Advances in Applied Mathematics 应用数学进展, 2022, 11(4), 1657-1667
Published Online April 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.114180

Hans汉斯

基于Bagging集成方法的互联网金融信用风险评估

陈凯玥

河北工业大学理学院，天津

收稿日期：2022年3月14日；录用日期：2022年4月8日；发布日期：2022年4月18日

摘　要

互联网金融是对传统金融模式的延伸，但由于部分借款人在借款后无法按期、足额还款，使得互联网金融平台面临着信用风险。对借款人的信用风险进行准确评估，可以降低风险，并且能够在一定程度上为互联网金融行业的稳定发展提供保障。数据分析方法在信用风险评估领域已有广泛应用。本文从国内某互联网金融平台借款人的个人、资产、借款信息三类数据提取特征，研究了数据分析方法中Logistic回归的衍生方法逐步Logistic回归、弹性网络和Bagging集成方法的代表Bagging、极端随机树和随机森林。研究发现随机森林与逐步Logistic回归分别在F1-score、Accuracy、FPR和AUC指标下效果最优，且筛选出的重要特征也保持一致。

关键词

信用风险评估，Logistic回归，Bagging集成，特征重要性

Internet Financial Credit Risk Assessment Based on Bagging Ensemble Method

Kaiyue Chen

School of Science, Hebei University of Technology, Tianjin

Received: Mar. 14th, 2022; accepted: Apr. 8th, 2022; published: Apr. 18th, 2022

Abstract

Internet finance is an extension of the traditional financial model. As some borrowers are unable to repay in full or on time, the platform faces credit risks. Accurate assessment of the credit risk can reduce losses and provide a guarantee for the stable development of the Internet finance in-

文章引用: 陈凯玥. 基于 Bagging 集成方法的互联网金融信用风险评估[J]. 应用数学进展, 2022, 11(4): 1657-1667.
DOI: 10.12677/aam.2022.114180

dustry. Data analysis methods have been widely used in the field. This paper extracted features from the three types of personal, asset and loan information. Stepwise Logistic regression and Elastic Net which are the derivative methods of Logistic regression, and the representative of Bagging ensemble method, Bagging, Extremely Randomized Trees and Random Forest were studied. It is found that Random Forest and Stepwise Logistic regression have the best results under F1-score, Accuracy, FPR and AUC indicators respectively, and the important features selected are also consistent.

Keywords

Credit Risk Assessment, Logistic Regression, Bagging Ensemble, Feature Importance

1. 研究背景及意义

互联网的飞速发展为金融机构提供了新的工作方法和服务方向。互联网金融是一种利用互联网技术和服务，将金融和互联网有效结合的模式，通常定义为互联网公司从事金融活动，如微信支付、蚂蚁金服、京东白条等。截至 2020 年底，我国手机网络支付用户规模达 8.54 亿，占手机网民 86.4%，全国数字支付交易规模突破 200 万亿元[1]。

互联网金融在国内实现跨越式发展的同时也面临着风险，目前最主要的是信用风险，表现为借款人无法按期、足额还款等。造成这种现象的主要原因是：互联网金融平台的主要服务对象通常是传统金融模式无法服务的企业和个人，他们通常收入不稳定、自身贫困、偿还能力不强；目前我国的信用登记机制还不够完善，平台并不能对借款人的真实信用状况做出准确的评估。信用风险一旦发生会对互联网金融平台造成不可逆的影响，给投资者和借贷平台带来严重损失，因此需要对借款人的信用风险进行准确评估。

数据分析方法已在信用风险评估领域有深入研究。基于互联网大数据个人信用风险评分系统能够预测网络借款人的违约风险[2]。利用数据挖掘技术对借款人的交易数据进行探索，来评估借款人的信用风险，可以降低信用风险，有效保障平台和投资者的利益，并且能够在一定程度上为互联网金融行业的稳定发展提供保障。

2. 文献综述

构建信用风险评估模型的方法主要包括统计方法和非统计方法。统计方法是通过构建统计模型描述信用风险问题中的函数关系，从而实现风险评估的量化分析方法。Moscatelli M 等(2020)发现线性判别分析(LDA)在预测非金融公司的破产等方面表现良好并应用广泛[3]。方匡南等(2014)将 Lasso-Logistic 模型引入个人信用评估[4]，韦勇凤等(2019)利用 Group-Lasso 方法对某商业银行信用卡数据进行变量选择，构建基于 Logistic 回归的信用评分模型[5]。由于 Logistic 回归没有对非线性或复杂交互进行考虑，并且对异常值和缺失数据不敏感，而对 Logistic 回归的弹性网络正则化可改善此问题[3]。王小燕等(2021)将弹性网络(Elastic Net)的惩罚项与 Logistic 结合构建了 PIPL 模型评估贷款信用风险[6]，Dayu Xu 等(2020)在特征选择时使用弹性网络来减少弱相关或不相关的变量[7]。

因为实际应用中的数据无法达到统计方法需要的严格假定条件，所以具有局限性。非统计方法是利用计算机技术的不需要严格假定条件的机器学习技术。Bekhet H A 等(2014)利用人工神经网络(ANN)和 Logistic 回归对约旦商业银行贷款决策进行信用风险评估，发现在识别违约用户方面人工神经网络优于 Logistic 回归[8]。王程龙等(2016)发现决策树在构建 P2P 平台信用评级体系方面表现出适用性强、精度高、可解释性强的优势[9]。单一的方法会因数据结构、特征选择、研究问题等不同而表现出不同的精度，此问题可通过集成学习方法改进[10]。Gang Wang 等(2010)以 Logistic 回归、决策树等为基学习器，对 Bagging、Boosting 和 Stacking 集成思想进行比较性能评估[10]，Li Yiheng 等(2020)进一步研究随机森林、AdaBoost、XGBoost、LightGBM 和 Stacking 在信用风险评估领域的表现，结果表明除 AdaBoost 外集成学习优于单个学习器[11]，Zhenya Tian 等(2020)选取梯度提升决策树(GBDT)进行信用风险评估[12]。莫赞等(2019)针对 UCI 中数据集将 GBDT 分别与 Logistic 回归和支持向量机(SVM)结合后利用 Bagging 集成[13]。白鹏飞等(2017)等以 Logistic 回归为基准，选用随机森林、XGBoost 和 SVM 建立信用预测模型并进行投票加权融合[14]。为减少相关性较弱的变量对模型效果的影响，操玮等(2018)发现利用随机森林选取相对重要性较高的变量比用全变量构建模型的精度高[15]，周永圣等(2020)构建 XGBoost-RF 模型，先利用 XGBoost 筛选重要特征后再利用随机森林建模，预测效果在 AUC 值的表现上有所改进[16]。

综上所述，Logistic 回归在构建信用预测模型中不仅是应用广泛的单一学习器，而且也可以作为集成方法的基学习器[3] [10] [11]。在该方法的基础上，加入惩罚项的改进算法 Lasso-Logistic、弹性网络等也是经典的统计方法[3] [4] [5] [6]。集成方法可以将多个单一学习器的缺点进行改进，其中的研究重点随机森林是 Bagging 方法中 Random Patches 的代表[11] [14] [15] [16]。Gang Wang 等(2010)研究表明 Bagging 在文中数据集上表现均优于 Boosting [10]，而在已有的文献中，学者多是研究 Bagging 方法的某一分支。本文以逐步 Logistic 回归、弹性网络为基准，探究 Bagging 集成方法的三种算法思想 Bagging、Random Sbuspace 和 Random Patches 在互联网金融的信用风险评估领域的表现。

3. 模型建立

3.1. 逐步 Logistic 回归

逐步 Logistic 回归(Step-LR)是将 Logistic 回归和逐步回归思想相结合的算法。Logistic 回归是一种最常见的广义线性模型，以违约概率本身的两个值作为因变量 Y，信用良好的借款人 $Y=0$，违约借款人 $Y=1$。借款人的特征变量为 $X=\left(x_1,x_2,\cdots,x_p\right)^{\mathrm{T}}$，违约概率表示为

$$P\left(Y=1\middle|X\right)=\frac{\exp\left\{X^T\beta\right\}}{1+\exp\left\{X^T\beta\right\}} \tag{1}$$

假设选择了 n 个借款人作为样本，$\left(x_i,y_i\right)$，$i=1,2,\cdots,n$ 表示第 i 个样本的观测值，其中 $x_i=\left(x_{i1},x_{i2},\cdots,x_{ip}\right)^{\mathrm{T}}$。则 Logistic 回归的损失函数为

$$J\left(\beta\right)=\sum_{i=1}^{n}\left[-y_i\left(x_i^T\beta\right)+\log\left(1+\exp\left\{x_i^T\beta\right\}\right)\right] \tag{2}$$

逐步回归的基本思想是将自变量逐个引入回归方程，引入的条件是其偏回归平方和经检验后是显著的。每引入一个自变量后要对已经选入的解释变量逐个进行 t 检验，剔除偏回归平方和不显著的自变量。此过程迭代至回归方程中既无新变量引入也无旧变量删除为止。

3.2. 弹性网络

弹性网络(Elastic Net)是一种使用 L1、L2 范数作为先验正则项训练的线性回归模型，损失函数为

$$J(\beta)=\sum_{i=1}^{n}\left[-y_i\left(x_i^T\beta\right)+\log\left(1+\exp\left\{x_i^T\beta\right\}\right)\right]+\lambda\sum_{j=1}^{p}\left(\alpha\left|\beta_j\right|+(1-\alpha)\beta_j^2\right) \tag{3}$$

α 用来调节 L1 和 L2 范数的凸组合，λ 用来调节模型复杂度的惩罚项系数。

3.3. Bagging 方法

Bagging 方法是并行集成的代表，根据抽样策略不同分为 Bagging (仅对样本集抽样)、Random Subspace (仅对特征集抽样)和 Random Patches (既对样本集抽样也对特征集抽样)，而极端随机树(Extremely Randomized Trees, ET)、随机森林(Random Forest)分别是 Random Subspace 和 Random Patches 在决策树上的算法特例。

Bagging 算法使用全部特征，通过一次并行采样获得大量数据子集训练基学习器，并组合基学习器预测结果进行输出。

当数据集有限但特征较多时，为了保证基学习器之间的差异性，选择对特征进行采样而使用全样本集进行训练，即为 Random Subspace 算法，其中的算法特例是极端随机树(Extremely Randomized Trees, ET)。极端随机树在分支阈值选择时用的方法并非递归二叉分裂，而是随机选取分叉的阈值。

当样本量与特征量都有限时，只对样本集或特征集采样都无法获得有足够差异的基学习器，此时需要同时对二者进行采样，即为 Random Patches 算法，随机森林(Random Forest, RF)为代表算法。随机森林在特征集中随机选择一定规模的特征子集，再从中选择最优的特征进行划分。

4. 数据选择与处理

本文使用的数据集来自于中国某互联网金融平台，共有 13,681 条样本，包含 12,151 个正常还款用户和 1530 个有违约记录用户。廖理等(2015)借助 P2P 数据研究发现高学历借款者按期还款概率更高[17]，说明个人信息可以作为预测违约的特征。本文通过指标转换、数值型指标标准化，选择了包含个人信息、资产信息、借款信息三大类的 18 个特征作为解释变量；特征“标的状态”作为被解释变量，取“0”代表“好”用户(正常还款用户)，“1”代表“坏”用户(有违约记录用户)，具体特征的指标说明如表 1 所示。

Table 1. Description of characteristic indicators
表 1. 特征指标说明

一级指标	序号	二级指标	数据类型
个人信息	X_1	出生年代	0-50 后；1-60 后；2-70 后；3-80 后；4-90 后
	X_2	性别	0-男；1-女
	X_3	学历	0-高中或以下；1-大专；2-本科；3-研究生或以上
	X_4	婚姻	0-已婚；1-未婚；2-离婚或丧偶
	X_5	收入	0-1000 元以下；1-1000~2000 元；2-2000~5000 元；3-5000~10,000 元；4-10,000~20,000 元；5-20,000~50,000 元；6-50,000 元以上
	X_6	公司规模	0-10 人以下；1-10~100 人；2-100~500 人；3-500 人以上
	X_7	工作性质	0-网商；1-工薪阶层；2-私营企业主
	X_8	工作时间	0-1 年(含)以下；1-1~3 年(含)；2-3~5 年(含)；3-5 年以上
	X_9	是否外籍工作者	0-否；1-是

Continued

资产信息	X_{10}	房产	0-无；1-有
	X_{11}	房贷	0-无；1-有
	X_{12}	车产	0-无；1-有
	X_{13}	车贷	0-无；1-有
借款信息	X_{14}	标的总额	定量变量
	X_{15}	借款成功率	定量变量
	X_{16}	还清比率	定量变量
	X_{17}	年利率	定量变量
	Y	标的状态	0-好用户；1-坏用户

在该互联网金融平台上，信用额度是指借款人单笔借款的上限，则(信用额度 × 成功借款次数)为借款人借款总额上限，而数据集中 776 条样本的借款总额大于其上限，本文将此类样本定义为异常样本，其余为正常样本。

5. 结果分析

本文选取不同的训练集和测试集比例，分别对全样本数据集和正常样本数据集进行分类测试。为降低实验误差，对每一种算法在不同集合划分比例下重复实验 50 次，选取 F1-score、Accuracy、FPR 和 AUC 为指标对结果进行分析，得到以下结论。

5.1. 全样本数据集结果

对于全样本数据集进行建模，得分如表 2~5 所示。

从前三种指标观测，Bagging 集成方法优于 Logistic 回归，其中随机森林最佳，Logistic 回归中弹性网络得分略低于逐步 Logistic 回归。回归。从 AUC 值观测，逐步 Logistic 回归得分最高，其次是弹性网络，Bagging 集成方法略低于与逐步 Logistic 回归。在训练集与测试集划分比例为 8:2 时，各算法表现最好。

Table 2. F1-score of full sample dataset
表 2. 全样本数据集 F1-score

	5:5	6:4	7:3	8:2	9:1
Step-LR	0.9822	0.9828	0.9832	0.9842	0.9824
EN	0.9743	0.9739	0.9746	0.9761	0.9763
Bagging	0.9962	0.9965	0.9970	0.9974	0.9971
ET	0.9977	0.9976	0.9980	0.9980	0.9973
RF	**0.9984**	**0.9986**	**0.9989**	**0.9988**	**0.9984**

Table 3. Accuracy of full sample dataset
表 3. 全样本数据集 Accuracy

	5:5	6:4	7:3	8:2	9:1
Step-LR	0.9960	0.9962	0.9963	0.9965	0.9961
EN	0.9943	0.9942	0.9944	0.9947	0.9947
Bagging	0.9991	0.9992	0.9993	0.9994	0.9993
ET	0.9995	0.9995	0.9996	0.9996	0.9994
RF	**0.9996**	**0.9997**	**0.9998**	**0.9997**	**0.9996**

Table 4. FPR of full sample dataset
表 4. 全样本数据集 FPR

	5:5	6:4	7:3	8:2	9:1
Step-LR	1.590×10^{-3}	1.387×10^{-3}	1.383×10^{-3}	1.276×10^{-3}	1.547×10^{-3}
EN	1.214×10^{-3}	1.119×10^{-3}	1.131×10^{-3}	1.268×10^{-3}	1.070×10^{-3}
Bagging	6.317×10^{-4}	5.764×10^{-4}	5.923×10^{-4}	4.598×10^{-4}	5.594×10^{-4}
ET	1.020×10^{-4}	1.070×10^{-4}	6.044×10^{-5}	8.193×10^{-5}	9.850×10^{-5}
RF	$\mathbf{4.928 \times 10^{-5}}$	$\mathbf{4.938 \times 10^{-5}}$	$\mathbf{2.739 \times 10^{-5}}$	$\mathbf{2.462 \times 10^{-5}}$	$\mathbf{3.242 \times 10^{-5}}$

Table 5. AUC of full sample dataset
表 5. 全样本数据集 AUC

	5:5	6:4	7:3	8:2	9:1
Step-LR	**0.9994**	**0.9995**	**0.9995**	**0.9995**	**0.9995**
EN	0.9993	0.9993	0.9993	0.9994	0.9993
Bagging	0.9984	0.9985	0.9991	0.9990	0.9990
ET	0.9981	0.9980	0.9982	0.9983	0.9977
RF	0.9986	0.9987	0.9990	0.9989	0.9985

5.2. 正常样本数据集结果

去除异常样本，对正常样本数据集进行建模，各算法的违约用户识别效果如表 6~9 所示。

从 AUC 指标观测，Logistic 算法仍然比 Bagging 集成方法表现优秀，其中得分最高的是逐步 Logistic 回归，Bagging 集成中表现最好的是 Bagging 和随机森林，其在不同划分数据集比例下的均值与逐步 Logistic 回归相差 0.0007。从其余指标来看，Logistic 回归明显比 Bagging 集成效果差。按 8:2 划分时训练集与测试集比例时，各算法得分最高。

Table 6. F1-score of normal sample dataset
表 6. 正常样本数据集 F1-score

	5:5	6:4	7:3	8:2	9:1
Step-LR	0.9841	0.9847	0.9855	0.9850	0.9842
EN	0.9753	0.9756	0.9765	0.9759	0.9757
Bagging	0.9963	0.9965	0.9969	0.9972	0.9969
ET	0.9971	0.9972	0.9977	0.9976	0.9972
RF	**0.9982**	**0.9983**	**0.9988**	**0.9989**	**0.9990**

Table 7. Accuracy of normal sample dataset
表 7. 正常样本数据集 Accuracy

	5:5	6:4	7:3	8:2	9:1
Step-LR	0.9963	0.9965	0.9966	0.9965	0.9963
EN	0.9943	0.9944	0.9945	0.9944	0.9944
Bagging	0.9991	0.9992	0.9993	0.9993	0.9993
ET	0.9993	0.9994	0.9995	0.9994	0.9993
RF	**0.9996**	**0.9996**	**0.9997**	**0.9998**	**0.9998**

Table 8. FPR of normal sample dataset
表 8. 正常样本数据集 FPR

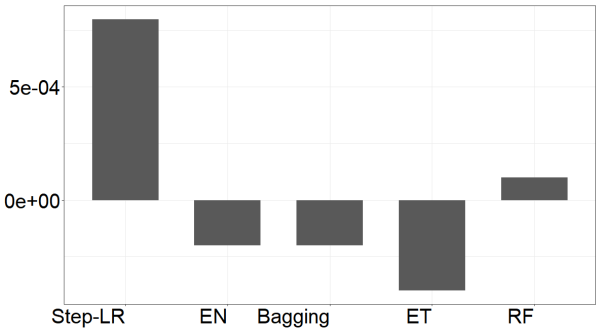
	5:5	6:4	7:3	8:2	9:1
Step-LR	1.462×10^{-3}	1.276×10^{-3}	1.148×10^{-3}	1.186×10^{-3}	1.511×10^{-3}
EN	1.374×10^{-3}	1.381×10^{-3}	1.317×10^{-3}	1.299×10^{-3}	1.580×10^{-3}
Bagging	6.042×10^{-4}	6.535×10^{-4}	6.148×10^{-4}	5.702×10^{-4}	6.842×10^{-4}
ET	1.861×10^{-4}	2.061×10^{-4}	1.407×10^{-4}	1.667×10^{-4}	2.285×10^{-4}
RF	$\mathbf{5.616 \times 10^{-5}}$	$\mathbf{7.017 \times 10^{-5}}$	$\mathbf{4.679 \times 10^{-5}}$	$\mathbf{2.630 \times 10^{-5}}$	$\mathbf{3.540 \times 10^{-5}}$

Table 9. AUC of normal sample dataset
表 9. 正常样本数据集 AUC

	5:5	6:4	7:3	8:2	9:1
Step-LR	**0.9995**	**0.9995**	**0.9996**	**0.9995**	**0.9995**
EN	0.9994	0.9993	0.9994	0.9993	0.9994
Bagging	0.9983	0.9986	0.9989	0.9990	0.9991
ET	0.9978	0.9979	0.9982	0.9982	0.9980
RF	0.9984	0.9986	0.9989	0.9990	0.9991

5.3. 不同数据集的预测精度比较

由于各算法之间的 Accuracy、FPR 和 AUC 得分相近，且上述结果表明在两种数据集下，训练集与测试集按照 8:2 划分最准确，所以本文选择在训练集与测试集划分比例为 8:2 时的 F1-score 作为观测两种数据集测精度的指标，如图 1 所示。正常样本数据集在逐步 Logistic 回归和随机森林上的 F1-score 比全样本数据集的表现好，而在弹性网络、Bagging、极端随机树上表现不如全样本数据集。由于在 F1-score、Accuracy 和 FPR 上随机森林是最为突出的算法，逐步 Logistic 回归在 AUC 值最高，且随机森林和逐步 Logistic 回归在去掉异常样本建模后精度有所提高。总体来看，识别异常样本后，利用正常样本数据集进行分析对违约用户识别的预测效果有所提升。



注：此图中展示的是正常样本集与全样本集下 F1-score 的差值。

Figure 1. F1-score of different datasets
图 1. 不同数据集的 F1-score

5.4. 特征重要性

由于在 F1-score、Accuracy 和 FPR 指标下，Bagging 集成方法中的随机森林得分最高，逐步 Logistic 回归的 AUC 比 Bagging 集成方法表现优秀，且对正常样本数据集以 8:2 划分训练接和测试集来建模的随机森林和逐步 Logistic 回归最为突出，所以本文针对随机森林和逐步 Logistic 回归，观察模型结果，分析其关于重要特征的筛选，如图 2、表 10 所示。

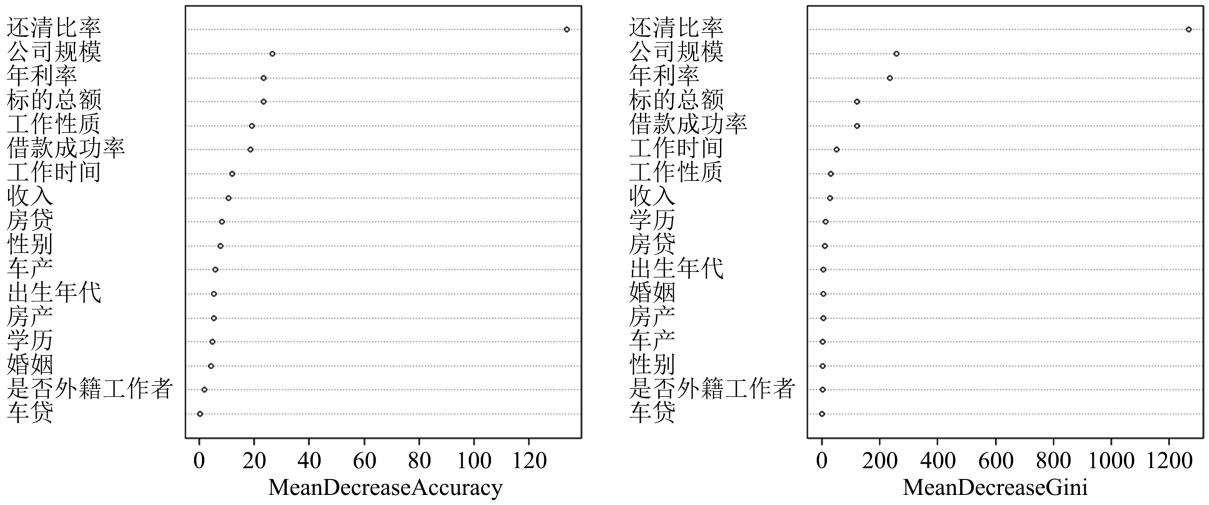


Figure 2. Feature importance of random forest
图 2. 随机森林特征重要性

从图 2 可以看出，随机森林根据均值精度和均值节点纯度所得到的特征重要性排序中，前 8 个自变量相同，第 9 个自变量在均值精度观测下为房贷，在均值节点纯度观测下为学历。综合来看，逐步 Logistic 回归和随机森林得到的重要特征基本保持一致，筛选出的特征为还清比率、公司规模、年利率、标的总额、工作性质、借款成功率、工作时间，说明在构建模型中上述变量起到重要作用。

从表 10 结果看出，AIC 值相比于原来的 Logistic 回归模型有所减小，且 Fisher 评分算法的迭代次数在合理区间内，说明逐步 Logistic 回归模型结果符合预测要求。变量显著性方面，性别、公司规模、工作性质、工作时间、车产、标的总额、借款成功率、还清比率、年利率 9 个自变量在 95%的显著性水平上不能拒绝原假设，说明其对识别违约用户具有重要影响。

Table 10. The summary of step Logistic regression results
表 10. 逐步 Logistic 回归模型结果摘要

	Estimate	Error	z value	Pr (>|z|)
(Intercept)	15.5170	2.1344	8.207	2.27e−16***
性别-1	−2.1374	0.6728	−3.177	0.00149**
公司规模-1	6.1349	0.7413	8.275	<2e−16***
公司规模-2	7.0044	0.8587	8.157	3.43e−16***
公司规模-3	6.5243	0.8497	7.679	1.61e−14***
工作性质-1	−8.2298	1.1323	−7.268	3.65e−13***
工作性质-2	−6.2449	1.1955	−5.224	1.75e−07***
工作时间-1	3.7579	0.8739	4.300	1.71e−05***
工作时间-2	3.9575	0.9105	4.346	1.38e−05***
工作时间-3	4.6857	0.8772	5.341	9.22e−08***
车产-1	1.1617	0.4321	2.688	0.00718**
标的总额	−1.2870	0.3011	−4.274	1.92e−05***
借款成功率	−2.2979	0.9648	−2.382	0.01723*
还清比率	−25.9019	2.3002	−11.261	<2e−16***
年利率	1.5311	0.3408	4.492	7.06e−06***
无效偏差		7548.8		
残差		205.1		
AIC		235.1		
Fisher 迭代次数		12		

注：*、**、***分别表示在 5%、1%和 0.1%的水平上显著。

6. 结论与展望

本文以中国某互联网金融平台数据集作为样本集，对经典的信用风险评估模型 Logistic 回归和集成方法 Bagging 进行分析研究，得出了以下结论与启示：

1) 本文对全样本数据集、正常样本数据集划分为不同比例的训练集与测试集，通过重复实验，发现在训练集与测试集划分比例为 8:2，且去除异常样本后，预测效果达到最优；

2) Bagging 集成方法在 F1-score、Accuracy 和 FPR 观测下效果比逐步 Logistic 回归和弹性网络有明显提升，其中随机森林表现最优且最稳定，但在 AUC 指标下略低于逐步 Logistic 回归，这说明 Bagging 集成方法在信用风险评估中可以对预测精度提高起到一定作用；

3) 本文研究 Bagging 集成方法中效果最佳的算法随机森林的特征重要性，并与逐步 Logistic 回归结果进行对比，发现其结果保持一致，筛选出的重要特征为还清比率、公司规模、年利率、标的总额、工作性质、借款成功率、工作时间。

本文选择我国互联网金融数据，研究逐步 Logistic 回归、弹性网络以及 Bagging 集成方法在信用风险评估领域的应用效果，发现逐步 Logistic 回归相比于弹性网络更能提高预测精度，Bagging 集成方法普遍优于传统方法，且其中随机森林精度最高、稳定性最优，说明 Bagging 集成方法在该问题上有一定的研究价值。研究逐步 Logistic 回归和随机森林的特征重要性，发现其筛选出的特征一致。基于目前的结果，进一步的计划是将本文研究方法在互联网金融的更多应用场景上进行测试，修改模型，以提高泛化能力。

参考文献

[1] 粟麟, 杨伟明. 数字金融: 发展现状、未来趋势与监管启示[J]. 北方金融, 2021(6): 8-12.

[2] 李焱文, 蒋文华, 王纯洁. 网络大数据信用风险评分能有效预测信贷违约风险吗? [J]. 经济问题, 2021(7): 70-77.

[3] Moscatelli, M., Parlapiano, F., Narizzano, S., *et al*. (2020) Corporate Default Forecasting with Machine Learning. *Expert Systems with Applications*, **161**, Article ID: 113567. https://doi.org/10.1016/j.eswa.2020.113567

[4] 方匡南, 章贵军, 张惠颖. 基于 Lasso-Logistic 模型的个人信用风险预警方法[J]. 数量经济技术经济研究, 2014, 31(2): 125-136.

[5] 韦勇凤, 向一波. 基于 Group-Lasso 方法的非均衡数据信用评分模型[J]. 中国科学院大学学报, 2021, 38(2): 181-188.

[6] 王小燕, 张中艳, 马双鸽. 基于文本先验信息的贷款信用风险评估模型[J]. 中国管理科学, 2021, 29(5): 34-44.

[7] Xu, D.Y., Chen, J.H., Zhang, X.Y. and Hu, J.G. (2020) A Novel Ensemble Credit Scoring Model Based on Extreme Learning Machine and Generalized Fuzzy Soft Sets. *Mathematical Problems in Engineering*, **2020**, Article ID: 7504764. https://doi.org/10.1155/2020/7504764

[8] Bekhet, H.A. and Eletter, S. (2014) Credit Risk Assessment Model for Jordanian Commercial Banks: Neural Scoring Approach. *Review of Development Finance*, **4**, 20-28. https://doi.org/10.1016/j.rdf.2014.03.002

[9] 王程龙, 陈程. 基于决策树的 P2P 网贷平台信用评级体系研究[J]. 农村金融研究, 2016(12): 45-50.

[10] Wang, G., Hao, J.X., Ma, J. and Jiang, H.B. (2010) A Comparative Assessment of Ensemble Learning for Credit Scoring. *Expert Systems with Applications*, **38**, 223-230. https://doi.org/10.1016/j.eswa.2010.06.048

[11] Li, Y.H. and Chen, W.D. (2020) A Comparative Performance Assessment of Ensemble Learning for Credit Scoring. *Mathematics*, **8**, 1-19. https://doi.org/10.3390/math8101756

[12] Tian, Z., *et al*. (2020) Credit Risk Assessment Based on Gradient Boosting Decision Tree. *Procedia Computer Science*, **174**, 150-160. https://doi.org/10.1016/j.procs.2020.06.070

[13] 莫赞, 张灿凤, 魏伟, 游德创, 张舒. 基于 Bagging 集成的个人信用风险评估方法研究[J]. 系统工程, 2019, 37(1): 143-151.

[14] 白鹏飞, 安琪, Nicolaas Fransde Rooij, 李楠, 周国富. 基于多模型融合的互联网信贷个人信用评估方法[J]. 华南师范大学学报(自然科学版), 2017, 49(6): 119-123.

[15] 操玮, 李灿, 贺婷婷, 朱卫东. 基于集成学习的中国 P2P 网络借贷信用风险预警模型的对比研究[J]. 数据分析与知识发现, 2018, 2(10): 65-76.

[16] 周永圣, 崔佳丽, 周琳云, 孙红霞, 刘淑芹. 基于改进的随机森林模型的个人信用风险评估研究[J]. 征信, 2020, 38(1): 28-32.

[17] 廖理, 吉霖, 张伟强. 借贷市场能准确识别学历的价值吗?——来自 P2P 平台的经验证据[J]. 金融研究, 2015(3): 146-159.