**基于逻辑回归和随机森林模型的信用评分方法研究**

**摘 要**：信用评分通过分析个体信息预测信用违约，能够帮助金融机构科学决策，确保金融系统稳健运作。本文运用逻辑回归和随机森林模型，以信用卡申请审批为例进行建模，比较两种方法的建模效果和预测准确性。综合分析两种模型的结果，本文为金融机构提供更加科学、准确的信用评估工具，有助于降低信贷风险，提高金融机构的信贷审批效率。

**关键词：**逻辑回归；随机森林；信用评分

**一、引言**

**（一）研究背景**

信用评分模型通过分析个体的年龄、债务、收入等多维信息，预测其信用违约等的概率，对于降低金融机构的信用风险、提高贷款决策的科学性具有重要意义。在当今信息化社会背景下，统计分析和机器学习技术的广泛应用为决策提供了更为科学、准确的支持，提高了信用评分模型的预测效率和预测能力。

**（二）研究问题**

本文以是否通过信用卡申请的审批为例，借助统计分析和机器学习的方法，研究在信用评分领域逻辑回归和随机森林两种不同方法在建模过程中的表现以及对信用违约预测准确性的差异。通过对比两种算法的结果，旨在为金融决策提供更为可靠的建模工具。

**（三）研究意义**

通过本文的实证分析，可以深入了解逻辑回归和随机森林在信用评分中的应用效果，为金融机构提供更科学、更精准的信用评估工具。同时，对比分析不同算法的性能差异，能够启发我们在实践中选择合适的建模方法，提高模型的实用性。

**二、文献综述**

信用评分模型是经济金融领域的热门话题，许多研究文献都对这一主题进行了深入探讨。模型选择方面，个人信用评分模型的主要有判别分析、回归分析和数学规划几类（石庆焱、靳云汇，2003）。变量选择方面，美国FICO评分系统综合考虑偿还历史、账户数、使用年限、使用类型和新开立账户，能够较为科学地对个人信用进行评估（姜琳，2006）。在大数据和互联网技术高速发展背景下，基于逻辑回归等的传统评价方法存在局限性，随机森林、GBDT、BP神经网络等机器学习方法的优势逐渐凸显（周毓萍,陈官羽，2019）。另外，将传统方法与机器学习方法相结合组成新算法进行预测也是一种可行的思路（刘开元，2016）。

本文准确把握时代需求，在对当前技术背景深刻理解的基础上尝试定量比较不同方法的差异，探究上述文献模型选择和构建背后的内在逻辑，强调信用评分方法更新的合理性和必要性，为以后继续深入研究和探讨信用评分领域创新方法提供有力支持，具有实用性、前瞻性、创新性。

**三、数据介绍**

本文采用的信用卡审批数据主要来自开源网站UCI Machine Learning。具体来说，这是一个名为“Credit Approval”的多元数据集，记录了一些信用卡申请者的情况，并以最终是否通过审批划分为批准和未批准两类。该数据集包含16个变量，690个实例，变量有定量和定性两种。这16个变量的概要如下表1所示。

表1 研究所涉及变量表

|  |  |  |  |
| --- | --- | --- | --- |
| 变量类型 | 变量名称 | 变量含义 | 变量水平 |
| 因变量 | Approved | 是否批准 | 0=未批准，1=已批准 |
| 自变量  自变量 | Gender | 性别 | 0=女性，1=男性 |
| Age | 年龄 | 连续型 |
| Debt | 未偿债务 | 连续型，特征已缩放 |
| Married | 婚姻情况 | 0=单身/离异等，1=已婚 |
| BankCustomer | 银行客户 | 0=无银行账户，1=有银行账户 |
| Industry | 所处行业 | 当前或最近一份工作的工作部门 |
| Ethnicity | 族裔 | 白人、黑人、亚裔、拉丁裔、其他 |
| YearsEmployed | 就业年数 | 连续型 |
| PriorDefault | 先前违约 | 0=无先前违约，1=先前违约 |
| Employed | 就业状态 | 0=未就业，1=已就业 |
| CreditScore | 信用分数 | 连续型，已按比例调整 |
| DriversLicense | 驾驶执照 | 0=无驾照，1=有驾照 |
| Citizen | 公民身份 | 出生公民、其他方式公民、临时公民 |
| ZipCode | 邮政编码 | 可视为连续型 |
| Income | 收入 | 连续型，已按比例调整 |

**四、研究方法**

**（一）逻辑回归模型**

逻辑回归（Logistic Regression Model）属于广义线性回归，与多元线性回归分析存在诸多相似之处。二者在模型构建上基本一致，皆为的形式，它们的差异在于所处理的因变量类型。多元线性回归直接将作为因变量，即，而逻辑回归通过对表达式进行logit转换，最后依据阈值的设定来确定因变量的取值。

Logistic模型的一般形式为：，其中p为设定的阈值。通过调整阈值的取值来确定因变量二元取值的结果。由于信用卡审批的问题中因变量的取值只有两种，即是（用1代表）和否（用0代表），非常适合使用Logistic模型研究，因此逻辑回归是一种可选的研究方法。

**（二）随机森林模型**

随机森林（Random Forests） 是由Leo Breiman于2001年提出的一种基于决策树的机器学习方法。随机森林的算法构造如下：

1.数据采样：从训练集中进行有放回的随机抽样，抽取n个样本组成一个数据集。

2.决策树生成：从抽取的n个样本组成的数据集中创建决策树。在每个节点处，无放回地随机选择d个特征，根据这些特征拆分节点。

3.重复抽样：重复步骤1和2若干次，以生成多个决策树。

4.预测与集成：对于分类问题，通过投票的方式统计各个决策树预测结果的多数类别作为最终预测结果。

是否批准信用卡审批的问题是一种典型的分类问题，即审批是（用1代表）或否（用0代表）通过，因此也非常适合使用随机森林的方法进行研究。

**（三）模型比较**

在信用卡审批等分类问题的研究中，当前有两种主流模型——统计学模型和机器学习模型。上述逻辑回归和随机森林分别对应其中的统计学模型和机器学习模型。本文在对信用卡审批问题进行合理评估的基础上，希望能以逻辑回归和随机森林为例，比较两者的效果。下表2列出逻辑回归模型和随机森林模型的部分优缺点，这些特征也可以为建模提供启发。

表2 逻辑回归模型和随机森林模型的优缺点

|  |  |  |
| --- | --- | --- |
| 模型 | 优点 | 缺点 |
| 逻辑回归模型 | 基于线性方程，易于理解和解释模型的决策过程；  系数可以表示特征对响应变量的影响程度，有助于分析特征的重要性；  训练过程较快，计算资源需求较低。 | 对于非线性关系问题表现可能较差；  对于异常值较为敏感，可能导致模型预测不稳定；  特征选择对于模型的性能至关重要，但特征选择过程繁琐且耗时。 |
| 随机森林模型 | 具有较好的泛化能力和较高的预测准确性；  对噪声和异常值具有较好的容忍度，能够处理高维数据；  在预测过程中可以采用并行处理，提高计算效率；  可以应用于分类和回归多种问题。 | 需要构建多个决策树，计算成本较高；  特征选择是在构建决策树过程中随机进行的，可能未能充分利用所有有效特征；  当特征数量较少时，模型性能可能受到影响。 |

**五、研究结果**

**（一）数据预处理**

来自UCI Machine Learning的原始数据存在一些缺失值，并且对于变量的描述不够清晰。为解决上述问题，使数据分析能够顺利进行，需对原始数据进行预处理。本文通过填充缺失值，确保数据集的完整性，使得每个样本都包含等量信息。填充缺失值的方法根据变量类型进行差异化处理，保证了对连续型和分类型变量的合理补全。此外，本文还通过推断补充了变量的分类名称，使得数据集更容易理解和使用。

**（二）描述性统计**

为了初步探寻因变量是否批准与各解释变量之间的关系，可以观察在因变量的两种水平0和1下各特征呈现出的分布是否具有显著差异。由此绘制如下图1。

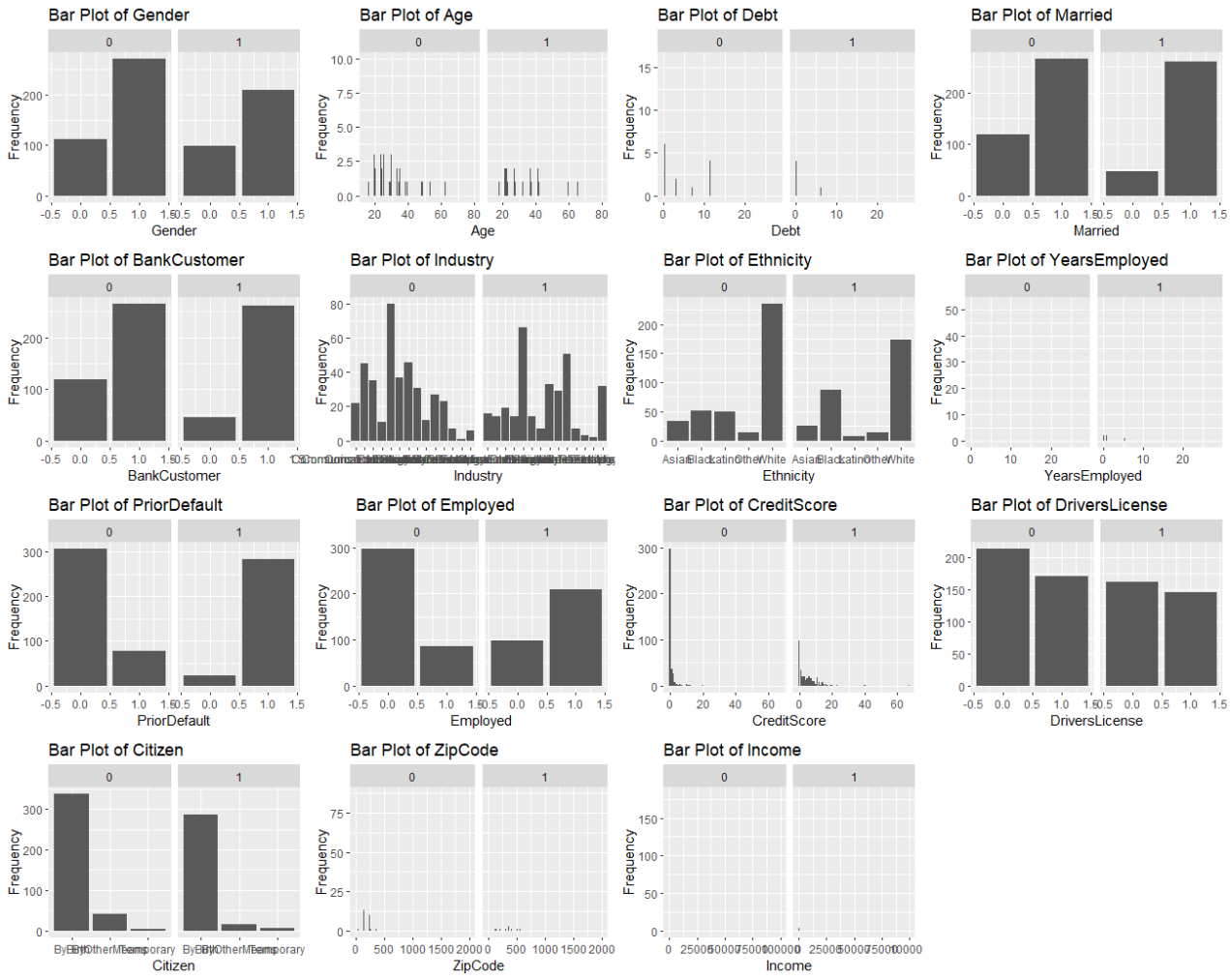


图1 按是否批准（Approved）分组柱状图

观察上图1，有些特征在因变量不同水平下呈现出较大差异，例如，拥有银行账户的人获批比例更高，失业者和有先前违约记录的人大概率不能获批，而就业者和没有违约记录的人反之等。连续型数据方面，获得批准的人未偿债务相对较少，未获批准的人信用评分较低且呈现强烈的右偏等。由上面的分析可以初步得出，该数据集中因变量是否批准与许多解释变量之间均存在一定的相关关系，选择其中一些用于预测具备合理性。

为了更清晰地了解其中连续变量的情况，绘制箱线图2。在其中为了强调箱体分布和减少图形复杂性，对离群点，即落在和之外（其中IQR表示四分位距）的点做隐藏处理。

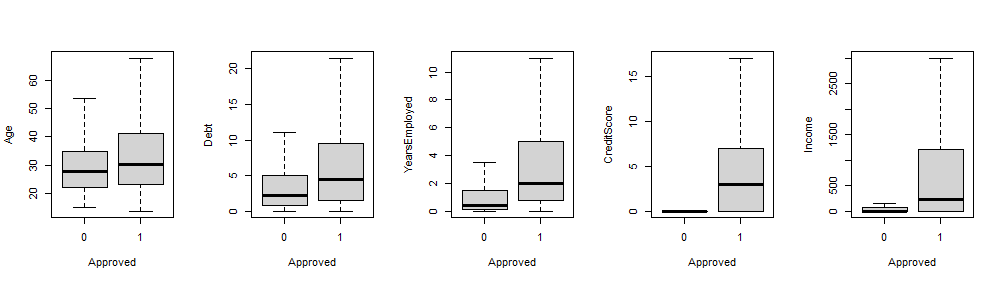


图2 按是否批准（Approved）分组箱线图

观察上图2，年龄、未偿债务、就业年数、信用分数、收入在批准的两个水平下均呈现显著差异，获批群体更为年长、负债更多、工龄更长、信用评分更高、收入更多。这基本与经验认知一致。另外，年龄与工龄、收入等通常有正相关关系，在数据分析时可能需要考虑共线性问题。

**（三）逻辑回归模型**

1.变量选择

在面对具有多个水平的分类变量时，可以采用哑变量编码，这种编码方式将原始的分类变量拆分成多个二元（0或1）的虚拟变量，每个虚拟变量代表原始变量的一个水平。例如对于本数据集中的特征Citizen，它具有三个水平ByBirth、ByOtherMeans和Temporary，可以将其用两个虚拟变量和来表示。其中：

根据和的不同取值可以体现Citizen的取值。

利用R语言中广义线性回归工具，将以上15个特征经过哑变量编码处理后全部代入作为解释变量，进行全模型逻辑回归。观察p-值，IndustryUtilities、PriorDefault、CreditScore、CitizenTemporary、ZipCode、Income 6个变量的，对于解释因变量的变异起着重要作用。

回归结果还显示，零模型偏差为948.16，残差偏差为401.64，残差偏差明显小于零模型偏差，说明模型中的解释变量对因变量的变异起到了一定作用，模型的拟合效果相对较好。除上述6个变量以外的变量在95%的置信水平下不显著。这可能是受多重共线性的影响，导致联合显著性较好，但单个变量不显著。

为了尽可能减小多重共线性的影响，对全模型进行VIF检验。结果显示，Married和BankCustomer两个变量的VIF值大于5。在处理多重共线性时，可以删除VIF较高的自变量以减少冗余信息，提高模型的稳定性和解释性。同时根据经验，婚姻情况和有无银行账户在本问题中并无特别重要的解释能力，可以考虑在后续建模中删除。

逐步回归是一种通过逐步添加或删除预测变量改进模型的方法，以达到最小化信息准则（本文选择AIC最小信息准则）的目标。对全模型做逐步回归，最终得到当选择Married、BankCustomer、Ethnicity、PriorDefault、Employed、CreditScore、Citizen、ZipCode、Income为解释变量时可使AIC取到最小值452.77。又因此时回归结果显示Married、BankCustomer两者不显著，再结合上文对VIF的分析，最终确定将它们删除。

结合上面的分析，考虑将Ethnicity、PriorDefault、Employed、CreditScore、Citizen、ZipCode、Income或其对应的虚拟变量作为解释变量进行逻辑回归。

2.模型构建

在模型建立前，首先进行数据的分割和划分，为建立和评估模型提供独立的训练和测试数据，从而更准确地评估模型的性能，提高模型的泛化能力和在实际应用中的预测能力。本文将80%的数据划分为训练集，剩余的20%划分为测试集，接着运用选择的变量进行逻辑回归，得到回归结果见表3。

表3 逻辑回归结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 估计值 | 标准误 | z-值 | p-值 |
| 常数项 | -3.44 | 0.57 | -6.03 | <0.05 |
| EthnicityBlack | 0.75 | 0.54 | 1.40 | 0.16 |
| EthnicityLatino | -1.36 | 0.79 | -1.72 | 0.09 |
| EthnicityOther | 0.51 | 0.84 | 0.60 | 0.55 |
| EthnicityWhite | 0.34 | 0.47 | 0.72 | 0.47 |
| PriorDefault | 3.58 | 0.34 | 10.41 | <0.05 |
| Employed | 0.70 | 0.38 | 1.84 | 0.07 |
| CreditScore | 0.10 | 0.06 | 1.73 | 0.08 |
| CitizenByOtherMeans | 0.28 | 0.49 | 0.57 | 0.57 |
| CitizenTemporary | 3.12 | 0.90 | 3.47 | <0.05 |
| ZipCode | 0.00 | 0.00 | -1.26 | 0.21 |
| Income | 0.00 | 0.00 | 3.08 | <0.05 |

对表中结果进行解读，截距项p值远小于0.05，表明在其他解释变量保持不变的情况下，模型中存在显著的截距项。这意味着当所有其他解释变量为零时，是否批准的对数几率为-3.44。在族裔方面，黑人、拉丁裔、其他、白人各族裔的系数分别为0.75、-1.36、0.51、0.34，这些系数表示与参考族裔（亚裔）相比，每个族裔对是否批准的对数几率的影响。先前违约的系数为3.58，且p值极小，表明在其他变量不变的情况下，有违约历史的个体更有可能被批准。就业状态、信用分数的系数为正且p值较小，说明在其他条件相同的情况下，就业状态、信用分数对是否批准的对数几率有一定的正向影响。邮政编码和收入的系数在保留两位小数的情况下均为0，这表明在其他条件相同的情况下，邮政编码和收入对获批准的影响极小。

综上，利用逻辑回归模型，构建关于是否批准的函数：

其中，是被批准的概率，表示对数几率，可以通过 转换为概率，从而得到模型对每个样本被批准的估计概率。该函数可用以预测新数据点被批准概率。

**（四）随机森林模型**

随机森林通过观察在构建每棵树时每个特征被用来进行分割的频率为每个特征分配一个相对的重要性分数，因此即使数据中存在较多特征，模型也可以自动识别对于目标变量的预测最为关键的特征，无需进行变量选择。为使两种模型预测效果可比，采用构建逻辑回归模型时的训练集和测试集。图3可视化了随机森林模型的结果。

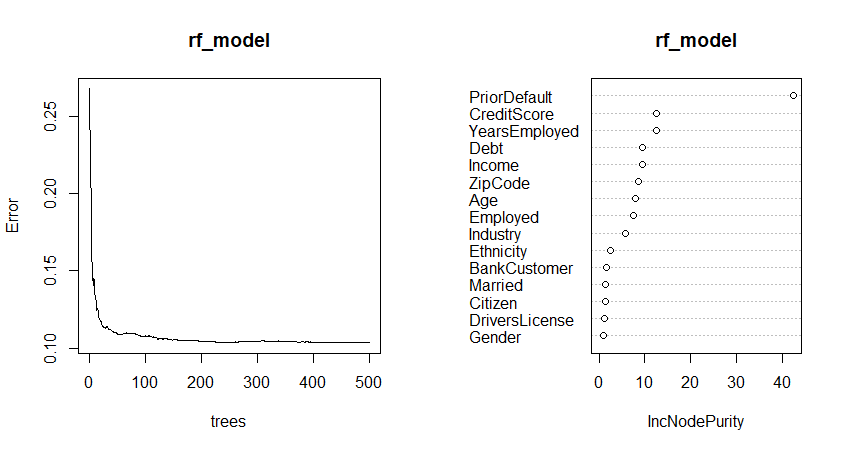


图3 随机森林模型袋外误差随树数量变化和特征重要性图

袋外误差随树数量变化图呈现了完全无监督的随机森林方法得到的训练数据结果，它展示了随着模型中树的数量的增加，模型错误率的变化情况。如图所示，随机森林算法的误差逐渐减小且趋于平缓，表明随机森林算法的泛化能力逐渐增强。特征重要性图展示了15个解释变量的贡献率，其中空心点在纸带区域分布在最右边表示该特征对是否批准的贡献作用最大。贡献率数值见表4。

表4 变量贡献率（递减顺序）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 变量 | PriorDefault | CreditScore | YearsEmployed | Debt | Income |
| 贡献率 | 42.506519 | 13.203137 | 12.463145 | 10.027807 | 9.538538 |
| 变量 | ZipCode | Age | Employed | Industry | Ethnicity |
| 贡献率 | 8.660741 | 8.040899 | 6.835598 | 5.879687 | 2.333236 |
| 变量 | BankCustomer | Citizen | Married | DriversLicense | Gender |
| 贡献率 | 1.372140 | 1.358598 | 1.355205 | 1.227700 | 1.016765 |

以上结果与逻辑回归有一定相似性，例如先前违约对最终能否通过审批影响极为显著；也存在不同之处，例如随机森林模型中贡献率较高的就业年数甚至并未出现在逻辑回归最终选择的变量集中。这表明两种方法存在差异，需结合两种模型的结果进行综合分析。

**（五）模型比较**

本文引入了逻辑回归和随机森林两种分类算法，以对样本进行分类和预测。逻辑回归作为一种经典的线性模型，被广泛用于二元分类问题。其简单有效的特性使其应用于许多场景；随机森林采用集成学习的策略，通过整合多个决策树的预测结果来提升分类性能。这两种方法在模型背后的理论和计算机实现上存在显著差异。为将差异定量呈现，本文引入混淆矩阵和ROC曲线对这两种算法在处理数据集上的性能进行评估。首先，计算并整理出逻辑回归和随机森林在测试集上的混淆矩阵如图4。

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 逻辑回归 | | 实际类别 | |  | 随机森林 | | 实际类别 | |
| 0 | 1 |  | 0 | 1 |
| 预测类别 | 0 | 68 | 6 |  | 预测类别 | 0 | 73 | 12 |
| 1 | 10 | 54 |  | 1 | 5 | 48 |

图4 逻辑回归和随机森林模型混淆矩阵

混淆矩阵是用于评估分类模型性能的表格，其中包括了模型的真正例（记作TP）、真负例（TN）、假正例（FP）和假负例（FN）的计数，通过衡量模型对不同类别的分类准确性评估其整体性能。使用如下性能指标来评估性能：

其中，准确率表示模型整体上的分类准确率，精确率表示模型在被判定为正例的样本中实际为正例的比例，召回率表示模型能够较好地捕捉到实际为正例的样本，F1分数是对模型精确率和召回率的综合考虑。经计算，得逻辑回归和随机森林模型的各性能指标的值如表5。

表5 逻辑回归和随机森林模型性能

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 准确率 | 精确率 | 召回率 | F1分数 |
| 逻辑回归 | 88.41% | 90.00% | 84.38% | 0.87 |
| 随机森林 | 87.68% | 80.00% | 90.57% | 0.85 |

表中数据显示，在该测试集下，逻辑回归综合性能略优。就具体问题而言，假正例是需要关注的量，因为信用卡的审批是一种保守行为，若过多的实际中不应通过审批的个体获得了批准，将给金融机构带来较大信用风险。从这个角度来说，随机森林较小的FP更具优势。

ROC曲线显示了模型在不同阈值下真正例率（TPR）与假正例率（FPR）之间的权衡关系。其中真正例率和假正例率的计算公式为：

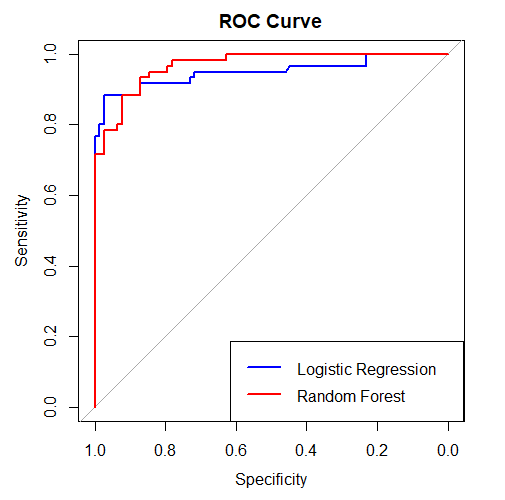


图5 ROC逻辑回归 v.s.随机森林

以FPR为横轴，TPR为纵轴，得到如图5所示ROC曲线。ROC曲线向图形左上方凸起程度反映了在不同阈值下区分正负例的能力。图形显示逻辑回归和随机森林的性能均较为优异。曲线下方的面积AUC可以用于定量地表示分类预测能力。计算得逻辑回归和随机森林模型ROC曲线对应的AUC值分别为0.95和0.97，这进一步反映了它们在预测上的强大能力。

值得注意的是，图5的ROC曲线反映了当阈值较小时，逻辑回归呈现出更好的性能，而当阈值较大时，随机森林的性能更加优异。对于信用卡审批的问题来说，较高的阈值是更加合理的选择。此时模型更为保守，倾向于提高精确率，以减少误报的风险，符合降低风险的目标。

**六、研究结论与建议**

在对信用卡审批数据集进行分类和预测时，本文引入逻辑回归和随机森林两种分类算法，通过对混淆矩阵和ROC曲线的评估，发现逻辑回归在综合性能上略优于随机森林，但在特定问题场景中，随机森林在减少假正例方面具有优势。两种模型均具有高AUC值，表明它们在预测上表现出强大的性能。另外，研究中发现随机森林的性能相对逻辑回归略显不足，这可能与数据集提供的样本量较小有关。随机森林作为一种集成学习方法，通常在大规模数据集上表现更为出色，为了更准确评估随机森林的性能，可考虑分析样本量更大的数据集，以发掘随机森林模型在信用评分领域的潜力。

基于上述研究结果，可在实际应用中根据具体问题场景选择适当的分类算法。若降低假正例对于问题的重要性较大，可以优先考虑随机森林模型；在对整体性能有更高要求的情况下，逻辑回归或是更合适的选择。此外，对于信用卡审批等需要保守决策的问题，往往在选择模型时考虑较高的阈值。本文研究发现，随机森林在较高阈值下表现出色，这与研究问题的特殊需求相吻合。采用高阈值有助于确保模型更为保守，符合降低风险的整体目标。

**参 考 文 献**

[1]石庆焱,靳云汇.个人信用评分的主要模型与方法综述[J].统计研究,2003(08):36-39.DOI:10.19343/j.cnki.11-1302/c.2003.08.008.

[2]姜琳.美国FICO评分系统述评[J].商业研究,2006(20):81-84.DOI:10.13902/j.cnki.syyj.2006.20.025.

[3]周毓萍,陈官羽.基于机器学习方法的个人信用评价研究[J].金融理论与实践,2019(12):1-8.

[4]刘开元.随机森林与逻辑回归模型在违约预测中的应用[J].信息与电脑(理论版),2016(21):111-112.