E-Commerce Letters 电子商务评论, 2024, 13(4), 3527-3538
Published Online November 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.1341551

Hans汉斯

基于机器学习的信贷风险量化研究

吴佳尧

贵州大学经济学院，贵州 贵阳

收稿日期：2024年8月30日；录用日期：2024年11月12日；发布日期：2024年11月19日

摘 要

随着互联网金融的发展，对于商业银行来说网络信贷业务变得越来越重要，而随之而来的信贷风险控制也日益凸显其重要性。本文通过对机器学习相关知识的研究和学习，在对金融机构的信贷数据进行相应的预处理以及数据集拆分之后，构建了基于逻辑回归、SVM、随机森林等方法的多个风险量化决策模型。在进行特征指标的选取、模型参数等细节的研究和设置之后，基于训练集数据来构建风险量化决策模型并对信贷客户的违约行为进行判断，然后将测试集数据代入模型中并把预测值与客户实际还款情况进行对比来验证模型的有效性。通过本文的研究和实验结果表明，通过构建风险量化决策模型来预测信贷客户的还款情况，特别是优化后的随机森林模型和SGD Classifier模型拥有较好的预测效果，具有较高的可行性和准确率。在客户申请贷款业务时，只需要输入对应的特征信息到预测模型中，就能立即对客户的违约情况进行预测。这对信贷风险的控制起着较大的促进作用，也对我国金融信贷市场的稳健发展有着积极的意义。

关键词

信贷风险量化，信贷违约预测，机器学习

Quantitative Analysis of Credit Risk Based on Machine Learning

Jiayao Wu

School of Economics, Guizhou University, Guiyang Guizhou

Received: Aug. 30th, 2024; accepted: Nov. 12th, 2024; published: Nov. 19th, 2024

Abstract

With the development of internet finance, online credit business has become increasingly important for commercial banks, and the accompanying risk control of online credit has also become

文章引用: 吴佳尧. 基于机器学习的信贷风险量化研究[J]. 电子商务评论, 2024, 13(4): 3527-3538.
DOI: 10.12677/ecl.2024.1341551

increasingly important. Through the research and learning of machine learning related knowledge, after the corresponding pre-processing of credit data of financial institutions and the splitting of data sets, this paper constructs multiple risk quantitative decision-making models based on logical regression, SVM, random forest and so on. After studying and setting the selection of feature indicators, model parameters, and other details, a risk quantification decision model is constructed based on the training set data to judge the default behavior of credit customers. Then, the test set data is substituted into the model and the predicted values are compared with the actual repayment situation of customers to verify the effectiveness of the model. The research and experimental results of this paper show that the optimized random forest model and SGD Classifier model have good prediction effect, high feasibility and accuracy by building a risk quantitative decision-making model to predict the repayment of credit customers. When a customer applies for loan business, they only need to input the corresponding feature information into the prediction model to immediately predict the customer's default situation. This plays a significant role in promoting the control of credit risks and has a positive significance for the stable development of China's financial credit market.

Keywords

Credit Risk Quantification, Credit Default Prediction, Machine Learning

1. 引言

近些年，随着互联网技术的蓬勃发展，各类诸如 ChatGPT、元宇宙等新技术不断涌现，各种基于互联网技术的金融创新也层出不穷。网络信贷成为了商业银行进行业务创新的一个重要方向。但是，由于世界经济下行以及互联网金融本身就具有信息不对称、监管难等问题，对信贷业务中信用风险的管理就成为了银行金融风险控制的重要部分。在当前的经济背景下，如何在最大限度地满足信贷需求并促进经济复苏的同时，以完备的风险控制模型来尽可能地降低信贷违约率进而增强金融经济的安全性就成为了一个十分重要的课题。

在互联网金融领域，商业银行拥有着天然的大数据优势，如何通过海量数据来实现对信用风险的定量控制具有非常重要的意义。目前，关于网络信贷的研究多集中于 P2P 等形式的信贷风险研究，并且较为缺少实证研究。基于此，本文在金融机构发布的信贷数据的基础上，构建了基于逻辑回归、SVM、随机森林等多个风险量化决策模型，并对特征指标的选取、模型参数等细节进行设置和调整，最终通过对比挑选出性能最好的基于随机森林方法来构建风险量化决策模型来对信贷客户的违约行为进行预测判断，并将预测值与客户实际还款情况对比来验证模型的有效性。这对商业银行在网络信贷业务风险控制方面有着较大的参考作用。

2. 文献综述

2.1. 支持向量机在金融应用方面的研究

在应用支持向量机用于信用风险评估研究方面，2019 年李健等人以汽车供应链为研究对象，构建了一个基于 SVM 的供应链金融预警模型。该模型为识别和预警供应链金融模式下的信用风险提供了有效工具[1]。2020 年 Goran Martinovic 等建立了以支持向量机为基础方法的客户化信用评分模型[2]。2020 年

申晴等在多家上市企业数据样本的基础上，提出了一种混合了 SVM 和 KNN 的组合模型[3]。2021 年冯昊等人对比了多种算法后发现，基于遗传算法优化的 SVM 在信用评分领域表现突出。他们的研究证实了这种混合模型在实际应用中的高效性[4]。2021 年李佩霏利用 SVM 和 GARCH 模型对某公司的股价数据进行了深入分析，并结合两者的优点进行了预测，其研究成果为投资者决策提供了有价值的参考[5]。2022 年向实运用支持向量机方法对债券违约问题进行了研究，并提出了有效的检测手段。他的工作为理解和应对债券违约问题提供了新的视角[6]。2024 年蔡毅等人提出了一种结合反向混频数据抽样和机器学习算法的新模型，用于预测股市走势。该模型在实证检验中表现出色，优于其他传统模型[7]。2024 年李昕等人结合农村供应链金融的特点，利用支持向量机(SVM)构建了一个针对农业中小企业的风险评估模型。通过实证检验，该模型能够有效地评估这些企业的信用风险[8]。

2.2. 随机森林在金融应用方面的研究

在随机森林应用于信用风险评估方面，2018 年胡蝶对债券违约从多个角度进行归因分析，然后利用随机森林方法来构建模型。通过实证分析，作者发现在特征之间存在着交互作用[9]。2019 年陈标金等人利用随机森林算法对多个宏观经济和技术指标进行了预测，并成功构建了一个有效的国债期货量化投资模型。他们的研究为国债期货投资提供了新的思路[10]。2020 年方若男等人结合随机森林算法构建了一个风险预警机制，并以拥有支付牌照的企业为样本进行了实证检验。该研究为第三方支付行业的风险管理提供了重要参考[11]。2021 年周亮利用随机森林模型对多个股票因子进行了拟合和预测，并发现了其良好的投资性能。该研究为股票多因子投资策略提供了新的方向[12]。2021 年闫政旭等人为了提高股票价格预测的准确性和降低噪声影响，构建了一个基于 Pearson 系数的随机森林组合模型。研究发现该模型在股价预测方面具有显著优势[13]。2021 年孙玲莉等人结合 Benford 定律改进了随机森林模型，并发现这一改进显著提升了模型的实用性，该研究为财务风险预警领域提供了新的方法[14]。

2.3. 逻辑回归在金融应用方面的研究

在逻辑回归应用于信用风险评估方面，2019 年杨睿哲等人以不完全信息下的保理公司授信评估为背景，构建了一个基于逻辑回归和神经网络等算法的授信额度估算模型[15]。2019 年郝婷婷等采用统计学建模方法，将逻辑回归模型应用于农村商业银行客户信用评级中[16]。2020 年边玉宁等从商业银行信贷违约问题出发，建立了以逻辑回归为核心的违约预测模型[17]。2021 年刘荣珍在逻辑回归和机器学习的基础上，提出了两个违约预测模型。在经过利用贝叶斯参数调优之后，模型的适用能力得到了有效提升[18]。2021 年曹杰等人以石油企业在采购选商为研究背景，利用石油企业的进项发票数据和供应商主数据构建了一个数据集。基于该数据集，他们采用逻辑回归模型来拟合模型参数并构建评分卡。通过实证分析，作者发现基于逻辑回归评分卡的方法能够显著提升石油企业对供应商的风险管理效果，并为采购管理者提供有力的决策支持[19]。2023 年张媛媛以金融欺诈为背景，选取了广告点击欺诈和银行信用卡欺诈两个典型的欺诈场景。作者将逻辑回归、均值不确定逻辑回归、XGBoost 和 LightGBM 四个模型以不同的方法构建出四个混合模型。通过实证分析，发现引入新衍生特征的组合模型在广告和银行领域的欺诈场景中具有更好的欺诈识别能力[20]。

2.4. 文献评述

由上述的研究成果可以看出，机器学习相关算法在金融领域的应用非常广泛，与传统的方法相比，采用机器学习相关方法来对分类问题进行研究时能得到更高的预测精度，并且模型的泛化能力也更优秀。不过，国内外大多数文献在信用风险评估方面的研究大都是以训练单一的模型为基础的，并且对于模型

的参数的设置与优化方面比较模糊，文章中模型的效果同其他模型进行对比分析也不多。

基于此，本文以国内外相关文献为基础，构建出基于逻辑回归、支持向量机、随机森林等多种方法的多个预测模型，并对模型进行相应的参数优化，同时也将多个模型进行对比评估。最后，将效果最好模型用于信贷客户违约行为预测中。

3. 数据处理与模型评价指标选择

3.1. 数据处理与拆分

本次研究数据来源于美国 LendingClub 公司官网的信贷数据，包含了将近 5 千个样本量，共有 151 个变量。本次研究将用户还款情况(loan status)这一变量作为本次研究的目标变量。对原始数据进行以下处理：将因变量即用户还款情况(loan status)中的未全部偿还设为 1，全部偿还取值为 0；对只有唯一值的变量进行去除；对于缺失值大于 90%的变量进行去除；经过以上步骤，处理后的数据剩下 4952 行，88 列。

在数据集的拆分中，考虑到样本量数量，将研究样本总量的 90%划分为研究的训练集，样本总量的 10%划分为测试集。具体的数据集拆分情况见表 1：

Table 1. Dataset splitting demonstration table
表 1. 数据集拆分示意表

	样本量	变量数
X_train	4456	87
y_train	4456	1
X_test	496	87
y_test	496	1
总数据集	4952	88

3.2. 描述性统计分析

3.2.1. 客户分级的统计分析

经过对客户分级这一特征进行统计性分析之后，分析结果见图 1。从该图可以看出各个等级上的贷款数量以及该等级上的贷款最终偿还情况的分布情况。从分析结果可以得出以下结论：

在 7 个大类的客户中，贷款业务在 B 和 C 类更集中，大部分的贷款都发放给了这两个类别的客户。而对于分类中最为靠后的 F 和 G 类客户，这两类客户得到的贷款数量是最少的。

除此之外，还能从分析结果中看出，随着客户的分级从 A1 到 G5 递减时，对应级别上的贷款业务数量的分布是先增加后减小的。

贷款偿还结果在各个客户分级上的分布是不一样的。最终偿还结果为全部偿还的贷款分布更加集中在 B3 和 B4 上，随后向两边递减。而最终结果为违约的贷款分布更加集中在 C1 和 C2 上，随后向两边递减。二者的分布有所偏差，结果为违约的贷款分布相较于未违约的贷款，其分布更加往分类级别低的方向偏移。

客户的违约率随着客户分级的降低而上升。也就是说，随着客户分级从 A 向 G 变化时，客户发生贷款违约行为的概率是在递增的。

3.2.2. FICO 评分的统计分析

通过分别绘制各个 FICO 评分区间上的违约和未违约客户的 FICO 评分的概率密度分布图，可以看出违约和未违约客户的 FICO 评分概率密度分布在各个评分区间上的分布状况，具体的分布情况见图 2。

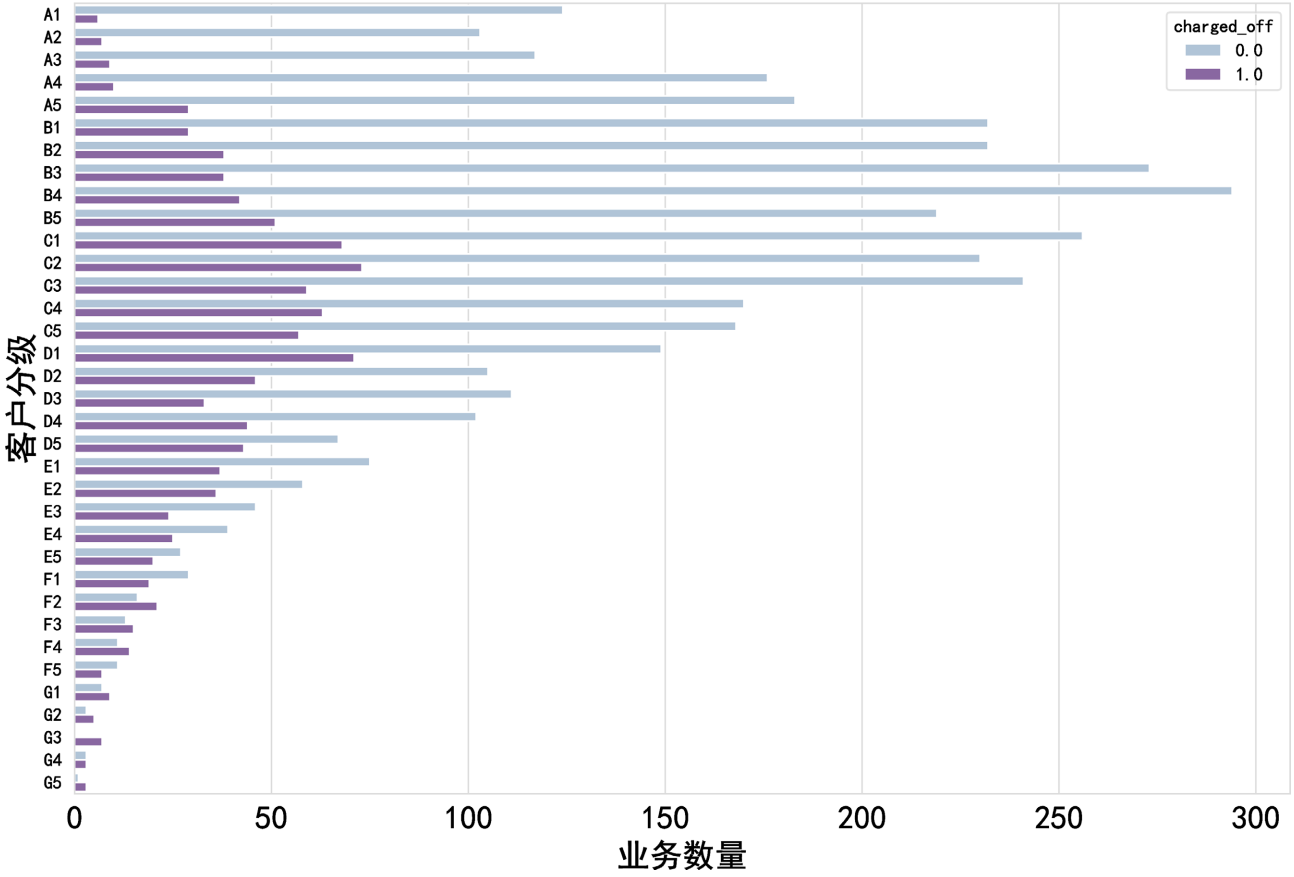

Figure 1. Customer tier statistical analysis chart
图 1. 客户分级统计分析图

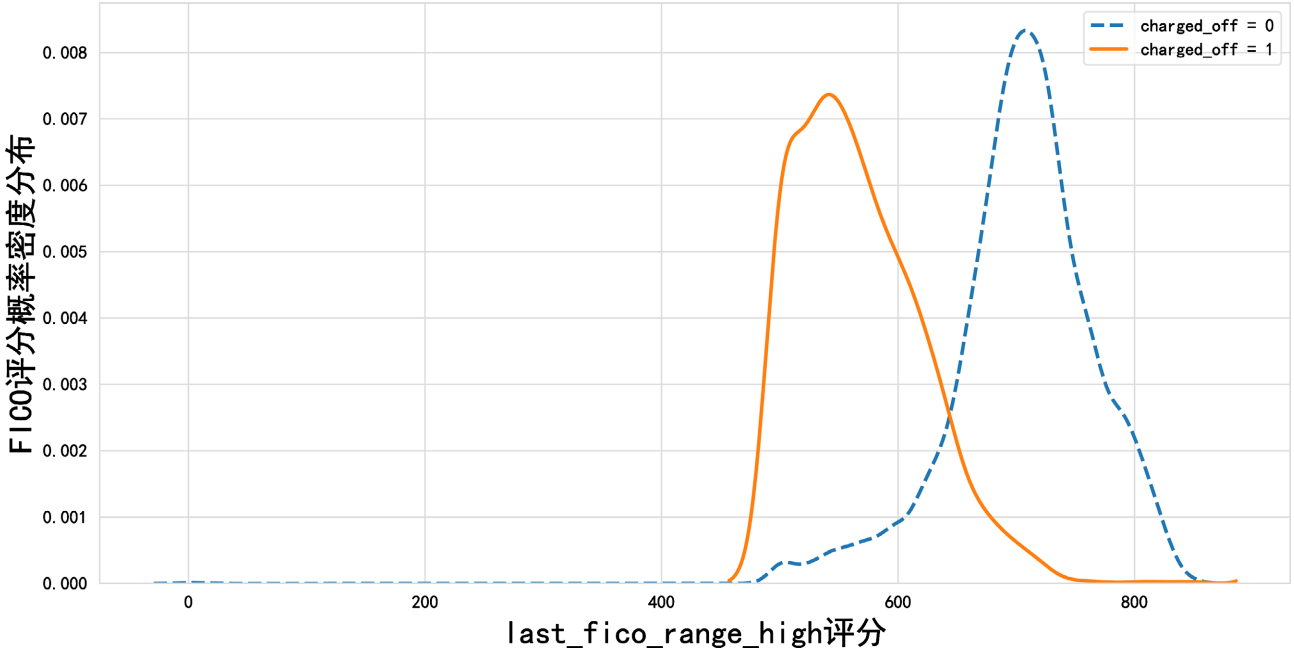

Figure 2. FICO score probability density distribution chart
图 2. FICO 评分概率密度分布图

可以看出，违约的和未违约的客户的 FICO 都呈现出了较为典型的正态概率分布。未违约客户的 FICO 评分(蓝色虚线表示)均值在 700 左右，而违约客户的 FICO 评分(黄色实线表示)均值在 530 左右。

3.2.3. 模型评价指标选择

由于在此信贷数据中，违约人数与未违约人数比值在 1:3.9 左右，数据存在失衡，因此，本次研究使用的模型的评价指标，主要是查准率(Precision)、查全率(Recall) (也称为召回率，命中率，TPR)、假正率(假警报率，FPR)、和 AUC 得分，本次研究主要依据这些指标来进行对模型的评价。

4. 信贷风险量化模型构建与效果评估

4.1. 逻辑回归模型构建与效果评估

在数据拆分工作完成后，将划分好的训练集数据带入到逻辑回归模型(模型参数选择为默认参数)，然后开始进行模型训练。模型训练好之后，将测试集带入模型进行预测。最后，将测试集中的目标变量与模型依据测试集中的自变量预测出来的预测变量进行对比，得到的结果见表 2 和表 3。

Table 2. Confusion matrix table for logistic regression model
表 2. 逻辑回归模型混淆矩阵表

	0 (预测为不违约)	1 (预测为违约)
0 (实际不违约)	378	21
1 (实际违约)	24	73

Table 3. Experimental results of logistic regression model
表 3. 逻辑回归模型实验结果

	precision	recall	f1-score	support
0	0.94	0.95	0.94	399
1	0.78	0.75	0.76	97
macro avg	0.86	0.85	0.85	496
weighted avg	0.91	0.91	0.91	496
Accuracy		0.9092741935483871		
Area under the curve		0.849973		

从数值上看，该模型的准确率为 91%，表现比较不错。此外，AUC 得分为 0.85，说明该模型预测的效果也不错。而该模型的查准率为 0.78，意味着在所有的模型预测会贷款违约的人数中，实际违约的人数所占的比例为 78%。而该模型的查全率为 0.75，在所有的实际违约人数中，被模型预测违约的人数所占的比例为 75%。总的来说，该逻辑回归模型的效果比较不错。

4.2. 随机森林模型构建与效果评估

在随机森林模型的构建中，对于随机森林模型参数的选择，通过查阅文献和相关理论，这里选取的参数为普遍使用的参数。将特征选择标准取值为“entropy”信息熵，森林中决策树的数量取成 1000，决策树最大深度取值为 None，子节点往下划分所需的最小样本数取为 10。

在模型参数设置好后，将之前划分好的训练集数据带入随机森林模型，然后进行模型的训练。在模型训练好之后，将测试集中的自变量带入模型中进行预测。最后，再将测试集中的目标变量与模型预测

出来的变量预测值进行对比，得到的结果见表 4 和表 5。

Table 4. Confusion matrix table for random forest model
表 4. 随机森林模型混淆矩阵

	0 (预测为不违约)	1 (预测为违约)
0 (实际不违约)	373	26
1 (实际违约)	19	78

Table 5. Experimental results of random forest model
表 5. 随机森林模型实验结果

	precision	recall	F1-score	support
0	0.95	0.93	0.94	399
1	0.75	0.80	0.78	97
macro avg	0.85	0.87	0.86	496
weighted avg	0.91	0.91	0.91	496
Accuracy		0.9092741935483871		
Area under the curve		0.869480		

从数值上来看，随机森林模型具有很高的准确率(0.91)。此外，该模型的 AUC 为(0.87)，说明该模型预测的效果也不错。而模型的查准率为 0.75，意味着在所有的模型预测会贷款违约的人数中，实际违约的人数所占的比例为 75%。查全率为 0.80，在所有的实际违约人数中，被模型预测违约的人数所占的比例为 80%。从查准率、查全率的调和平均数 F1-score 来看，该模型相比与前面的模型来说表现略低(0.78)。模型的整体预测效果是比较好的。

4.3. 随机森林模型参数调优与新模型的构建

本次研究使用的参数调优方法是网格搜索，通过使用穷举的方式来将所有的候选参数遍历，并通过循环建立模型且在每一次建立模型时都进行模型有效性和准确性的评估，最终选取表现最好的参数来作为模型的最终结果。在此处的参数调优工作中，需要进行调优的超参数有：特征选择标准(criterion)、决策树最大深度(max_depth)、子节点往下划分所需的最小样本数(min_samples_split)。

各个超参数的候选范围分别将其设置为：决策树最大深度的范围从 5、7、9、11、13 中选择，特征选择标准从信息熵、基尼系数中进行选择，子节点往下划分所需的最小样本数范围为 5，7，9，11，13，15。

在参数范围设定好之后，便开始进行参数遍历，并进行交叉验证。经过穷举之后，最终得到的最优参数如下：{'criterion': 'entropy', 'max_depth': 11, 'min_samples_split': 15, 'random_state': 123}。在经过参数调优之后，将上面得到的最优参数带入模型进行模型的构建。然后，将之前划分好的训练集数据带入随机森林模型，然后进行模型训练。在模型训练好之后，再将测试集中的自变量带入模型进行预测。最后，再将测试集中的目标变量与模型预测出来的变量预测值进行对比，模型得到的结果见表 6 和表 7。

从数值上来看，改进后的随机森林模型具有比较高的准确率(0.915)，与之前的随机森林模型的 0.90 相比有所提升。此外，该模型的 AUC 得分也很高(0.869)，说明该模型也是较为完善的。而模型的查准率为 0.78，意味着在所有的模型预测会贷款违约的人数中，实际违约的人数所占的比例为 78%。查全率为 0.79，在所有的实际违约人数中，被模型预测违约的人数所占的比例为 79%。从查准率、查全率的调和平

均数 F1-score 来看，相较于改进前的模型，改进后的随机森林模型略有提升(0.79)。总的来说，改进后的随机森林模型具有很好的预测效果。

Table 6. Optimized confusion matrix for random forest model
表 6. 优化随机森林模型混淆矩阵

	0 (预测为不违约)	1 (预测为违约)
0 (实际不违约)	377	22
1 (实际违约)	20	77

Table 7. Experimental results of optimized random forest model
表 7. 优化随机森林模型实验结果

	precision	recall	F1-score	support
0	0.95	0.94	0.95	399
1	0.78	0.79	0.79	97
macro avg	0.86	0.87	0.87	496
weighted avg	0.92	0.92	0.92	496
Accuracy		0.9153225806451613		
Area under the curve		0.869338		

通过此方法，我们还能知道在 87 个特征中各个特征的重要性。在 87 个变量中，最重要的前 7 个特征为：last_fico_range_low、last_fico_range_high、sub_grade、int_rate、term、dti、fico_range_low。而排名最后的几个特征为 disbursement_method、hargeoff_within_12_mths、acc_now_delinq、application_type、tax_liens、delinq_amnt、collections_12_mths_ex_med，这些特征重要性为 0，对于模型的预测根本未提供有用信息，在模型的简化中可将这些特征剔除。

4.4. 支持向量机模型构建与效果评估

在支持向量机模型的构建中，先将支持向量机中 SVC 函数的参数 kernel 设置为 sigmoid，其他参数用默认参数。然后将之前划分好的训练集数据带入支持向量机模型，然后进行模型训练。在模型训练好之后，将测试集中的自变量带入模型进行预测。最后，再将测试集中的目标变量与模型预测出来的变量预测值进行对比，得到的结果见表 8 和表 9。

Table 8. Confusion matrix for support vector machine model
表 8. 支持向量机模型混淆矩阵

	0 (预测为不违约)	1 (预测为违约)
0 (实际不违约)	356	43
1 (实际违约)	29	68

从数值上来看，支持向量机模型虽然也具有较高的准确率(0.85)，但与前面优化过的随机森林模型相比还是有差距。此外，该模型的 AUC 值为 0.80，虽然也能说明该模型较为完备，但与前面的模型相比也有较大差距。而模型的查准率为 0.61，意味着在所有的模型预测会贷款违约的人数中，实际违约的人数所占的比例为 61%。查全率为 0.70，在所有的实际违约人数中，被模型预测违约的人数所占的比例为 70%。从查准率、查全率的调和平均数 F1-score 来看，该模型相比于前面的模型来说表现较差(0.65)。

Table 9. Experimental results of SVM model
表 9. SVM 模型实验结果

	precision	recall	F1-score	support
0	0.92	0.89	0.91	399
1	0.61	0.70	0.65	97
macro avg	0.77	0.80	0.78	496
weighted avg	0.86	0.85	0.86	496
Accuracy		0.854839		
Area under the curve		0.796631		

4.5. SGD Classifier 模型构建与效果评估

在 SGD Classifier 模型的构建中，将 param_grid 的参数设置两份参数来候选：

候选参数组合 1：参数 loss 设置为 hinge，参数 class_weight 设置范围为(None, balanced)，参数 warm_start 设置为 True，其他为默认参数。

候选参数组合 2：参数 loss 设置为 log，参数 class_weight 设置范围为(None, balanced)，参数 warm_start 设置为 True，参数 penalty 设置范围为：[l3, l1]，其他为默认参数。

在模型参数设置好以后，将之前划分好的训练集数据带入 SGD Classifier 模型中，然后进行模型训练。在模型训练好之后，将测试集中的自变量带入模型中进行预测。最后，再将测试集中的目标变量与模型预测出来的变量的预测值进行对比，得到的结果见表 10 和表 11。

Table 10. Confusion matrix for SGD classifier model
表 10. SGD classifier 模型混淆矩阵

	0 (预测为不违约)	1 (预测为违约)
0 (实际不违约)	362	37
1 (实际违约)	8	89

Table 11. Experimental results of SGD classifier model
表 11. SGD classifier 模型实验结果

	precision	recall	F1-score	support
0	0.98	0.91	0.94	399
1	0.71	0.92	0.80	97
macro avg	0.84	0.91	0.87	496
weighted avg	0.93	0.91	0.91	496
Accuracy		0.9092741935483871		
Area under the curve		0.912397		

从数值上来看，SGD Classifier 模型具有很高的准确率(0.91)。此外，该模型的 AUC 为(0.91)，与前面的模型相比提升非常明显，这表明这一模型是非常完备的。而模型的查准率为 0.71，意味着在所有的模型预测会贷款违约的人数中，实际违约的人数所占的比例为 71%。查全率为 0.92，在所有的实际违约人数中，被模型预测违约的人数所占的比例为 92%。从查准率、查全率的调和平均数 F1-score 来看，该模

型相比与前面的模型来说表现是最好的(0.80)。总的来说，该预测模型的预测效果是十分优秀的。

4.6. 模型对比与效果评估

本次研究通过分别使用逻辑回归模型、原始随机森林模型、优化随机森林模型、SGD Classifier 模型等多个模型，对同一组数量为 4456 的样本数据进行模型训练，在模型训练完成后再对 495 个测试集样本进行预测，并最终将预测结果与实际违约情况进行对比。最终得到的实验结果见表 12：

Table 12. Comparison table of experimental results evaluation metrics
表 12. 实验结果评价指标对比表

模型	准确率	AUC 得分	查准率	查全率	f1-score
逻辑回归模型	0.91	0.85	0.78	0.75	0.76
随机森林原始模型	0.91	0.87	0.75	0.80	0.78
随机森林优化模型	0.92	0.87	0.78	0.79	0.79
支持向量机模型	0.85	0.79	0.61	0.70	0.65
SGD Classifier 模型	0.91	0.91	0.71	0.92	0.80

从准确率这一指标来看，以上五个模型均具有较好的预测价值和研究价值，其中表现最好的是优化以后的随机森林模型(准确率为 92%)，除了支持向量机模型表现不足外(准确率为 85%)，其他三个模型也均具有较高的预测价值。

从 AUC 得分这一指标来看，五个模型的表现都是非常不错的。尽管支持向量机模型表现略差(AUC 得分为 0.79)，但也是略大于 0.75，属于能接受的范围。而其他 4 个模型的 AUC 得分均达到 0.85 以上，均是非常不错的模型。四个模型中，表现最好的是 SGD Classifier 模型，其得分为 0.91。其次便是随机森林模型，其 AUC 值为 0.87。

从查准率来看，除了支持向量机模型表现略差(查准率为 61%)外，其他 4 个模型均在 70%以上；从查全率来看，除了支持向量机模型表现略差(查全率为 70%)外，其他 4 个模型均在 75%以上。

从 F1-score 来看，除了支持向量机模型表现略差(65%)外，其他 4 个模型均在 80%左右。总的来说，逻辑回归模型、优化前后的随机森林模型以及 SGD Classifier 四个模型的精度是较高的。

由于样本数据没违约与违约客户的比列为 4:1，数据有较大的失衡。因此，准确率这一指标并不能很好地反映模型预测的效果。而对于查准率、查全率这两个指标，可以用它们的调和平均数 F1-score 来衡量。综合以上，应当以 AUC、F1-score 两个指标来衡量以上模型的好坏。依据这两个指标，可以知道 5 个模型中预测效果最好的模型是 SGD Classifier 模型和优化以后的随机森林模型。

对于 SGD Classifier 模型，它采用了一系列梯度下降来求解参数，因此该模型的效果很好。而对于优化以后的随机森林模型，因为采用了网格搜索的方法来穷举出了最优参数，因此优化后的随机森林模型各方面的表现都优于逻辑回归模型和原始的随机森林模型。由此可见，参数的选择对一个模型来说是至关重要的。

5. 结论与展望

5.1. 研究结论

本文通过对机器学习相关理论的研究和学习，构建了基于逻辑回归、SVM、随机森林等 5 个风险量化决策模型。通过对特征指标的选取、模型参数等细节的研究和设置，以构建风险量化决策模型的方式

对信贷客户的违约行为进行判断，并将预测值与客户实际还款情况进行对比来验证了模型的有效性。最终，本文通过对五个模型进行对比评价，挑选出了效果最好的优化之后的随机森林模型和 SGD Classifier 模型来作为信贷风险量化决策模型。本文所得到的研究结论如下：

(1) 在本文所构建的所有模型中，基于随机森林的信贷风险量化模型和基于 SGD Classifier 模型的预测效果最好。均有较高的 AUC 值和 F1-score，说明采用该模型能够有助于商业银行对借贷风险的判定，也能有助于辅助贷款人做出科学的借贷决策。

(2) 本文通过将各模型分析结果进行对比，发现随机森林方法和 SGD Classifier 方法与其他方法相比，在对信贷违约行为进行预测方面具有相对优越性。

(3) 通过特征重要性排序表可以看到，在 87 个变量中，最重要的前 7 个特征为：last_fico_range_low、last_fico_range_high、sub_grade、int_rate、term、dti、fico_range_low。这说明这 7 个特征对于判断借款人是否做出违约行为有着非常大的影响。

(4) 本文通过使用得到的特征重要性排序表来对模型进行简化，发现能够有效地降低模型的计算量并明显地提升运算速度。

(5) 本文通过利用网格搜索的方法对随机森林模型进行了参数优化，通过对比优化前后的模型评价指标，发现参数的优化能给模型的性能带来较大的提升效果。

5.2. 不足与展望

本文通过数据预处理、数据集拆分、参数优化等重要步骤来调试随机森林模型和 SGD Classifier 模型，构建了信贷风险量化决策策略，保证较高程度的预测效果。但在实际应用中，仍有以下几个方面需要改善：

在目标变量处理时，只考虑了到期全额偿还(0)和到期未全额偿还(1)。因为其他情况数量不多，为了简化模型就将其他情况忽视了。在实际的运用中，应将其他情况考虑进来。

在以后的研究中应采取更多的方法进行参数的优化。由于数据量不大，因此本文的研究中仅仅采用了网格搜索和交叉验证的方法来确定超参数，这种方法存在占用资源过大和运行耗时等问题。在之后的研究中应使用更有效和便捷的贝叶斯搜索、随机搜索的方法来进行参数调优。

在数据预处理中，因为在本文使用的数据中违约与没违约的人数比例是 1:4，数据是存在失衡情况的。尽管曾尝试过使用 SMOTE 算法来进行失衡处理，但由于 SMOTE 处理过后的模型在训练集和测试集上的效果差异太大，模型的泛化能力大幅下降，因此只能放弃对失衡数据进行处理。这也就使得本文并未将准确率作为主要的评价标准来评估模型的预测效果。在之后的研究与应用中，可以从该方面进行改进。

参考文献

[1] 李健, 张金林. 供应链金融的信用风险识别及预警模型研究[J]. 经济管理, 2019, 41(8): 178-196.

[2] Nalić, J. and Martinovic, G. (2020) Building a Credit Scoring Model Based on Data Mining Approaches. *International Journal of Software Engineering and Knowledge Engineering*, **30**, 147-169. https://doi.org/10.1142/s0218194020500072

[3] 申晴, 张连增. 一种新的银行信用风险识别方法: SVM-KNN 组合模型[J]. 金融监管研究, 2020(7): 23-37.

[4] 冯昊, 李树青. 基于多种支持向量机的多层级联式分类器研究及其在信用评分中的应用[J]. 数据分析与知识发现, 2021, 5(10): 28-36.

[5] 李佩霏. 基于支持向量机和 GARCH 模型的股价预测[D]: [硕士学位论文]. 大连: 大连理工大学, 2021.

[6] 向实, 曾银球, 闫新国, 等. 基于支持向量机方法的债券违约风险监测预警研究[J]. 金融经济, 2022(1): 40-50.

[7] 蔡毅, 唐振鹏, 吴俊传, 等. 基于灰狼优化的混频支持向量机在股指预测与投资决策中的应用研究[J]. 中国管理

科学, 2024, 32(5): 73-80.

[8] 李昕, 谢昊伦. 基于支持向量机的农业中小企业供应链金融信用风险评价[J]. 物流科技, 2024, 47(5): 146-149.

[9] 胡蝶. 基于随机森林的债券违约分析[J]. 当代经济, 2018(3): 28-30.

[10] 陈标金, 王锋. 宏观经济指标、技术指标与国债期货价格预测——基于随机森林机器学习的实证检验[J]. 统计与信息论坛, 2019, 34(6): 29-35.

[11] 方若男, 骆品亮. 基于随机森林的第三方支付违规风险预警研究[J]. 技术经济, 2020, 39(9): 11-21.

[12] 周亮. 基于随机森林模型的股票多因子投资研究[J]. 金融理论与实践, 2021(7): 97-103.

[13] 闫政旭, 秦超, 宋刚. 基于 Pearson 特征选择的随机森林模型股票价格预测[J]. 计算机工程与应用, 2021, 57(15): 286-296.

[14] 孙玲莉, 杨贵军, 王禹童. 基于 Benford 律的随机森林模型及其在财务风险预警的应用[J]. 数量经济技术经济研究, 2021, 38(9): 159-177.

[15] 杨睿哲, 王智敏. 客户信息不完全下的授信评估问题——基于逻辑回归、神经网络等模型[J]. 现代商业, 2019(36): 91-92.

[16] 郝婷婷, 俞俊杰, 陈燕. 基于逻辑回归的商业银行客户信用评级研究[J]. 科技资讯, 2019, 17(3): 255-256.

[17] 边玉宁, 陆利坤, 李业丽, 曾庆涛, 孙彦雄. 基于逻辑回归的金融风投评分卡模型实现[J]. 计算机科学, 2020, 47(S2): 116-118.

[18] 刘荣珍. 基于逻辑回归和机器学习的个人信用风险研究[D]: [硕士学位论文]. 兰州: 兰州大学, 2021.

[19] 曹杰, 张岩松, 刘速, 等. 基于逻辑回归评分卡的石油企业供应商风险模型研究[J]. 油气与新能源, 2021, 33(5): 51-57.

[20] 张媛媛. 基于特征工程和均值不确定逻辑回归在广告和银行领域欺诈识别[D]: [硕士学位论文]. 济南: 山东大学, 2023.