Modeling and Simulation 建模与仿真, 2021, 10(3), 890-897
Published Online August 2021 in Hans. http://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2021.103088

Hans汉斯

基于机器学习方法的企业贷款违约风险预测

陈旭岚[1]，韩苏皖[2]，庞建华[1*]

[1]广西科技大学理学院，广西 柳州
[2]南京审计大学统计与数学学院，江苏 南京
Email: *pjh968@126.com

收稿日期：2021年7月16日；录用日期：2021年8月18日；发布日期：2021年8月25日

摘　要

研究企业的贷款违约风险不仅对金融机构解决“惜贷”问题和防范信用风险具有重要的现实意义，而且能为企业规范自身经营和改善财务状况提出有针对性的建议及措施。本文根据某机构的企业贷款违约数据对贷款违约风险进行研究，首先对原始数据进行缺失值处理、特征选择和不平衡数据处理，然后利用逻辑回归、随机森林、XGBoost和LightGBM四种机器学习方法对数据进行建模和分析并比较模型优劣，最后利用GBDT模型计算特征重要性。结果表明：1) 三种集成模型的预测效果显著优于单一模型，2) 在集成模型中LightGBM模型表现出了最好的预测性能，3) 企业的纳税情况和曾经获得的授信情况可以作为判断该企业是否会发生贷款逾期现象的重要参考。

关键词

机器学习，企业贷款，违约风险

Corporate Loan Default Risk Prediction Based on Machine Learning Method

Xulan Chen[1], Suwan Han[2], Jianhua Pang[1*]

[1]Faculty of Science, Guangxi University of Science and Technology, Liuzhou Guangxi
[2]School of Statistics and Mathematics, Nanjing Audit University, Nanjing Jiangsu
Email: *pjh968@126.com

Received: Jul. 16th, 2021; accepted: Aug. 18th, 2021; published: Aug. 25th, 2021

Abstract

The research on the loan default risk of enterprises not only has important practical significance

*通讯作者。

文章引用: 陈旭岚, 韩苏皖, 庞建华. 基于机器学习方法的企业贷款违约风险预测[J]. 建模与仿真, 2021, 10(3): 890-897. DOI: 10.12677/mos.2021.103088

for financial institutions to solve the problem of "reluctant to lend" and prevent credit risks, but also can put forward targeted suggestions and measures for enterprises to standardize their own operation and improve their financial situation. This paper, based on the enterprise loan default data of an organization studies the default risk of the enterprise, first of all to the original data missing value processing, feature selection and unbalanced data processing, and then uses four machine learning methods of logistic regression, random forests, XGBoost and LightGBM for data modeling and analysis model, and advantages and disadvantage are compared. Finally, GBDT model is used to calculate the importance of features. The results show that: 1) The prediction effect of the three integrated models is significantly better than that of the single model; 2) LightGBM model shows the best prediction performance among the integrated models; 3) The tax payment and the credit obtained by the enterprise can be used as an important reference to judge whether the enterprise will have the loan overdue phenomenon.

Keywords

Machine Learning, Enterprise Loans, The Risk of Default

1. 引言

长期以来，贷款难的问题一直阻碍着企业的发展。虽然国家出台了一系列的支持政策，但企业自身存在的经营风险以及贷款违约现象的频繁发生使得银行等金融机构越来越“惜贷”。这种情况产生的根本原因在于金融机构缺乏具体有效的信用风险评估方法和防控措施，特别是随着大数据时代的来临，基于统计学方法的传统违约风险预测模型在数据处理方面的不足日益明显，且不具备自学习能力，已经无法适应如今的要求。因此，有关机构能否获得更加高效又切实可行的信用风险评估模型成为决定其未来发展的关键。

2. 文献综述

随着人工智能理论的深入发展，众多学者在基于机器学习算法的信用风险评估方法上取得了丰硕的研究成果。例如，朱景文以上市公司的债券违约风险为研究对象，证明了经过遗传算法优化的 BP 神经网络模型在公司债违约风险评估应用中的高准确性[1]。王晓菲等通过将 AI 模型与传统信用风险度量模型相结合的方法，从多角度对房地产企业的信用违约风险进行研究[2]。李占玉将 SMOTE 算法与随机森林相结合构建了互联网金融公司的财务风险评估模型，并取得了良好的预测效果[3]。潘永明等利用引入信息增益的 SVM 算法构建供应链中小企业信用风险分类预测模型，结果显示该模型比单一 SVM 模型精度提高 8.97% [4]。郑建国等将 PCA、SMOTE 与通过网格搜索法进行参数寻优的 SVM 模型相结合对企业信用风险进行研究，结果表明新构建的模型具有更高的稳定性和预测能力[5]。胡贤德等针对传统 BP 神经网络在实际应用中学习速度慢、易陷入局部解以及运算结果误差较大等缺陷，提出一种基于改进离散型萤火虫(IDGSO)算法的 BP 神经网络集成算法，并以此建立了小微企业信用风险评估模型，该模型有效提高了预测准确性[6]。

3. 基本方法

本文选择使用逻辑回归、随机森林、XGBoost (Extreme Gradient Boosting)和 LightGBM (Light Gradient Boosting Machine)四种模型对企业贷款违约数据进行建模研究，并根据研究结果选出预测性能最优的模

型，为有关机构进行风控管理提供参考。其中，逻辑回归模型是一种单一模型，具有更加简单的算法原理和内部结构，且该模型的输出形式易于理解、建模效率较高，有十分广泛的应用；随机森林是一种由大量分类回归树组成的集成算法，其应用十分灵活，在特征较多、数据量较大的数据集上表现良好；XGBoost 是一种基于 GBDT (Gradient Boosting Decision Tree)的集成算法，它不仅对损失函数进行二阶泰勒展开，而且将正则化项加入到损失函数中，有效控制了模型的复杂程度并大大提高了模型的预测性能；LightGBM 也是一种基于 GBDT 的改进集成算法，主要用于处理样本容量大且特征维度高的数据，其具有运算速度快、内存占用率低、准确率高以及支持并行计算等诸多优点。

4. 实证分析

4.1. 数据来源

本文基于某机构的企业贷款违约数据集进行实证研究。整个数据预处理、建模对比、特征分析过程均在软件 Python3.8 上实现，使用的安装包为 pandas、numpy、matplotlib、sklearn、xgboost、lightgbm 等。

4.2. 数据描述

该数据集共包含 8 个变量，分别是：企业 ID (脱敏)、当年是否产生逾期贷款、当年地税纳税总额、当年电量、当年国税纳税总额、当年增值税应纳总额、当年获得授信总额、当年已用授信总额。该数据集具体格式如表 1 所示。

Table 1. Basic information of variables
表 1. 变量基本情况

变量名称	类型	示例
企业 ID	int64	00010426
当年是否产生逾期贷款	int64	1
当年地税纳税总额	float	76,959.7
当年电量	float	20,758
当年国税纳税总额	float	32,400
当年增值税应纳总额	float	40,487.53
当年获得授信总额	float	45,000,000
当年已用授信总额	float	15,000,000

Table 2. Statistical description of variables
表 2. 变量的统计描述

变量名称	最大值	最小值	均值	标准差	中位数
地税纳税总额	2,040,000,000	−540	22,296,550.39	122,608,799.27	432,583.65
电量	45,000,000	0	783,581.21	2,701,843.58	176,223
国税纳税总额	12,800,000	−243,798.8	656,452.05	1,544,412.93	87,458.04
增值税应纳总额	19,100,000	0	372,694.25	1,537,349.98	8305.58
获得授信总额	3,590,000,000	0	119,289,179.91	330,324,069.82	16,500,000
已用授信总额	1,650,000,000	0	71,390,498.14	179,437,531.86	11,750,000

根据表 2 的统计性描述可知，数据集中包含的企业既有每年纳税金额过亿的大中型企业，也有经营

规模较小的小微企业，还有经营不善导致无法及时缴纳税费的濒临破产的企业。这表明该数据集包含的企业类型广泛，能较好的反映某地区总体的企业经营状况。

4.3. 数据预处理

根据本文的研究内容选择“是否产生贷款逾期”为响应变量，其有三个分类：有贷款未逾期(负类，由 0 表示)、有贷款且逾期(正类，由 1 表示)和没有贷款信息。因为本文研究主题为企业贷款的违约风险，所以应选择有贷款行为的企业为研究对象，根据该要求选择数据集中“当年是否产生逾期贷款”变量不为空值的记录，将该变量为空值的记录进行删除，删除后数据集中剩余 1215 条相关记录。接着，观察原始数据集可以发现存在数据缺失的情况，因此对各个特征的数据缺失情况进行统计(结果如图 1)，删除缺失值比例大于 40%的特征，即删除当年电量、当年国税纳税总额和当年增值税应纳总额三个特征。然后对缺失值进行处理，先算出每条记录中含有缺失值的个数，再将每条记录中缺失值个数大于等于 2 个的记录删除，此时数据集中只有地税纳税总额中还有缺失值，我们用其均值进行填补。

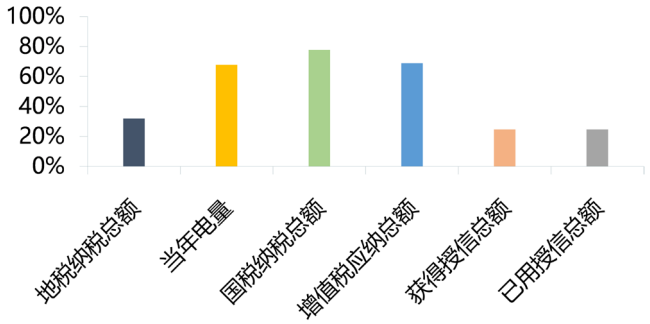

Figure 1. The proportion of missing values for each feature
图 1. 各个特征的缺失值比例

经过预处理后的数据集还有 3 个特征，552 条记录，其中有贷款且逾期的记录有 534 条，有贷款未逾期的记录有 18 条，显然这是一个非平衡数据集。因此，本文使用 SMOTE 上采样方法来增加负类样本数量，即随机复制已有的负类样本使负类样本总数达到 534 个，从而使数据集变为平衡数据集。

4.4. 建立模型

1) 数据集划分。按照 7:3 的比例将数据集依据随机原则划分为互斥的训练集和测试集，其中训练集有 747 条记录，测试集有 321 条记录。

2) 模型训练。将训练集分别代入四种机器学习模型进行训练，其中 XGBoost 模型和 LightGBM 模型需要调参。我们选择使用贝叶斯优化来寻找最优参数，因为该方法迭代次数少、耗时短，是一种非常实用的优化算法，并将 ROC-AUC 值(ROC 曲线下面积)作为目标函数，该函数值越高表示模型分类效果越好。调参后发现，对于 XGBoost 模型，没有调参之前的效果更好，因此使用默认参数；对于 LightGBM 模型，调参后模型取得了更好的效果，因此使用获得的最优参数。

XGBoost 模型的最优参数：

Booster = gbtree, objective = binary:logistic, learning_rate = 0.3, max_depth = 6, min_child_weight = 1, n_estimators = 100, n_jobs = 16, reg_lambda = 1, subsample = 1.

LightGBM 模型的最优参数：

bagging_fraction = 0.97, feature_fraction = 0.94, learning_rate = 0.03, min_data_in_leaf = 10, num_boost_round = 1500, num_leaves = 60, max_depth = 7, num_threads = 8, bagging_freq = 5, num_class = 2.

3) 模型预测。将测试数据代入训练好的模型得到相应的预测值，并根据评价指标对四个模型的预测效果进行评估。

4.5. 模型评估

本文根据混淆矩阵[7]的各项指标对四个模型的预测效果进行评估，结果如图 2 所示。

	precision	recall	f1-score	support
0.0	0.6727	0.2242	0.3364	165
1.0	0.5188	0.8846	0.6540	156
accuracy			0.5452	321
macro avg	0.5958	0.5544	0.4952	321
weighted avg	0.5979	0.5452	0.4907	321

(a)

	precision	recall	f1-score	support
0.0	0.8701	0.9333	0.9006	165
1.0	0.9236	0.8526	0.8867	156
accuracy			0.8941	321
macro avg	0.8968	0.8929	0.8936	321
weighted avg	0.8961	0.8941	0.8938	321

(b)

	precision	recall	f1-score	support
0.0	0.8800	0.9333	0.9059	165
1.0	0.9247	0.8654	0.8940	156
accuracy			0.9003	321
macro avg	0.9023	0.8994	0.9000	321
weighted avg	0.9017	0.9003	0.9001	321

(c)

	precision	recall	f1-score	support
0.0	0.8869	0.9030	0.8949	165
1.0	0.8954	0.8782	0.8867	156
accuracy			0.8910	321
macro avg	0.8912	0.8906	0.8908	321
weighted avg	0.8910	0.8910	0.8909	321

(d)

Figure 2. Comparison of evaluation indexes of each model. (a) Logistic model; (b) Random forest model; (c) XGBoost model; (d) LightGBM model

图 2. 各模型评价指标对比。(a) 逻辑回归模型；(b) 随机森林模型；(c) XGBoost 模型；(d) LightGBM 模型

图 2 中，“accuracy”表示模型分类的准确率，即预测正确的样本占总样本的比例，四种模型的准确率分别为 0.5452、0.8941、0.9003、0.8910，可以看出逻辑回归模型的准确率最低，XGBoost 模型的准确率最高，但 XGBoost 与随机森林和 LightGBM 的准确率相差很小。“precision”表示精确率，即预测类型为正类(负类)的样本中实际类型也为正类(负类)的比例，“recall”表示召回率，即实际类型为正类(负类)的样本中

预测类型也为正类(负类)的样本比例，而“f1-score”值同时考虑了模型的精确率和召回率，其值越接近于1表明模型的综合性能越好。观察正类样本(有贷款且逾期，用1.0表示)的f1-score值并将其从大到小排序的结果为0.9059 > 0.9006 > 0.8949 > 0.3364，这表明四种模型对正例的识别能力从强到弱分别为XGBoost、随机森林、LightGBM、逻辑回归，且前三种模型识别能力相近。总体而言，单一算法(逻辑回归模型)的分类效果并不理想，其各项指标都明显低于集成算法(随机森林、XGBoost和LightGBM)，这展现出集成算法所具有的强大的分类性能；从总体准确率和负类的f1-score值来看，三个集成算法在测试集上的预测效果相差不大，针对本次划分的测试集数据，XGBoost模型的预测效果略优于随机森林和LightGBM。

图3为四种模型的ROC曲线图，ROC曲线是辅助确定概率分割值的有效工具，它是反映Sensitivity (灵敏度)和Specificity (特异性)连续变量的综合指标，曲线越偏向于左上角，即曲线下方围成的面积(AUC值)越大，则表明模型的预测效果越好。可以看出，随机森林、XGBoost和LightGBM的ROC曲线明显更接近于坐标图的左上角，他们的AUC值分别为0.94、0.94、0.95，显著大于逻辑回归模型的AUC值0.74。这表明三个集成算法的预测性能要显著优于单一模型，该结果与之前其他评价指标得出的结论一致；但从AUC值来看，LightGBM模型的预测效果要略优于随机森林和XGBoost，这与之前的结论不同。

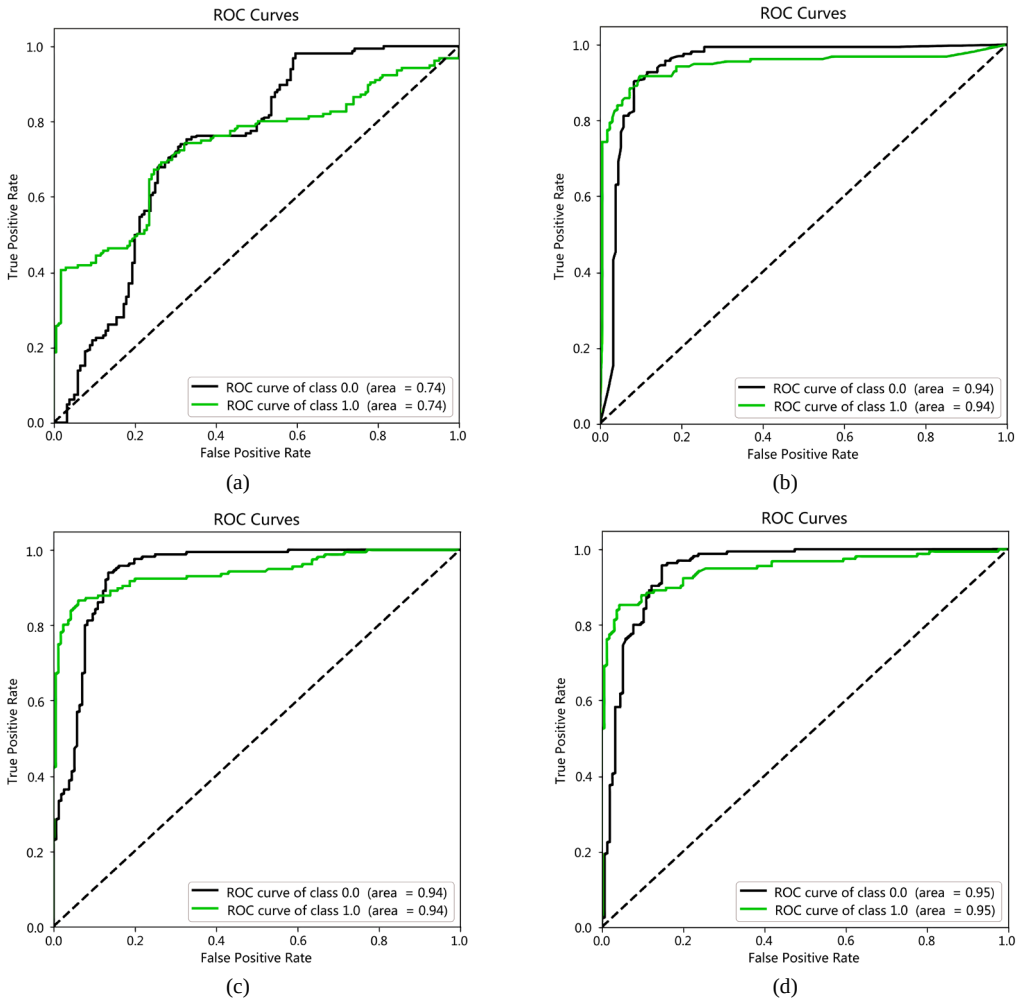

Figure 3. ROC curve of each model. (a) Logistic model; (b) Random forest model; (c) XGBoost model; (d) LightGBM model

图3. 各模型的ROC曲线。(a) 逻辑回归模型；(b) 随机森林模型；(c) XGBoost模型；(d) LightGBM模型

为了进一步判断三个集成算法中哪个最优，同时考虑到之前的模型结果仅仅是针对一次数据集划分而来的，所以我们在整个数据集上使用以分层采样为基础的 5 折交叉验证进行实验，以便充分利用已有的数据集获取尽可能多的有效信息，并计算三种集成模型在交叉验证下的五次 f1-score 值均值和五次准确率均值，结果如表 3 所示。可以看出，在三个集成模型中 LightGBM 模型的 f1-score 值和准确率都最大，这表明 LightGBM 模型的预测性能要略优于其他两个模型，该模型在企业贷款违约风险预测问题上有更强的适用性。虽然针对本文的数据集三种模型的预测效果相差不大，但这主要是因为本文数据集特征纬度低且样本量小。在特征维数很高且样本量很大的情况下，LightGBM 模型的不仅能极大的提升计算效率同时还能保证计算精度[8]。

Table 3. The five-fold cross-validation mean of the three integrated models
表 3. 三种集成模型的五折交叉验证均值

	随机森林	XGBoost	LightGBM
f1-score 值	0.8873	0.8814	0.8932
准确率	0.8839	0.8783	0.8895

4.6. 特征重要性

在选出最优模型后，本文接着使用 GBDT 模型来评估三个特性的重要性。主要使用 Python 软件中的 plot_importance 函数来获取相应数值，数值越大表明该特征越重要，结果如图 4 所示。

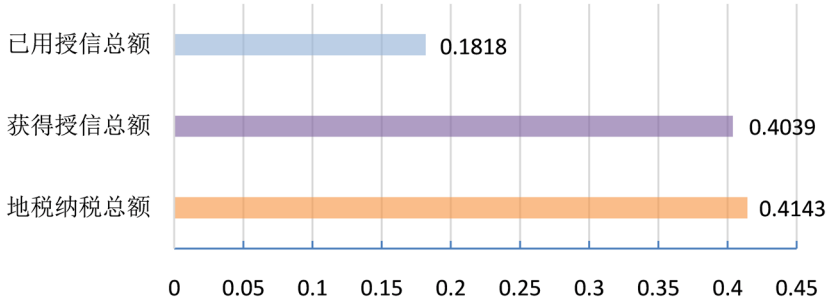

Figure 4. Characteristic importance
图 4. 特征重要性

结果表明，三个特征的重要性由大到小依次为地税纳税总额、获得授信总额和已用授信总额，其中地税纳税总额和获得授信总额的重要性相差不大，都为 0.4 左右。分析该结果我们可以得到如下启示：相关金融机构在对企业的贷款申请进行审查时，除了关注该企业的各项财报、企业主个人征信等情况外，还可以重点关注该企业过去几年的纳税情况和曾经的授信情况。纳税情况可以间接反映该企业的经营状况、资产状况、主营方向、公司规模等信息，有助于从不同角度辅助分析一个企业的还贷能力和其真正需要的贷款金额。曾经的授信情况可以反映其他机构对该企业的信任程度以及该企业的还款意愿和诚信程度，有助于为相关金融机构确定是否发放贷款以及发放多少贷款提供重要参考。

5. 结论

针对当今社会众多的企业贷款违约问题，本文利用 Python 软件基于机器学习方法构造违约风险评估模型，并以某机构的企业贷款违约数据为例进行实证分析。首先将原始数据进行缺失值处理、特征选择和不平衡数据处理，然后将数据集划分为训练集和测试集，并分别代入逻辑回归、随机森林、XGBoost

和 LightGBM 四种模型进行训练和预测。结果表明：综合混淆矩阵和 ROC 曲线的多个指标，可以发现后三种集成模型(随机森林、XGBoost 和 LightGBM)对分类问题的预测效果明显优于单一模型(逻辑回归模型)，并且在集成模型中 LightGBM 模型拥有最好的表现，其两项五折交叉验证均值均大于另外两个集成模型，这与其运算效率高、内存消耗低、支持并行计算等优点密切相关。因此，使用基于 LightGBM 算法的评估模型对实际贷款违约问题进行研究和预测是一个明智的选择，它能帮助有关机构更好地规避信用违约风险、加强风控措施。最后，本文利用 GBDT 模型对数据集中的特征重要性进行评估，发现企业的纳税情况和曾经获得的授信情况的重要性最大，分别为 0.414 和 0.403，这两个特征的相关情况应该作为判断该企业是否会发生贷款逾期现象的重要参考。

本文的不足之处在于有效数据集较小，数据集涵盖的特征种类较少，无法充分体现出不同集成算法间的差别。同时，LightGBM 模型的预测效果还有进步的空间，可以考虑将它和其他模型再次融合，进一步提升预测的准确率和稳定性。

基金项目

本论文获广西自然科学基金(2018GXNSFBA281185)资助。

参考文献

[1] 朱景文. 基于遗传算法的上市公司债券违约风险识别方案策划[D]: [硕士学位论文]. 上海: 上海师范大学, 2020.

[2] 王晓菲, 刘继端, 詹梓雯, 刘彦清, 张燕玲, 周燕. 基于 AI 与传统风险度量模型下房地产企业信用风险度量分析[J]. 商讯, 2021(20): 89-91.

[3] 李玉占. 基于 SMOTE-随机森林的互联网金融公司财务风险预警模型[J]. 经济研究导刊, 2020(33): 79-80.

[4] 潘永明, 王雅杰, 来明昭. 基于 IG-SVM 模型的供应链融资企业信用风险预测[J]. 南京理工大学学报, 2020, 44(1): 117-126.

[5] 郑建国, 李新. 基于 SVM 模型的企业信用风险评估研究[J]. 企业科技与发展, 2020(5): 220-221+224.

[6] 胡贤德, 曹蓉, 李敬明, 阮素梅, 方贤. 小微企业信用风险评估的 IDGSO-BP 集成模型构建研究[J]. 运筹与管理, 2017, 26(4): 132-139+148.

[7] 孔英会, 景美丽. 基于混淆矩阵和集成学习的分类方法研究[J]. 计算机工程与科学, 2012, 34(6): 111-117.

[8] Ke, G., Meng, Q., Finley, T., *et al*. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, **12**, 3149-3157.