Modeling and Simulation 建模与仿真, 2024, 13(2), 1797-1813
Published Online March 2024 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2024.132169

Hans汉斯

一种基于分类算法集成学习模型的金融信贷违约预测

肖家鑫，于莲芝

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2024年1月6日；录用日期：2024年3月21日；发布日期：2024年3月29日

摘　要

随着金融市场的不断发展，金融信贷业务的激增也导致了信用风险的不断增加。为了应对这一挑战，传统的风险评估方法已经不能满足实际需求。目前集成模型成为违约问题研究的热点，通过整合多个分类算法的预测结果，充分利用各个算法的优势，以提高预测准确性和鲁棒性。本文研究了双阶段异构堆叠集成模型(DH-SEM)在金融信贷违约中的应用。该模型包括两个关键阶段，在第一阶段，选择了SVM、KNN、朴素贝叶斯作为三个监督基础学习器；在第二阶段，采用了随机森林作为元学习器来预测分类结果。对于金融信贷违约预测，DH-SEM模型预测准确率为0.886，相比传统的模型预测的更加准确。

关键词

集成学习，金融信贷，分类算法，元学习器

An Ensemble Learning Model for Predicting Financial Credit Default Based on Classification Algorithm

Jiaxin Xiao, Lianzhi Yu

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Jan. 6th, 2024; accepted: Mar. 21st, 2024; published: Mar. 29th, 2024

Abstract

With the continuous development of the financial market, the proliferation of financial credit

文章引用: 肖家鑫, 于莲芝. 一种基于分类算法集成学习模型的金融信贷违约预测[J]. 建模与仿真, 2024, 13(2): 1797-1813. DOI: 10.12677/mos.2024.132169

business has also led to a continuous increase in credit risk. In order to cope with this challenge, traditional risk assessment methods can no longer meet the actual demand. Currently integrated models have become a hotspot in the study of default problems, which make full use of the advantages of each algorithm by integrating the prediction results of multiple classification algorithms in order to improve the prediction accuracy and robustness. In this paper, we study the application of two-stage heterogeneous stacked integration model (DH-SEM) in financial credit default. The model consists of two key phases; in the first phase, SVM, KNN, and plain Bayes are selected as the three supervised base learners; in the second phase, Random Forest is employed as the meta-learner to predict the classification results. For financial credit default prediction, the prediction accuracy of DH-SEM model is 0.886, which is more accurate compared to the traditional model prediction.

Keywords

Integrated Learning, Financial Credit, Classification Algorithm, Meta-Learner

1. 引言

发放贷款是全球银行的核心业务。巨大的违约损失和激烈的竞争要求金融中介机构准确有效地区分申请人。因此，银行应在申请筛选时决定是否提供信贷。数据主要来自申请表、客户人口统计以及过去借款和还款行为的大量记录。通常，信用评分问题会转化为二元或多类分类。换言之，利用信用数据开发分类器，构建决策支持系统，从而协助银行决定是否向特定申请发放贷款。应用于信用评分的预测模型大致可分为两组：统计方法和人工智能(AI)方法。许多研究都集中在建模方法上，这些方法提供了一种新颖的算法来提高信用评分的准确性。这些方法包括统计方法，如线性判别分析(Linear Discriminant Analysis, LDA) [1]和逻辑回归(Logistic Regression, LR) [2]，以及人工智能方法，如人工神经网络(Artificial Neural Networks, ANN) [3]支持向量机(Support Vector Machines, SVM) [4]和决策树(Decision Tree, DT) [5]。尽管基于人工智能的方法取得了突破，但 LDA 和 LR 等简单型仍然是流行的信用评分方法，因为它们易于实施且准确[6]。

国内在信用风险评估领域起步较晚，最初主要依赖专家经验进行评估，容易存在主观性误差。然而，自 2003 年起，机器学习方法开始引入信用风险评估。李萌[7] (2005)利用主成分分析构建 Logistic 回归模型评估商业银行信用风险。郑昱[8] (2009)通过 Probit 模型发现职业稳定性和过去信贷状况是影响个人信用的主要因素。支持向量机在 1995 年提出，适用于二分类问题，可以通过核方法进行非线性分类。姚潇和余乐安[9] (2012)将支持向量机应用于信用风险评估，并证明其具有良好的判别效果。任潇等[10] (2016)比较了四种常见的单一模型，发现 SVM 方法效果最好。随着算力提升和机器学习的发展，集成学习在个人信用风险评估中得到广泛应用。集成学习通过训练多个弱分类器并组合其结果来提高预测效果。集成学习包括 Bagging 和 Boosting 两种类型。随机森林是一种常用的 Bagging 集成算法，在商业银行和贷款机构应用广泛。方匡南、吴见彬等[11] (2010)利用随机森林评估信用风险，并取得良好表现。Boosting 是一种串行集成方法，代表算法有 AdaBoost、XGBoost 和 LightGBM。白鹏飞[12] (2017)研究发现 XGBoost 在互联网信贷风险评估中表现优于其他模型。LightGBM 是微软团队在 2017 年提出的轻型 GBDT 梯度提

升框架，具有较低的运算量和内存消耗。朱丽云[13] (2020)构建了 LightGBM 模型，并应用于个人信用风险评估，取得了良好的效果。刘晓晨[14] (2020)发现 Boosting 集成方法在个人信用评估中表现较好。

本文以阿里天池公开的贷款违约数据，包括用户基本信息，首先对数据进行一些统计性分析，然后对数据进行数据清洗，特征编码，特征衍生，特征选择等一些预处理，最终提取了 20 维的特征变量。构建了双阶段异构堆叠模型，并对模型进行贝叶斯调参优化，通过 K 折交叉验证的方式利用 AUC 值，准确率进行模型评估。本文主要研究了双阶段异构堆叠集成模型在金融信贷违约中的应用，已经有很多学者证明集成学习相比传统的模型预测的更加正确，但对于双阶段异构堆叠集成模型[15]在金融信贷违约风险预测的应用还很少。本文将多个较为前沿的集成方法应用到了金融信贷风险预测中并进行比较，分析得出了相比传统的模型预测，双阶段异构堆叠集成模型效果更好。

2. 集成学习的基础

任何集成学习系统的一般框架都是使用聚合函数 G 来组合一组基线分类器 $c_1, c_2, \cdots, c_h$ 用于预测单个输出。给定大小为 n 的数据集和维度的特征：

$$m, D = \left\{\left(x_i, y_i\right)\right\},\ 1 \le i \le n,\ x_i \in R^m \tag{1}$$

基于该集成方法的输出预测由式(1)所示：

$$y_i = \varphi\left(x_i\right) = G\left(c_1, c_2, \cdots, c_h\right) \tag{2}$$

图 1 说明了集成学习的一般抽象框架。所有集成都由一组基线分类器(分类器集成)组成，这些分类器已在输入数据上进行训练，这些输入数据产生预测，这些预测组合在一起以产生聚合预测。

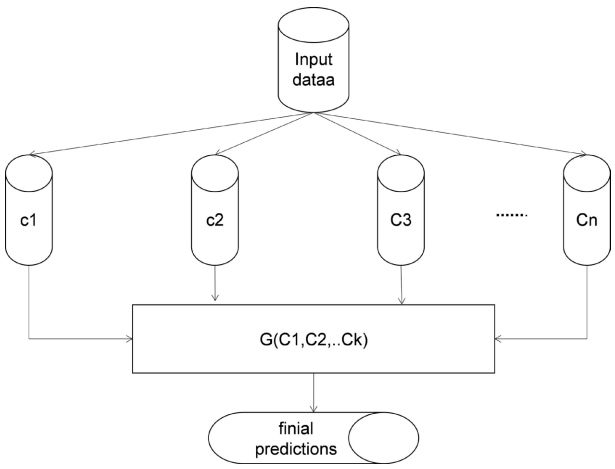

Figure 1. General framework of Ensemble
图 1. Ensemble 的一般框架

2.1. 基础分类器

2.1.1. 支持向量机

SVM 的基础是由 Vapnik 开发的，由于其各种吸引人的特性和有前途的性能而获得认可。该公式体现了结构风险最小化原则，并优于传统机器学习方法所采用的传统经验风险最小化原则。

SVM 主要描述使用支持向量方法进行分类。在分类问题中，目标是通过从可用示例中诱导的函数来分隔两个类，并且分类器在看不见的示例上工作得很好，即它能很好地泛化。2 中的一个简单的例子阐

明了这一概念。请注意，许多可能的线性分类器可以分离数据，但只有一个分类器可以最大化边距。这种线性分类器被称为最佳分离超平面(图 2)。

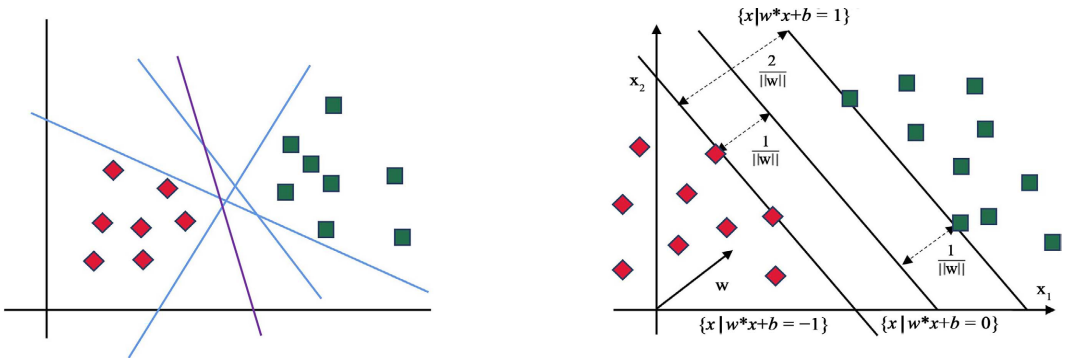

Figure 2. SVM classification
图 2. SVM 分类

2.1.2. KNN

K 最近邻(KNN)算法是一种基本的分类和回归方法。该算法的核心思想是通过测量不同特征值之间的距离来进行分类。在分类问题中，KNN 算法会寻找与新样本数据最接近的训练样本，然后根据这些邻居的类别进行投票来确定新样本的类别。在回归问题中，KNN 算法会采用邻居样本的平均值来对新样本进行预测。KNN 算法的关键参数是 K 值的选择。K 值的选择会直接影响模型的性能，较小的 K 值会使模型更加复杂、容易受到噪声干扰，而较大的 K 值会使决策边界更加平滑，但可能忽略了样本局部特征。KNN 算法简单易用，无需估计参数，适用于多分类问题。该算法的缺点是计算成本高，在数据量大时性能较差；对异常值敏感；需要事先确定 K 值。针对 KNN 算法的缺点，研究者们提出了很多改进版本，如加权 KNN、基于树结构的 KNN (如 KD 树、Ball 树)等。KNN 算法在实际中被广泛应用于模式识别、图像处理、推荐系统等领域。

2.1.3. 朴素贝叶斯

朴素贝叶斯算法是一种基于概率统计的分类算法，它基于贝叶斯定理和特征条件独立假设。该算法常被用于文本分类、垃圾邮件过滤、情感分析等领域。朴素贝叶斯算法的核心思想是通过计算给定类别下特征的条件概率来进行分类。它假设每个特征与其他特征之间相互独立，这也是“朴素”一词的来源。具体而言，对于一个待分类的样本，朴素贝叶斯算法会计算其在每个类别下的后验概率，并选择具有最高后验概率的类别作为最终的分类结果。朴素贝叶斯算法的关键是学习每个类别下特征的概率分布。通常采用极大似然估计或平滑技术来估计这些概率。常见的朴素贝叶斯算法包括多项式朴素贝叶斯、伯努利朴素贝叶斯和高斯朴素贝叶斯等。朴素贝叶斯算法具有以下优点：简单高效，易于实现和理解；对小规模数据表现良好；适用于多分类问题。然而，朴素贝叶斯算法的缺点是对特征之间的依赖关系做了较强的假设，可能导致分类性能下降。综上所述，朴素贝叶斯算法是一种基于概率统计的分类算法，具有简单高效、易于实现等优点，在文本分类、垃圾邮件过滤等领域有广泛应用。

2.2. 元学习器

RF (Random Forest, RF)分类器是一种集成分类器，它使用一组 CART (Classification and Regression Tree)进行预测[16]。树是通过替换(装袋方法)绘制训练样本的子集来创建的。这意味着可以多次选择相同的样品，而其他样品可能根本无法选择，如图 3 所示。

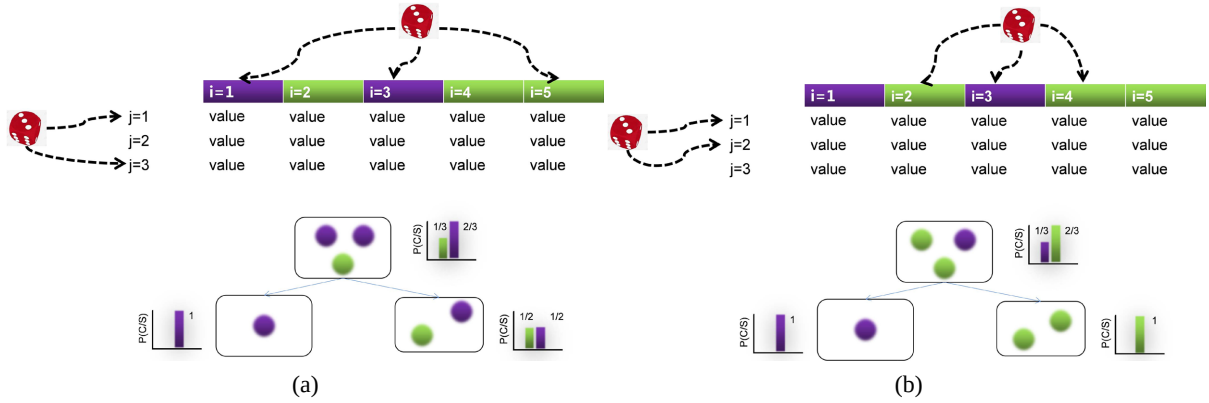

Figure 3. Training phase
图 3. 训练阶段

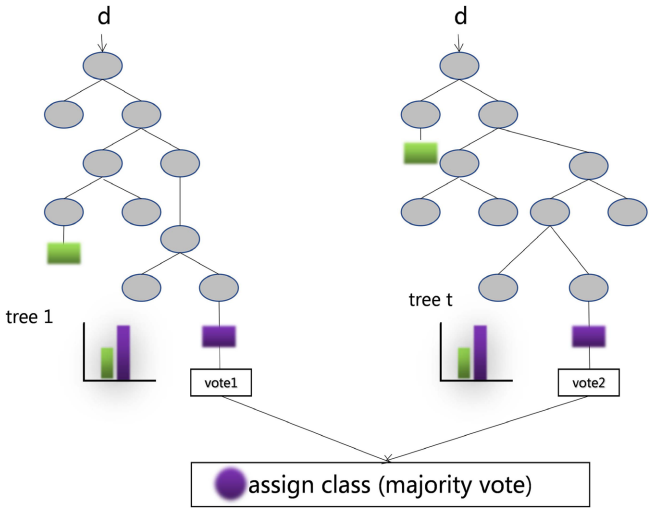

Figure 4. Classification phase
图 4. 分类阶段

随机森林分类器的训练和分类阶段：i = 样本，j = 变量，p = 概率，c = 类，s = 数据，t = 树数，d = 要分类的新数据，*value* = 变量，j 可以具有的不同值。大约三分之二的样品(称为袋内样品)用于训练树木和剩余的三分之一(称为袋外样品)用于内部交叉验证技术，以估计所得 RF 模型的性能。

此误差估计称为袋外(OOB)误差。每个决策树都是独立生成的，无需任何修剪，每个节点都使用用户定义的特征数量(Mtry)进行拆分，这些特征是随机选择的。通过将森林增加到用户定义的树数(Ntree)，该算法创建了具有高方差和低偏差的树。最终的分类决策是通过平均(使用算术平均值)所有生成的树计算的类分配概率来做出的。因此，根据集成中创建的所有决策树评估新的未标记数据输入，并且每个树都投票选出一个类成员。得票最多的会员等级将是最终选择的等级(图 4)。

3. 双阶段异构集成模型

集成方法通过训练多个分类或者回归算法以获得比任何基本算法具有更高精度的结果模型。它使用不同的技术来实现，如投票、堆叠和混合等。叠加是本文实现的一种集成算法。它的工作原理是将选定的基本算法分为两层，其中第一层形成模型堆栈，第二层是原学习分类器。堆叠组合了三个或者更多不

同模型的优点，以产生高度的准确性和控制过拟合。当它与不同的基本算法混合使用时效果最好，这些基本算法独立产生非常不同的结果。图 5 为堆叠集成算法的 2 个阶段；训练数据首先被反馈送到模型堆栈，然后将预测结果输入到第二阶段(元学习器)得到结果，元学习器与基础学习器的预测相匹配，从而做出更准确的预测。下列是建议模型的步骤。

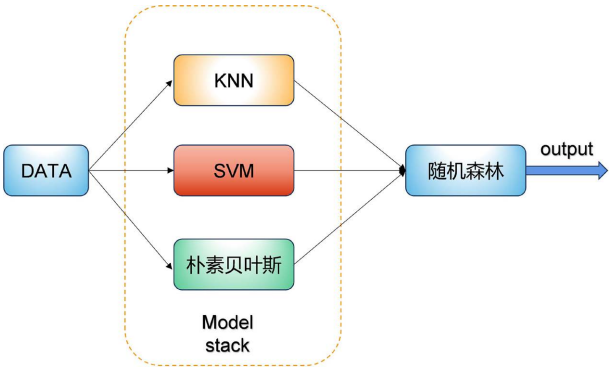

Figure 5. Stacked integrated learning algorithm
图 5. 堆叠集成学习算法

第一步是确定 N 个基本算法，这些算法构成堆栈模型(以及模型的参数)，并最适合训练数据集。接下来我们选择一个元学习算法，它构成了集成的第二阶段，在这一步中，我们在数据集上训练所选的 N 个基本算法，接下来将数据集分为 K 个部分，对每个分类算法进行交叉验证，并收集 N 个基本算法的预测结果，得到结果后，编译 M*N 矩阵，作为下一阶段的输入，这里 M 是 N 个基本算法的预测值，这些结果被称为第一阶段的结果。接下来我们在第一阶段分类器得到的结果上训练元学习算法，这些来自作为元学习器输入的基本模型的预测是“样本外”数据，这意味它们没有用于训练基本模型。元学习器的最好地结合了基础模型的预测，并对数据集的测试训练分割的不同比例得到结果。最后，我们在不同的测试数据上对集合进行测试，并评估预测精度。

4. 实验数据采集及预处理

4.1. 数据字段说明

在本研究中，使用天池比赛贷款数据集进行建模和测试，该数据集包括 47 列变量信息，如信用等级、贷款金额、最新支付信息、信用评分、财务查询数量和地址，其中 15 列为匿名变量。删除了一些没有建模价值的数据，包括当前贷款流程不完整和特征严重缺失的记录，删除率为 3.7%。最后，保留 200,000 条比较完整的贷款记录作为初始数据集。数据集所包含的部分字段如表 1 所示。

Table 1. Partial field descriptions
表 1. 部分字段描述

字段	字段描述
loanAmnt	贷款金额
term	贷款期限(year)
interestRate	贷款利率
installment	分期付款金额

续表

grade	贷款等级
employmentlength	就业年限
dti	债务收入比

4.2. 分类变量与违约状态可视化

在明确数据的来源和变量含义后，进行数据探索是研究的第一步。我们首先查看了数据集中各变量的描述统计量，并对标签的分布进行了初步探索。数据集包含 20 万条记录，其中违约为 0 的记录有 16,052 条，占比 80.5%；违约为 1 的记录有 38,948 条，占比 19.5%。可以看出，正负样本的比例接近 4:1，存在样本不平衡问题，这在金融风险评估中是常见的情况。大多数人并不会发生违约。

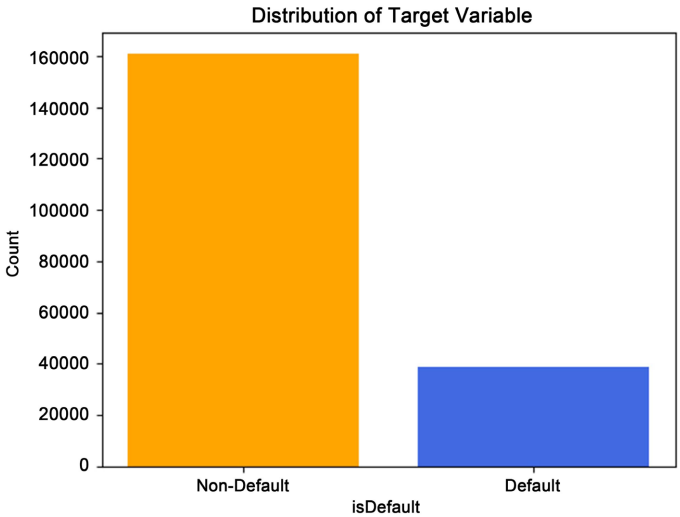

Figure 6. Histogram of target features
图 6. 目标特征直方图

针对本文研究的二分类问题，通过可视化变量与标签之间的关系进行了初步探索。图 6 展示了标签的取值分布，数据共 20 万条记录，违约为 0 的有 16,052 条记录，违约为 1 的有 38,948 条记录，正负样本接近 4:1，存在样本不平衡现象，这是金融风险评估的常见现象。图 7 为工作年限与违约的关系示意图，可以看出，工作年限超过 10 年的贷款记录最多，可能是因为工作年限越长，工作越稳定，可以稳定偿还贷款，违约的可能性比较小。此外，图 7 中可以看出各个工作年限的违约情况相差不大，一般情况下工作年限为 1 年的应该是应届毕业生，但是违约的占比和工作多年的违约占比差别不大，可能是因为个人消费贷款，金额小，所以还款没有压力。

图 8 为贷款等级占比，数据集中将贷款等级分为 A~G 七个等级。其中，贷款等级为 B 和 C 的数量最多，而贷款等级为 G 的数量最少。可能由于经过银行对借款人的综合考评，此类借款客户由于自身信用或其他原因导致其违约可能性很大，因此该等级的人能较难申请到银行贷款。图 9 为贷款等级与违约状态的关系示意图，可以看出，贷款等级为 F 和 G 的违约可能性最大，已经超过了 40%；而贷款等级为 A 的违约可能性最低，仅占 5%。随着贷款等级的提高，违约风险逐渐降低。因此，我们可以假设贷款等级对违约有一定的影响。

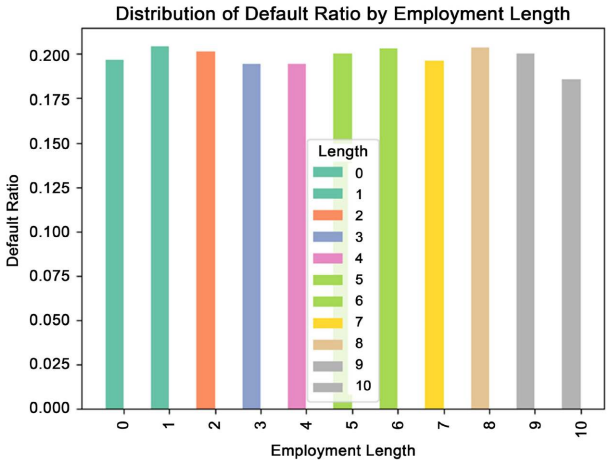

Figure 7. Percentage of work-years in default
图 7. 工作年限违约占比

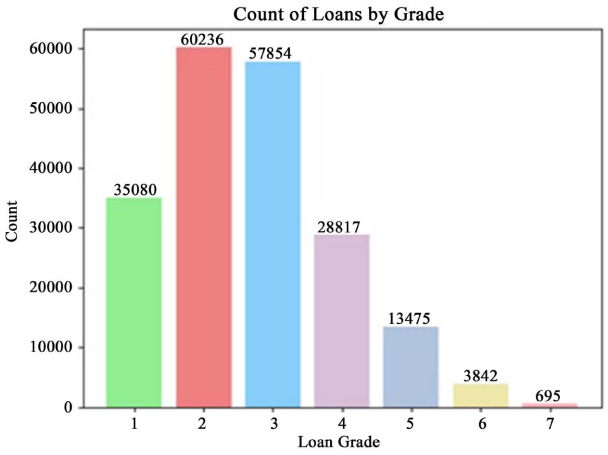

Figure 8. Percentage of loan grades
图 8. 贷款等级占比

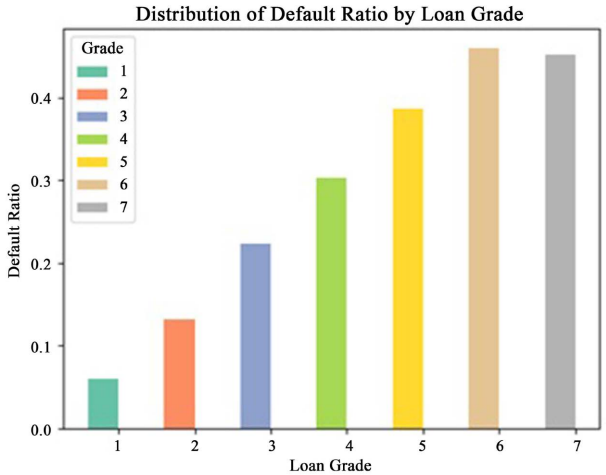

Figure 9. Percentage of loan-grade defaults
图 9. 贷款等级违约占比

4.3. 连续变量与违约状态可视化

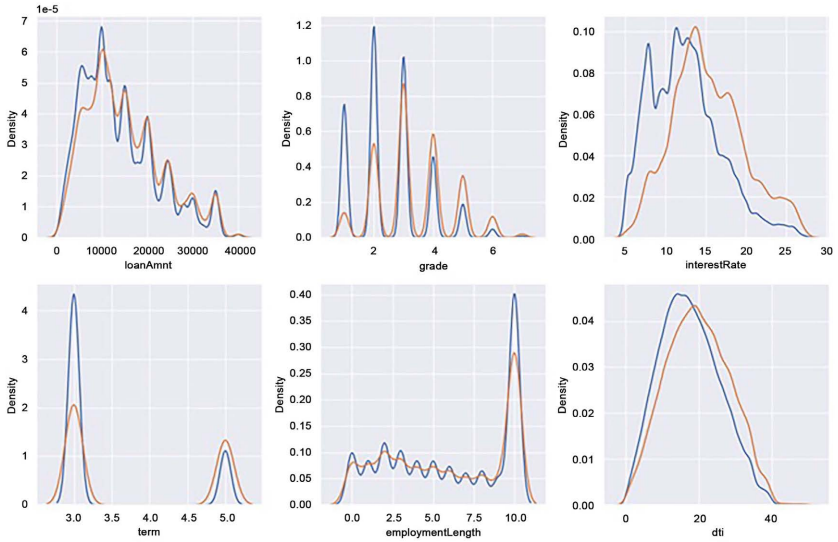

Figure 10. Partial characteristic kernel density curves
图 10. 部分特征核密度曲线

图 10 是部分特征的核密度曲线，在违约和不违约的条件下 KDE 核密度曲线存在差异，说明其在违约预测中起到一定的作用，可以初步假设它对贷款违约预测有影响。像 interestRate 中在违约和不违约状态下的分布有明显差异，利率越高的违约用户最多。而对于核密度曲线几乎完全重合的特征，可以假设它对贷款违约预测没有影响。

图 11 表示了数据集部分特征分布可视化图，这些直方图是使用 python 中的数据可视化库绘制的，可以看出这些特征分布较为均匀，不存在特别不平衡的情况。从图中可以看出贷款金额多数集中在 10,000~20,000 之间，最高贷款 4 万左右，可知样本的贷款金额范围不大。贷款年限大多数都分布在 10 年及以上。

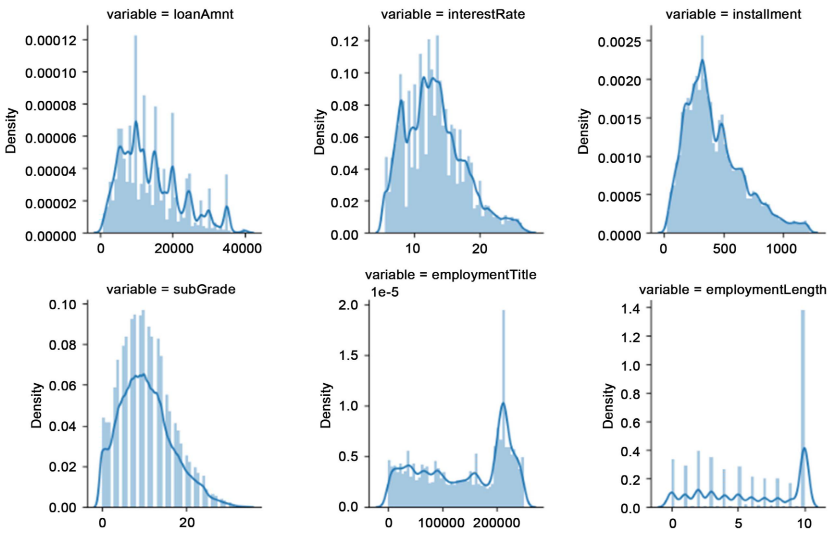

Figure 11. Visualisation of the distribution of some features
图 11. 部分特征分布可视化

4.4. 数据清洗

在对实际的研究问题中，我们挖到的数据往往会出现数据缺失，数据不平衡等问题，在进行模型训练之前需要对数据进行处理，处理结果会影响模型的建立和数据分析的质量。在本研究中，id 是唯一标识的，对实验没有帮助，将它删除，并且 grade 和 subgrade 是完全相同的特征，这里将 subgrade 特征删除。

4.4.1. 文本处理

步骤一：对贷款等级 grade 进行特征编码，将 A~F 改为 1~6；

步骤二：对贷款发放日期 issuedata 转化为时间格式；

步骤三：将贷款年限 employmentLength 中的缺失值用中位数补充，并将<1 year 用 0 替换，最后将 year 关键词删除，得到相应的工作年限；

步骤四：求出各个特征与目标的相关系数，综合考虑排除相关性小于 0.01 的特征 initialListStatus，n5，n11，n12，n8，postCode；

步骤五：将相关性过高 installment，interestrate 进行删除；

步骤六：对于高基数特征，为了更好地表示这种分类特征与目标变量之间的关系，使用一些特征编码技术来将其衍生出新的特征列 grade_to_mean_n0，grade_to_mean_n1 等，将数据特征扩展到 80 列。

4.4.2. 缺失值处理

数据集 train 共有 47 列，有 22 个特征含有缺失值，部分特征缺失比例如表 2，可以看出存在缺失值的特征有 employmentTitle、employmentLength、匿名变量 n0、n1、n3、n11 等等，其中 n11 的缺失比例最大，employmentTitle 的缺失比例最小，对匿名变量 n0、n1、n3、n11 的分析可知 n0、n1、n3 缺失的数量一样，有可能缺失的都是相同的行，经代码验证得，n0 缺失的行，n1、n3、n11 均缺失；n11 缺失的行，n0、n1、n3 部分缺失，说明 n11 缺失与 n0、n1、n3 缺失存在某种对应关系，虽然 n0、n1、n3、n11 字段含义并不清楚，但观察它们取值可以发现，取值均为整数，且均呈右偏分布，考虑用中位数来填充缺失值；同样 employmentTitle 取值也为整数，且呈右偏分布，因此对 employmentTitle 也使用中位数来补齐。对于 employmentLength 这样的特征，因为数据集存在与之相关的特征，所以对 employmentLength 用决策树填补就业年限，对于分类变量采用众数填补缺失值，对于连续变量采用中位数填补缺失值。

Table 2. Proportion of partial features missing
表 2. 部分特征缺失比例

特征名称	缺失值所占比例(%)
employmentTitle	0.000125
employmentLength	5.0849875
dti	0.029875
n0	5.033750
n1	5.033750
n3	5.033750
n11	8.719000
pubRecBankruptcies	0.050625

4.4.3. 异常值处理

通过描述统计发现 policyCode 具有唯一值 1，将其删除。为检测是否存在其他异常值，绘制箱线图如图 12 所示，对于其他存在异常值的变量，不作处理，这是因为在金融风控领域中，异常值的存在是有意义的，可能会包含其他有用的信息，有时常被看成是一种风险，所以不可轻易处理。

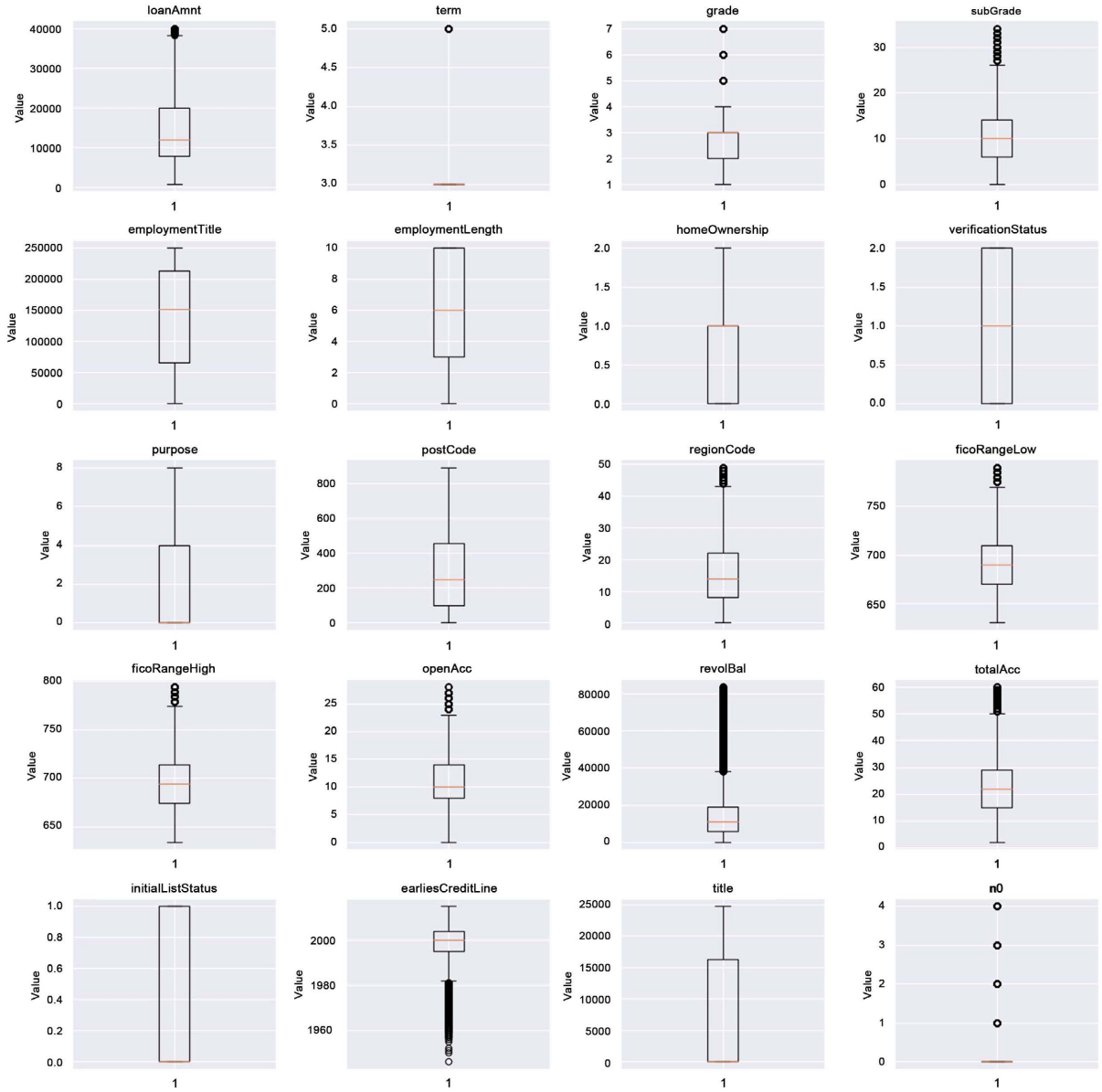

Figure 12. Box plot
图 12. 箱线图

4.5. 特征工程

在数据清洗之后，便可以进行特征构建了。特征构建也可以称为特征衍生，是根据数据集的现有特征结合业务理解去衍生出新的具有意义的特征，由于我们所得到的数据维度较小，且都是简单的单一特征，不包含由公式计算得到的复杂特征，其含有的有用信息较少，需要通过寻找特征之间的关系构建衍

生特征来体现，通常是已有数据的组合。通过对数据集中特征的观察并结合前期的分析，有以下几点思考。第一，可以根据贷款发放时间 issueDate 和信用额度开立时间 earliesCreditLine 相减来构造特征 Issue_Earlies_Diff：审批发放时长 = 贷款发放时间 − 信用额度开立的时间 = issueDate − earliesCreditLine；

第二，贷款评分与是否违约之间具有一定的关联，可以根据上下限分数加权平均构造特征 fico_average，并将评分上限和下限删掉；fico 所属平均范围=(借款人在贷款发放时的 fico 所属的下限范围 + 借款人在贷款发放时的 fico 所属的上限范围)/2 = (ficoRangeLow + ficoRangeHigh)/2；

第三，从数据缺失情况来看，由于 n0、n1、n3、n11 的缺失之间似乎存在某种关联，尝试以 n0 和 n11 为例构建特征 n0_n11：取值为 0 时表示 n0 和 n1 均不缺失，取值为 1 表示 n0 和 n11 均缺失，取值为 2 表示 n0 缺失，n11 不缺失；

第四，贷款结束的年份可能会受外部环境或市场变化的影响而间接影响还款意愿，因此考虑通过贷款发放时间 issueDate 和贷款年限 term 相加来构造特征 end_year：贷款结束的时间 = 贷款发放时间 + 贷款年 = issueDate + term。

4.5.1. 特征选择

在原有特征的基础上构建完衍生特征后，需要对所有的特征进行筛选，选择对违约预测有用的特征加入到模型中。并不是所有的特征都能提高模型的精度，特征筛选是机器学习中的一个关键步骤，它可以帮助我们选择具有重要预测性能的变量，并在模型训练中减少噪音和复杂性。常见的特征筛选方法包括相关系数法、模型自身特征筛选和 IV 值筛选。相关系数法是一种基于 Pearson 相关系数计算变量之间线性相关性的方法，如果相关系数高于给定的阈值，则删除其中一个变量。这种方法适用于处理高度相关的变量，但并不适用于非线性关系的变量。模型自身特征筛选是指在建立机器学习模型的过程中，同时进行特征选择和模型训练。这种方法可以根据评价指标自动筛选变量，从而提高模型的准确性和效果。集成学习模型如随机森林和 LightGBM 就是利用这种方法进行特征选择。这种方法在处理非线性关系的变量时表现良好。IV 值筛选特征是一种基于变量的预测能力进行量化的方法。IV 值越大，代表变量的预测能力越高。IV 值的计算是以 WOE 分箱为基础的。逻辑回归常用 IV 值进行特征筛选，其他机器学习方法则采用机器学习本身的自动筛选特征功能。本实验采用的是相关性系数法，选用特征重要性最高的 20 个特征进行训练，这样既可以降低模型的复杂度也可以减少训练时间，图 13 是特征重要性图，特征重要性图可以帮助我们识别哪些特征对于预测结果最为重要，这有助于我们选取最优的特征来进行建模，减少模型训练的时间和计算成本。我们将数据分为两组：训练集、和测试集，权重分别为 80%、20%。

4.5.2. 平衡数据集

观察数据集中的 isDefault 标签的数量分布显示，其中取值为 0 的样本数占总体样本的 19.5%，而取值为 1 的样本数占总体样本的 80.5%，呈现出明显的不平衡状态。考虑到正样本数量较少，欠采样会导致整体数据量减少，不利于模型的训练和泛化能力。因此，针对不平衡数据集，我们选择使用过采样方法来平衡数据分布。具体而言，我们尝试了随机过采样、SMOTE 过采样以及 SMOTE 与 Tomek Link 结合使用的综合采样三种方法。通过比较这三种方法在评估指标上的表现，我们可以选择最适合的方法来处理不平衡数据。在进行过采样时，我们需要保留部分原始数据用于评估采样效果。因此，我们首先将数据集按照 8:2 的比例划分成训练集和测试集，然后分别采用以上三种过采样方式增加正样本数量。接着，将三种方法得到的样本分别放入随机森林、SVM 和朴素贝叶斯模型中进行拟合，并利用之前划分出来的测试集来评估三种方法在三个模型上的表现。经过验证后发现，对于该数据集，随机过采样后的数据在三个模型上的测试集 AUC 值均高于 SMOTE 采样和综合采样。具体而言，在随机森林模型上，测试集 AUC 值为 0.837877；在 SVM 模型上，测试集 AUC 值为 0.817957；而在朴素贝叶斯模型上，测试集

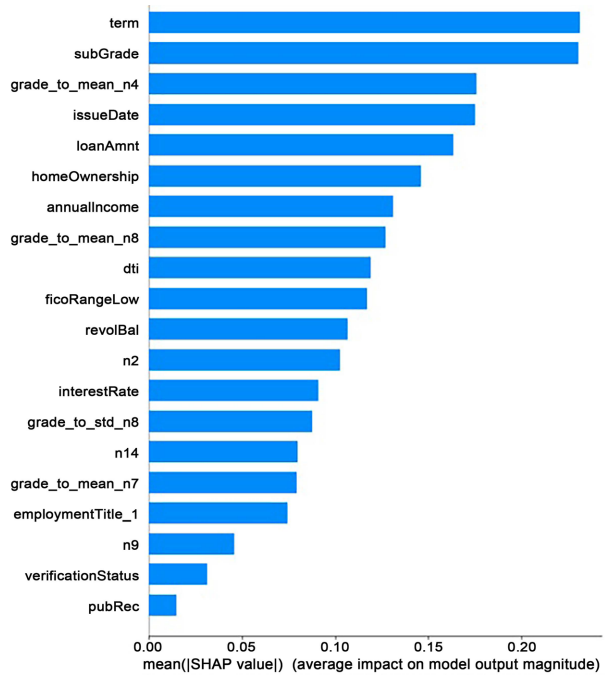

Figure 13. Feature importance
图 13. 特征重要性

AUC 为 0.805264。综上所述，根据实验结果，我们选择随机过采样方法来平衡数据集，以提高模型性能和预测准确度。随机过采样方法能够在处理不平衡数据时取得良好的效果，有助于提升模型的泛化能力和预测性能。

图 14 描述了 DH-SEM 的完整流程图，其中包括对预处理数据集、集成训练和预测分类。

步骤 1：该数据集由多个特征组成，经数据处理和最由特征选择。

步骤 2：数据被划分为 80:20 的训练测试比。

步骤 3：然后由三个基础学习器和 1 个元学习器组成的堆叠集成对这些数据进行训练和测试。

步骤 4：预测结果将数据分为两类，即违约和不违约。

集成模型的时间复杂度计算为 SVM、KNN 和朴素贝叶斯的最大时间复杂度与随机森林元学习器的时间复杂度之和。

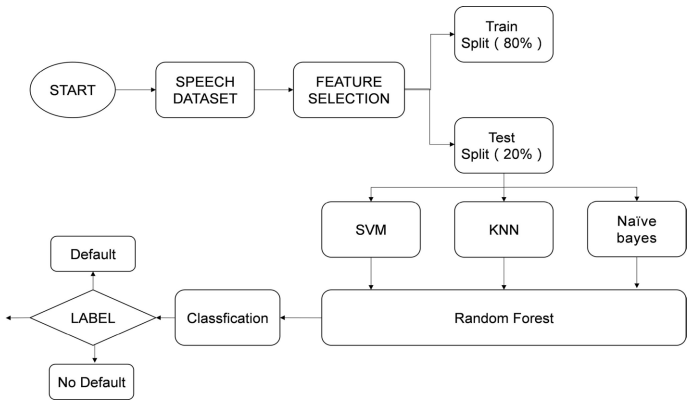

Figure 14. DH-SEM protocol flow chart
图 14. DH-SEM 方案流程图

4.6. 模型参数设置

首先，用随机过采样平衡后的数据集作为训练集，使用默认参数构建随机森林模型，用前期划分的未参与采样处理的原始数据作为测试集，用于评估随机森林的模型效果。其次，通过网格搜索来对随机森林的超参数进行调整。调整得到的最优超参数如表 3 所示。KNN 和 SVM 的模型参数如表 4，表 5 所示。

Table 3. Random forest parameter settings
表 3. 随机森林参数设置

参数	值
n_estimators	180
min_sample_split	10
min_sample_leaf	5
max_feature	10

Table 4. KNN parameter settings
表 4. KNN 参数设置

参数	值
n_neighbors	5
weights	uniform
algorithm	auto
leaf_size	30
p	2
metric	minkowski

Table 5. SVM parameter settings
表 5. SVM 参数设置

参数	值
penalty	12
loss	squared_hinge
tol	0.0001
C	6
Intercept_scaling	1
kernel	RBF
max_iter	1000

4.7. 评价方法

为了评估我们的模型，我们使用了四个最常用的评估指标，即准确率、精确率、召回率和 F1 分数(宏)，这是用于评估文本分类的典型方法。

它们的计算方法如下：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$F1(\text{Macro}) = \frac{1}{n}\sum 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

TP (真阳性)表示将阳性样本预测为阳性，*TN* (真阴性)表示将阴性样本预测为阴性，*FP* (假阳性)表示将阴性样本预测为阳性，*FN* (假阴性)表示将阳性样本预测为阴性。

5. 实验结果与讨论

本节讨论了金融信贷违约问题的建议技术的结果，以及其性能的评价与原有研究的比较。首先，将数据集分为 80%的训练数据点和 20%的测试数据点，并将类标记为违约为 1，不违约为 0。用于训练模型的计算系统有一个 2.4 GHz 的英特尔四核 i5 处理器和 12 gb 的随机存取存储器。Python 语言用于编写 Jupyter 笔记本中的算法脚本。

采用支持向量机、KNN、朴素贝叶斯作为基础学习器，随机森林作为元学习器，建立堆栈集成模型。选择 SVM 作为集成堆栈中的一种算法，因为它在小数据集和两类明显分离的情况下表现良好。使用 KNN 是因为它随着新的数据点不断进化，只有一个超参数需要调整，并且训练时间复杂度最小。朴素贝叶斯是一个理想的选择作为分类器，由于其快速的实时预测和高准确性。该算法中使用的随机森林元学习器对基础学习器的预测进行拟合，并试图克服基础学习器的局限性。在对数据集进行堆叠集成模型训练之前，先对堆叠中使用的每个模型和元学习器的测试精度进行了测量。这些结果以及堆叠集成学习算法的测试成绩如表 6 所示。该表的结果显示，与随机森林元学习器相比，前三个基本模型的性能相对较差。但是，堆叠集成能够最好地将这些结果结合起来，形成一个高度精确的模型，这是它与其他机器学习技术的区别。

Table 6. Comparison of stacked models with individual algorithms
表 6. 堆叠模型与单个算法的比较

Model	Accuracy
随机森林	0.853
KNN	0.804
朴素贝叶斯	0.753
SVM	0.822
集成模型	0.886

ROC 曲线

绘制了 ROC 曲线来评估算法的工作效果。ROC 曲线下的面积显示了分类器区分类别之间差异的程度。AUC 越高，表明模型在区分正类和负类。如果一个模型显示出 100%的准确性，那么它的 AUC 为 1，而无法正确预测任何类别的模型的 AUC 为 0。图 15 显示了考虑模型的 ROC 曲线。

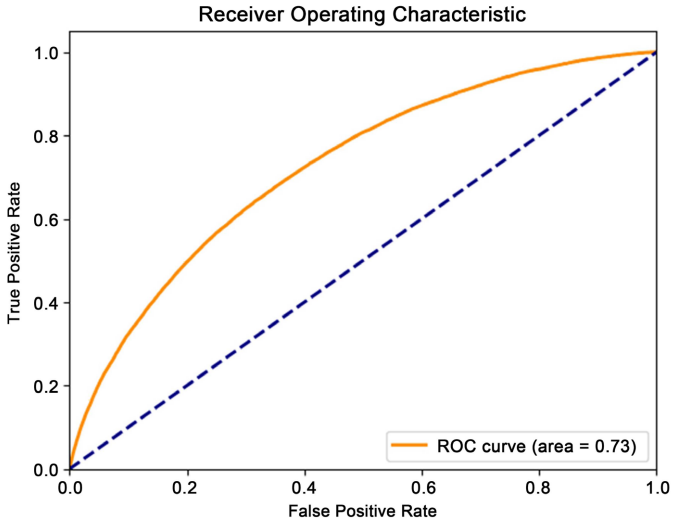

Figure 15. ROC curve
图 15. ROC 曲线

6. 结论

这项工作探索了基于阿里天池数据集的个人信息特征来确定是否违约的集成学习。研究了 DH-SEM 技术在金融信贷违约的分类性能。随机森林元学习器提高了预测精度，控制了过度拟合。利用 ROC 曲线对该方法进行了性能评价。与现有方法相比，该集成模型的性别识别准确率达到 88.6%。

参考文献

[1] Altman, E.I. (1968) Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, **23**, 589-609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

[2] Wiginton, J.C. (1980) A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *Journal of Financial and Quantitative Analysis*, **15**, 757-770. https://doi.org/10.2307/2330408

[3] West, D. (2000) Neural Network Credit Scoring Models. *Computers & Operations Research*, **27**, 1131-1152. https://doi.org/10.1016/S0305-0548(99)00149-5

[4] Huang, C.L., Chen, M.C. and Wang, C.J. (2007) Credit Scoring with a Data Mining Approach Based on Support Vector Machines. *Expert Systems with Applications*, **33**, 847-856. https://doi.org/10.1016/j.eswa.2006.07.007

[5] Lee, T.S., Chiu, C.C., Chou, Y.C. and Lu, C.J. (2006) Mining the Customer Credit Using Classification and Regression Tree and Multivariate Adaptive Regression Splines. *Computational Statistics & Data Analysis*, **50**, 1113-1130. https://doi.org/10.1016/j.csda.2004.11.006

[6] Finlay, S. (2011) Multiple Classifier Architectures and Their Application to Credit Risk Assessment. *European Journal of Operational Research*, **210**, 368-378. https://doi.org/10.1016/j.ejor.2010.09.029

[7] 李萌. Logit 模型在商业银行信用风险评估中的应用研究[J]. 管理科学, 2005(2): 33-38.

[8] 郑昱. 基于 Probit 模型的个人信用风险实证研究[J]. 上海金融, 2009(10): 85-89.

[9] 姚潇, 余乐安. 模糊近似支持向量机模型及其在信用风险评估中的应用[J]. 系统工程理论与实践, 2012, 32(3): 549-554.

[10] 任潇, 姜明辉, 车凯, 等. 个人信用评估组合模型选择方案研究[J]. 哈尔滨工业大学学报, 2016, 48(5): 67-71.

[11] 方匡南, 吴见彬, 朱建平, 等. 信贷信息不对称下的信用卡信用风险研究[J]. 经济研究, 2010, 45(S1): 97-107.

[12] 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.

[13] 朱丽云. 基于 LightGBM 算法的个人信用风险评估研究[D]: [硕士学位论文]. 广州: 华南理工大学, 2020.

[14] 刘晓晨. 基于集成策略的个人信用评估模型[D]: [硕士学位论文]. 湘潭: 湘潭大学, 2020.

[15] Taran, S. and Pandey, A. (2023) A Dual-Staged Heterogeneous Stacked Ensemble Model for Gender Recognition Using Speech Signal. *Applied Acoustics*, **205**, Article ID: 109271. https://doi.org/10.1016/j.apacoust.2023.109271

[16] Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324