率0.88、精确率0.9、召回率0.97和F1分数0.93均大于0.9，表明模型能够很好地识别Default类别的样本。欠采样在Paid类别上表现最佳。 值得注意的是过采样和欠采样在极小类别Late上的性能较差，还需进一步分析原因并改进模型或采样策略。过采样在Paid类别上的准确率为0，表明模型完全无法识别Paid类别的样本，这可能是因为过采样策略没有保留足够的Paid类别样本。根据采样结果可知总的来看随机过采样综合表现优于欠采样，因此选择随机过采样后的数据进行进一步分析。

## 变量选取和特征工程

### 变量选取

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二个方面的因素：

贷款信息因素：这包括借款的月还款、债务水平、贷款金额和贷款利率，贷款等级等。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。

客户信息角度，又可细分为三个方面的因素：1、短期难以更改的硬性指标：包括借款人的年收入、债务收入比、住房情况。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。2、借款人本身因素：包括工作年限，贷款用途。3、信用指标：过去一年内除医疗外的逾期次数，开放信用账户数量，公开贬损公共记录数量，循环信用余额，收入验证，循环信用利用率。清洗后的有效变量与测度汇总表见表格11。

**表格 6 清洗后的有效变量与测度汇总表**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 指标归类 | 字段名 | 字段含义 | 数据类型 |
|  | 因变量 | loan\_status | 借款当前状态 | object |
| 平台放贷角度 | 贷款信息 | installment | 按月支付贷款 | float64 |
| loan\_amnt | 贷款金额，即借款人申请的贷款金额 | float64 |
| int\_rate | 贷款利率，通常以年利率表示 | float64 |
| grade | 贷款等级，根据借款人的信用评分和还款能力评定 | float64 |
| sub\_grade | 贷款子等级，进一步细分贷款等级 | float64 |
| term | 贷款期限，以月为单位 | float64 |
| 客户信息角度 | 硬指标(短期内不会改变的指标) | annual\_inc | 借款人的年收入 | float64 |
| dti | 借款人的债务收入比，即每月债务支付与月收入的比率 | float64 |
| home\_ownership | 借款人的住房状况，如租赁、自有等 | object |
| 基本信息 | emp\_length | 借款人的工作年限 | float64 |
| purpose | 贷款用途，如债务合并、个人消费等 | object |
| 信用指标 | collections\_12\_mths\_ex\_med | 过去12个月内除医疗外的逾期次数 | float64 |
| delinq\_2yrs | 过去两年内逾期次数 | float64 |
| open\_acc | 借款人的开放信用账户数量 | float64 |
| pub\_rec | 借款人的公开记录数量贬损公共记录数 | float64 |
| revol\_bal | 循环信用余额，即信用卡或类似信用产品的未结余额 | float64 |
| verification\_status | 收入是否通过平台验证 | object |
| revol\_util | 循环信用利用率，即循环信用余额与信用额度的比例 | float64 |

### 特征工程

特征工程是数据科学中的关键步骤，从原始数据中提取，构造和选择有效和高性能的特征有助于模型性能的提升。特征衍生，特征抽象和特征缩放是特征工程中的关键步骤。

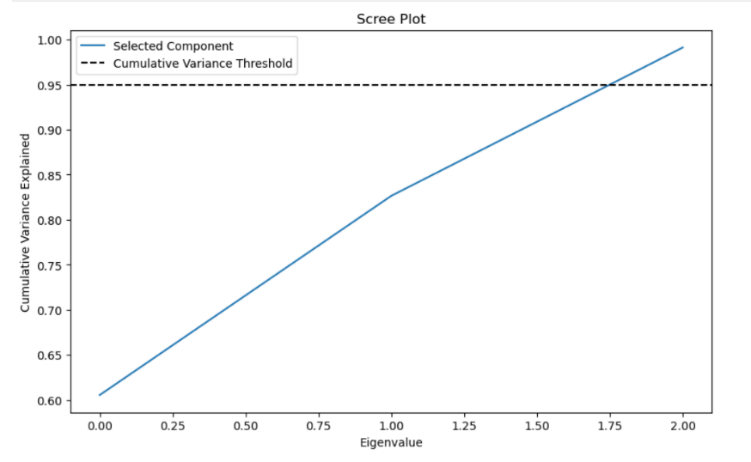
2.1特征衍生- PCA数据降维

特征衍生是基于原始数据创建新的特征，基于本数据集的贷款信息类别数据中，存在共线性非常显著的情况，因此需要进行数据降维处理。

主成分分析方法是常用的数据降维算法。主要思想是将n维特征映射到k维上，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征。

对于含有p个变量、n个数据的样本，可以计算出其协差阵。根据该协差阵可以求出p个特征值并从大到小排序得与相应的p个特征向量。则第i主成分为见式10：

选取主成分的个数需要依据主成分贡献率与累计贡献率。第k个主成分的贡献率见式11：

在实际应用中，从6个变量中提取3个主成分，需要对这新的1个变量做合理的解释。而进行解释往往需要根据具体问题来判断，并结合该主成分中保存原各个变量的比重，即求解该主成分与原始各个变量的因子载荷量见式12：

**图 10 碎石图**

表示第k主成分在原第i变量的因子载荷量（它表示变量对主成分的重要程度）。表示协差阵的第i行i列元素，表示第k个特征向量的第i行。根据处理的碎石图（见图片10），选择3个主成分可以解释超95%的方差。

2.2特征抽象

特征抽象需要把数据转化成算法可理解的数据，通常涉及到从原始数据中提取关键信息。如文本数据集中无法直接对文本数据进行数据分析，本数据集需要对文本数据列进行独热编码和等级映射处理。

2.2.1独热编码

根据本研究用到的深度学习网络库和其他模型对比需要的数据转换，需要对文本特征和原始目标变量进行独热编码处理。独热编码是一种将离散型特征转换为二进制向量的方法，可以使特征之间的距离或相似度更合理。

**表格 7 文本列独热编码表**

|  |  |  |
| --- | --- | --- |
| 变量名 | 文本特征 | 描述 |
| home\_ownership | Rent | 借款人的住房状况，如租赁、自有等，出现为1，未出现为0。 |
| Own |
| Mortgage |
| Other |
| None |
| Any |
| verification\_status | Verified | 收入是否通过平台验证，出现为1，未出现为0。 |
| Not verified |
| Source verified |
| purpose | Credit\_card | 贷款用途，如债务合并、个人消费等，出现为1，未出现为0。 |
| Small\_business |
| Other |
| Wedding |
| Debt\_consolidation |
| Car |
| Home\_improvement |
| Major\_purchase |
| medical |
| Moving |
| Vacation |
| House |
| Renewable\_energy |
| educational |

基于数据集中房屋所有权、身份认证、贷款目的文本列的数据分布特征，转化为二进制变量，对不同的文本特征进行独热编码处理得到多维特征列。表格7为文本特征独热编码规则，每一个文本变量有且仅有一个文本特征，针对每一个出现的文本特征，每出现在数据集对应的数据行就被标记为1，否则为0。

根据本文的研究目标，目标变量仅有三个类别，以表格8所示形式进行编码：

**表格 8 目标变量独热编码表**

|  |  |
| --- | --- |
| Status | 编码 |
| Paid | 0 |
| Default | 1 |
| Late | 2 |

2.2.2数值映射

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，每个等级包含1、2、3、4、5 五个子级。有序特征映射是常用的特征抽象方法，多用于保持数据的顺序性。考虑到lending club平台提供的是不可量化的抽象等级数据，为方便后续的建模，更直观量化等级，将官方评定的等级按以下规则进行赋分处理：

**表格 9 Grade数据映射规则表**

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

**表格 10 Subgrade数据映射规则表**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| A1 | A2 | A3 | A4 | A5 |
| 0 | 1 | 2 | 3 | 4 |
| B1 | B2 | B3 | B4 | B5 |
| 5 | 6 | 7 | 8 | 9 |
| C1 | C2 | C3 | C4 | C4 |
| 10 | 11 | 12 | 13 | 14 |
| D1 | D2 | D3 | D4 | D5 |
| 15 | 16 | 17 | 18 | 19 |
| E1 | E3 | E4 | E5 | E6 |
| 20 | 21 | 22 | 23 | 24 |
| F1 | F2 | F3 | F4 | F5 |
| 25 | 26 | 27 | 27 | 29 |
| G1 | G2 | G3 | G4 | G5 |
| 30 | 31 | 32 | 33 | 34 |

表格9为根据lending club官方网站对客户的特征画像主等级划分的等级分数映射。表格10为根据lending club官方网站对客户的特征画像子等级划分的等级分数映射，基于主等级的细分子等级，本文将根据下表依次赋分将等级转化为连续的数值变量。

2.3特征缩放：归一化

特征缩放是将特征值通过数学方法调整到一个统一的范围，以消除不同特征之间巨大数值差异对模型训练的影响。本研究用到的数据集的变量差异较大（见图），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。

本研究采用的归一化公式如式13所示：

为样本数据的最大值，为样本数据的最小值。将本研究中用到的数据集进行量级归一化处理后，能够避免某些特征值对模型训练的影响，同时加速模型训练速度。

# 模型建立和评估

## 预测模型构建

使用Keras调谐器进行超参数调优，添加一个可变数量的Dense层和一个可变数量的unit，经数据集采样和多次模型拟合实验，将隐藏层可变数量设置为4到20，神经元数量设置为32到512（步长32），设置学习率为0.01，0.001，0.0001三个可选项。根据三分类的特点，输出层采用Softmax激活函数单元采用准确率作为模型性能指标。以多类交叉熵作为损失函数，输出层指定为3个，根据验证精度搜索最佳超参数。

使用从调谐器获得的最佳超参数，定义num\_layers, num\_nodes和activation用于构建具有给定层数和每层激活函数的模型。在层上运行一个循环，将密集层添加到模型中，并制定优化器。

模型的整体准确率达到了84.52%，这表明模型在分类任务中表现出较高的准确性，能够识别出大部分的正类样本。但模型的精确率、召回率和F1值表明还有进一步优化的空间。精确率（85.6%）和F1值（78.2%）都相对较高，表明模型在预测为正类的样本中，大部分确实是正类样本。而召回率（73.8%）则相对较低，这可能意味着模型在识别正类样本方面存在一定的改进和调整空间。

## 结果分析和讨论

为了验证模型的有效性，本研究使用了大量的真实贷款数据进行实验。基于数据集大小与计算效率的考虑，在多次试验后确定超参数的搜索范围以达到理想的预测效果，下面为预测结果分析。

### 网络模型分析

* 1. 预测准确度分析

根据自定义的最佳超参数搜索模型，得到全连接层数量为3，分别有288，320，64个神经元。由于本研究使用的数据集较大，在超参数搜索过程中，调谐器尝试的模型构建和训练的总次数仅为3次。但从数值结果上看，准确率和召回率为84.52%的实验结果表明模型在预测的总样本中有84.52%的预测是正确的，模型成功识别出的正类样本占所有真实正类样本的比例为84.52%；精确率是模型预测为正类的样本中模型预测为正类的样本占的比例，真正为正类的样本所占的比例仅为81.04%，在预测贷款违约风险方面准确性达到了理想的预测效果。82.40%的F1值也显示了模型整体性能的良好平衡。

总的来看模型的准确率和召回率较高，但精确率仍有欠缺，模型可能会倾向于预测出现频率最高的类别，对于某些类别的边界不够敏感。或对于某些边界过于敏感，倾向于将某些样本预测为少数类别，从而提高了召回率，但由于敏感性导致的错误分类，又造成了精确率的较低数值。与传统的风险评估方法相比，该模型基于更大的数据集，更广的自变量列，能够更好地捕捉到隐藏在数据中的复杂关系，为金融机构提供风险评估结果参考。

* 1. 预测模型优势

该模型最大的优势在于超参数调优的灵活性、多分类处理、交叉熵损失函数的应用、多性能指标的评估、特征和样本平衡可调整性。

1.2.1调优灵活性

通过使用Keras调谐器进行超参数调优，可以有效地搜索到最佳的模型参数。调优策略可以帮助找到隐藏层的数量、神经元的数量、学习率最优的模型配置，提高模型的性能。

1.2.2概率输出预测

模型的输出层采用Softmax激活函数，将输出转换为概率分布，通过比较不同类别之间的概率值，并根据这些值选择概率最高的类别作为预测结果。

1.2.3损失函数处理

多类分类问题使用多类交叉熵作为损失函数，多类交叉熵考虑了所有类别的预测概率，能够惩罚与真实标签概率分布不一致的预测，因此可以有效地衡量模型的预测误差。这种损失函数可以确保模型在优化过程中能够准确地识别每个类别的预测，通过计算验证集上的多类交叉熵损失，可以了解模型在多类分类任务上的表现，并据此进行调整和优化。

1.2.4多性能指标评估

使用准确率、精确率、召回率、F1值作为模型的性能指标，准确率关注模型预测正确的样本数与总样本数的比例；精确率关注模型预测为正例的样本中，真正为正例的样本所占的比例；召回率关注在所有真正的正例样本中，模型能够预测为正例的样本所占的比例；F1值是精确率和召回率的调和平均值，可以综合精确率和召回率，更全面地评价和解释模型的性能。

总的来说，虽然模型的整体准确率达到了84.52%，但样本平衡仍有局限，通过持续的调整特征或样本平衡，模型的性能可以得到持续的改进。这表明模型具有良好的可调性，可以针对其他具体业务需求具体问题进行优化。

* 1. 模型对比分析

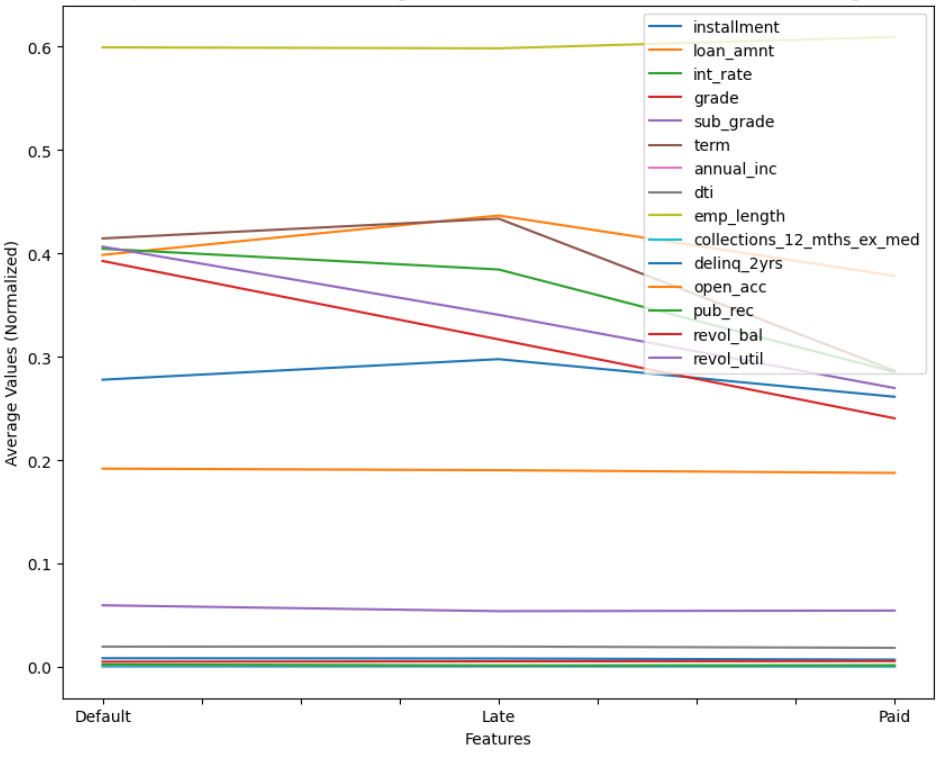
为了检验本研究中ANN自动寻找最优超参数模型的效率提升程度，简单调用了sklearn库中的逻辑回归模型LR，XGboost模型，决策树模型对数据集进行借款客户类别预测和评价值对比。其中准确率84.5%相对于其他模型（见表格11所示），ANN的整体性能最优越。从准确率和Recall值分析，准确率84.52%均远大于XGboost与逻辑回归LR模型的64.94%，56.81%；可以初步根据得到的数值确定本文的人工神经网络模型具有良好的适用价值，在该数据集上优于XGboost与逻辑回归LR模型。

**表格 11 模型评价值对比表**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1 |
| ANN | 84.52% | 81.04% | 84.52% | 82.40% |
| XGboost | 64.94% | 88.80% | 64.94% | 72.01% |
| LR | 56.81% | 57.65% | 56.81% | 57.06% |

这个模型具有较好的预测性能和较强的实际应用价值，可以为贷款者提供一定程度的预警和检测。同时，模型还具有良好的可调性，可以通过调整特征和样本平衡来进一步提高性能。

### 特征结果分析

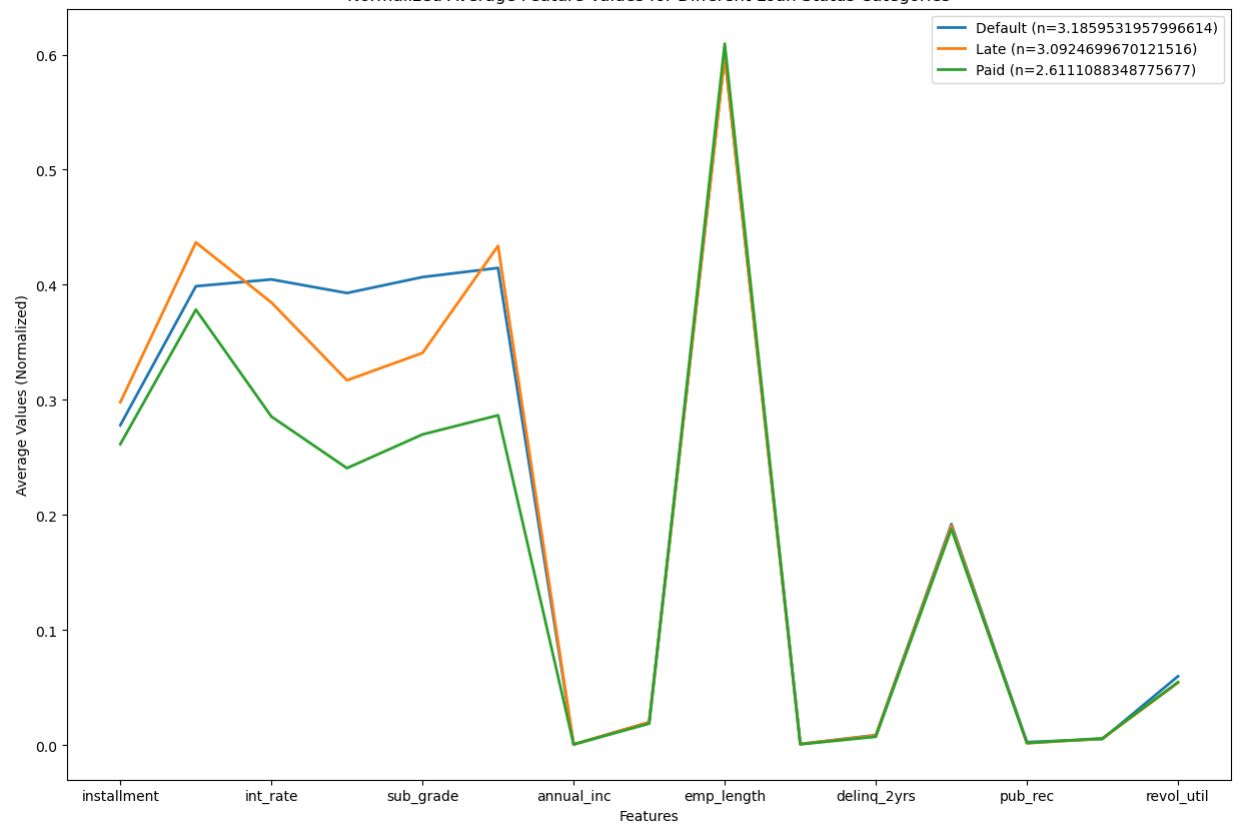
降维后的三个主成分分别解释了60.56%，22.10%，16.50%的方差，共计99.16%的方差被捕获，能够保留原数据集的大部分关键信息。以数据处理后归一化后的特征均值为衡量标准，数据集特征图（见图11）显示出现相对明显差异的特征为term（以月为单位的贷款期限），open\_acc（借款人的开放信用账户数量），int\_rate（贷款利率），sub\_grade（贷款子等级），revol\_bal（循环信用余额），installment（月付贷款）。这些特征是区分不同类别贷款客户的主要特征，因此未来的个人信用评估可以加强贷款申请信息（利率，月付）权重的同时，注意客户历史的信用行为（开放信用账户数量，循环信用余额），另外贷款子等级需要中介谨慎把握评分规则。

**图 11 不同贷款类别的标准化平均特征值比较图**

从不同贷款类别的标准化平均特征值图（见图12）结果可知，Paid（已全部归还贷款的）类型贷款客户的int\_rate（贷款利率），sub\_grade（贷款子等级），installment（月付贷款）均值显著小于另外两种类别，这类客户通常有良好的信用记录和还款能力，可见低等级、低利率、低月付的贷款更有良性贷款的倾向。例如一个客户申请了一笔贷款，他的贷款利率为5%，贷款子等级为A，月付贷款为1000元，这表明他属于Paid类型客户。对于这类型客户，平台可以考虑减少这类客户的贷款门槛，降低贷款手续信息的繁琐程度，增加此类贷款的投放规模，以达到“薄利多销”的效果。平台也可以考虑对这类客户提供更多的优惠政策，如减免手续费、降低利率等，以进一步吸引他们使用平台贷款。

对于Late（延期归还贷款）类型客户，平台的评级明显较低，这表明他们在还款方面可能存在一定的困难。他们所面临的月付贷款和利率较高，这可能会增加他们的还款压力。建议平台对这类客户进行定期或实时的分析与监测，控制成本的前提下还可以安排专人进行还款规划提醒，根据客户的实际情况调整还款期限与月付贷款，以减轻他们的还款压力。再贷款后续的运营过程中，平台还需要加强贷后的管理，确保客户的贷款能够按时还款。

Default（欠款）类客户特征相对均衡，利率保持在一个最高水平，这表明，这类客户可能存在一定的风险。平台需要对此类客户加强风险评估与特征权重的分配，平衡放贷营收与坏账的风险。建议平台对此类客户增加个人身份或其他平台的身份绑定，联合其他企业对不守信用的违约账户进行冻结或限制活动，同时制定针对Default类客户的有效的催收策略，以提高催收成功率。可能的坏账损失也需要被纳入考虑范围，以便进行充分的准备。

总的来说，即对于低风险的Paid类型客户，贷款平台可以考虑降低贷款门槛，简化申请流程，并增加贷款规模；对于高风险的late和Default类型客户，需要定期监测和调整还款计划，并加强贷后管理，同时采取严格的身份验证和有效的催收策略。

**图 12 不同贷款类别的标准化平均特征值图**

# 结 语

## 研究结论

本文首先梳理了数据挖掘相关理论和技术，获取数据集后对数据集进行数据预处理，根据数据分布进行数据列的初步选择，结合空值行属性以及在数据集中的占比，对空值行和重复行选择进行删除处理；根据箱线图分布对离群异常值进行去除；针对原始数据集中带有单位、符号、特殊标记的列，使用正则化方法并进行数据归一化处理，去除计量单位、数学符号在内的干扰信息。剔除无关变量以后分析贷款决策的影响因素，整合出与网络借贷客户借贷前有关的特征，并分为文本变量与数值变量。

针对初步预处理后的数据进行包括特征衍生、特征抽象、特征缩放的特征工程处理。特征衍生是基于原始数据创建新特征处理步骤，本研究对共线性极高的客户贷款申请信息进行主成分降维处理，根据碎石图将原有的6个高共线性特征合并为3个，能够表示超99%以上的方差。特征抽象是将是原始数据表示为机器能够识别的数据，文本和类别数据难以直接量化，官方定义的等级层次不好区分，本研究中主要包括特征编码和数值映射，对文本数据进行了分类和重新独热编码处理，针对客户的两种等级进行了进一步的数值映射二次赋分处理，以达到量化信息与模型数据准备的效果 。特征缩放是为消除数据量纲，避免过拟合等影响，是数据处理的常规步骤。

最终结合数据集的不平衡特征，进行过采样和欠采样实验，结果证明过采样88%的准确度远大于13%的欠采样准确度，能够平衡late数据极少的影响。构建风险预测模型，通过相关库和重复实验寻找最优超参数的寻找范围，确定最终超参数设置。建立ANN模型预测全额结清、拖欠、超时三类客户的类别，准确性达到84.52%，综合预测效果远超逻辑回归模型与xgboost模型，根据特征重要性讨论贷款客户的重要预测特征，总结包括全额结清,拖欠、超时的三类贷款风险客户画像。最终针对不同类客户的画像得出结论并基于平台运营角度提出针对性建议。

## 研究展望

### 模型改进

针对上述模型的准确率与召回率较高，但准确率不足的情况。在模型训练过程中，可以尝试调整不同类别样本的权重，在进行抽样之后，对不同类别样本进行加权处理，使模型更关注精确率较低的类别。针对模型精确度不高且可能出现的敏感性问题，由于本研究超参数范围设置与试验次数的局限原因，最终获得的数值仅为一个合适的理想范围，未来可以根据实际情况尝试调整模型的决策阈值多次实验来尽可能地提高预测精度。

### 研究局限性

2.1数据指标局限

本文的借款风险预测完全基于Lending club数据集，在消费人群的地理分布，数据指标的完善层面都有很大的平台局限性。如样本人群为美国公民，可能受政治文化因素影响，各国家地区p2p贷款人群画像差异明显，模型预测结果可能与东亚地区的数据结果有差异，需要进一步的进行特定地区的针对性研究和拓展。

2.2社会环境变化

本研究使用2007-2020长达13年的数据集，在这个时间段存在各类通货膨胀，利率调整，政策环境变化等多种因素影响。同时，自2020年开始的新冠肺炎疫情对社会和当地经济因素的影响，以及疫情期间公民的心理预期、消费观念带来的网络借贷行为变化，在进行数据预处理和建模分析时暂未考虑这些因素。

2.3不平衡样本集处理

样本集不平衡是常见的问题，实际生产生活中可以获得的原始数据集大多不完全平衡，需要根据数据集的特点进行针对性方法的比较和选择。本文目前只选取了随机过采样和基于距离的过采样进行对比，随机过采样仅达到88%的准确度。另外还有过采样，修改阈值，集成学习，成本敏感学习等方法处理不平衡数据方法可以尝试。同时可能存在采样后的数据集与原有数据分布不一致导致的偶然性信息丢失和过拟合问题。

### 未来方向

贷款违约预测一直是社会的热点问题，随着机器学习技术的拓展，各种基于大数据大模型的应用也在不断发展推进。未来的研究可以探索更多高效的集成学习方法：如Stacking，bagging，boosting，选择、训练、评估和融合不同的机器学习模型。特征工程是提高模型性能的关键，目前已经有非常多的学者研究不同种类机器学习方法在数据集中的预测效果和对比精度，自动化的特征选择和特征构造比人为的主观性选择更有科学性，研究可解释性模型和特征主导的客户分类、风险量化评级将是未来的主导研究方向。

预计未来还可以结合图像，声音，IP地址等多模态数据对借款人进行建模，结合人文社会环境，个体物理性因素等可以进行更全面的风险评估。第三方消费平台数据也可以纳入指标构建，推动共享付款的额度、购物品牌及金额等信息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。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。

总的来说，未来的贷款违约预测研究将更加侧重于模型的准确性和稳定性，以及模型的可解释性[[29]](#footnote-29)和实用性。通过多次训练和评估后采用最新的机器学习技术和大数据处理方法，可以更有效地评估借款人的违约风险，为金融机构提供有力的支持。

# 主要参考文献

1. 陈健巧. 普惠金融背景下P2P投资者的借贷行为影响因素研究[D].上海交通大学,2022.DOI:10.27307/d.cnki.gsjtu.2020.003769.
2. 蔡青松,吴金迪,白宸宇.基于可解释集成学习的信贷违约预测[J].计算机系统应用,2021,30(12):194-201.DOI:10.15888/j.cnki.csa.008220.
3. 戴杨文意. 大学生网络借贷行为问题现状及其社会工作介入策略[D].华南理工大学,2021.DOI:10.27151/d.cnki.ghnlu.2020.004597.
4. 姜晨. 我国P2P网贷风险识别研究[D].北方工业大学,2022.DOI:10.26926/d.cnki.gbfgu.2021.000656.
5. 梁晓颖.P2P网络借贷风险的现状及对策研究[J].中国商论,2020(19):61-62.DOI:10.19699/j.cnki.issn2096-0298.2020.19.061.
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. 刘雨杭. 股市投资者情绪对借贷市场的影响[D].吉林大学,2023.DOI:10.27162/d.cnki.gjlin.2022.005316.
8. 刘翱,邓旭东,童泽平等.P2P网络借贷研究进展[J].系统工程学报,2020,35(03):402-415.DOI:10.13383/j.cnki.jse.2020.03.010.
9. 雷佳. 基于机器学习的P2P个人信贷风险研究[D].对外经济贸易大学,2022.DOI:10.27015/d.cnki.gdwju.2020.000538.
10. 明依东. 基于P2P在线借贷产品违约风险评估的个性化投资推荐方法研究[D].天津理工大学,2021.DOI:10.27360/d.cnki.gtlgy.2021.000372.
11. 秦启雯.大数据在互联网金融风控中的应用研究[J].商展经济,2022(03):92-95.DOI:10.19995/j.cnki.CN10-1617/F7.2022.03.092.
12. 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.
13. 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.
14. 翟金阳. 网贷平台欺诈行为影响因素研究[D].苏州大学,2023.DOI:10.27351/d.cnki.gszhu.2021.001275.
15. 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.
16. 张丽改. P2P网络借贷平台信用评级的有效性分析[D].山东大学,2023.DOI:10.27272/d.cnki.gshdu.2023.007825.
17. Artem Bequé, Stefan Lessmann. Extreme learning machines for credit scoring: An empirical evaluation[J]. Expert Systems With Applications, 2017, 86: 24-27.
18. LYÓCSA, Štefan, Petra VASANICOVA and Oleg DEEV. Peer-to-peer loan returns: heterogeneous effects across quantiles. Applied Economics Letters. UK: Taylor & Francis, 6 pp. ISSN 1350-4851. doi:10.1080/13504851.2023.2298412. 2024.
19. Ruqoyah, R., Afriani, Y., Fahleti, W. H., Arifin, N. Y., Zarvianti, A. A., & Ramadhan, A. R. (2023). Analysis of Public Opinion on the Utilization of Online Loan Applications. Tambusai Journal of Education, 7(3), 22820–22825. https://doi.org/10.31004/jptam.v7i3.10210
20. Senney, G.T., Lhost, J.R. Balancing Attraction and Risk Revelation: The Optimal Reservation Price in Peer-to-Peer Loan Auctions. Rev Ind Organ 64, 289–314 (2024). <https://doi.org/10.1007/s11151-023-09914-0>
21. ŠPANKO, Martin. Performance of credit risk models in P2P lending Online. Master's thesis. Praha: University of Economics, Prague. 2024. Available from: https://theses.cz/id/sbxnv8/. [cit. 2024-04-08].
22. Wijaya, C., Nugroho, B. Y., & Arkanuddin, M. F. (2022). The Analysis of Risk Effect Towards Fintech Ecosystem On P2P Lending Industry in Indonesia. Indonesian Journal of Business and Entrepreneurship. <https://doi.org/10.17358/ijbe.8.3.481>
23. WijayaC., NugrohoB. Y., & ArkanuddinM. F. (2024). The Important Role of Financial Architecture Regulation Toward Fintech P2P Lending Ecosystem. Indonesian Journal of Business and Entrepreneurship (IJBE), 10(1), 13. https://doi.org/10.17358/ijbe.10.1.13

# 附 录

## 附录一

**表格 12 编程相关库版本表**

|  |  |
| --- | --- |
| 库名 | 版本 |
| Scikit-learn | 1.4.1post1 |
| Tensorflow | 2.15.0 |
| Keras | 2.15.0 |
| Xgboost | 2.0.3 |
| Keras-tuner | 1.4.6 |

# 后 记

作为一篇毕业论文，本论文一开始的目标是希望以金融领域的某个具体场景为主题，结合本专业所学内容，初步体现基础的数据处理流程，从获取数据开始，数据初步分析分布情况，数据预处理，数据特征工程，到选择具有一定适用模型或可解释性的模型，模型求解，模型评价值分析，结果呈现和分析。涉及到各类数据网站的数据源获取，数据处理软件的使用，各类脏数据的处理，模型建立的数据准备，目标为提高准确度的超参数调优，以及根据实际生产应用场景结合业务的决策建议。

从整个数据集的处理和分类到数据集中具体列的处理分析，怎么去找到数据的价值，考虑可能的影响因素，探索一个数据集给实际生产活动中带来的价值，如何根据理想化的计算结果进行对应决策的参考才是最复杂且重要的内容。以本研究中用到的数据集为例，它仅仅考虑的是从平台获取的贷款客户在申请网贷过程中的信息，对于如何发放贷款、客户定级、客户个性化决策制定等方面仅作为一个参考项，在实际的平台运营过程中，也需要更多的注意到贷款行为后续运行的过程中客户的行为变化，同时可以增加的参考想还有其他平台的行为、客户实时的消费行为等因素。

不严谨的部分可能来自于方法（如抽样、预测回归、填充空值、神经网络超参数的选择）太多，只是初步的选择了其中某个可行的方法，得到一个看似理想的预测结果。而不知道如何去进行定量和定性的解释和选择，也有可能是本身题目和数据集选择范围过大，问题立意不够小的原因。这也是我在完成论文的过程中遇到的最棘手的问题。

非常感谢祝老师在毕业论文修改过程中事无巨细地提示和引导，让我注意到论文中许多细节改进。同时也能在文章的主要结构和内容上靠近学术论文的写作方法和思路，我学会了如何根据找到的问题去细分各个小问题，并且根据拆分的问题，在论文章节中的对应部分进行问题抛出、分析和回答。

终于接近尾声，希望一切顺利，今后做一个快乐从容的人。

1. 来源：经济参考报《后P2P时代：平台转型面临重重困境》 [↑](#footnote-ref-1)
2. 《国务院关于推进普惠金融高质量发展的实施意见》 [↑](#footnote-ref-2)
3. 人民银行召开2023年金融市场工作会议提要 [↑](#footnote-ref-3)
4. P2P(Person-to-Person)网络借贷是指借助于互联网大数据等技术，完成个人对个人的资金借贷，通常是一种小额信贷的直接信贷方式。其主要是以P2P网络借贷平台作为中介，将资金借贷双方联系起来，同时实现借贷双方各自的借贷需求，以完成整个借贷交易过程的互联网金融模式。 [↑](#footnote-ref-4)
5. 秦启雯.大数据在互联网金融风控中的应用研究[J].商展经济,2022(03):92-95.DOI:10.19995/j.cnki.CN10-1617/F7.2022.03.092. [↑](#footnote-ref-5)
6. 翟金阳. 网贷平台欺诈行为影响因素研究[D].苏州大学,2023.DOI:10.27351/d.cnki.gszhu.2021.001275. [↑](#footnote-ref-6)
7. 明依东. 基于P2P在线借贷产品违约风险评估的个性化投资推荐方法研究[D].天津理工大学,2021.DOI:10.27360/d.cnki.gtlgy.2021.000372. [↑](#footnote-ref-7)
8. 陈健巧. 普惠金融背景下P2P投资者的借贷行为影响因素研究[D].上海交通大学,2022.DOI:10.27307/d.cnki.gsjtu.2020.003769. [↑](#footnote-ref-8)
9. 刘雨杭. 股市投资者情绪对借贷市场的影响[D].吉林大学,2023.DOI:10.27162/d.cnki.gjlin.2022.005316. [↑](#footnote-ref-9)
10. 戴杨文意. 大学生网络借贷行为问题现状及其社会工作介入策略[D].华南理工大学,2021.DOI:10.27151/d.cnki.ghnlu.2020.004597. [↑](#footnote-ref-10)
11. 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. [↑](#footnote-ref-11)
12. 张丽改. P2P网络借贷平台信用评级的有效性分析[D].山东大学,2023.DOI:10.27272/d.cnki.gshdu.2023.007825. [↑](#footnote-ref-12)
13. 姜晨. 我国P2P网贷风险识别研究[D].北方工业大学,2022.DOI:10.26926/d.cnki.gbfgu.2021.000656. [↑](#footnote-ref-13)
14. 雷佳. 基于机器学习的P2P个人信贷风险研究[D].对外经济贸易大学,2022.DOI:10.27015/d.cnki.gdwju.2020.000538. [↑](#footnote-ref-14)
15. 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. [↑](#footnote-ref-15)
16. 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. [↑](#footnote-ref-16)
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].北京科技大学,2023.DOI:10.26945/d.cnki.gbjku.2023.000344. [↑](#footnote-ref-17)
18. Wijaya, C., Nugroho, B. Y., & Arkanuddin, M. F. (2022). The Analysis of Risk Effect Towards Fintech Ecosystem On P2P Lending Industry in Indonesia. Indonesian Journal of Business and Entrepreneurship. https://doi.org/10.17358/ijbe.8.3.481 [↑](#footnote-ref-18)
19. Senney, G.T., Lhost, J.R. Balancing Attraction and Risk Revelation: The Optimal Reservation Price in Peer-to-Peer Loan Auctions. Rev Ind Organ 64, 289–314 (2024). https://doi.org/10.1007/s11151-023-09914-0 [↑](#footnote-ref-19)
20. LYÓCSA, Štefan, Petra VASANICOVA and Oleg DEEV. Peer-to-peer loan returns: heterogeneous effects across quantiles. Applied Economics Letters. UK: Taylor & Francis, 6 pp. ISSN 1350-4851. doi:10.1080/13504851.2023.2298412. 2024. [↑](#footnote-ref-20)
21. Ruqoyah, R., Afriani, Y., Fahleti, W. H., Arifin, N. Y., Zarvianti, A. A., & Ramadhan, A. R. (2023). Analysis of Public Opinion on the Utilization of Online Loan Applications. Tambusai Journal of Education, 7(3), 22820–22825. https://doi.org/10.31004/jptam.v7i3.10210 [↑](#footnote-ref-21)
22. ŠPANKO, Martin. Performance of credit risk models in P2P lending Online. Master's thesis. Praha: University of Economics, Prague. 2024. Available from: https://theses.cz/id/sbxnv8/. [cit. 2024-04-08]. [↑](#footnote-ref-22)
23. Senney, G.T., Lhost, J.R. Balancing Attraction and Risk Revelation: The Optimal Reservation Price in Peer-to-Peer Loan Auctions. Rev Ind Organ 64, 289–314 (2024). https://doi.org/10.1007/s11151-023-09914-0 [↑](#footnote-ref-23)
24. LYÓCSA, Štefan, Petra VASANICOVA and Oleg DEEV. Peer-to-peer loan returns: heterogeneous effects across quantiles. Applied Economics Letters. UK: Taylor & Francis, 6 pp. ISSN 1350-4851. doi:10.1080/13504851.2023.2298412. 2024. [↑](#footnote-ref-24)
25. Artem Bequé, Stefan Lessmann. Extreme learning machines for credit scoring: An empirical evaluation[J]. Expert Systems With Applications, 2017, 86: 24-27. [↑](#footnote-ref-25)
26. ŠPANKO, Martin. Performance of credit risk models in P2P lending Online. Master's thesis. Praha: University of Economics, Prague. 2024. Available from: https://theses.cz/id/sbxnv8/. [cit. 2024-04-08]. [↑](#footnote-ref-26)
27. 刘翱,邓旭东,童泽平等.P2P网络借贷研究进展[J].系统工程学报,2020,35(03):402-415.DOI:10.13383/j.cnki.jse.2020.03.010. [↑](#footnote-ref-27)
28. WijayaC., NugrohoB. Y., & ArkanuddinM. F. (2024). The Important Role of Financial Architecture Regulation Toward Fintech P2P Lending Ecosystem. Indonesian Journal of Business and Entrepreneurship (IJBE), 10(1), 13. https://doi.org/10.17358/ijbe.10.1.13 [↑](#footnote-ref-28)
29. 蔡青松,吴金迪,白宸宇.基于可解释集成学习的信贷违约预测[J].计算机系统应用,2021,30(12):194-201.DOI:10.15888/j.cnki.csa.008220. [↑](#footnote-ref-29)