基于双层 Stacking 融合模型的 P2P 贷款研究

李婷

（东北大学 数学与统计学院）

摘　要：本文构建了双层 Stacking 融合模型，旨在研究影响借款人当前贷款状态的因素。数据预处理阶段，根据选定的两个指标将原始的 145 个特征列降维至 59 维。通过网格搜索对 SVM、支持向量机、决策树和逻辑回归等四个模型进行调参，并使用 5 折交叉验证查看过拟合程度，发现神经网络的分类性能最为出色。随后，通过因子降维，提取 30 个关键特征，累计方差解释率达 90.56%，并结合 AdaBoost、GBDT、CatBoost 和 LightGBM 这四个集成分类器，形成 Stacking 的第一层模型，第二层使用逻辑回归整合四个集成模型的输出，最后的结果显示 Stacking 的分类准确度达到 92.54%。结合互信息特征重要性选择，发现最终支付总额在评估借款人的贷款状态中具有关键作用。

关键词：Stacking；网格搜索；因子降维；互信息

中图分类号：请查阅中图分类号　**文献标志码：**A

Research on P2P Loan Based on Double Stacking Fusion Model

LI Ting

(School of Mathematics & Statistics, Northeastern University at Qinhuangdao, Qinhuangdao, Hebei 066004, China. E-mail: 1424743982@qq.com)

Abstract: This article presents a double-layer Stacking ensemble model, aiming to investigate the factors influencing the current loan status of borrowers. During the data preprocessing stage, the original 145 feature columns were dimensionally reduced to 59 dimensions based on the selection of two key indicators. A grid search was conducted to fine-tune four models, including SVM, Support Vector Machine, Decision Tree, and Logistic Regression. Overfitting was evaluated using 5-fold cross-validation, and it was found that the neural network exhibited the best classification performance.Subsequently, through factor dimensionality reduction, 30 key features were extracted, explaining a cumulative variance of 90.56%. These features were integrated with four ensemble classifiers: AdaBoost, GBDT, CatBoost, and LightGBM, forming the first layer of the Stacking model. In the second layer, Logistic Regression was utilized to combine the outputs of the four ensemble models. The final results demonstrated that Stacking achieved a classification accuracy of 92.54%. Furthermore, in conjunction with mutual information feature importance selection, it was discovered that the total payment amount played a crucial role in assessing the borrower's loan status.

Key words: stacking; grid search; dimensionality reduction with Factors; mutual information

P2P（Peer-to-Peer）金融是指个体之间的小额借贷交易，其模式版图可简要概括为借款者主动发布借款信息，包括借款金额、利率、还款方式和期限，实现自助式借款；而出借人则根据借款者的信息自主决定借出的金额，实现自助式借贷[1]。这种 P2P 金融模式是互联网技术和金融业的结合，以低门槛、低成本、高透明度、交易简便等特点为依托，为个人借贷提供了一种全新的选择[2]。Lending Clubs 是其中的一家知名 P2P 平台，它以其高度透明的操作和广泛的参与度，为借款者和出借人提供了安全、便捷的借贷环境。同时，这种模式也为中小微企业的融资需求提供了更多机会，对经济发展起到了积极作用，备受借贷者广泛关注和参与。

本文基于 Lending Club 公布的部分借贷数据，构建多个模型分析数据的内在关联，剖析影响借款人当前贷款状态的因素，帮助金融机构优化相关的评估机制，提供决策支持。

1　数据预处理

根据所提供的数据集，不难发现，2016 年、2017 年、2018 年这四个季度的数据中，唯有第二季度的数据是完整的，而其他季度的数据在某些年份均存

收稿日期：2023-10-25
基金项目：机器学习结课论文。
作者简介：李婷(2003-),女,湖南省娄底人, 东北大学学生,本科生.

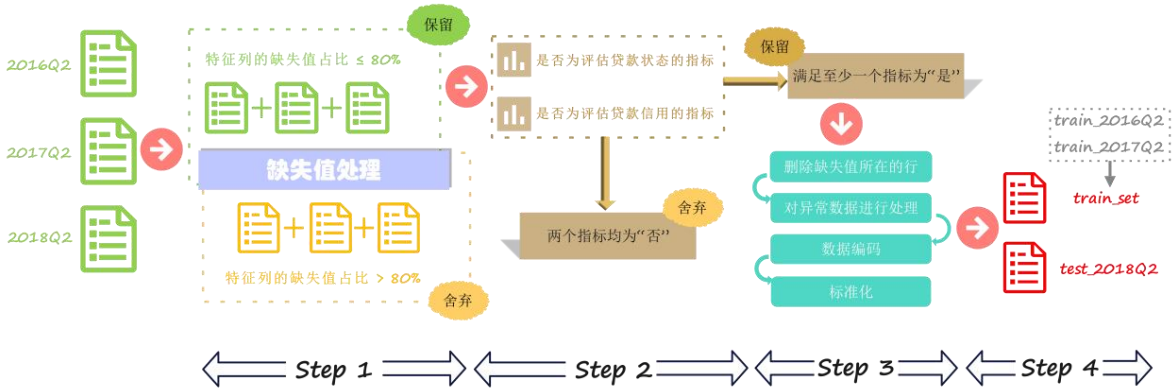

图 1 数据预处理流程图

Fig.1 Flow chart of data preprocessing

在不同程度的缺失。在本研究中，依据第二季度，单独选取 2016 年到 2018 年的数据，并将“loan_status”（贷款的当前状态）列作为标签列。经统计，选取的每个数据集中均有 145 个字段。

在初步审阅原始数据时，易发现文件中的数据类型不一致，若要进行下一步研究，首先就需要对原始数据进行系统地筛选和处理。数据预处理的基本流程如图 1 所示，主要分为四大步骤。

🕮 步骤一：缺失值处理

首先，针对三个文件中的每一列，进行缺失值统计。为了确保三个文件中字段的一致性，只要其中任何一列的缺失比例超过 80%，视该列数据的参考价值有限，本文不予保留；

🕮 步骤二：特征列筛选

原文件的特征字段较为繁多，且包含冗余信息。本文依据两关键指标“是否为评估贷款状态的指标”和“是否为评估贷款信用的指标”，若字段列满足其中两个指标中至少有一个为“是”，则保留该字段。筛选后，共留下 59 个字段；

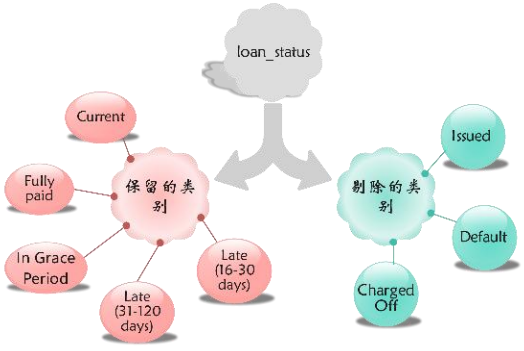

图 2 loan_status 的类别删减

Fig.2 Category deletion of loan_status

🕮 步骤三：特征工程

先剔除每个文件中存在缺失值的行，处理后对比发现，三个文件中的“loan_status”标签列存在类别差异，这将不利于后续的模型评估，因此本文剔除了“Charged Off”、“Default”、“Issued”这三种还款状态（如图 2 所示），但是不同列的数据类型仍存在差异。本文将特征分为类别型和数值型，针对类别型，进行标签编码，将特征列的 n 个类别映射成 n 个不同取值的数值；针对数值型，进行了 Z-Score 标准化处理，消除量纲的影响。

🕮 步骤四：数据集划分

经过上述的数据处理，所获得的第二季度数据中，2016 年包含了 15600 条记录，2017 年包含了 18899 条记录，2018 年包含了 15320 条记录。本文将 2016 年和 2017 年的数据合并作为一个训练集，而 2018 年的数据则单独作为测试集，用于评估模型的性能。

2 模型的初步构建与评估

网格搜索是机器学习中广泛用于超参数调优的技术，通过穷尽性搜索预定义参数组合来提高模型性能和泛化能力。本文着重探讨数据分类问题，在进行支持向量机、神经网络、决策树和逻辑回归等四个模型的训练时，运用网格搜索进行参数优化。

2.1 模型的初步搭建

➔ 支持向量机

支持向量机（SVM）的核心原理在于寻找一个最佳的超平面，将不同类别的样本有效分离，以实现高效的分类。通过网格搜索，确定了 SVM 的最佳参数配置：**正则化系数 C 为 100，核函数为径向基函数（RBF）。**

➔ 神经网络

神经网络是利用多层神经元构建模型，其中每层神经元将输入特征进行加权组合，并通过激活函数来输出结果。通过网格搜索，确定了神经网络的最佳参数配置：**隐藏层个数为 2（第一个隐藏层包含 15 个隐节点，第二个隐藏层包含 10 个隐节点），激活函数为 ReLu 函数。**

➔ 决策树

决策树是根据输入特征选择最佳分裂点，递归划分数据，以构建一棵树，使其能够对数据进行有效分类和预测。通过网格搜索，确定决策树的最佳参数配置：**树的最大深度为 10，每个节点的最少样本数为 10**。

➔ 逻辑回归

逻辑回归针对多分类采用一对多策略，即通过训练多个独立的二元逻辑回归分类器，每个分类器负责区分一个类别与其他类别，最终通过综合各个分类器的预测结果来实现多类别分类。通过网格搜索，确定逻辑回归的最佳参数配置：**正则化系数为 100，求解器为 liblinear**。

2.2 评估方法

本文选用均方根误差、分类准确度、F1 分数、Kappa 统计量评估模型分类性能，选用交叉验证精度评估模型的过拟合程度。

- 均方根误差（RMSE）

用于衡量预测值与真实值之间差异。

$$RMSE\left(y,\hat{y}\right)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y_i}\right)^2}$$

- 分类准确度（Accuracy）

用于衡量模型的分类性能。

$$Accuracy=\frac{\text{正确分类的样本数}}{\text{总样本数}}$$

- F1 分数（F1 score）

为精确度（Precision）和召回率（Recall）的调和平均，衡量了模型的准确性和完整性。

$$F1=\frac{2\cdot Precision\cdot Recal}{Precision+Recall}$$

- Kappa 统计量

用于度量分类模型的准确性，并考虑了随机分类的因素，取值范围[-1,1]，正值表示模型的准确性优于随机猜测。

$$\mathcal{K}=\frac{P_0-P_e}{1-P_e}$$

其中，$\mathcal{K}$ 表示 Kappa 统计量，P_0 表示观测准确性，P_e 表示预期的准确性。

- 交叉验证精度

侧重于评估模型是否过拟合，在不同的子样本上进行多次训练和测试，取结果的平均值。

$$CV_{(k)}=\frac{1}{k}\sum_{i=1}^{k}MSE_i$$

本文取 k 值为 5，即 5 折交叉验证。

3 模型的结果分析

表 1 模型的评估结果

Tab.1 Evaluation results of the model

模型	RMSE	F1	Kappa 系数	准确度
SVM	0.67	0.53	0.16	0.80
神经网络	0.70	0.78	0.39	0.92
决策树	0.97	0.49	0.08	0.64
逻辑回归	0.90	0.48	0.08	0.66

表 2 各模型交叉验证的精度

Tab.2 Accuracy of cross-validation of each model

模型	SVM	神经网络	决策树	逻辑回归
交叉验证精度	0.91	0.95	0.95	0.93

表 3 调参后模型的训练时长（单位：秒）

Tab.3 Training duration of the model after parameter adjustment (unit: seconds)

模型	SVM	神经网络	决策树	逻辑回归
运行时间	93.74	79.11	1.21	61.16

根据表 1 中模型的评估结果，**以神经网络分类性能最为出色**，其 F1 分数、Kappa 系数和分类准确度显著优于其它模型，而相对较低的 RMSE 值则表明其预测误差相对较小。相比之下，决策树和逻辑回归在此数据集中的分类效果相对不佳，仍有改进的空间；值得注意的是，尽管四个不同的模型在评估结果上存在差异，但根据表 2 的数据，各模型在不同的数据子集上表现稳定，交叉验证精度均达到了 90%以上，这表明各模型没有明显的过拟合现象；在时间的花销上，表 3 中决策树的速度量级明显优于其他三个模型，尽管神经网络先前展现出较高的分类性能，但其时间复杂度较高。

4 属性的重要性分析

在进行属性重要性分析时，需要评估 58 个特征列与标签列“loan_status”之间的关联程度，这有助于确定每个特征的相对重要性。因此，本研究采用了**基于互信息的特征选择方法**，以评估特征与目标变量之间的相关性。互信息度量了两个变量之间的信息共享程度，即一个变量包含关于另一个变量的信息量，且能同时考虑线性和非线性关系[3]。在特征选择过程中，互信息值越高，表明特征与标签列之间的依赖性越强，该特征的重

要性也越就大。

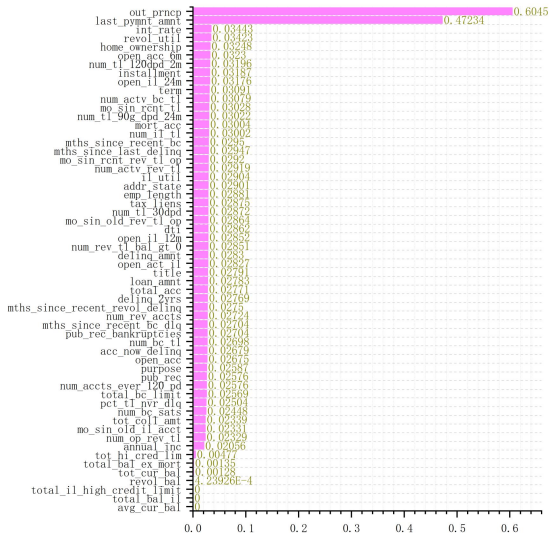

图 3 58 个特征的相对重要性

Fig.3 Relative importance of 58 characteristics

图 3 中展示了各特征与标签列“loan_status”之间的重要性程度。由图 3 可知，“out_prncp”和“last_pymnt_amt”这两个特征的互信息分数分别为 0.6045 和 0.4723，明显高于其他特征，表明其对预测借款人的贷款状态具有较大影响。其中，“out_prncp”反映了借款人尚未清偿的债务，可作为评估借款人债务负担和偿还能力的重要指标；“last_pymnt_amt”反映了最后收到的总还款金额，与借款人是否会按时还款存在较大关联。以上结果证明了借款人当前的财务状况对其还款能力具有直接影响。

通过分析特征重要性，建议在贷款违约风险评估和信贷决策过程中，应及时监控和调查借款人的资金状况。尤其是金融机构和贷款机构应更加重视借款人未偿还的本金和总还款金额，以更准确地评估潜在的违约风险，并采取相应的措施来降低不良贷款的风险。

5 特征降维

5.1 因子降维的可行性

因子降维的目标是从大量观测变量中提取具有代表性的因子，这些因子能够最大程度地保留原始数据的信息，从而减少数据中的噪声和冗余信息。在进行因子分析之前，通常需要验证 Bartlett’s 球状检验和 KMO 检验是否达标。在本研究中，Bartlett’s 球状检验的 p 值接近 0，这表明可以拒绝球状性假设，从而验证特征之间存在相关性；KMO 检验得出的值为 0.6577，超过了 0.6 的阈值，这符合进行因子分析的前提条件，证明了本文的数据集适合进行因子降维。

5.2 因子旋转

经过 Varimax 方差最大化因子旋转，获得了 30 个累计方差解释率高达 90.56%的特征。图 4 展示了这 30 个特征之间的相关性，可以看到，不同因子之间的相关性较小，在一定程度上能够表示相对独立的特征，进一步验证了本文进行的因子降维是有效的。

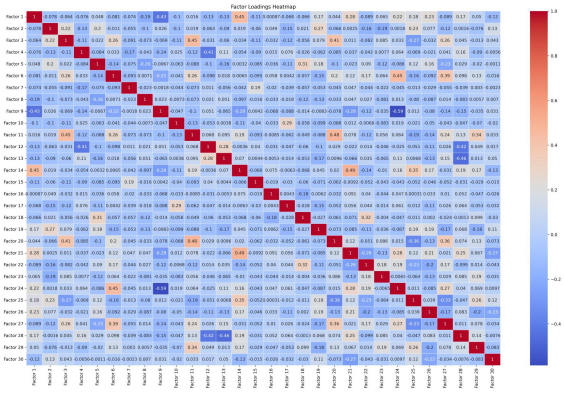

图 4 因子旋转后的载荷图

Fig.4 Load diagram after factor rotation

6 多模型融合

此部分，本文构建了双层 Stacking 融合模型。

6.1 Stacking 第一层模型

- AdaBoost

AdaBoost 通过串行训练多个弱分类器并根据它们的性能调整权重，从而构建一个强大的分类器。在每一轮训练中，AdaBoost 聚焦于之前分类错误的样本，提高这些样本的权重，使得模型更加关注难以分类的样本。

- GBDT

GBDT 直接采用当前决策树损失函数的负梯度作为残差的近似值，用来构建新的决策树。核心思想在于使用一种最速下降的近似方法，即通过朝着损失函数梯度的反方向不断减小损失函数的值。

- CatBoost

CatBoost 采用对称树结构和基于贪婪学习的方法以提高模型的训练速度和性能，在处理大规模数据集时表现出色。

- LightGBM

LightGBM 是一种基于直方图的梯度提升框架，不仅支持类别特征的自动处理，还具有出色

的并行学习能力，能够处理大规模数据集并且在保持准确性的同时，提高训练速度。

6.2 Stacking 第二层模型

在第一层模型训练的基础上，本文采用逻辑回归作为 Stacking 模型的第二层[4]映射输出，并采用 **2.2** 中的评估指标对 Stacking 双层模型的各分类器进行评估，如图 5 所示。

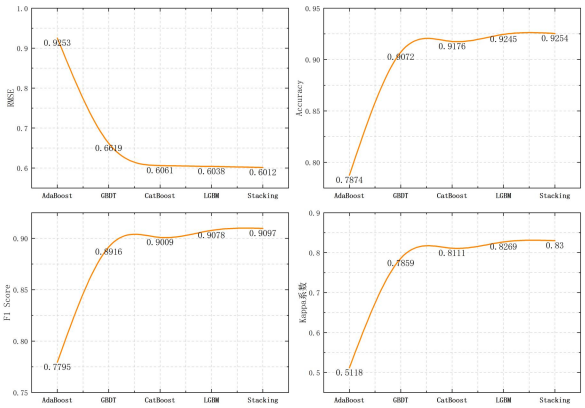

图 5 Stacking 子模型的评估可视化结果

Fig.5 Visualization results of Stacking sub-model evaluation

由图 5 可以看出，Stacking 第二层模型的评估优度均高于第一层各分类器，且通过了交叉验证。对比表 1 所呈现的基础模型性能，可见经过优化提升后的 Stacking 融合模型在数据集上的表现更为出色。根据 Stacking 最后一层的输出结果，对 30 个特征的重要性进行降序排序，如图 6 所示。

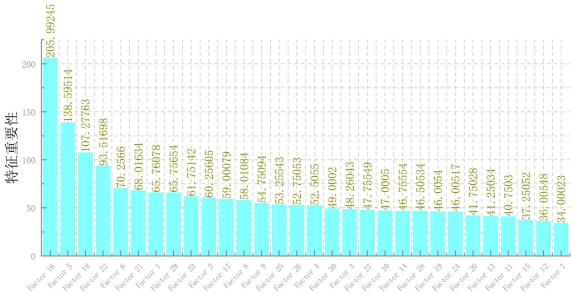

图 6 30 个新特征的重要性排名

Fig.6 Importance ranking of 30 new features

根据因子降维中 Factor16 所包含的信息，last_pymnt_amnt（最后收到的总付款金额）的影响因子显著高于其他特征，而图 6 中 Factor16 的重要性又显著高于其余 29 个 Factor 因子。由此证明了借款人最后支付的总金额在一定程度上会影响当前的贷款状态，这与之前互信息属性重要性分析中得出的主要结论不谋而合。

7 结论

1) 不同模型在同一数据集上的表现存在差异，分析具体问题时要选择合适的模型，不能一概而论。

2) 在分析多个特征与标签之间的关联时，剔除冗余的特征列会更加凸显主要特征列对标签的影响。

3) 通过因子降维得到的各个特征，更具有代表性，在双层 Stacking 融合模型中，有利于揭示对借款人贷款状态产生影响的主要因素，进而为金融决策提供可靠的数据支持。

参考文献

[1] 马子岳. 互联网金融对中小企业融资的影响及优化策略 [J]. 商场现代化, 2023(19):140-142. DOI:10.14013/j.cnki.scxdh.2023.19.002.
(Ma Ziyue. The impact of Internet finance on SME financing and optimization strategies [J]. Shopping Mall Modernization, 2023 (19): 140-142. Doi: 10.14013/j.cnki.scxdh.2023.19.002.)

[2] 倪雪莉, 马卓, 王群. 区块链 P2P 网络及安全研究 [J/OL]. 计算机工程与应用:1-15[2023-10-21].
(Ni Xueli, Ma Zhuo, Wang Qun. Research on P2P network and security of blockchain [J/OL]. Computer Engineering and Application: 1-15[2023-10-21].)

[3] 张瑾, 姜浩, 金秀章. 基于互信息变量选择的燃煤机组 SCR 脱硝系统 PSO-ELM 建模[J]. 网络安全与数据治, 2023, 42(09):88-95. DOI:10.19358/j.issn.2097-1788.2023.09.013.
(Zhang Jin, Jiang Hao, Jin Xiuzhang. PSO-ELM modeling of SCR denitration system of coal-fired units based on mutual information variable selection [J]. Network Security and Data Governance, 2023,42 (09): 88-95.DOI: 10.19358/J.ISSN.2097-1788.2007.2009)

[4] 陈旭然. 基于集成学习的 P2P 网贷用户的违约预测研究 [D]. 重庆大学, 2022. DOI:10.27670/d.cnki.gcqdu.2022.001921.
(Chen Xuran. Research on the Default Prediction of P2P Online Lending Users Based on Integrated Learning [D]. Chongqing University, 2022. DOI: 10.27670/d.cnki.gcqdu.20022.200100000006)