基于 R 语言的个人贷款违约模型的预测与分析

***

（东北大学秦皇岛分校 数学与统计学院，河北 秦皇岛）

摘　要：近几年，金融组织和机构以及银行迫切需要研究信用风险的产生机制和影响因素，为是否向贷款人发放贷款提供可靠依据。因此研究个人贷款违约数据对预测个人贷款信用以及构建征信系统具有重要意义。R 是统计建模的重要数据挖掘软件。基于 R 软件建立个人贷款违约的决策随机森林、逻辑回归和支持向量机模型，通过剪枝、平衡数据和调节参数提高了预测精度。评估三种模型后，我们认为，去参数 gamma=0.01，coat=0.1 时的支持向量机模型的预测精度最高，对贷款者信用评估有较好的指导意义，剪枝后的决策随机森林模型次之，平衡数据后的逻辑回归模型表现依旧不佳，不太具有参考意义。

关键词：随机森林；剪枝；逻辑回归；支持向量机；特征重要性；违约判别

中图分类号：请查阅中图分类号　**文献标志码：**A

Prediction and Analysis of a Personal Loan Default Model Based on R Language

***

(School of Mathematics and Statistics, Northeastern University of China, Qinhuangdao Branch. Corresponding author: ***, email: ***@qq.com)

Abstract: In recent years, there has been an urgent need for financial organisations and institutions as well as banks to study the mechanisms and influencing factors of credit risk in order to provide a reliable basis for whether to grant loans to lenders. R is an important data mining software for statistical modelling. The decision forest, logistic regression and support vector machine models for personal loan defaults based on R software have improved prediction accuracy by pruning, balancing data and adjusting parameters. After evaluating the three models, we concluded that the support vector machine model with deparameterised gamma=0.01 and coat=0.1 had the highest prediction accuracy and was a good guide for lender credit assessment, the pruned decision forest model was the next best, and the logistic regression model with balanced data continued to perform poorly and was less informative.

Key words： random forest algorithm; prune; logistic regression algorithms; support vector machine models; eigenvalue importance; default judgement

随着经济的飞速发展，货币流通速度的加快，产生信用风险因素的原因变得愈加复杂。为了预测借款人贷款违约的可能性，并为是否向贷款人发放贷款提供参考依据，减少借款违约的可能性和带来的损失。很多金融相关组织和机构都在深入研究如何评估借款人的信用风险来规避可能的损失。而统计建模分析是数据分析的一种常见手段，其简单明了、易于实现的优点给统计建模提供了便利，但也存在获取有效信息比较有限的弊端[1]。但是，机器学习弥补了这一不足，它是一种不再局限于人脑层面的思考，而是在机器层面分析，使数据中的潜在信息更容易被挖掘。其中决策树、逻辑回归等是常用的机器学习算法[2]。

在数据挖掘领域有许多功能强大的分析工具，如 R、Python、SPSS 等。其中 R 是一种流行的 GNU 开源数据挖掘工具，是针对编程语言和软件环境进行统计与绘图的软件[3]。此外，R 还可提供统计建模功能，包括线性和非线性建模、时间序列分析、分类、预测等等[4]。针对个人贷款违约数据，R 软件能够实现有效建模与精准预测。

1　贷款信用数据模型概述

1.1　决策树模型

模型是树结构预测模型，能够用于数据样本的分类和回归分析。其核心算法较为成熟，是最广泛的分析预测方法之一。利决策树进行模型构建的过程是从决策树的根节点开始，按照待测数据与决策树中特征节点的比较结果选择下一个分支，直到叶

收稿日期：2022-10-30

作者简介：***(2002−),*,东北大学秦皇岛分校本科生

子节点作为最终的决策结果[5]。决策树算法的本质是找出每列最佳划分与不同划分的先后顺序与排列布局[6]。决策树可以根据目标变量的形式分为分类树和归回树。目标变量采用离散值，即因子变量的树模型成为分类树，而目标变量采用连续值，即数值变量的决策树成为回归树[7]。分类树和回归树分别对应分类预测与回归预测模型，分别用于分类型和数值型输出变量的预测。

1.2 逻辑回归模型

在日常学习或工作中经常会使用线性回归模型对某一事物进行预测，例如预测房价、身高、GDP、学生成绩等，发现这些被预测的变量都属于连续型变量。然而有些情况下，被预测变量可能是二元变量，即成功或失败、流失或不流失、涨或跌等，对于这类问题，线性回归将束手无策。这个时候就需要另一种回归方法进行预测，即 Logistic 回归。广义线性回归是探索“响应变量的期望”与“自变量”的关系，以实现对非线性关系的某种拟合。这里面涉及到一个“连接函数”和一个“误差函数”，“响应变量的期望”经过连接函数作用后，与“自变量”存在线性关系。选取不同的“连接函数”与“误差函数”可以构造不同的广义回归模型。当误差函数取“二项分布”而连接函数取“logit 函数”时，就是常见的“logistic 回归模型”，在 0-1 响应的问题中得到了大量的应用。在实际应用中，Logistic 模型主要有三大用途：

1）寻找危险因素，找到某些影响因变量的"坏因素"，一般可以通过优势比发现危险因素；

2）用于预测，可以预测某种情况发生的概率或可能性大小；

3）用于判别，判断某个新样本所属的类别。

1.3 支持向量机模型

支持向量机(SVM)是一种深入学习理论的数据挖掘方法，在解决小样本、非线性和高维回归分析和分类问题上有很多优点[8]。支持向量机可用于分类分析和回归分析，对应于支持向量分类机和支持向量回归机。支持向量分类机可用于研究输入变量与二分类型输出变量的关系及新数据预测，简称为支持向量分类；支持向量回归机用于研究输入变量与数值型输出变量的关系及新数据预测，简称支持向量回归[9]。支持向量机适用于大多数学习任务，包括分类和数值预测。

1.4 随机森林模型

随机森林在 2001 年由 Breiman 提出，其解决了 logistic 回归容易出现共线性的问题，它包含估计缺失值的算法，如果有一部分的资料遗失，仍可以维持一定的准确度。随机森林中分类树的算法自然地包括了变量的交互作用(interaction)，所以它也不需要检查变量的交互作用和非线性作用是否显著。在大多数情况下模型参数的缺省设置可以给出最优或接近最优的结果。

随机森林可以简单的理解为很多的决策数通过分类投票。原理大致是：对训练集进行有放回随机抽样，获得的 多个个样本形成训练集的一个子集作为新训练集。然后在生成的新训练集中随机抽取训练集的 p 个特征形成子集，利用该子集训练一棵决策树，并且不对其进行剪枝。不断的重复这个过程直至训练出 n 棵决策树，把待分类的测试样本给每棵决策树进行分类，并对每棵决策树的分类结果进行统计，以最多决策树认同的类别作为最终的分类结果。

2 数据准备及模型构建

2.1 数据抓取与整理

(1) 抓取个人贷款违约数据

本文中使用的数据来源于直接提供的数据集。其中选用 Japandata、Germandata、Australiadata 三个 xlsx 表格文件作为本次探究的数据集。在 Japandata 数据集中有 15 个特征变量和 1 个分类指标变量，这个分类指标变量用于显示贷款者是否发生了违约行为。数据集中共 690 个样本，其中最终未发生违约的样本数为 307 个(分类因子为 0)，发生违约的样本数为 383 个(分类因子为 1)。Germandata 数据集中有 20 个特征向量和 1 个分类指标变量。数据集中共 1000 个样本，其中最终未发生违约的样本数为 700 个(分类因子为 0)，发生违约的样本数为 300 个(分类因子为 1)。Australiadata 数据集中有 14 个特征向量和 1 个分类指标变量。数据集中共 690 个样本，其中最终未发生违约的样本数为 307 个(分类因子为 0)，发生违约的样本数为 383 个(分类因子为 1)。

(2) 确定训练样本和测试样本

利用 R 软件编程将所有样本随机分为训练集和测试数据集。其中 70%的数据随机划分成训练数据集，剩下的 30%数据则放在测试数据集中。

2.2 构建个人贷款违约预测模型

(1) 决策随机森林模型

分别对三个数据集，利用 R 软件基于 rf.train 训练样本集和 randomForest 包构建模型。首先调用 library 函数加载 randomForest 包，将 Y 变量作为分类变量，选入所有的特征变量，建立决策随机森林

模型。

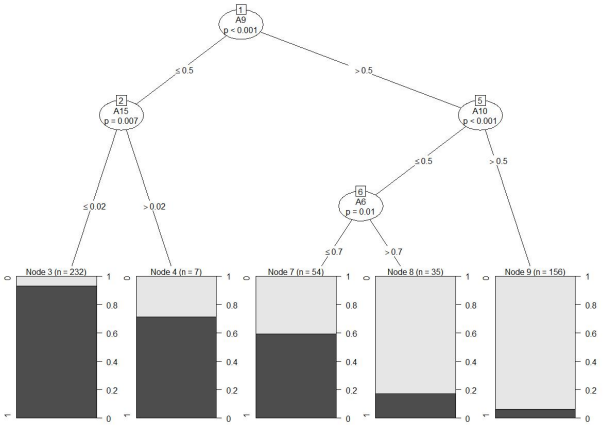

图 1 Japandata 的决策森林效果图

Fig.1 A rendering of Japandata's decision forest model

在输出结果中，我们可以得到，对于 Japandata 数据集，预测错误率为 15.73%，即分类变量为 0 的错了 38 个，分类变量为 1 的错了 38 个。利用 plot 函数画出预测效果图。

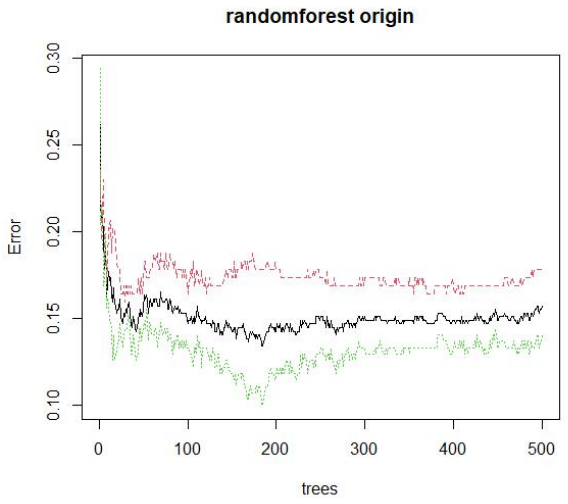

图 2 Japandata 的决策森林模型预测效果图

Fig.2 Japandata's decision forest model prediction results

根据决策森林模型对 Japandata 数据集进行预测，得到准确率为 0.8985507，kappa 值为 0.7876。即对于分类变量为 0 的预测错了 31 个，分类变量为 1 的预测错了 35 个。查看 AUC 值，得到 AUC 值为 0.9341。故 Japandata 数据集用决策森林模型预测结果较为理想。

由于本数据采集有 15 个特征向量，故需要输出变量重要性。为了直观表现，此处采用柱状图和散点图的形式表现。

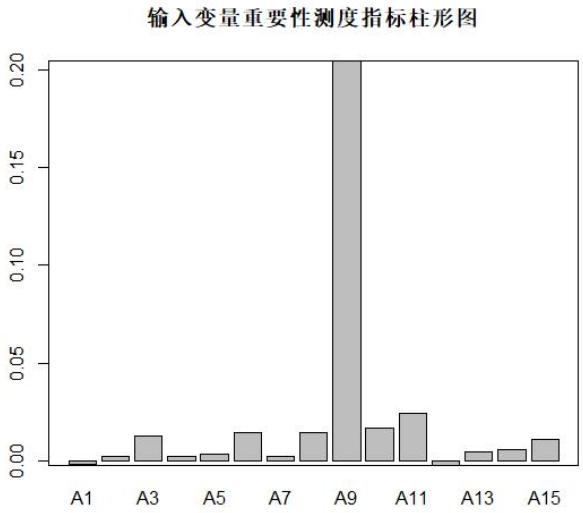

图 3 Japandata 输入变量重要性测度指标柱状图

Fig.3 Histogram of importance measures for Japandata input variables

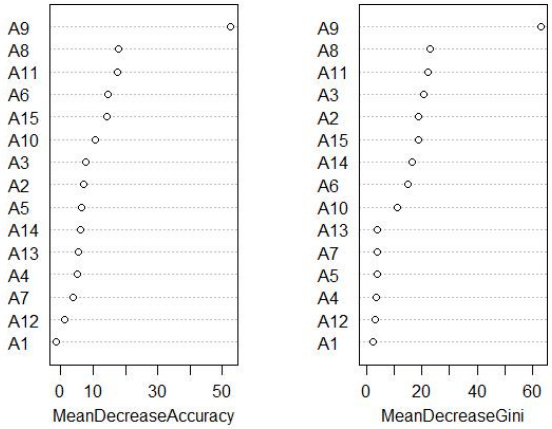

图 4 Japandata 输入变量重要性测度指标散点图

Fig.4 Japandata input variable importance measure scatter plot

最后展示随机森林模型中每棵决策树的节点数和 Japandata 数据集在二维情况下各类别的具体分布情况，如图 5 和图 6。

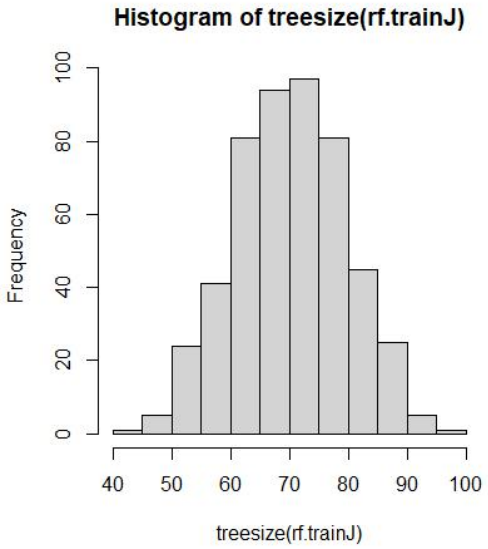

图 5 Japandata 随机森林模型中每棵决策树的节点数柱状图

Fig.5 Histogram of the number of nodes per decision tree in the Japandata random forest model

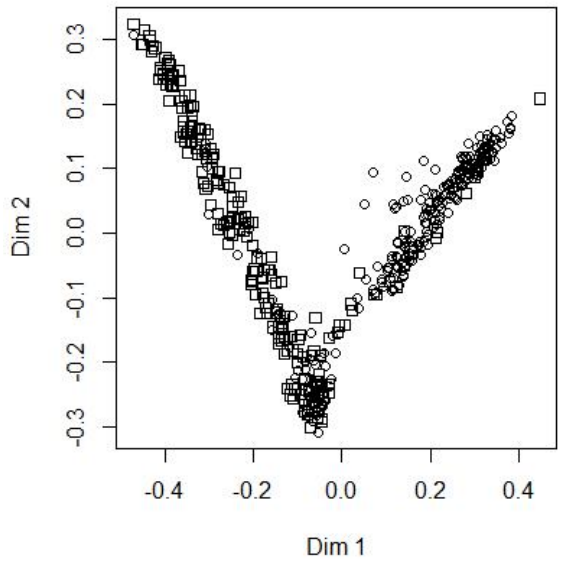

图 6 Japandata 各类别的具体分布情况

Fig.6 Specific distribution of Japandata categories

对于 Australiadata 数据集，建立决策森林模型。

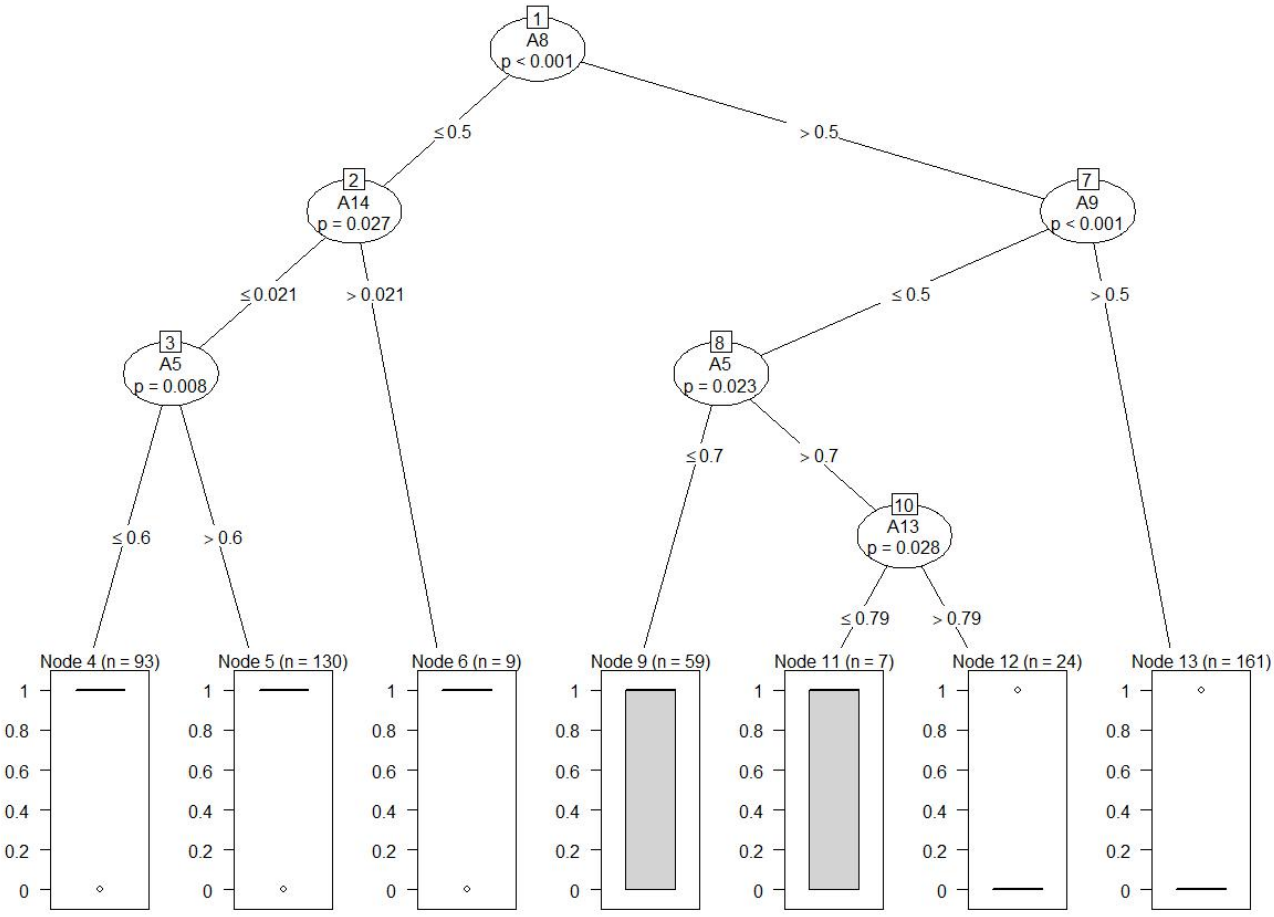

图 7 Australiadata 数据集决策森林模型效果图

Fig.7 Australiadata dataset decision forest model rendering

预测错误率为 14.7%，即分类变量为 0 的错了　　32 个，分类变量为 1 的错了 39 个。利用 plot 函数

画出预测效果图。

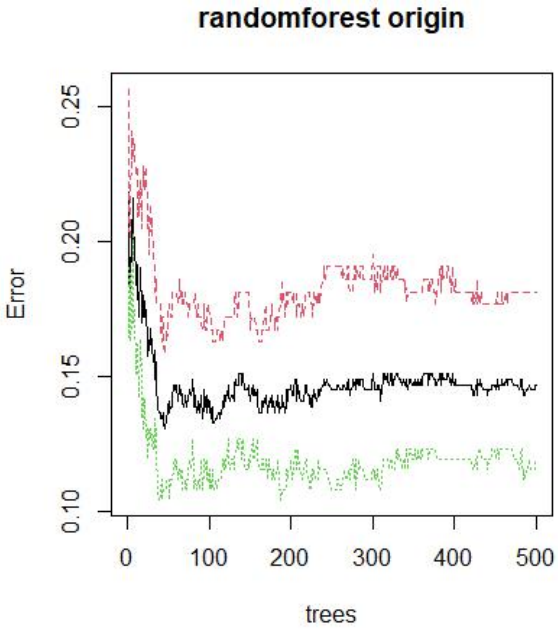

图 8 Australiadata 的决策森林模型预测效果图

Fig.8 Australiadata's decision forest model prediction results

根据决策森林模型对 Australiadata 数据集进行预测，得到准确率为 0.8454，kappa 值为 0.6908。即对于分类变量为 0 的预测错了 15 个，分类变量为 1 的预测错了 17 个。查看 AUC 值，得到 AUC 值为 0.934。故 Australiadata 数据集用决策森林模型预测结果较为理想。

由于本数据采集有 15 个特征向量，故需要输出变量重要性。为了直观表现，此处采用柱状图和散点图的形式表现。

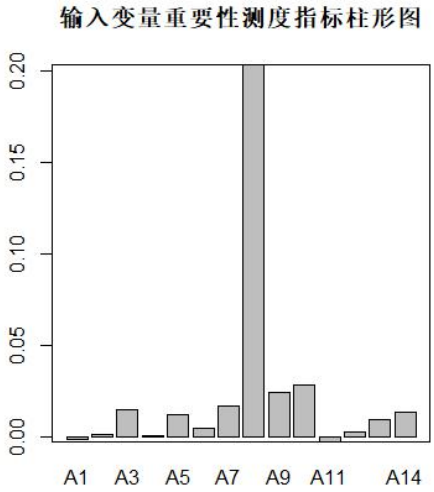

图 9 Australiadata 输入变量重要性测度指标柱状图

Fig.9 Histogram of importance measures for Australiadata input variables

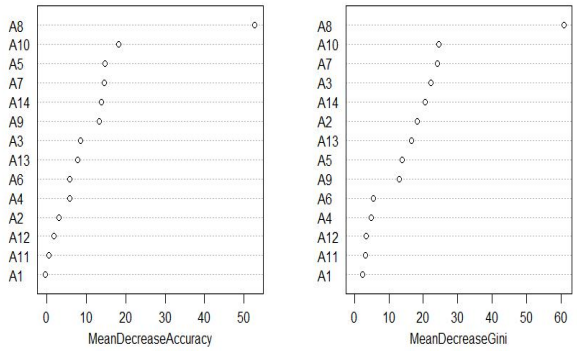

图 10 Australiadata 输入变量重要性测度指标散点图

Fig.10 Australiadata input variable importance measures scatterplot

最后展示随机森林模型中每棵决策树的节点数和 Australiadata 数据集在二维情况下各类别的具体分布情况，如图 10 和图 11。

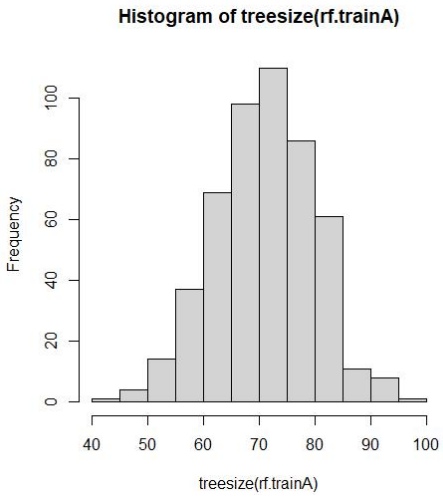

图 11 Australiadata 随机森林模型每棵决策树节点数柱状图

Fig.11 Australiadata random forest model histogram of the number of nodes per decision tree

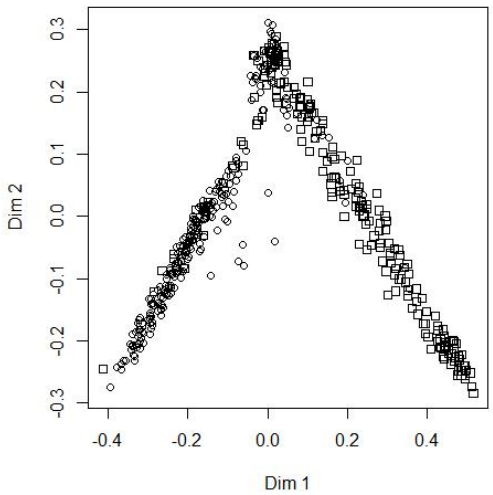

图 12 Australiadata 各类别的具体分布情况

Fig.12 Australiadata category specific distribution

对于 Germandata 数据集，预测错误率为 22.29%，即分类变量为 0 的错了 117 个，分类变量

为 1 的错了 39 个。根据决策森林模型对 Germandata 数据集进行预测，得到准确率为 0.7433，kappa 值为 0.3364。即对于分类变量为 0 的预测错了 21 个，分类变量为 1 的预测错了 56 个。查看 AUC 值，得到 AUC 值为 0.7979。故 Germandata 数据集用决策森林模型预测结果不理想，不采用随机森林模型对其预测。观察 Japandata 和 Australiadata，我们发现 Japandata 更适合决策森林。

(2) 逻辑回归模型

对 Australiadata 数据集，利用 R 软件基于 rf.train 训练样本集和 car 包构建模型。首先调用 library 函数加载 car 包，将 Y 变量作为分类变量，选入所有的特征变量，建立逻辑回归模型，并进行拟合和回归分析。

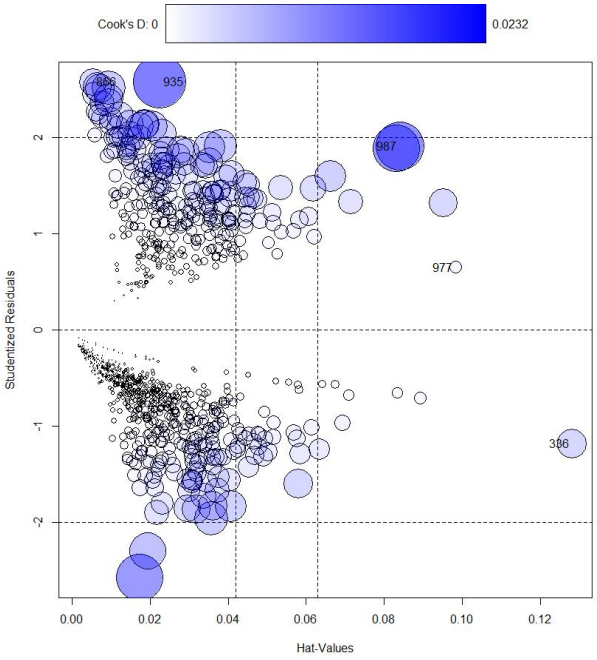

图 13 Australiadata 拟合和回归分析情况

Fig.13 Australiadata fit and regression analysis

通过编程计算我们得出此模型的精确度为 77.6%，模型精度较好。

(3) 支持向量机模型

在构建支持向量机模型时，SVM 从输入数据到输出结果的过程并不清晰，因此 SVM 属于黑盒方法。其优势在于利用了面向工程问题的核函数，能够提供准确率非常高的分类模型。针对 rf.train 集，使用支持向量机 e1071 包中 svm 函数来训练得到一个支持向量机。其中，svm()函数中 kernel 参数表示支持向量机的核函数，通过将 kernel 参数设置为“radial”来实现。当使用径向基函数作为核函数时，与其他核函数相比，针对相同的训练数据具有更高的精确度。gamma 值决定了分离超平面形状，通常与支持向量的数量相关，默认为数据维度的倒数，cost 的值默认设置为 1。使用 predict 函数和 table 函数对测试数据集进行分类预测并建立分类表。最后预测结果中，有 82 个违约用户被正确预测为违约用户，有 21 个违约用户被错误识别为未违约用户，93 个未违约用户被正确预测为未违约用户，10 个未违约用户被错误预测为违约用户。进一步调用 confusionMatrix 函数完成模型性能预测，并绘制 ROC 曲线图表示，绘制结果如下。

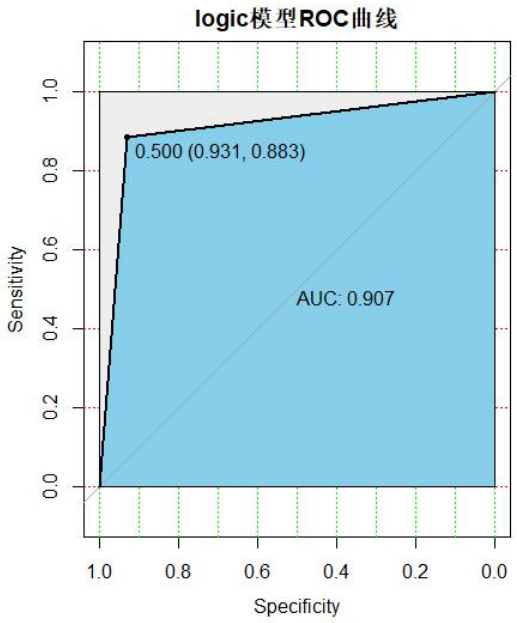

图 14 Australiadata 数据集 ROC 曲线

Fig.14 ROC curves for the Australiadata dataset

3 预测模型的改进与分析

本节将通过剪枝、平衡数据、调节模型参数等操作，基于个人贷款违约数据对于决策森林、逻辑回归和支持向量机三种模型进行改进，提高模型预测精度。

3.1 决策随机森林模型的改进

针对决策随机森林模型，为防止过度拟合，需要去除一些分类描述能力比较弱的因素，以提高模型的预测效果。对决策随机森林模型进行剪枝操作，主要步骤：①找到决策随机森林模型的最小交叉检验误差；②定位最小交叉检验误差；③计算最小交叉检验误差的成本复杂度参数值；④设置 prune 函数中的 cp 值并对决策随机森林模型进行剪枝；⑤利用测试数据集计算剪枝前后的模型预测精度。

应用 R 软件编程实现剪枝操作，结果发现，对于 Australiadata 数据集，精确度从之前的

84.54%，增加到了 86.96%。有 82 个样本被正确预测为未违约用户，98 个样本被正确的预测为违约用户。剪枝前后精度对比如表 1 所示。

表 1 剪枝前后决策森林模型精度对比

预测模型	预测精度
剪枝前的随机森林模型	84.54%
剪枝后的随机森林模型	86.96%

由表 1 可以看出，剪枝后的决策随机森林模型的预测准确率有所提高。

3.2 支持向量机模型的改进

支持向量机模型中，不改变核函数时，如果惩罚因子 cost 较小，分类间隔较大，将会产生较多错分样本；若增加惩罚因子的值，则会缩小分类间隔，但并不意味着导致错分样本的减少。gamma 是选择径向基核函数作为 kernel 后，该函数自带的一个参数。隐含地决定了数据映射到新的特征空间后的分布，gamma 越大，支持向量越少，gamma 值越小，支持向量越多。支持向量的个数影响训练与预测的速度。在 2.2 节中，支持向量机模型的预测精度在 82.5%。利用 SVM 中 tune.svm 函数对支持向量机参数进行调整，将 gamma 的取值设置为 10^{-6} 至 10^{0}，惩罚因子 cost 设为 10^{-2} 至 10^{2}。对参数多次调试后发现，当 gamma=0.01，cost=0.1 时，支持向量机模型的性能最优，对于对应模型预测精度可以达到 86.9%。

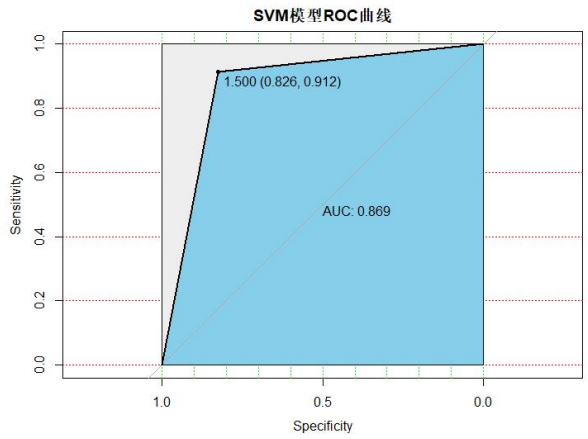

图 15 参数调整后支持向量机模型 ROC 曲线

Fig.15 Support vector machine model ROC curve after parameter adjustment

参数调整前后的支持向量机模型精度对比如表 2 所示。

表 2 参数调整前后决策森林模型精度对比

预测模型	预测精度
调整前的支持向量机模型	82.5%
调整后的支持向量机模型	86.9%

由表 2 可以看出，参数调整后的支持向量机模型的预测准确率有所提高。

3.3 逻辑回归模型的改进

对 Germandata 数据集，利用 R 软件基于 rf.train 训练样本集和 car 包构建模型。首先调用 library 函数加载 car 包，将 Y 变量作为分类变量，选入所有的特征变量，建立逻辑回归模型，并进行拟合和回归分析。

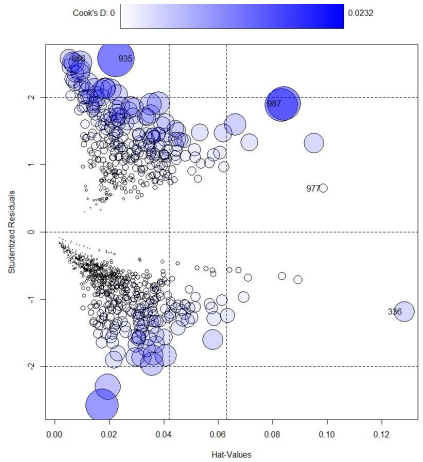

图 16 Germandata 拟合和回归分析情况

Fig.16 Germandata fit and regression analysis

通过编程计算我们得出此模型的 AIC 值为 968.52，我们知道 AIC 值越小说明模型越好，所以我们还需要对数据进行处理。通过观察我们发现此组数据不平衡，于是我们需要对本组数据进行平衡。R 语言处理不平衡数据的方法有四：欠采样法、过采样法、人工数据合成法和代价敏感学习法。此处因为数据集较大，因此我们采用欠采样方法对数据进行平衡，使指示变量为 0 的个数变为 174，指示变量为 1 的个数为 190。平衡后重新建模时，AIC 值明显降低，最后画出 ROC 曲线。

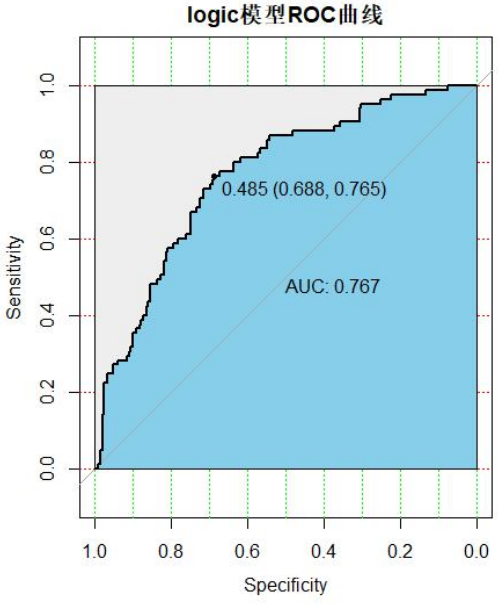

图 17 Germandata 数据集 ROC 曲线

Fig.17 ROC curves for the Germandata dataset

可以看出，平衡数据对于单一数据集来说是有预测精确度的提升的。

最后我们发现改进后的模型中决策随机森林模型的预测精度是最为理想的。它可以同时适配不同的数据集并能在每个数据集的预测上都有不错的表现。特别的，对于 Australiadata 数据集来说，逻辑回归模型的预测精度可以达到惊人的 90.7%，是我们分析统计过程中精度最高的。但是，针对个人贷款违约数据而言，无论是已有的 C5.0 算法构建的决策树和随机森林，还是本文构建的决策随机森林、逻辑回归、支持向量机模型，其预测精度都很难达到很高的精度，但是仍然能够为贷款者的信用评估提供依据。

4 结 论

以个人贷款违约数据为依托，基于 R 语言编程技术，详述了数据的抓取、分割和输入输出变量的特征含义。构建决策随机森林、逻辑回归、支持向量机模型，并对这三种模型进行改进与分析。对决策随机森林模型进行剪枝防止出现过拟合，对逻辑回归模型进行数据平衡，对支持向量机模型改变 gamma 和 cost 参数数值得到性能最优的模型。这些举动在一定程度上提高了模型预测能力。

此外，与现有经典 C5.0 决策树模型和随机森林模型比较，可以发现支持向量机模型不仅避免了决策树中过拟合现象和迭代次数的限制、随机森林中特征变量选择和度量的问题，而且模型拟合效果较好、预测精度较高，模型复杂度小、预测性能优势明显。故在预测个人贷款违约情况时，可以优先选择支持向量机模型进行预测。但同时并不是所有数据集的预测结果准确度都在 90%以上，故可能由于训练集较小导致。

因此，在过后的研究中我们可以利用爬虫技术，更多收集相关数据，增加训练样本量，对已经构建的统计模型进行多次仿真训练，以进一步达到提高模型预测精度的目的。从而为金融机构组织以及银行评估个人贷款信用以及构建完善征信系统提供科学有力的技术算法支撑和重要的实践指导依据。

参考文献

[1] 高歌.数据挖掘分类技术在商业银行贷款信用风险类别预测中的应用[D].北京:对外经济贸易大学，2011.

[2] 丘祐玮.机器学习与 R 语言实战[M].北京：机械工业出版社,2017:15-16.

[3] 李凤玲,徐力生,申群太.基于快速支持向量机算法的灌浆地层识别[J].中南大学学报(自然科学版),2019(2):478-483.

[4] 王星.R 语言在数据挖掘中的应用及其算法分析[J],2017(7):209-210.

[5] 薛涛,解蕾.R 在环境监测中的数据挖掘处理和应用分析[J].信息通信,2018(2):116-118.

[6] 张婧.基于 Logit 模型的商业银行个人贷款业务风险成因实证研究[J].经济师,2014(5):166-167,169.

[7] 周芸韬.基于 R 语言的大数据处理平台的设计与实现[J].现代电子技术,2017(2):53-56,59.

[8] 陶超,李超,李杰,等.数据挖掘在个人信用评估中的研究[J].商丘师范学院学报,2016(12):12-15.

[9] 刘建,李勇.借款人的还贷行为分析及与征信系统的关系[J].征信,2015(8):39-40.