**摘要**

随着政策环境的改善，个人资信评估工作也开始逐步显现成效，但我国社会信用体系建设起步较晚，人们对于保持良好信用记录的重要性认识不够，距离形成系统化和标准化的个人信用制度目标还很远。目前随着技术的进步，人工审核发展起来的信用决策方式正被以数据驱动为核心的、更加科学合理的、自动化的决策过程所代替。因此，研究个人信用风险评估技术对推动我国信用卡业务的持续稳定发展，推进我国个人信用制度的建设具有十分深刻的现实意义。

本文以个人信用评估方法为研究对象，通过在对Logistic回归模型、广义加性模型以及随机森林算法三种不同模型比较的基础上，建立个人信用评估组合模型，并在实证中进行了比较分析。本文从考虑自变量间的交互作用角度采用了分类树和广义加性模型的组合模型，首先分类树提取自变量之间的交互作用，然后引入广义加性模型，从而提升了广义加性模型的预测精度；从考虑信用市场的个体差异性角度，先利用随机森林的聚类算法进行划分，并用聚类有效性指标确定聚类中心数，再利用动态组合预测结果的方式进行预测。利用实际数据进行实证分析，结果显示两种组合模型都提高了预测精度，而通过聚类技术的组合方式，将在不同样本区域上建立的Logistic回归模型进行距离加权，保留了Logistic回归模型可解释性强，稳定性高的一些优点，具有良好的可操作性，有实际应用的价值。 同时也证明了变量之间非线性假设的存在性。

**关键词**：个人信用评估 组合模型 随机森林算法

**Abstract**

With the improvement of the policy environment, personal credit evaluation begins to have an effect gradually. But the social credit system is late to construct. People do not realize the importance of maintain a good credit record. We're currently so far from our goals that constructing a systematic and standardized personal credit system. With the development of technology, the credit decisions approved in a manual way is replaced by the credit decisions which are made in a more scientific, reasonable and automated way. So the research on credit risk assessment techniques has very profound practical significance to promote the sustainable and stable development of China's credit card business, and also promote the construction of China's personal credit system.

In this paper, the combined model is constructed on the basis of comparison of the logistic model, general additional model and random forest model. The model combined decision tree and general additional model is constructed from the perspective of considering the interaction between variables. Decision tree extracts the interactions between independent variables firstly, and then the interactions are added into generalized additive model, which improves the prediction accuracy of generalized additive models. The other combined model is constructed according to consider the individual differences in the credit market. The clustering algorithm based on random forest is applied to the training dataset, and the number of cluster centers is determined by the cluster validity index. The outcome is predicted by a dynamic ensemble method. Using an empirical analysis of actual data, the results showed two kinds of combination models improve the prediction accuracy. The second model keeps the advantages of Logistic regression models, such as interpretability and robustness of model. So the second model can be applied in a practical way.

Key Words: Personal credit evaluation Combined model Random forest

目录

[第一章 引言 1](#_Toc416085121)

[第一节 选题背景和意义 1](#_Toc416085122)

[一、国外相关研究综述 3](#_Toc416085123)

[二、国内相关研究综述 4](#_Toc416085124)

[第三节 本文的研究思路与框架 5](#_Toc416085125)

[第四节 本文的创新点 6](#_Toc416085126)

[第二章 相关理论回顾 6](#_Toc416085127)

[第一节 Logistic回归模型 6](#_Toc416085128)

[一、Logistic回归原理 6](#_Toc416085129)

[第二节 广义加性模型 7](#_Toc416085130)

[第三节 随机森林算法 11](#_Toc416085131)

[一、随机森林定义 11](#_Toc416085132)

[二、决策树的基本思想 12](#_Toc416085133)

[三、随机森林的泛化误差 14](#_Toc416085134)

[四、随机森林的多种应用 15](#_Toc416085135)

[第三章 个人信用单一模型 16](#_Toc416085136)

[第一节 数据来源及预处理 16](#_Toc416085137)

[一、数据来源 16](#_Toc416085138)

[二、数据预处理 19](#_Toc416085139)

[三、模型评估指标 20](#_Toc416085140)

[第二节 基于Logistic回归的个人信用评估模型 21](#_Toc416085141)

[一、数据的主成分分析过程 21](#_Toc416085142)

[二、Logistic回归建模 22](#_Toc416085143)

[第三节 基于广义加性模型的个人信用评估模型 24](#_Toc416085144)

[一、广义加性模型的实证研究 24](#_Toc416085145)

[第四节 基于随机森林的个人信用评估模型 26](#_Toc416085146)

[一、随机森林模型的变量重要性 26](#_Toc416085147)

[二、随机森林模型的参数选择 26](#_Toc416085148)

[第四章 个人信用评分组合模型 28](#_Toc416085149)

[第一节 基于决策树和广义加性模型的组合模型 28](#_Toc416085150)

[一、建模依据 28](#_Toc416085151)

[二、实证分析 28](#_Toc416085152)

[第二节 基于随机森林聚类的组合模型 30](#_Toc416085153)

[一、建模依据 30](#_Toc416085154)

[二、基于随机森林的聚类算法 30](#_Toc416085155)

[三、基于随机森林聚类的组合模型 31](#_Toc416085156)

[第三节 模型对比分析 34](#_Toc416085157)

[第五章 总结与展望 36](#_Toc416085158)

[参考文献 36](#_Toc416085159)

# 

# 第一章 引言

## 第一节 选题背景和意义

在传统的商业银行业务中，信用风险是银行作为信用中介机构所面临的最主要、最复杂的风险，它具有潜藏性和滞后性的特点。从广义上讲，信用风险可以是指任何合同或协议的一方没有履行其应承担的责任时给另一方带来的潜在损失。商业银行与债务人之间存在的信息不对称现象使得银行无法直接观察信用风险的变动情况，这也加大了信用风险评估、控制的难度。商业银行信用风险管理的基础是贷款，而贷款的风险主要来源于客户还款能力和还款意愿的变化。据银监会发布的数据，截止2014年二季度末，商业银行不良贷款余额为6944亿元，较上季末增加483亿元。商业银行不良贷款率1.08%，较上季末上升0.04个百分点。结合我国商业银行的发展状况来看，信用风险仍然是其所面临最主要的风险。因此在国内经济增速放缓、国际市场瞬息万变的背景下，进一步提高我国商业银行信用风险科学管理水平已成为当务之急。

近年来，伴随着中国经济的持续稳定增长，人们的物质生活需求不断加大，信用消费、超前消费的观念得到广泛普及，包括住房抵押贷款、信用卡、汽车贷款在内的消费贷款在我国正迅速扩大规模。根据中国人民银行发布的《2014年第四季度支付体系运行总体情况》：截至2014年第四季度末，全国累计发行银行卡49.36亿张，环比增长3.99%，增速较上季度放缓0.56个百分点。其中信用卡累计发卡4.55亿张，环比增长4.38%，全国人均持有信用卡0.34张。信用卡授信总额已达到5.60万亿元，信用卡逾期半年未偿信贷总额357.64亿元，信用卡逾期半年未偿信贷总额占期末应偿信贷总额的1.53%。从中国人民银行公布的数字来看，国内信用卡市场在取得迅速发展的同时各大发卡银行所面临的信用风险问题也日益突出。另一方面，信用风险所导致的损失占整个信用卡业务损失的90%以上。因此，商业银行在扩展信贷业务和控制风险之间取得平衡是信用风险管理的重要工作，商业银行的管理人员应该对信用卡信用风险问题保持高度关注。

信用卡的发展离不开风险管理的发展。在商业银行经营信用卡业务的实践中，信用卡评估技术是商业银行个人金融业务开展及信贷审批的关键环节,也是个人信用风险管理的核心。只有通过对客户的信用风险进行科学、准确的评估预测，才能更好地为其贷款、调整信用额度等决策提供科学依据，使得商业银行更有效地降低信用风险。个人信用评估模型作为量化信用风险的重要工具，是指通过获取客户的个人基本信息（年龄、职业、家庭收入、性别、和居住信息等）、贷款申请信息（贷款用途、已有贷款信息）、交易记录、征信机构数据和外部数据以及个人消费记录等相关信息，通过统计学、运筹学、机器学习等方法，找出客户实际还款行为与客户信息之间存在的关联性，并通过模型加以量化，从而对借款人履行经济承诺的能力进行综合的判断和评估。如果通过模型可以实现对贷款申请人的信用风险的准确评估，将有利于减少审核过程中由于审核人员主观认识的干扰，从而做出更科学的信贷决策，同时也提高了信贷决策的效率，从而降低授信成本。

随着技术的进步，从人工审核发展起来的信用决策方式让位于以数据驱动为核心的、更加科学合理的、自动化的决策过程。个人信用评估技术在国外已经经历了不断的发展和创新，特别是英国、美国等发达国家已经建立了完善有效的信用管理体系，征信行业发展良好，一些政府机构和商业组织能够提供信用评分，这些信用评分概况了个人在履行其信贷义务方面的具体表现。在市场经济最发达的美国，以Experian、Equifax、Trans Union等为代表的机构是提供这种信用评分服务的典型供应商。从20世纪90年代早期开始，大多数的消费信贷决策是通过自动化的信用评估系统进行的。在我国，信用经济从20世纪90年代中后期才逐渐发展起来，征信行业仍是一个新兴的行业，在各地区之间发展不平衡。与英美等欧美国家不同，我国主要使用公共的征信系统进行商业贷款评估，因此在征信指标设置上也只能是一些常规性的基础指标，因此信息的覆盖面较窄，难以达到全面和精细化程度。根据中国人民银行在2014年4月发布的《中国金融稳定报告（2014）》显示：截至2013年末，中国人民银行个人征信数据库已累计收集近8.4亿个个人信用信息，2013年个人查询累计总量达3.4亿次。在2013年3月15日，《征信业管理条例》颁布实施，这解决了征信业发展中无法可依的问题。从以上信息可以看出，随着政策环境的改善，个人资信评估工作也开始逐步显现成效，但我国社会信用体系建设起步较晚，人们的违约成本过低，缺乏严格的监督管理机制，对于保持良好信用记录的重要性认识不够，距离形成系统化和标准化的个人信用制度目标还很远。

在信用消费市场飞速发展的今天，建立科学有效的信贷风险防控机制，对提高商业银行的风险分析和应对能力和加强人们的信用意识都十分重要。商业银行风险管理系统要针对信贷风险防控工作的实际特点，通过收集客户交易信息以及其他可获取的信息来加强对客户的监测。因此，研究个人信用风险评估技术对推动我国信用卡业务的持续稳定发展，推进我国个人信用制度的建设具有十分深刻的现实意义。

**第二节 文献综述**

### 一、国外相关研究综述

随着信用消费市场的不断发展，越来越多的统计学、数据挖掘等方法应用于个人信用评估领域。传统的统计模型主要基于多元统计分析方法，主要的模型有多元判别分析、Logistic回归方法、近邻法和贝叶斯网络，Ohlson (1980)[3]第一次将Logistic回归方法应用于金融风险预警。Wiginton(1980)[1]指出线性判别分析建立多元正态的假设上，而实际数据中很多离散型的变量显然无法满足假设，实证研究表明Logistic 回归在分类正确率上要优于判别分析法。由于Logistic 回归法并不建立在正态性假设上，而且具有良好的稳健性，因此被广泛用于建立信用评分模型。统计模型最大的优点在于其就有明显的解释性，其缺陷是现实数据难以满足过于严格的前提假设。近年来，随着人工智能、机器学习和组合分类的发展，人工神经网络、支持向量机等人工智能方法在信用评估问题上进行了有益的探索和创新。神经网络通过模拟大脑的基本特性，将大量的处理单元互相连接，具有很强的非线性关系处理能力，能有效解决非正态分布、非线性的信用评估问题。Hussain（2014）[5]的研究表明Logistic回归模型在总体准确度上要优于径向基函数神经网络，但是径向基函数神经网络在识别违约客户方面表现更好。West(2000)[2]比较了多层感知器、混合专家系统、径向基函数、学习向量量化神经网络和模糊适度感知器这五种人工神经网络的分类精度，实证结果表明径向基函数神经网络的精度要高于常用的多层感知器方法。Tian-Shyug等(2002)[4]将传统的判别分析方法与BP神经网络进行组合，简化了人工神经网络的结构，从而加快了收敛速度，实证结果表明组合方法与判别分析方法和Logistic回归方法相比提高了预测精度。为了克服神经网络在个人信用评估问题上缺乏可解释性的缺点，人们开始探索增强神经网络的可解释性的方式。神经模糊系统将神经网络与模糊系统相结合，利用模糊逻辑可见的特点增强可解释性，将两种模型的优缺点进行了互补。Selwyn(1999)[8]建立基于Mamdani规则的神经模糊系统，该模型与神经网络相比增强了可解释性，但同时也损失了一定的精度。Huysmans等(2006)[9]利用自组织映射的可视化能力进行探索性数据分析，通过自组织映射与有监督分类器组合的方式进行分类。在神经网络方法在信用评估领域获得成功之后，很多研究开始尝试遗传算法。遗传算法是模拟进化论中自然选择的机理，用于搜索最优解的算法。Hoffman等(2002)[6]提出将遗传算法与模糊逻辑相结合，并与神经模糊分类器进行对比。Jih-Jeng等(2005)[14]提出了两阶段遗传规划的方法，首先利用遗传规划方法生成IF-THEN规则，再对规则处理后的数据用遗传规划的方法自动生成判别函数，该方法与决策树、粗糙集等方法相比更具灵活性且精度更高。支持向量机由于具有良好的实证效果，广泛应用于分类和回归问题。Li等(2006)[12]运用支持向量机模型对消费贷款申请进行识别。实证结果表明支持向量机比神经网络具有更好的泛化能力。Huang等(2007)[15]将遗传算法应用于支持向量机的特征选择和参数优化的过程中，用三种策略对支持向量机模型进行组合。Zhou等(2008)[16]采用直接搜索法确定支持向量机模型的参数，并与格点搜索、试验设计、遗传算法三种方法进行了比较。Ricardo等(2009)[18]用生存分析的方法研究个人违约概率，采用不同的条件分布刻画违约概率，用Cox回归模型、缺失情况下的广义线性模型、非参数核估计三种方法估计违约概率。Baesens等(2003)[19]对线性回归、包含二次项的线性回归、Logistic回归、线性规划、四种支持向量机模型，四种决策树算法，两种最近邻算法、神经网络、朴素贝叶斯、树增强型简单贝叶斯分类器(tree augmented naive Bayes classifier, TAN)这17种分类算法基于8个不同的数据集进行了比较。分类结果表明径向基函数、最小二乘支持向量机和神经网络方法在正确率和AUC方面都有良好表现。Mariola(2009)[20] 采用了Bagging和Adaboost的集成方法，Adaboost算法在训练过程中不断调整样本权重，而Bagging算法采用bootstrap的方式抽取训练样本，两种集成方法都可以提高信用评分模型的预测准确度以及泛化能力。

### 二、国内相关研究综述

我国的个人信用制度建设起步于“九五”末期，关于个人信用评估技术的研究仍然处于不断发展的阶段。很多学者根据中国信贷消费发展的实际情况提出了多种模型。陈为民(2009)[25]建立了混合支持向量机信用评分模型，采用分类回归数、多元自适应样条回归进行特征选择，利用格点搜索方法确定惩罚系数和核函数的参数，实证研究表明模型具有优异的性能。张佳维(2014)[26]利用模糊逻辑处理信息透明度髙、擅长逻辑推理的优点，将模糊神经网络理论应用到个人信用风险评估过程当中，与传统的人工神经网络相比模型具有更高的精确度。翟万里(2013)[27]增加了动量项和学习速率自调整的策略对标准BP神经网络进行了改进，保证收敛性的同时加快收敛速度，通过与标准BP模型以及Logistic回归模型的结果进行对比分析，验证了模型的有效性。肖智等[28]针对信用评估指标维数较高的问题,先采用主成分分析方法对数据进行降维，再对降维后的数据建立支持向量机模型，取得了较好的预测正确率。陈之远等(2010)[29]利用遗传算法计算Logit模型中的每个财务指标的分配系数的最优解，得到了可解释性强的模型。邹雅莹(2014)[30]构建了一种基于k-means聚类算法的马田系统模型,在筛选特征变量时采用以聚类结果上正确率作为信噪比的方法，优化了马田系统的基准空间，达到了很好的降维效果。王会军(2014)[32]引入稀疏贝叶斯学习方法建立个人信用评估模型，实证结果表明在分类精度方面较传统的信用评估模型有了一定提升。由于单一的分类模型往往精度不高，且容易出现过拟合的问题。因此，很多学者通过将单一模型进行集成的方式提高模型精度，这种方式成为组合(ensemble)或分类器组合(classifier combination)。王纯麟等(2006)[35]选择C4.5决策树作为基分类器，利用Adaboost算法将基分类器进行组合，克服了单一分类器对样本的敏感性问题，提升了分类的准确率。王昱(2015)[37]基于动态组合分类器的思想，首先通过有监督聚类方法对数据进行聚类，再利用不同违约总体两两配对的方式产生具有差异性的训练样本，最后采用样本邻域分类性能的加权投票作为组合预测接货，实现不同分类模型之间的优势互补。贾开瑜(2012)[38]考虑了变量之间存在的非线性关系，通过引入核函数，基于核主成分分析的方法对数据进行降维，通过线性规划的方法进行分类。史宁(2009) [39]对组合模型采用二次规划和遗传算法分别求解非负权重，两种确定权重方法得到的模型均优于单一模型。徐鑫柱(2013)[40]用支持向量机方法来调整BP神经网络中隐层到输出层之间的权值和阈值，改进的神经网络模型在第二误判率上组合模型要优于单一模型。杨海江[41]增加了惩罚因子，对AdaBoost算法进行了改进，有效降低了由于模型错判而导致的损失。向晖(2011)[42]从建立组合模型的角度出发，介绍了多种组合方式，构建了串行结构、异态并行结构、同态并行结构的个人信用评分组合模型，在提高模型精确性的同时也具有稳健性。姜叶飞(2014)[43]从变量选择方法上进行研究，将惩罚变量选择方法应用到信用卡信用风险评估问题的研究。

## 第三节 本文的研究思路与框架

第一章是引言部分，介绍了论文的研究背景和研究意义，也即论文的前提与必要性，综合阐述个人信用评估技术的国内外研究现状以及发展方向，给出了论文的主要创新点。

第二章对本文涉及到的基本方法和理论进行系统性的介绍,以便研究工作的开展。这些方法有Logistic回归模型、广义加性模型以及随机森林算法，并且对基于随机森林算法的聚类方法进行了介绍。

第三章通过实际数据对第二章中的三种方法进行实证分析，结果表明广义加性模型以及随机森林算法由于考虑了自变量的非线性关系，在预测精度上要优于Logistic回归模型，而随机森林算法不仅在预测精度上由于其他两种方法，而且提供了预测变量重要性的度量，适用于个人信用风险评估。

第四章建立个人信用评估组合模型，从考虑自变量间的交互作用角度采用了分类树和广义加性模型的组合模型，首先分类树提取自变量之间的交互作用，然后引入广义加性模型，从而提升了广义加性模型的预测精度；从考虑信用市场的个体差异性角度，先利用随机森林的聚类算法进行划分，并用聚类有效性指标确定聚类中心数，再利用动态组合预测结果的方式进行预测，提高了模型的准确度。

## 第四节 本文的创新点

本文的主要创新点在于：（1）由分类树方法获取自变量交互项，并引入广义加性模型。（2）提出了新的个人信用风险评估组合模型，利用随机森林的聚类算法进行划分，并用聚类有效性指标确定聚类中心数，再利用动态组合预测结果的方式进行预测，提供了另一种组合模型的思路。(3)实践的指导价值。论文的研究特色在于理论联系实际，研究成果具有一定的实用价值。目前我国商业银行的信贷决策还是过多的依赖于银行信贷人员的经验与专业技能，主观性较强，难以保证信贷决策过程的科学性、客观性和公正性，使得商业银行的信贷风险难以有效的预防与控制。本文所构造的基于聚类技术的组合模型具有良好的可操作性，风险识别准确率高，模型运算结果解释性强，可以应用于银行信贷决策的实务领域。

# 第二章 相关理论回顾

## 第一节 Logistic回归模型

### 一、Logistic回归原理

线性回归模型要求因变量是连续的正态分布变量，且自变量和因变量之间呈线性关系。当因变量是分类变量且并不满足线性关系时，由于无法满足线性回归模型的假设，线性回归模型失效，而Logistic回归模型非常巧妙地避开了因变量的分布问题，从因变量取某个值的概率角度进行分析。Logistic回归模型是广义线性模型的一种，由于具有可解释性强、对变量分布无要求等优点，因此是商业银行在信用评估中应用最多的模型。Logistic回归可以用于一个名义或顺序因变量的建模，在信用评分卡开发中起到核心作用。不同于神经网络、支持向量机等方法的“黑箱”运行模式，Logistic 回归具有很好的可解释性，在实际使用过程中可以通过对自变量进行证据权重转换，以标准评分卡的形式展现Logistic回归的结果。Logistic回归模型的自变量预测变量可以是连续变量，也可以是离散变量或者虚拟变量，并且不需要联合正态分布的假设，因此Logistic回归模型成为目前商业银行应用最广泛的模型。

假设Y是一种取值为0或1的随机变量，则Logistic回归方程为：



其中，为Y为1的概率，为与自变量无关的截距项，成为回归系数，它们分别反映自变量对于的影响程度。可以表示成的形式，它衡量的是Y为1的概率相对于Y为1的概率，也成为优势比(odds ratio)。从式（2-1）可以看出，自变量每变化一个单位，优势比就变化各单位，因此优势比是衡量因素作用大小的一个重要指标。当其他因素保持不变的情况下，变化引起的优势比变化越大，因变量对越敏感。对式（2-1）做简单变换可以得到如下的概率形式的Logistic回归模型：



Logistic回归的参数估计通常采用极大似然估计方法。Logistic回归的输出结果是分到某一类的概率，对于训练样本来说，每个样本是独立的，输出值为，样本的似然函数就是将所有样本输出值对应的概率值相乘，可以得到如下的对数似然函数：



在Logistic回归中，由于对数似然函数仍然是非线性的形式，无法直接进行求解，往往通过迭代的方式进行求解。Newton-Raphson多次逼近迭代算法是常用的一种方式，并可同时得到参数估计的渐进方差。

在样本量足够大的情况下，可以利用回归系数的渐进正态性质构造如下形式的置信区间：



## 第二节 广义加性模型

广义线性模型是一般线性模型的推广，而广义加性模型则是广义线性模型的推广。广义加性模型是将广义线性模型中链接函数的线性组合部分扩展为平滑函数，使得模型具有更好的非线性处理能力。



其中。g是已知的连接函数，同时可由加性模型的的不同选取分别得到一些特殊的模型:半参数模型,混合模型,非参数模型。为因变量的非参数估计的各种平滑函数，如光滑样条函数、局部回归函数、基样条函数、惩罚基样条函数和多项式函数等，这些非参数形式使模型更加灵活，也刻画了自变量的非线性效应，为截距项。从式(2-5)可以看出，广义加性模型也可以看做线性部分为自变量的光滑函数的广义线性模型。

当选取三次样条光滑函数作为平滑函数时，我们通过引入调和系数定义惩罚最小二乘函数为：为调和参数，与核光滑函数中的带宽以及loess中回归中的窗宽参数意义相同，当较小时意味着插值法估计，用光滑性可能很高的函数进行拟合，当值较大时，为了使式(2-6)达到最小，需要使趋向于0，因此得到的估计更倾向于最小二乘估计。从式(2-6)可以看出，式(2-6)的第一项是为了保证对数据的拟合效果，第二项避免了函数形式的过拟合。本文采用惩罚回归光滑函数作为平滑函数。

对于原始数据建立广义加性模型：



其中，是光滑函数，可以通过惩罚三次回归样条函数表示成如下形式：

其中和是三次样条基函数，因此最小二乘函数形式如下：

广义加性模型拟合的问题同样可以表示为如下的矩阵形式：

 (2-10)

其中X为如下形式：



矩阵，的形式分别如下：





限制参数矩阵C的形式如下：



其中代表第k个光滑函数的第j个节点的位置。限制矩阵C的前四行是对于三次样条基函数的限制，最后两行是对于光滑性的限制，从而使得对于光滑函数，有，这是对于模型可识别性的限制。因此对于给定的光滑系数和，可以通过最小化最小二乘函数估计参数。

上述模型实质是加性模型，只需要将推广到指数族分布就得到了广义加性模型，考虑以下模型：



其中X的是上述加性模型的设计矩阵，g是单调的连接函数，最小二乘函数同样加上惩罚系数，设是广义加性模型的对数似然函数，可以通过最小化负惩罚对数似然函数估计参数：



式(2-12)可以通过迭代加权的方法求解最小值。设为的方差，定义第k次迭代的结果如下：



其中是的对角矩阵，W是对角加权矩阵，对角线元素为

。因此惩罚最大似然估计是通过迭代求解最小化问题：

 (2-14)

换句话说，在给定光滑参数的情况下，可以用过迭代最小二乘的方法求解未知参数，而光滑参数通常通过广义交叉验证方法确定。广义交叉验证是交叉验证的加权形式。

本文的目的是为了通过发现对于个人客户违约有影响的解释变量和通过建立模型预测客户的违约状态，通过对广义加性模型的理论介绍，我们知道广义加性模型的优点是有良好的适应性。与线性模型不同之处在于广义加性模型通过引入平滑函数，反映了因变量和自变量之间复杂的非线性关系，因此应用广义加性模型更能捕捉实际信用市场中的变量关系。

## 第三节 随机森林算法

### 一、随机森林定义

随机森林(Random Forest, RF)是由Leo Breiman于2001年提出的一种统计学习理论，它用Bagging(bootstrap aggregating)方法形成不同的训练集,利用随机选取特征的方式生成决策树，且决策树在生成过程中不进行剪枝，最后以简单多数投票（分类问题）或简单平均（回归问题）作为其组合规则确定输出值。Bagging和随机选择特征分裂的结合使随机森林算法能较好地容忍异常值和噪声，具有良好的稳健性。通过大量的实证研究都证明了随机森林算法与传统统计方法相比在预测准确度上具有优势，而且对噪声数据具有很好的稳健性。通过理论推导可以证明随机森林存在与决策树强度和相关性有关的泛化误差，因此不容易出现过拟合情况。随机森林只需要对样本数据进行不断训练，捕捉数据间线性和非线性的客观关系，有良好的自适应性，适用于解决先验知识不清、无规则多约束条件和数据不完全的实际应用问题，具有较快的学习速度，可以利用计算过程中生成的相似度矩阵作为聚类分析中的相似度度量，是一种高效的预测和分类算法。

### 二、决策树的基本思想

在了解随机森林算法之前需要清楚决策树算法的原理，决策树是一种非参数的学习方法，不需要假定总体满足任何先验分布具有简单易操作的特点。决策树可以视为一系列IF-THEN规则的集合，其基本思想是通过二元递归的方式分割原始数据，决策树中包含3类节点：根节点、内部节点，终节点（叶节点）。每棵决策树只有一个根节点，根节点包含所有的数据样本。决策树中每个内部节点都对应一个IF-THEN规则，它将到达该节点的数据样本按照最优分割的方法进行分裂。每个终节点（叶节点）对应的就是该数据的最终对应的标签。根据确定最优分割的方式不同，决策树的算法有很多种，常见的有ID3算法，C4.5算法，分类回归树(CART)算法等。这些算法均采取自上而下进行递归的方式，采用贪心算法，在每个内部节点选择分类效果最好的属性进行分裂，直到决策树能够将所有训练样本全部打上标签值，或者所有的变量都已经出现在决策树中。

规则1

规则1

规则2

图2.1 决策树示意图

分类与回归树(CART)算法是决策树中经典的算法，它既可以用于分类也可以用于回归，CART算法生成的决策树结构简洁，与ID3和C4.5算法相比具有更好的抗噪声性能。给定一个样本量为N的训练样本，分类就是通过数学方法得到一个目标函数(target function)C，利用该函数将属性空间X映射到因变量Y所构成的空间。误分率就可以定义为：



决策树的构造过程就是寻找一个分类树使误分率达到最小，其中表示节点，为训练样本D分割后的各数据集，表示变量分割方法。具体来说，CART算法由上而下的递归算法可以表示为：

（1）创建根节点，在节点T处应用变量分割方法对训练样本D寻找最优分割变量X。

（2）如果节点T是属于同一类别或只剩下一个样本则返回节点T为叶节点，为其分配标签。

（3）假设节点T的子节点数为n，训练样本D分割为，对于分割变量X标记上T。

（4）构建子节点T的节点，对于，在节点处应用变量分割方法对数据集进行分割，直到节点为叶节点。

从上述步骤可以看出，分类回归树是采用递归的方式搜索遍历可能的决策树空间的一种算法，在每个节点处选择最优分割，直到所有属性都已被使用过或全部训练数据准确分类。最优分割的方式可以采用不纯度 (impurity) 增益的度量方法。不纯度增益指的是父节点与子节点之间不纯度的差。分别表示父节点、左子结点、右子节点，表示不纯度函数，因此对于任意分割，分类与回归树是解决使不纯度增益达到最大化的问题：



其中和分别是左右子节点的概率。

不纯度函数有多种度量方法，常用的有Gini指数法、信息增益法、信息增益比率法、统计量、分类误差率和G统计量等。其中Gini指数法是目前应用最广的方法，适用于离散型和连续型变量。当决策树产生的节点过多时，容易出现过拟合的情况，影响预测的准确性，因此需要对决策树进行剪枝。根据剪枝发生在建立决策树的前后可以将剪枝方法分为预剪枝和后剪枝方法。错误率降低剪枝(Reduced-Error Pruning, REP)、悲观剪枝(Pessimistic Error Pruning, PEP)、代价复杂度(Cost-Complexity Pruning, CCP)是常用的后剪枝方法。

### 三、随机森林的泛化误差

随机森林是多个决策树的集成模型，每颗决策树之间互不影响，这保证了随机森林的泛化性能。随机森林通过Bagging方式选取训练样本，增加了分类模型之间的差异性。与boosting算法采用赋权重组合的方式不同，随机森林采用通过平均的方式汇总基学习器的结果。给定分类器集合，原始数据集为(X,Y)，定义余量函数：



其中表示示性函数，表示取均值。于是，间隔函数刻画了在正确分类Y下X的得票超过其他分类的最大平均得票数的程度。该值越大，表明分类器的预测可靠性越高。随机森林的泛化误差（外推误差）可以写成：



其中下标X,Y表明了概率的定义空间。

当随机森林中数的个数做够大时，满足强大数定律，泛化误差几乎处处收敛于：



式（2-18）说明了随机森林算法并不会随着决策树的增加产生过拟合的问题，但是值得注意的是会产生有限的泛化误差。对于随机森林，泛化误差的上边界可以到处两个参数，这两个参数用来测度单个分类模型的精度以及分类模型之间的独立性。

定义随机森林的余量函数为：



则基分类器集合的强度可以定义为：



通过切比雪夫不等式和原始余量函数的定义可以证明泛化误差的上界为：



其中是不同的两棵决策树的相关系数关于分布的均值，也反映了不同决策树之间的相关性。从式（2-21）可以看出，泛化误差的上界随着相关性的减少或分类模型的强度增加而减小，因此实际特征选取的方式可以减小决策树之间的相关性，即通过减小的方式使泛化误差的上界减小保证随机森林算法的效果。

在生成训练集时，由于是采用自助法重抽样的方法，样本容量为N的原始训练集D中每个样本不被抽中的概率为。当样本容量充分大时，收敛于，也就是说样本充分大时，原始训练集中有36.8%的数据没有进入训练样本，这部分数据称为袋外(out-of-bag, OOB)数据。利用袋外数据可以对建立的模型进行误差估计。对于每颗生成的决策树我们都可以用袋外数据计算误差估计，从而得到随机森林的泛化误差。本节所讨论的泛化性能体现了所建的评估模型对于新样本点的预测准确性，模型的泛化性能越好，对新客户的判断就越接近实际情况。

### 四、随机森林的多种应用

随机森林算法不仅可以对数据进行分类和回归，还可以借助于随机森林方法进行特征选择、聚类分析和异常值检测。随机森林的一个重要特点是能计算单个指标的重要性，常用的度量方式有两种：基于袋外数据误分率的增加量和基于分裂时的GINI下降量。基于袋外数据对于重要性的度量通过以下的方式实现：当一个重要指标（即对预测的准确率起重要作用的指标）加入噪声干扰之后，随机森林的预测准确率将显著降低。具体做法如下：

首先对训练好的随机森林模型用袋外数据测试其分类性能，得到一个袋外准确率；在通过随机改变袋外数据集中的某个变量值（即人为进行噪声干扰）的方式，再用经过噪声干扰的袋外数据测试所建立的随机森林模型的预测能力，得到一个新的袋外准确率；最后计算原始的袋外数据的准确率与加入噪声后的袋外准确率之差，可以将差值作为所选指标的重要性的度量值。这一值越大说明所选的特征对于预测正确性起到了重要作用，因此重要性越高。随机森林的这一性能可以用来寻找信用评估过程中起到重要作用的一些变量，也是将高维数据进行降维的一种方式，对于个人信用体系的指标体系建设有一定的借鉴意义。

随机森林与其他常见的分类算法（如决策树、神经网络、支持向量机等）相比的一个重要优点在于随机森林可以计算各个样本点之间的相似程度，从而得到样本的相似度矩阵。通过相似性矩阵我们得到样本之间距离矩阵，从而进行聚类分析。在随机森林中，当用一棵决策树对所有数据样本进行分类时，这些样本点最终将出现在该树的某一叶节点上，两个样本在全部决策树的同一叶子节点上出现的频率越大，说明这两个样本之间的相似程度越高，这两个样本属于同一类的概率也就越大。相似度矩阵（简记为Prox矩阵）的具体计算过程如下：

步骤1：对于样本量为n的训练样本，首先初始化相似度矩阵Prox为n行n列的零矩阵，记作。对于训练样本建立随机森林模型。

步骤2：利用随机森林中生成的一棵决策树对所有样本点进行判别，对于任意两个样本点和，如果它们同时出现在所建决策树的叶节点上，则对相似度矩阵中对应的元素进行更新：，由于相似度矩阵为对称矩阵，不需要计算矩阵中所有元素的值。

步骤3：重复进行上一步骤，直到遍历随机森林中所有的决策树，得到最终的矩阵。

步骤4：将遍历所有决策树后的矩阵中每一个元素除以总决策树的数目，进行归一化处理，得到最终的相似度矩阵Prox。

# 第三章 个人信用单一模型

本章介绍的三种模型都很有代表性，其中Logistic回归是目前商业银行使用最广泛的模型，也是建立信用评分卡的核心，广义加性模型则是非参数方法,而随机森林算法是目前数据挖掘领域较为成功的算法。

## 第一节 数据来源及预处理

### 一、数据来源

由于国内的信用体系建设起步较晚，基础数据和材料收集较为困难，并涉及隐私性等问题，本文采用国外网站Kaggle提供的数据集进行实证研究。该数据集包括10个特征变量和最终违约状态的变量，其中特征变量包括：账户余额比率(Revolving Utilization Of Unsecured Lines)、年龄(Age)、逾期30-59天次数(Number Of Time 30-59 Days Past Due Not Worse)、负债比率(Debt Ratio)、月收入(Monthly Income)、信用卡和分期付款数(Number Of Open CreditLines And Loans)、逾期90天次数(Number Of Time 90 Days Late)、抵押贷款和房贷数量(Number Real Estate Loans Or Lines)、逾期60-89天未还款次数(Number Of Time 60-89 Days Past Due Not Worse)、家属人数(Number Of Dependents)，具体的变量含义见表3.1：

表3.1初选特征变量含义或特征项

|  |  |
| --- | --- |
| 变量名 | **变量含义** |
| 账户余额比率(Revolving Utilization Of Unsecured Lines) | （信用卡余额+免抵押贷款额度（不含房贷和分期付款））/信用额度 |
| 年龄(Age) | 借款者年龄 |
| 逾期30-59天次数（Number Of Time 30-59 Days Past Due Not Worse） | 过去两年借款人逾期30-59天未还款次数 |
| 负债比率(Debt Ratio) | （还款额+赡养费+生活费）/月收入 |
| 月收入(Monthly Income） | 月收入 |
| 信用卡和分期付款数(Number Of Open CreditLines And Loans) | 分期付款（比如车贷、抵押贷款）和免担保贷款（比如信用卡）数量 |
| 逾期90天次数(Number Of Time 90 Days Late) | 借款者90天以上逾期未还款次数 |
| 抵押贷款和房贷数量(Number Real Estate Loans Or Lines) | 抵押贷款和房贷数量 |
| 逾期60-89天未还款次数(Number Of Time 60-89 Days Past Due Not Worse) | 过去两年借款人逾期60-89天未还款次数 |
| 家属人数(Number Of Dependents) | 家属（包含配偶，子女）人数 |

在对数据进行建模之前，我们需要对数据集的整体分布进行初步的了解和认识。整个数据集包含有15万个客户信息，其中包含139974个好客户和10026个坏客户。对于数值型变量，分别计算139974个好客户和10026个坏客户样本的各个变量的均值(Mean)、上四分位数(Q1)、下四分位数(Q3)、标准差(Std. D)，并进行比较，进而了解各个自变量与因变量之间的关系，其统计结果如表3.2所示：

表3.2 连续性变量描述性统计结果

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 变量名称 | 正常客户 | | | | 违约客户 | | | |
| Mean | Q1 | Q3 | Std.D | Mean | Q1 | Q3 | Std.D |
| 账户余额比率 | 6.17 | 0.03 | 0.49 | 256.13 | 4.37 | 0.40 | 1.00 | 131.84 |
| 年龄 | 52.75 | 42 | 63 | 14.79 | 45.93 | 36 | 54 | 12.92 |
| 负债比率 | 357.2 | 0.2 | 0.9 | 2083.28 | 295.12 | 0.19 | 0.89 | 1238.36 |
| 月收入 | 6748 | 3461 | 8333 | 14813.5 | 5631 | 2963 | 6800 | 6171.72 |

从违约客户和正常客户对比来看，正常客户的账户余额比率、年龄和月收入指标的平均水平都高于违约客户，与实际情况相符，表现出较强的信用风险预警能力，而账户余额比率、负债比率和月收入三个指标的标准差都比较大，说明实际数据中客户之间良莠不齐的现象较为明显，各个样本间的差异较大。通过对比违约客户和正常客户负债比率的上下四分位数发现，从总体上看两个样本的负债比率分布并没有很大差异，从均值和四分位数的对比可以看出存在部分负债比率高的样本，而且正常客户中的比例更高。

对于非数值型(描述型)变量,分别计算样本总体、正常客户样本和违约客户样本的在每个特征项下的频数和百分比,见表3.3。

表3.3离散性变量描述性统计结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 变量名 | 特征项 | 正常客户 | | 违约客户 | | 全部样本 | |
| 频数 | 百分比 | 频数 | 百分比 | 频数 | 百分比 |
| 逾期30-59天次数 | 0 | 120977 | 86.43 | 5041 | 50.28 | 126018 | 84.01 |
| 1 | 13624 | 9.73 | 2409 | 24.03 | 16033 | 10.69 |
| 2 | 3379 | 2.41 | 1219 | 12.16 | 4598 | 3.07 |
| >2 | 1994 | 1.43 | 1357 | 13.53 | 3351 | 2.23 |
| 信用卡和分期付款数 | 0-4 | 30556 | 21.83 | 3103 | 30.95 | 33659 | 22.44 |
| 5-9 | 60117 | 42.95 | 3590 | 35.81 | 63707 | 42.37 |
| 10-14 | 33067 | 23.62 | 2096 | 20.91 | 35163 | 23.44 |
| >14 | 16234 | 11.6 | 1237 | 12.33 | 17471 | 11.65 |
| 逾期90天次数 | 0 | 135108 | 96.52 | 6554 | 65.37 | 141662 | 94.44 |
| 1 | 3478 | 2.49 | 1765 | 17.6 | 5243 | 3.50 |
| 2 | 779 | 0.56 | 776 | 7.74 | 1555 | 1.04 |
| >2 | 609 | 0.43 | 931 | 9.29 | 1540 | 1.02 |
| 抵押贷款和房贷数量 | 0 | 51546 | 36.80 | 4672 | 46.60 | 56218 | 37.48 |
| 1 | 49590 | 35.43 | 2748 | 27.41 | 52338 | 34.89 |
| 2 | 29757 | 21.26 | 1765 | 17.60 | 31522 | 21.01 |
| >2 | 9081 | 6.51 | 841 | 8.39 | 9922 | 6.62 |
| 逾期60-89天未还款次数 | 0 | 135140 | 96.55 | 7256 | 72.37 | 142396 | 94.93 |
| 1 | 3954 | 2.82 | 1777 | 17.72 | 5731 | 3.82 |
| 2 | 557 | 0.40 | 561 | 5.60 | 1118 | 0.75 |
| >2 | 323 | 0.23 | 432 | 4.31 | 755 | 0.50 |
| 家属人数 | 缺失 | 3745 | 2.68 | 179 | 1.79 | 3924 | 2.62 |
| 0 | 81807 | 58.44 | 5095 | 50.82 | 86902 | 57.93 |
| 1 | 24381 | 17.42 | 1935 | 19.30 | 26316 | 17.54 |
| 2 | 17938 | 12.82 | 1584 | 15.80 | 19522 | 13.01 |
| >2 | 12103 | 8.64 | 1233 | 12.29 | 13336 | 8.89 |

从总体来看，84.01%的客户从未出现过逾期30-59天不还款的情况，其中49.72%的违约客户曾出现过逾期30-59天的情况，说明良好的还款记录表明更小的信用风险，而46.6%的违约客户没有抵押贷款或房贷，高于总体平均水平，说明这些客户的信用风险更高。

由于原始数据集样本量过大，影响之后建立模型的计算效率，因此本文采用了分层随机抽样的方法。将总体数据分为违约子总体和正常子总体，在每层中独立地采用“不放回”的方法抽样进行简单随机抽样。最终，本文从建模数据中随机抽取了6000个数据作为训练样本，其中包含1200个违约样本和4800个未违约样本。为了避免由于抽样的随机性造成的误差，本文采用多次抽样建模的方式，利用平均值作为参考。

### 二、数据预处理

个人信用数据有多种来源，不仅包括所在商业银行的交易记录信息，还包括政府部门和征信机构的信用资料，因此在集成多个数据源过程中容易出现个人信用数据缺失的状况。个人信用数据中记录出现缺失的原因大致分为两种：由于个人原因造成的数据缺失(用户漏填的信息等)；由于银行操作方面的疏忽（登记、录入误差）或者银行只针对部分用户收集相关信息等原因所造成的数据缺失。个人信用数据的缺失减少了信用评估过程中所能获取的信息量，给评估带来了一定的困难，因此需要根据实际缺失情况采取合理的方法对缺失数据进行插补或剔除，根据插补技术的不同可以分为以下几种方式插补:

（1）均值插补法。均值插补是单一插补法的一种，当缺失变量是无偏分布的时候，运用均值插补在一定程度上提高了估计精度。当缺失变量为连续型变量，根据现有数据的均值作为缺失数据的替代值；当缺失变量为名义型变量，可以根据现有数据的众数作为缺失数据的替代值。但是当缺失率比较高时，运用均值插补方法会人为的增大均值所在组的频率，偏离了样本的真实分布，因此需要根据变量的实际缺失率大小进行判断。

（2）回归插补法。回归插补法是根据对现有数据建立回归模型的方法来预测缺失值。使用这种方法的前提假设是有缺失值的变量和其它未缺失变量之间存在线性关系，从而使得回归方程具有预测能力，当有缺失值的变量过多时，需要通过多次建立回归模型进行插补。

（3）期望最大化算法。期望最大化算法（EM算法）是一种迭代算法，首先在现有数据和现有参数的条件下求缺失数据的条件期望，再对条件期望求最大值，经过不断迭代达到收敛之后得到最终的参数估计值，从而以缺失数据的条件均值进行插补。

（4）多重插补。多重插补通过多个单一插补的组合的方式既分享单一插补的优点，也纠正了其缺点，反映了缺失数据的不确定性。

对于初选变量，只有月收入和家属人数两个变量存在缺失，为了充分利用客户信息，本文在计算缺失率和比较多种方法后采用均值插补法。对于月收入变量，由于月收入的分布并不是标准的无偏分布，因此本文采用中位数进行插补，对于家属人数变量，由于缺失率较低，本文采用众数进行插补。

### 三、模型评估指标

对于两分类问题，一些分类器得到的结果往往不是0，1这样的标签，如神Logistic回归，得到诸如0.5，0.8这样的分类结果。这时，我们人为取一个阈值，如果选取阈值为0.5，则预测概率大于或等于这个临界值将被认定为违约。通过将模型预测结果与实际结果进行对比，我们可以得到混合矩阵，如表3.4所示：

表3.4 混合矩阵

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 实际结果 | |
| 正常 | 违约 |
| 预期结果 | 正常 | TN(True Negative) | FN(False Negative) |
| 违约 | FP(False Positive) | TP(True Positive) |

将实际违约的客户标为正类，记为P，实际正常的客户标为负类，记为N。则灵敏度(sensitivity)、特殊性(specificity)和分类错误的正常记录比例(false positive rate, FPR)可以分别定义为：







灵敏度反映的是通过模型准确分类实际违约的客户的能力，也被称为命中率或分类正确的违约记录比例。受试者工作特征(Receiver Operating Characteristic, ROC)曲线反映不同判定违约的阈值下灵敏度与分类错误的正常记录比例（1-特殊性）之间的关系。

鉴于ROC曲线的定义，随机预测方式的ROC曲线就是斜率为1的线段，对角线以上的图形越高，说明在相同的分类错误的正常记录比例下模型的准确分类实际违约的客户的能力越强，因此可以通过比较ROC曲线的方式来评估不同模型的效果。ROC曲线之下的面积被称为AUC统计量或c-统计量。AUC统计量越大，说明模型识别风险的能力越强。ROC曲线及AUC系数主要用来反映模型对客户进行正确排序的能力。

## 第二节 基于Logistic回归的个人信用评估模型

### 一、数据的主成分分析过程

在建立模型之前，需要考虑变量之间是否存在很强的相关性，原因在于Logistic回归模型是广义线性模型，自变量之间的强线性相关关系会使估计值的方差很大。另一方面，如果某些变量之间存在具有很强的相关性，这意味着其中某一变量可以近似地由其他变量线性表示，因此包含了相同的信息内容，即使建模算法允许，最后得到的模型的预测也要比使用更广泛的自变量数据建立的模型预测力弱。因此，通过判断初选自变量之间的相关性可以避免出现多重共线性的情况。由于三个逾期次数变量都是代表客户信用记录的变量，从样本的相关系数矩阵也可以看出这些变量之间存在强的线性关系，如果将所有变量纳入模型会产生多重共线性的问题，得到的参数估计的方差也很大，影响结果的有效性，而主成分分析方法既保证了进入模型的变量之间的正交性，也尽可能保留变量的全部信息，因此对训练样本先进行主成分分析。

通过主成分分析,我们得到特征值向量，其中各个特征值按由大到小的顺序排列，特征值代表的是各个成分的重要程度。前六个主成分的累计贡献率已经达到87.3%。我们提取前六个的成分作为主成分，这里的贡献率就是某个主成分的方差占全部数据方差的比重，实际也就是某个特征值占全部特征值的比重。累积贡献率越大，说明了这些主成分变量所包含的原始变量信息就越多，我们常提取累计贡献率高于85%的主成分变量集合。

经过主成分分析,得到的各个主成分表达式如下:











其中变量—分别为账户余额比率、年龄、逾期30-59天次数、月固定消费、月收入、信用卡和分期付款数、逾期90天次数、抵押贷款和房贷数量、逾期60-89天未还款次数、家属人数这些变量中心化后的数据，这6个主成分变量代表的经济意义分别是:

中的各原始变量分别为逾期30-59天次数、信用卡和分期付款数、逾期90天次数、逾期60-89天未还款次数，不考虑信用卡和分期付款数，其余三项都是与借款人的还款记录有关，所以我们可以将解释为与申请人有关的信用记录。而表达式中各个原变量之前的系数都为负,这也恰好解释了逾期次数越少,其成为好客户的可能就越大,而申请人的得分就越高。

中的各原始变量分别为年龄、月固定消费、月收入、信用卡和分期付款数、抵押贷款和房贷数量、家属人数，而信用卡和分期付款数、抵押贷款和房贷数量的系数较高，因此我们可以将解释为申请人的还款压力。信用卡和贷款数目越多,其还款压力越大,从而坏账的可能性就越大,因此在对其得分时,用负系数相应地减少其分值，这也是比较符合实际经济解释的。

中各分原始变量别为申请人年龄、月固定消费、月收入、家属人数，而年龄和家属人数的系数较高，因此我们可以将解释为申请人的家庭情况。已婚的客户比未婚的客户违约率更低，因为已婚客户可能有第二份经济来源，因此他们面对抵御失业、疾病等风险时违约的可能性更小。

中各分原始变量别为账户余额比率、年龄、月固定消费、月收入、抵押贷款和房贷数量，综合一下，我们可以将解释为用户的月收支盈余情况，月固定消费的系数为负、月收入系数为正，说明月盈余越大，其违约的可能性越小。

中各分原始变量别为账户余额比率、月固定消费、月收入、家属人数，其中账户余额比率的系数为0.971，而账户余额比率为信用卡余额占信用额度的比例，因此我们可以将解释为与申请人有关的资产成分，账户余额比率越高，说明账户内的可用资金越多，其违约的可能性就越小。

中各分原始变量别为年龄、月固定消费、月收入、信用卡和分期付款数、抵押贷款和房贷数量、家属人数，我们可以将解释为用户的消费习惯情况，月固定消费和月收入前的系数都为正，而贷款数量的系数为负，说明对于月固定消费、月收入都比较高的客户年轻客户，其所承担的贷款数量越少，其违约的可能性越小。

### 二、Logistic回归建模

在进行主成分分析之后，我们需要对这6个主成分变量建立Logistic回归模型。逐步回归法旨在得到一个用最少的解释变量去解释因变量变异性的回归模型，常用F检验来判断引入一个变量或者从回归方程中剔除一个变量, 以保证在引入新变量前回归方程中只含有对影响显著的变量,但由于在实际使用中F检验的结果跟预先设定的显著性水平有关，逐步回归分析的结果常因F检验临界值的不同而异,得到的解可能有多个。因此本文基于最小信息准则采用向后逐步回归法构建Logistic模型。向后逐步回归是指先将所有自变量纳入回归方程，每一步选择使AIC信息统计量达到最小的模型，从而达到变量选择的目的，利用AIC准则得到以下结果，保留6个主成分使得AIC达到最小，从而也说明这6个主成分对预测违约状态的必要性。

表3.5 Logistic回归结果

|  |
| --- |
| Estimate Std. Error z value Pr(>|z|)  (Intercept) -1.47283 0.03517 -41.880 < 2e-16 \*\*\*  Comp.1 -0.19945 0.04064 -4.907 9.23e-07 \*\*\*  Comp.2 0.16465 0.03189 5.162 2.44e-07 \*\*\*  Comp.3 0.26343 0.03152 8.357 < 2e-16 \*\*\*  Comp.4 -0.37376 0.04147 -9.013 < 2e-16 \*\*\*  Comp.5 -0.11673 0.03471 -3.363 0.000770 \*\*\*  Comp.6 -0.22420 0.06406 -3.500 0.000465 \*\*\* |

可以看到主成分变量对于模型都是显著的，得到的Logistic方程为：

 (3-4)

其中为发生违约事件的概率，根据Z统计量可以判断进入模型的所有自变量在统计上都是显著的。图3.1是Logistic回归模型训练样本的ROC曲线：

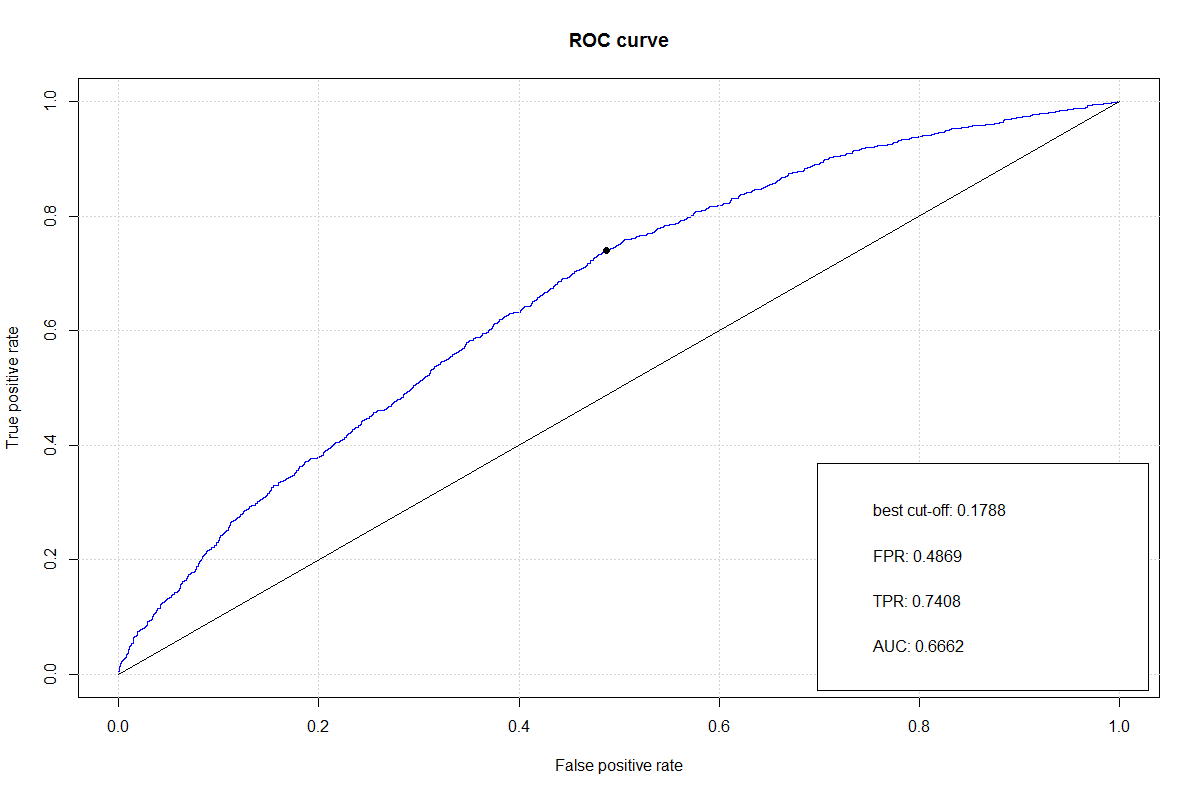


图3.1 Logistic回归的ROC曲线

## 第三节 基于广义加性模型的个人信用评估模型

### 一、广义加性模型的实证研究

广义加性模型(General Additional Model, GAM)是一种非参数模型，通过引入平滑函数代替广义线性模型中的线性项，对广义线性模型进行了扩展，广义加性模型在处理非但单调、非线性关系上的能力要强于广义线性模型，也更具灵活性。另外，广义线性模型很难处理连续型解释变量，在实际分析中通常是将连续型变量通过最优分群算法或等距分组方法进行分组，这种方法适用于信用评分卡的开发，但并不是所有的连续型变量都适合分组，而且临界值常常需要人工进行判断，不利于模型的推广。

与广义线性模型类似，广义加性模型也由随机部分，系统部分，联结函数三部分组成，不同的是广义加性模型的系统部分引入了非线性的部分，可以表达成如下解释变量的可加形式：



其中。g是已知的连接函数，为非参数估计的各种平滑函数，如光滑样条函数、局部回归函数、基样条函数、惩罚基样条函数和多项式函数等，是解释变量。

当将所有变量纳入模型时，账户余额比率系数和负债比率变量的系数并不显著，因此使用其他8个变量建立模型。其中对于调和参数的估计，本文采用广义交叉验证方法。最终得到的参数估计结果为表3.6：

表3.6 广义加性模型结果

|  |
| --- |
| edf Ref.df Chi.sq p-value  s(var2) 3.721 3.950 79.676 2.59e-16 \*\*\*  s(var3) 2.926 3.021 240.967 < 2e-16 \*\*\*  s(var5) 2.690 2.996 24.459 2.01e-05 \*\*\*  s(var6) 3.717 3.943 13.050 0.01067 \*  s(var7) 3.949 3.976 237.053 < 2e-16 \*\*\*  s(var8) 3.836 3.977 17.022 0.00190 \*\*  s(var9) 1.951 2.007 46.909 6.82e-11 \*\*\*  s(var10) 1.000 1.000 7.245 0.00712 \*\* |

在测试样本上的预测结果如图3.2：

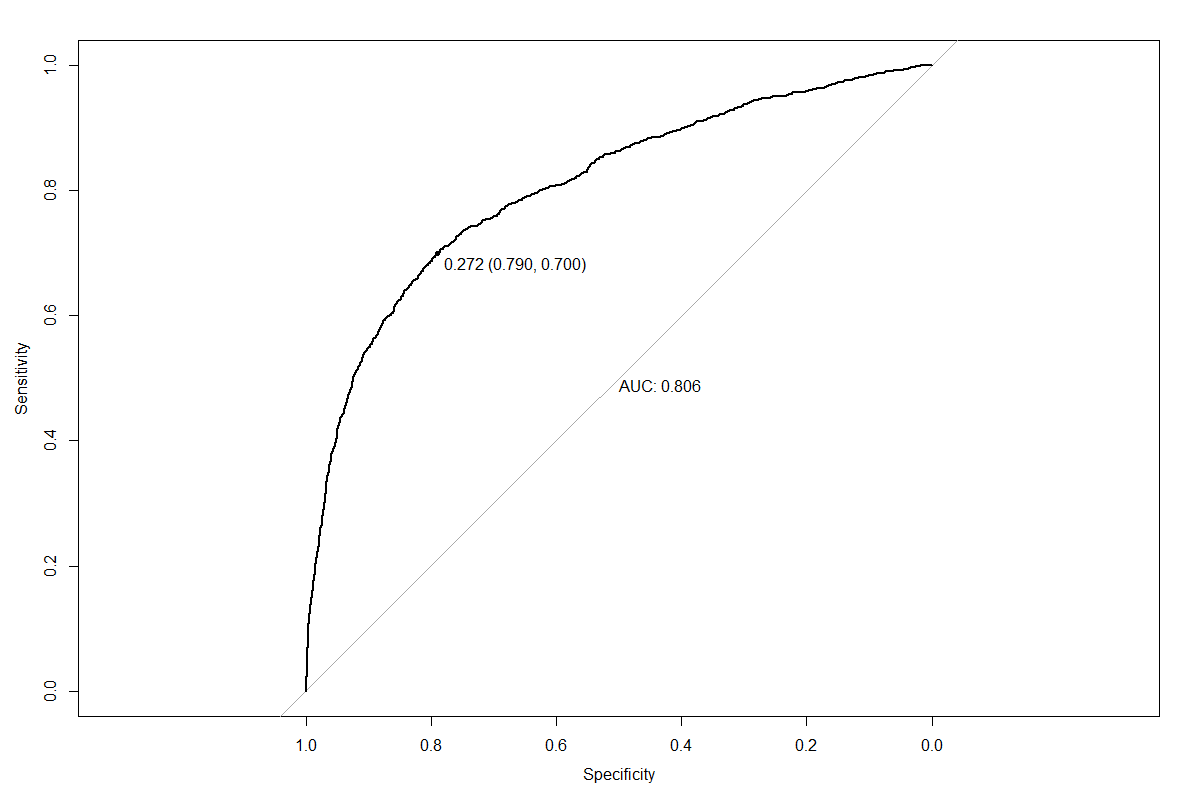


图3.2 广义加性模型的ROC曲线

通过与Logistic回归结果对比，GAM模型的AUC值得到了很大的提升，说明变量的非线性部分对预测违约起到了很好的作用。

## 第四节 基于随机森林的个人信用评估模型

### 一、随机森林模型的变量重要性

随机森林的一个重要特点是能计算单个指标的重要性，常用的度量方式有两种：基于袋外数据误分率的增加量和基于分裂时的GINI下降量。基于袋外数据对于重要性的度量通过以下的方式实现：当一个重要指标（即对预测的准确率起重要作用的指标）加入噪声干扰之后，随机森林的预测准确率将显著降低。

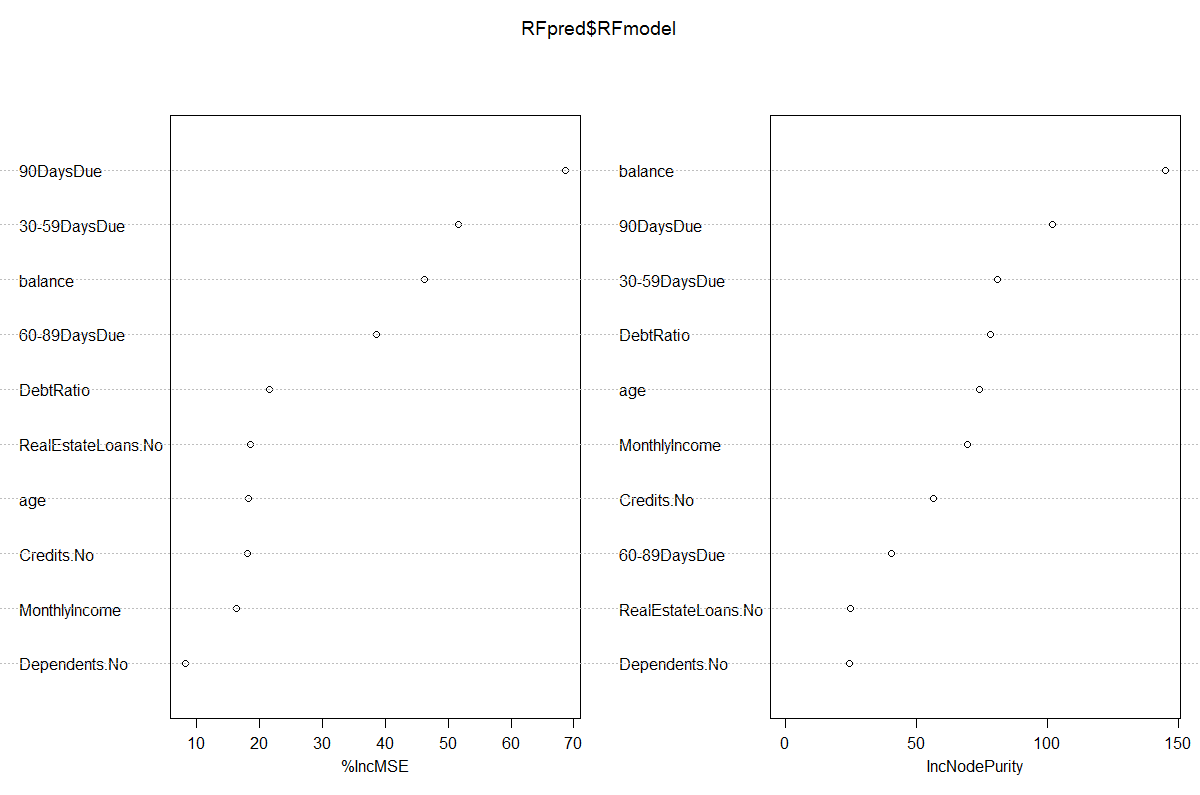


图3.3 随机森林的变量重要性测量

%IncMSE衡量的是把一个变量的取值变为随机数，随机森林预测准确性的降低程度，该值越大表示该变量的重要性越大。IncNodePurity是通过基尼(Gini)指数计算每个变量对分类树每个节点上观测值的异质性的影响，从而比较变量的重要性。该值越大表示该变量的重要性越大。从图3.3可以看出，虽然个别变量之间的顺序有所差别，但是两种计算方式都将账户余额、90天逾期次数和30-59天逾期次数放到前三位，说明这些变量对于最终预测提供了非常重要的信息。综合以上两种重要性排序，可以看出，在处理贷款申请时,要特别注意借款人的负债情况和月收入。

### 二、随机森林模型的参数选择

随机森林是一个包含多个决策树的分类器，在建立每一棵决策树的过程中，随机森林对要随机选择训练样本和分类属性，从而增加决策树之间的差异性。有了这两个随机性的保证，随机森林就不会产生过拟合的现象。因此需要设置的参数有（1）ntree，随机森林中决策树的数量；（2）mtry，建立决策树时每个节点处自变量的个数。在选择树的数量时需要使随机森林效果趋于稳定，一般选择是在计算负荷可以接受的情况下选择最大的分类树的数量。

通过实际比较，我们将树的数量设定为400，此时增减树的数量并不会引起分类效果的明显变化；在分类问题中mtry往往选取自变量总数的平方根；如果是回归分析，mtry选取是自变量总数的1/3。通过实际比较，我们将候选特征数设定为2，即在每棵树的分裂节点处随机选择两个特征进行分裂，这样充分体现了每棵树之间的差异性，从而在一定程度上提高了整个随机森林的分类准确率。图3.4为随机森林模型的ROC曲线：

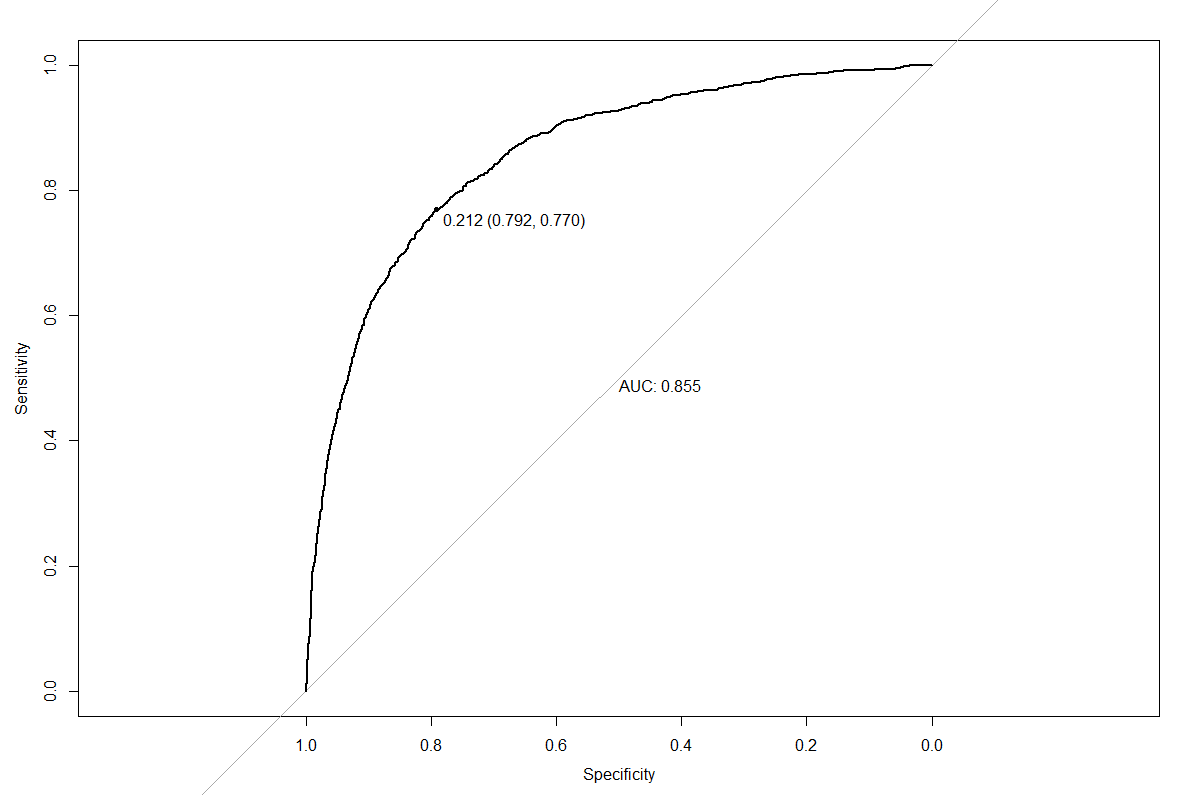


图3.4 随机森林的ROC曲线

从以上实证分析可以看出，广义加性模型和随机森林由于考虑了个人信用市场间复杂的非线性关系，因此分类精度要明显高于Logistic回归模型，由于在广义加性模型中并没有考虑变量之间的交互作用，与实际情况不相符，因此存在进一步提升的空间，而随机森林算法功能强大而且结构清晰，减少了类似于支持向量机中的调试过程，而且变量重要性的度量对于信用体系指标建设提供了借鉴意义，拥有广泛的应用前景。

# 第四章 个人信用评分组合模型

## 第一节 基于决策树和广义加性模型的组合模型

### 一、建模依据

在应用广义加性模型建立个人信用评估模型时，我们只考虑了单个变量的非线性形式，而实际中往往更加复杂，不同的变量之间存在着交互作用。在自变量个数较少时，可以通过将交互项纳入模型，检验交互项的系数就能判断是否存在交互作用。然而当自变量的个数增加时，交互项个数迅速增加，有时甚至大于样本个数，模型检验的方式变得难以实现。而分类树模型可以筛选可能出现的交互作用，从而大大减少试验次数。

基于以上考虑，本文借助于分类树模型筛选交互作用，再将交互作用纳入广义加性模型，并检验交互项的系数显著性，从而改善广义加性模型。

### 二、实证分析

首先对于训练样本建立分类树，从分类树中提取可能存在的交互作用，为了避免出现自变量过多的情况，本文只提取二阶和三阶的交互项，放入广义加性模型，得到的最终结果为表4.1：

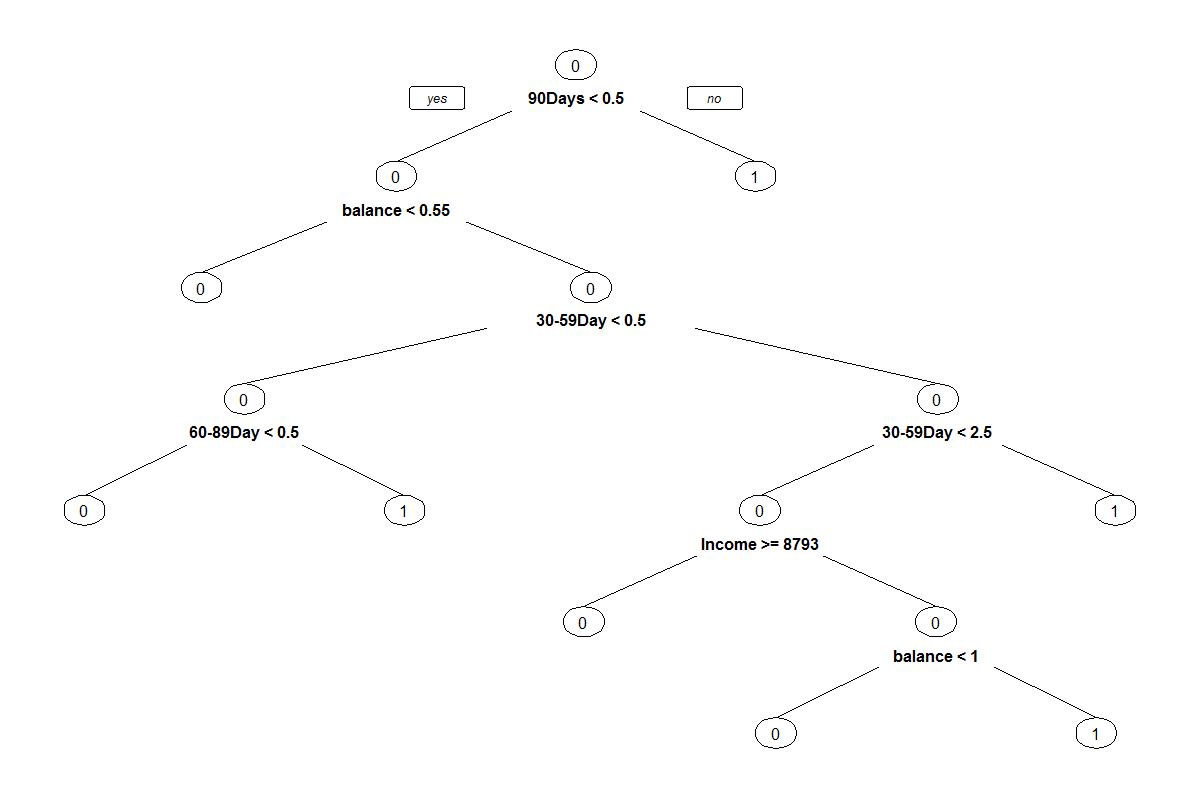


图4.1 基于分类树的实证结果

表4.1 引入交互项后广义加性模型结果

|  |
| --- |
| s(var1) 1.000 1.000 0.012 0.91339  s(var2) 3.300 3.719 35.315 4.18e-07 \*\*\*  s(var6) 1.051 1.101 11.041 0.00120 \*\*  s(var7) 1.024 1.039 0.604 0.44909  s(var8) 1.004 1.008 5.404 0.02037 \*  s(var9) 1.015 1.028 1.407 0.24115  s(var10) 1.018 1.036 2.347 0.13035  s(var7,var1) 1.514 2.000 31.430 2.28e-09 \*\*\*  s(var3,var5) 3.085 3.442 16.057 0.00194 \*\*  s(var1,var7,var3) 11.000 11.000 45.582 3.83e-06 \*\*\* |

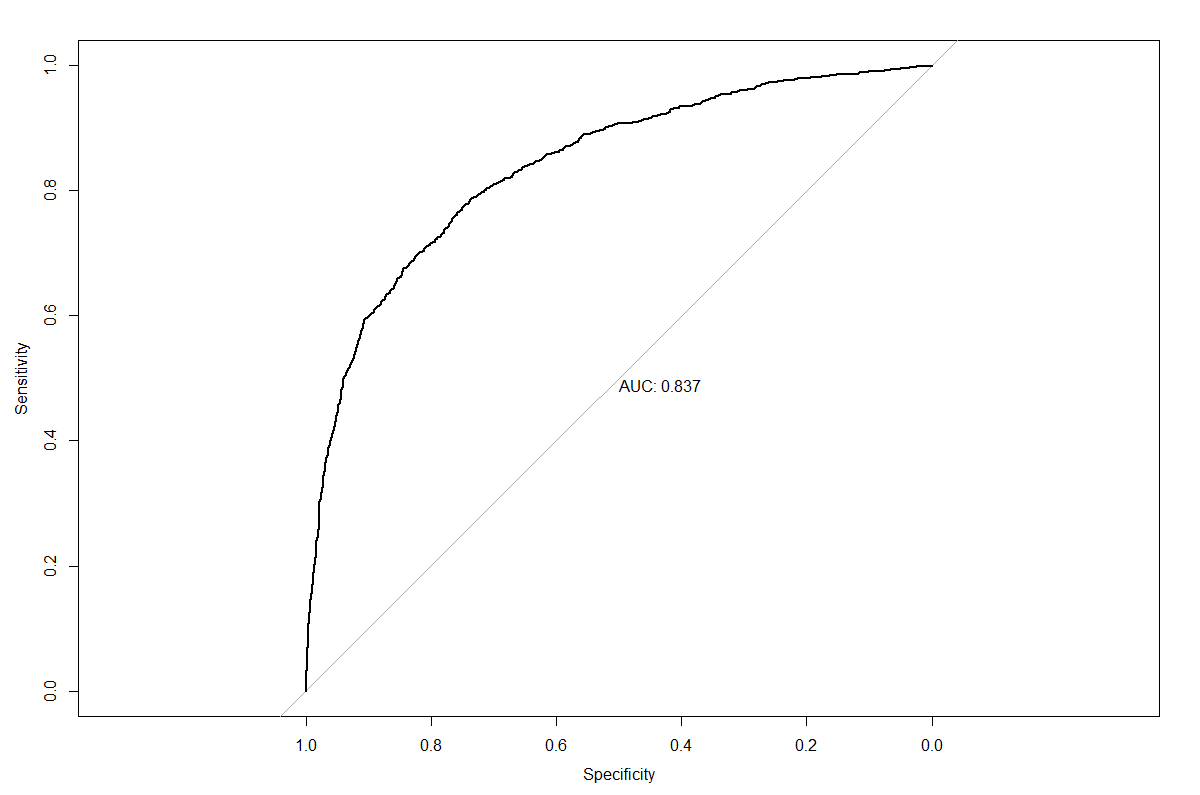


图4.2 组合模型的ROC曲线

为了降低因样本选择而产生的随机性，这里通过多次试验取平均值的方式计算AUC变化量。从组合模型和广义加性模型的对比来看，通过引入交互作用使得组合模型的AUC值平均上升了0.04。说明自变量之间的交互作用对于预测违约状态提供了信息，这一方法达到了提供分类准确度的目的。

## 第二节 基于随机森林聚类的组合模型

### 一、建模依据

对于信用评估问题，由于当前的信用市场中个体的个性化差异性较大，使用同一个分类模型进行预测，显然不能满足实际需求。而且由于分类器是基于训练样本训练得到的，因此训练样本的随机性是误差的主要来源之一。所以通过组合在不同训练样本上构建的分类模型,将有助于减少这种误差。对每个基分类器的预测值进行加权来得到分类结果，避免了样本间信息的损失,保证了每个分类器的分类性能可以在输出结果上反映出来。由于不同样本的特征不同,因此在样本空间中就有不同的区域，而同一分类器在样本空间不同区域的分类性能会有所变化。因此本文通过聚类技术将样本空间进行划分，从而在不同的区域建立模型，最终将各区域模型的结果从输出结果上反映出来。

### 二、基于随机森林的聚类算法

随机森林算法不仅可以对数据进行分类和回归，还可以借助于随机森林方法进行特征选择、聚类分析和异常值检测。由于随机森林在处理混合数据和对异常值的稳健性方面具有优势，因此借助于随机森林算法进行聚类分析。

随机森林与其他常见的分类算法（如神经网络、决策树等）相比的一个重要优点在于它可以计算样本点之间的相似程度，得到样本之间的相似度矩阵。当用一棵树对所有数据进行分类时，这些数据最终都将落于该树的某个叶节点上，我们可以用两个样本在每一棵树上同一叶子节点上出现的频率大小来度量这两个样本之间的相似程度,或者是这两个样本属于同一类的概率大小。相似度矩阵（简记为Prox矩阵）的计算过程如下：1.对于样本量为n的数据集，首先初始化相似度矩阵Prox为n行n列零矩阵，记为。2.利用生成的决策树对所有样本点进行判别，对于任意两个样本和，如果它们出现在所建决策树的同一叶节点上，则对Prox矩阵中相应元素进行更新：。3.对以上步骤进行重复，遍历随机森林中所有决策树，得到相应的矩阵。4.将相似度矩阵中的每一个元素除以总的决策树数目，从而进行归一化处理，得到最终的相似度矩阵。

聚类算法的一个核心问题就是选择合适的样本相似性度量。随机森林的相似度矩阵是基于决策树模型而产生的，因此它能很好得对属性和有序数据进行分析。相似度矩阵对于变量的单调变换具有不变性，而且对于异常值具有稳健性，因此本文基于随机森林的相似度矩阵对数据进行聚类分析。

在聚类算法的选取上，本文采用PAM（Partitioning Around Medoids，围绕中心点的划分）聚类算法，该算法利用反复替换中心点的方法提高聚类质量，克服了K-均值聚类算法容易陷入局部最优的缺点，也降低了对于离群点的敏感性。在确定使用PAM算法进行聚类后，另一个关键问题是如何确定最优的聚类数。由于Wu和Yang[45]提出的聚类有效性指标能够很好刻画类之间的差异性和类之间的相似度，并且对于噪声数据具有较强的稳健性，因此本文选取其作为聚类有效性指标:



式（4-1）中K为聚类中设定的聚类个数，为聚类有效性指标，为反映数据类内数据点相似度的指标，为反映不同数据类之间差异性的指标，和的定义如下：





在式（4-2）和式（4-3）中，为第k个数据类中的第i个数据点，为输入数据集中所有样本点的中心，和分别为第k个数据类的中心和第k个数据类的样本点个数。对于式（4-2），如果同一聚类内的数据点相似度越高，则样本离该数据类的中心点的距离越近，越小；对于式（4-3），如果不同聚类之间的样本差异度越大，则K个数据类中心点之间的平均距离就越大，越小。因此，最优聚类的结果使得聚类有效性指标达到最小。

### 三、基于随机森林聚类的组合模型

由于不同样本的存在差异性，因此表现在样本空间上就有不同的区域，而同一分类器在不同特征的训练样本上的分类性能会有所不同，而且使用单一模型可能无法刻画复杂的变量关系，因此通过聚类技术将样本空间进行划分，从而在不同的区域建立模型。使用基于随机森林相似度矩阵的聚类方法将样本分为k个数据类，根据聚类有效性指标选择合适的k，从而使同一分区中样本之间的相似度高，不同分区之间样本之间的差异度大。由于训练样本根据聚类分析结果划分为不同的聚类分区，而后续研究将在不同聚类分区上建立模型。因此，通过建立基于聚类技术的分区集成模型将实现样本的多分类器动态集成机制。模型训练过程的基本步骤如下：

步骤1：设训练样本为，对训练样本进行数据预处理，对于缺失的连续变量采用中位数进行插补，离散变量采用众数进行插补。

步骤2：基于随机森林相似度矩阵对训练样本进行聚类。

1. 在训练样本上构建随机森林模型。
2. 将训练样本中的所有样本投入到随机森林中，计算中每一个样本之间的相似度。
3. 利用随机森林算法计算得到的相似度矩阵Prox作为相似性度量，采用PAM聚类算法将训练样本分为K类，每一类的聚类中心为。其中K的大小根据聚类有效性指标确定。

步骤3：对于每个聚类，建立各分区样本的个人信用评估子分类模型。基本分类器训练学习算法可以选择Logistic回归模型、广义加性模型、决策树、支持向量机等分类算法。

数据分类过程的基本步骤如下：

步骤1：数据预处理，对缺失值进行插补。

步骤2：假设待测样本为x，计算x到所有各个聚类中心之间的欧氏距离，并记为。

步骤3：对于待测样本x，利用各聚类分区上的子分类模型预测违约概率，利用对该样本到各个聚类中心的距离进行加权，得到最终的预测值。

根据式（4-1）定义的聚类有效性指标，计算不同聚类树中心数下的具有有效性指标，为了保证各分区模型的有效性，限制最大聚类数为10，最终得到的结果如图4.3：

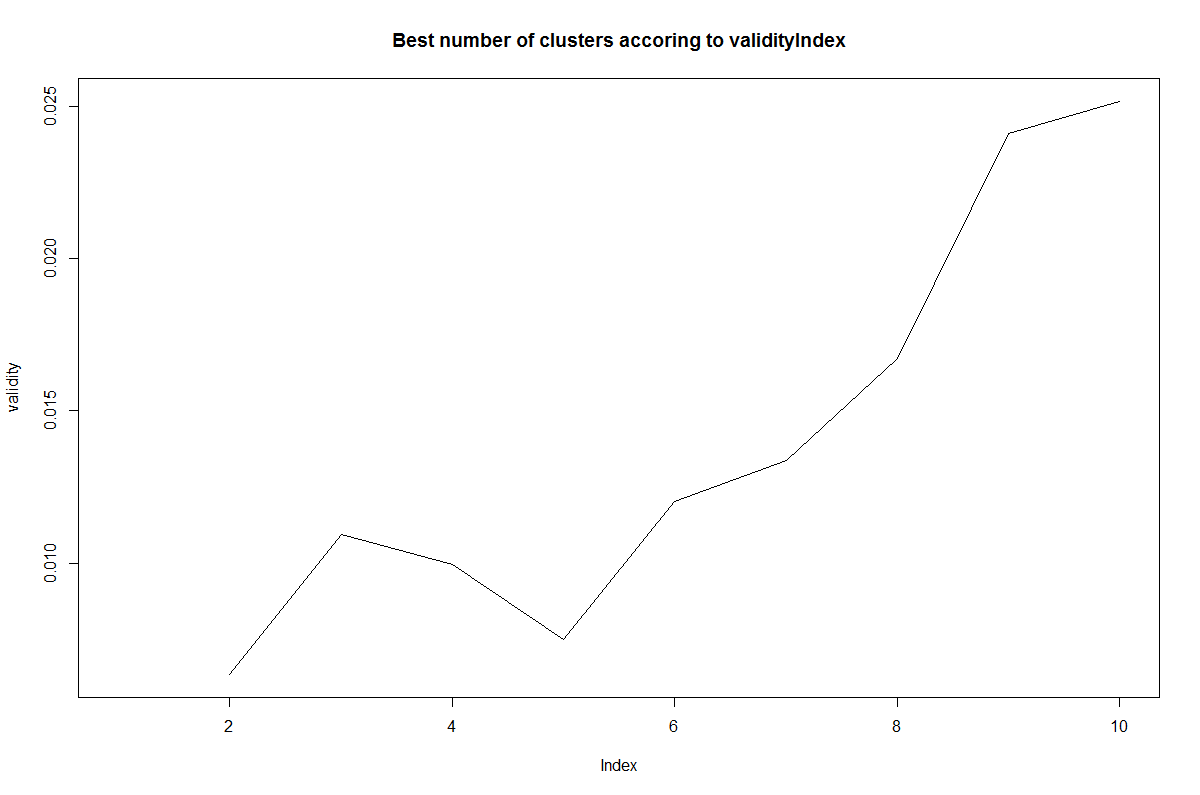


图4.3 最优聚类数

从得到的结果来看，当聚类数为2时，由式（4-1）定义的有效性指标达到最小，当聚类中心数大于6时，各个聚类分区之间的差异性并不明显。因此选择聚类中心数为2，对训练样本进行PAM聚类，在不同聚类分区上建立Logistic回归模型，最终得到基于预测样本的ROC曲线如图4.4：

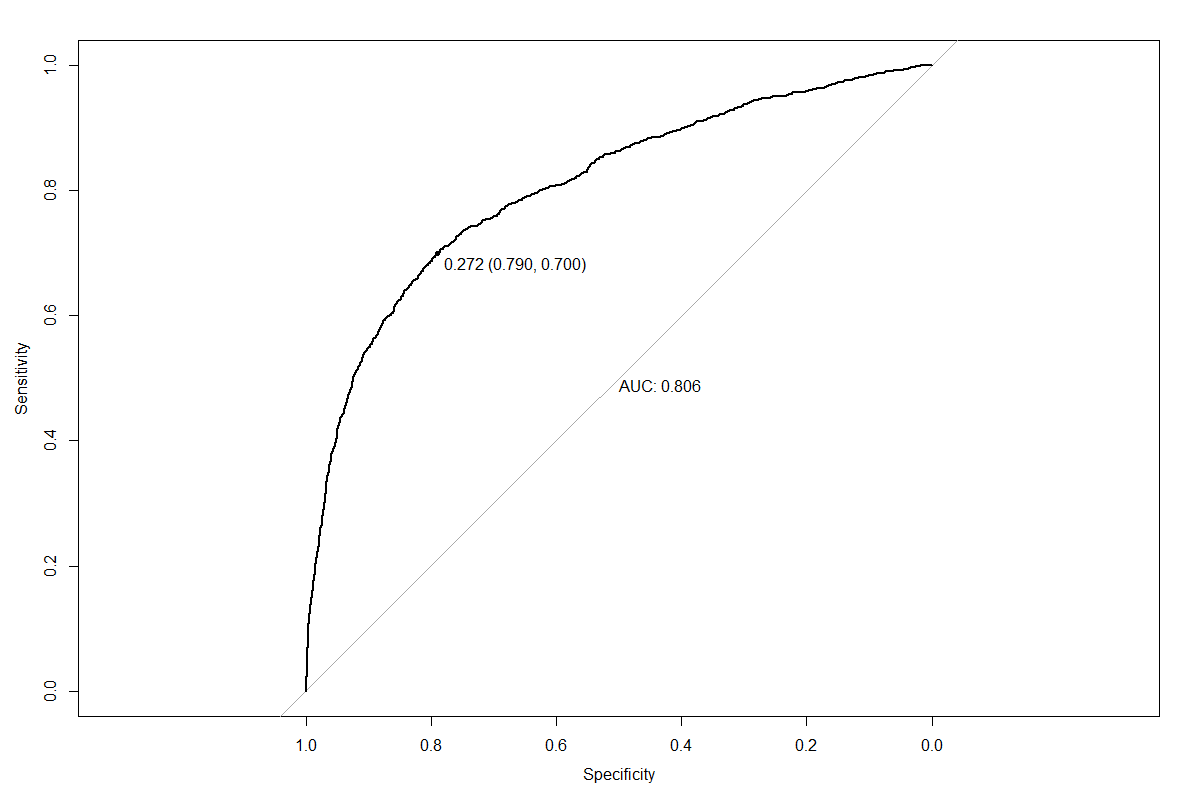


图4.4 组合模型的ROC曲线

通过对比Logistic回归模型的预测效果，经过随机森林聚类技术处理的模型提升了预测的准确性， 说明了原始样本间存在着不同的样本区域，而使用单一的Logistic回归模型不足以刻画全部样本的特征，缺乏预测的准确性。由于在不同样本区域上建立的仍然是Logistic回归模型，因此保留了Logistic回归模型可解释性强，稳定性高的一些优点。

## 第三节 模型对比分析

我们将Logistic回归、广义加性模型、随机森林模型、引入交互作用的广义加性模型和两种基于聚类算法的组合模型的预测效果汇总到表5.1中，我们定义两类错误：第一类错误是将“信用好”的个人误判为“信用差”；第二类错误是将“信用差”的个人误判为“信用好”的贷款个人，可以得到以下分析结论：

表5.1 模型的分类效果比较

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 第一类误判率 | 第二类误判率 | 总分类误判率 | AUC |
| Logistic | 27.69% | 39.83% | 30.11% | 0.6984 |
| GAM | 16.46% | 33.67% | 19.90% | 0.8195 |
| RandomForest | 21.60% | 22.08% | 21.70% | 0.8544 |
| Tree+GAM | 22.94% | 22.42% | 22.83% | 0.8452 |
| Cluster+LR | 21.96% | 31.17% | 23.80% | 0.7992 |
| Cluster+GAM | 27.63% | 17.67% | 25.63% | 0.8509 |
|  | | | | |

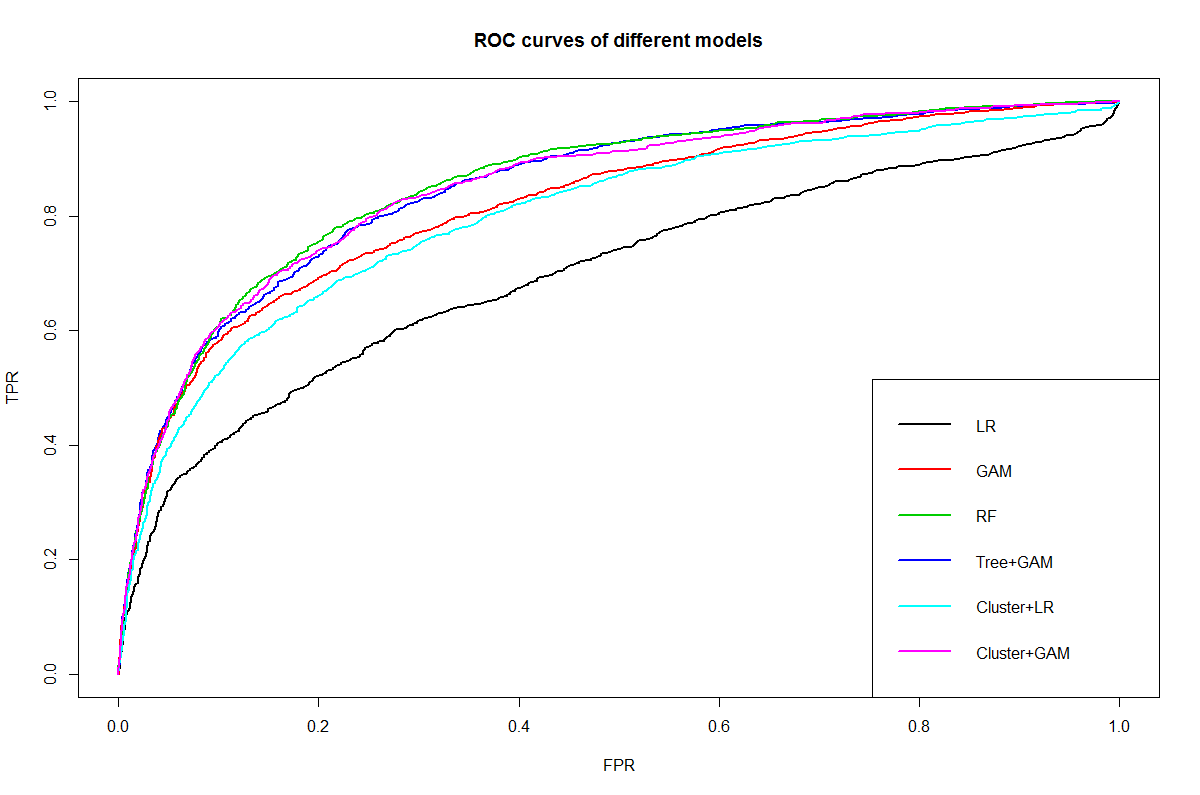


图5.1 多种模型的测试样本ROC曲线

1. 单一模型的比较

从图5.1可以看出，在测试样本上，广义加性模型和随机森林模型不管在第一类分类错误、第二类分类错误和AUC上都要优于Logistic回归模型，说明了变量之间非线性假设的存在性。在第一类分类错误率的对比上，广义加性模型的误判率最低，说明在广义加性模型对信用好的客户的区分能力要好于其他模型。在第二类分类错误率的对比上，随机森林模型的误判率最低，说明随机森林对于坏信用的客户区分能力最优。从横向对比来看，三个模型的第二类误判率都要高于第一类误判率，这说明坏信用的样本数据有着更大的变异性,即坏信用客户的个体特征差异更大。

1. 组合模型的比较

通过对比引入交互作用的广义加性模型和广义加性模型来看，通过引入交互作用大大降低了第二类误判率，与广义加性模型相比降低了11.25%，表明其能较好的处理变异程度更大的坏信用数据，从而对坏信用有较强的分类能力。从另一个角度来看，在银行个人贷款中，第二类错误的损失远远大于第一类错误带来的损失，因此引入交互作用能够更好的适应个人信用评估的实际要求。

通过对比基于聚类算法的组合分类模型和单一模型来看，对于Logistic回归模型，基于聚类的组合Logsitic回归模型与Logsitic回归模型相比在第一类误判率和第二类误判率上分别降低了5.73%，8.67%，而基于聚类的组合广义加性模型与广义加性模型相比大大降低了第二类误判率，说明这种组合方法具有很好的可行性及良好的判断结果，能有效反映商业银行企业客户的信用风险状况，具有较高的实用价值。在实际应用中，利用商业银行内部大量关于信用风险状况评估的历史数据对模型训练好后，就可以使用模型来评估具体的信用风险状况，从而为商业银行做出信贷决策提供预测基础，达到降低商业银行信用风险的目的。

# 第五章 总结与展望

本文通过在对Logistic回归模型、广义加性模型以及随机森林算法三种不同模型对比的基础上，建立个人信用评估组合模型，从考虑自变量间的交互作用的角度采用了分类树和广义加性模型的组合模型，首先利用分类树提取自变量之间的交互作用，然后将交互因素引入广义加性模型，从而提升了广义加性模型的预测精度，大大降低了第二类分类误判率；从考虑信用市场的个体差异性角度，先利用随机森林的聚类算法进行划分，并利用聚类有效性指标确定聚类中心数，再利用动态组合预测结果的方式进行预测。利用实际数据进行实证分析，结果显示两种组合模型都提高了预测精度，而通过聚类技术的组合方式，将在不同样本区域上建立的Logistic回归模型进行距离加权，保留了Logistic回归模型可解释性强，稳定性高的一些优点，在第一类误判率、第二类误判率和AUC上都优于Logistic回归模型，具有具有较高的实用价值。

# 参考文献

Ohlson, James A. "Financial ratios and the probabilistic prediction of bankruptcy." *Journal of accounting research* (1980): 109-131.

Wiginton, John C. "A note on the comparison of logit and discriminant models of consumer credit behavior." *Journal of Financial and Quantitative Analysis* 15.03 (1980): 757-770.

Bekhet, Hussain Ali, and Shorouq Fathi Kamel Eletter. "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach." *Review of Development Finance* 4.1 (2014): 20-28.

West, David. "Neural network credit scoring models." *Computers & Operations Research* 27.11 (2000): 1131-1152.

Lee, Tian-Shyug, et al. "Credit scoring using the hybrid neural discriminant technique." *Expert Systems with applications* 23.3 (2002): 245-254.

Piramuthu, Selwyn. "Financial credit-risk evaluation with neural and neurofuzzy systems." *European Journal of Operational Research* 112.2 (1999): 310-321.

Huysmans, Johan, et al. "Failure prediction with self organizing maps."*Expert Systems with Applications* 30.3 (2006): 479-487.

Huang C L, Chen M C, Wang C J. Credit scoring with a data mining approach based on support vector machines[J]. Expert systems with applications, 2007, 33(4): 847-856.

Huysmans J, Baesens B, Vanthienen J, et al. Failure prediction with self organizing maps[J]. Expert Systems with Applications, 2006, 30(3): 479-487.

Hoffmann, F., et al. "Comparing a genetic fuzzy and a neurofuzzy classifier for credit scoring." *international journal of intelligent systems* 17.11 (2002): 1067-1083.

Lee, Tian-Shyug, et al. "Credit scoring using the hybrid neural discriminant technique." *Expert Systems with applications* 23.3 (2002): 245-254.

Chrzanowska, Mariola, Esteban Alfaro, and Dorota Witkowska. "The individual borrowers recognition: Single and ensemble trees." *Expert Systems with Applications* 36.3 (2009): 6409-6414.

Huang, Jih-Jeng, Gwo-Hshiung Tzeng, and Chorng-Shyong Ong. "Two-stage genetic programming (2SGP) for the credit scoring model." *Applied Mathematics and Computation* 174.2 (2006): 1039-1053.

Li, Sheng-Tun, Weissor Shiue, and Meng-Huah Huang. "The evaluation of consumer loans using support vector machines." *Expert Systems with Applications* 30.4 (2006): 772-782.

Huang, Cheng-Lung, Mu-Chen Chen, and Chieh-Jen Wang. "Credit scoring with a data mining approach based on support vector machines." *Expert systems with applications* 33.4 (2007): 847-856.

Zhou L., Lai K., and Yu L., Soft Computing a Fusion of Foundations Methodologies and Applications, Springer Verlag, 2008.

Abad, Ricardo Cao, Juan Manuel Vilar Fernández, and Andrés Devia Rivera. "Modelling consumer credit risk via survival analysis." SORT: statistics and operations research transactions 33.1 (2009): 3-30.

Baesens, Bart, et al. "Benchmarking state-of-the-art classification algorithms for credit scoring." Journal of the Operational Research Society 54.6 (2003): 627-635.

陈为民. *基于支持向量机的信用卡信用风险管理模型与技术研究*. Diss. 湖南大学, 2009.

张佳维. *基于模糊神经网络的个人信用风险评估*. MS thesis. 内蒙古大学, 2014.

翟万里. *基于人工神经网络的商业银行信用风险评估模型研究.* MS thesis. 长沙理工大学,2013.

肖智, 李文娟. "基于主成分分析和支持向量机的个人信用评估." 技术经济 29.3 (2010): 69-72.

陈之远, 孟蕾. "基于Logit 模型与GA 算法的商业银行信用风险评估模型研究." 大众商务 14 (2010): 72-72.

邹雅莹. *基于 k-means 算法的马田系统研究及其在个人信用评价中的应用*. MS thesis. 南京理工大学, 2014.

王会军. *基于稀疏贝叶斯学习的个人信用评估*. MS thesis. 西南财经大学, 2014.

王纯麟, 何建敏, and 钱苏丽. "基于组合分类器的个人信用评估模型." *现代管理科学* 11 (2006): 13-14.

贾开瑜. *基于成分分析法的银行信用卡线性评分模型*. MS thesis. 山东大学, 2012.

史宁. "商业银行对个人信用评估的组合预测模型." *商业研究* 11 (2009): 154-156.

徐鑫柱. *基于支持向量机和 BP 神经网络的个人信用评估模型研究*. MS thesis. 内蒙古大学, 2013.

杨海江, 魏秋萍, 张景肖. "基于改进的 AdaBoost 算法的信用评分模型." *统计与信息论坛* 26.2 (2011): 27-31.

姜叶飞. *惩罚变量选择方法比较分析及其在信用卡信用风险中的应用*. MS thesis. 厦门大学, 2014.

向晖. *个人信用评分组合模型研究与应用*. Diss. 博士学位论文]. 长沙: 湖南大学, 2011.

Wu, Kuo-Lung, and Miin-Shen Yang. "A cluster validity index for fuzzy clustering." *Pattern Recognition Letters* 26.9 (2005): 1275-1291.