第 30 卷第 10 期
2013 年 10 月
统计研究
Statistical Research
Vol. 30, No. 10
Oct. 2013

个人住房贷款违约预测与利率政策模拟

方匡南　吴见彬

内容提要:本文首次构建了基于非参数随机森林(Random Forest, RF)的住房贷款违约风险评估模型,利用某大型银行个人住房贷款数据,研究了借款人特征、贷款特征、房产特征和经济文化特征等因素对贷款违约的影响。实证研究发现已偿还比例、利率、贷款收入比、额度等是贷款违约最重要的影响因素,并且 RF 方法的预测准确率明显高于 logistic 模型等其他方法。此外,本文还研究了利率调整对贷款违约的影响,发现利率对违约率的影响是负方向的,且呈不对称性和非线性。

关键词:个人住房贷款;违约预测;利率政策模拟
中图分类号:C812　　文献标识码:A　　文章编号:1002 - 4565(2013)10 - 0054 - 07

Default Forecasting on Housing Mortgage and Interest Rate Policy Simulation

Fang Kuangnan & Wu Jianbin

Abstract: This paper proposed a housing mortgage default risk forecasting model based on non-parametric random forest at first. Then by using the housing mortgage database from a big famous bank in China, this paper studied the effect of housing mortgage default according to borrowers' characteristics, loan characteristics, housing characteristics and local economic and cultural characteristics. The empirical study found that the proportion which had been repaid, interest rate, ratio of loan to income, loan amount were the most important factors. The results also showed the prediction accuracy of RF were much higher than other methods such as logistic regression. In addition, this paper also studied how the interest rate affected mortgage default, finding that interest rate had negative effect, which were asymmetry and nonlinear, on the mortgage default.

Key words: Housing Mortgage; Default Forecasting; Interest Rate Policy Simulation

一、引言

我国自 1992 年首次开展个人住房抵押贷款以来,由于其收益性高、安全性好的特点,发展非常迅速,业务规模快速扩大。根据央行的统计数据,截至 2012 年底,我国个人住房抵押贷款余额 7.36 万亿,比上一年增长 12.8%,且过去五年以年均 17.4% 的速度增长。在贷款额不断上升的同时,风险也逐渐暴露出来,目前个人住房按揭贷款所面临的风险主要有四类:信用风险、抵押物风险、流动性风险和银行的操作经营风险,其中信用风险又分为违约风险和提前还款风险。其中违约风险是银行面临的最大风险,也是最难处理的一种风险。

2008 年金融危机以来,美国政府控制的住房抵押贷款机构房利美和房地美公布的高额亏损备受关注,而贷款违约就是造成此巨额亏损的主要原因。相比之下,我国的抵押贷款风险更加难以掌控,目前国内银行在实务操作中,仍然以经验判断为主,很少进行定量分析。因此,在信贷信息不对称下,如何利用统计分析、数据挖掘等方法建立可靠的分析模型,对贷款用户的行为进行风险识别和预测,有效提高我国个人住房贷款市场的风险管理水平具有重要的理论和现实意义。另一方面,利率的调整也会对贷款风险有较大的影响,本文将进一步研究利率调整对违约的影响。

二、文献回顾

本文首先从研究方法的角度对国内外有关个人

住房贷款市场风险管理的主要研究成果进行了归纳、梳理。信用评分方法是以评价对象的相关指标为解释变量,运用数理统计方法建立模型,以模型输出的信用分值或违约预测来度量评价对象的风险大小。从研究方法来看,目前最常用的信用评估方法是 Logistic 回归模型。但该方法主要存在如下几点问题:第一,在 logistic 回归中,或多或少受到变量共线性的影响,从而导致估计出的系数存在一定的失真,而且不太稳健,即当在方程中剔除或加入其他自变量时,该系数都可能发生明显的变化。第二,自变量对违约率的影响往往存在不对称性、非线性的特征,而 logistic 回归难以对非线性的不对称数据进行拟合。第三,logistic 回归主要用于处理数值型变量,当遇到属性变量(如学历、职级、性别等)时只能通过加入虚拟变量来替代,但客户数据库中大多数变量都为属性变量,如学历、职级、婚姻、行业、性别、地区、是否有其他担保品、是否有担保人等,当自变量存在 J 个水平值,则需要添加 $J-1$ 个虚拟变量,假如要把所有的属性变量都转化为虚拟变量,则模型中要包含几十个甚至更多的自变量,即使是先进行了变量筛选,剔除其中一些不重要的变量,自变量的个数仍然很多,如此庞大的自变量个数必然导致严重的共线性,使参数检验失效,出现较大的偏差。第四,logistic 回归必须先假定模型符合一定的假设,如残差服从正态分布、相互独立等,这些在实际操作中通常难以得到保证。

近年,一些学者将决策树等非参数方法引入到信用评分中,比如 chen 等(2010)利用决策树方法来分析抵押贷款的取消赎回权;Medema 等(2009)利用决策树对违约评估的有效性进行了研究,将评估有效性划分为理论有效、数据有效和统计有效。

此外,Lee(2007)构建了基于支持向量机的信用风险评估模型,对违约率进行了预测。

Bellotti 等(2009)利用支持向量机(SVM)方法建立了信用评分模型,认为有房者的违约风险较低,且过去 6 个月申请贷款的次数越多,风险越大,同时年龄变量也有重要影响。Twala(2010)使用了组合机器学习的方法进行信用风险的预测,并与单算法 ANN、决策树、贝叶斯网络、k 近邻和 Logistic 模型进行了比较,研究发现组合学习方法在信用评估中更为有效。Bhekisipho(2010)用组合学习算法评估信用风险,研究表明该方法的准确率较之单个分类器有着显著的提高。

目前国内的相关研究较少,且以实证研究为主。马宇(2009)利用 logistic 对实地调查的 637 个住房抵押贷款样本进行违约风险影响因素的实证研究,发现对违约影响较大的因素依次是住房面积、月还款额占家庭收入比、是否期房、受教育程度。王福林等(2005)发现影响我国个人住房抵押贷款违约因素按重要性依次为:是否当地人、贷款价值比、是否期房、月还款额占家庭收入比、还款方式、家庭收入和住房面积。颜新秀(2009)利用线性回归的方法,运用 1994 – 2009 年第一季度美国有关季度数据,对违约率与宏观经济指标之间的关系进行了研究,并指出当 GDP 增长放缓或房价连续下跌时,或利率急剧下降且长期低位运行时,应该警惕违约风险上升等。平新乔和杨慕云(2009)利用回归法对 4510 个样本进行了实证研究,研究结果显示消费信贷违约的影响因素,包括性别、年龄、受教育程度、年收入和职业类型等借款人的个人特征对违约率的影响。

综上所述,目前个人住房贷款违约预测通常使用的还是 logistic 模型,近年来国外学者开始尝试将决策树、神经网络、贝叶斯网络和组合学习等非参数方法引入到个人住房贷款信用风险研究中。由于 logistic 模型的预测准确率低,难以拟合非线性数据、不适合处理属性变量以及模型依赖于理想假设等缺陷,不适合用来对客户信用数据库进行建模预警。而且 logistic 模型、决策树、神经网络等建立在单个模型基础上的方法往往容易出现过拟合问题,前瞻性研究效果差。本文首次构建了基于非参数随机森林(Random Forest, RF)的住房贷款违约风险评估模型,RF 是 Breiman 于 2001 年提出的一种非参数统计方法,具有预测精度高、不容易过拟合等特征(Breiman, 2001)。

三、基于非参数的住房贷款违约风险评估模型

本文借鉴了 Breiman 的 RF 思想,构建了基于非参数 RF 的信用风险评估模型,基本思路是:首先,从住房贷款数据库中导出包含客户有关资料的样本集 D,然后生成随机向量序列 $\Theta_i(i=1,\cdots,k)$,利用 bootstrap 重抽样方法从原始样本集 D 中抽取 k 个子样本集,记为 $D_i(i=1,2,\cdots,k)$;接着,对每个子样本集 $D_i(i=1,2,\cdots,k)$ 分别建立住房贷款决策

树模型 $\{h(X,\Theta_i),i=1,\cdots,k\}$，其中，$X$ 是从住房贷款数据库里筛选出来的用于研究住房贷款信用风险的自变量矩阵；最后，通过 k 轮训练，得到了分类模型序列 $\{h_1(X),h_2(X),\cdots,h_k(X)\}$，再用它们构成一个多分类模型系统，该系统的最终分类结果采用多数投票法，可用如下公式表示：

$$H(x)=\arg\max_{Y}\sum_{i=1}^{k}I(h_i(x)=Y) \quad (1)$$

其中，h_i 是单个决策树分类模型，$I(\cdot)$ 为示性函数，Y 表示输出变量（或称目标变量），$H(x)$ 表示组合分类模型。基于非参数 RF 的住房贷款信用风险模型示意图见图 1。

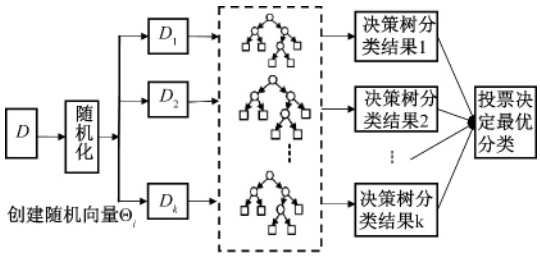

图 1 基于非参数随机森林的信用卡信用风险模型示意图

给定一组分类模型 $\{h_1(X),h_2(X),\cdots,h_k(X)\}$，每个分类模型的训练集都是从原始数据集 (X,Y) 随机抽样所得。则我们可以得到其余量函数（margin function）：

$$mg(X,Y)=av_kI(h_k(X)=Y)-\max_{j\neq k}av_kI(h_k(X)=j)$$

余量函数用来测度平均正确分类数超过平均错误分类数的程度。余量值越大，分类预测就越可靠。外推误差（泛化误差）可写成：

$$PE^{*}=P_{X,Y}(mg(X,Y)<0)$$

当决策树分类模型足够多，$h_k(X)=h(X,\Theta_k)$ 服从于强大数定律。可以证明，随着决策树分类模型的增加，所有序列 $\Theta_{1,\cdots}PE^{*}$ 几乎处处收敛于

$$P_{X,Y}(P_{\Theta}(h(X,\Theta)=Y)-\max_{j\neq Y}P_{\Theta}(h(X,\Theta)=j)<0)$$

这说明了为什么 RF 方法不会随着决策树的增加而产生过度拟合的问题。

RF 通过 bootstrap 抽样得到不同的训练集以增加模型间的差异，从而提高组合分类模型的外推预测能力。Breiman（2001）证明了 RF 具有很高的预测准确率，对异常值和噪声具有很好的容忍度，且不容易出现过拟合。可以说，RF 是一种自然的非线性建模工具，只需对样本信息不断进行训练，有很好的自适应功能，非常适合于解决先验知识不清、无规则多约束条件和数据不完备的问题。

RF 的另一个重要特点是还可以进行变量筛选。其基本思想是先对已生成的 RF，用 OOB 数据测试其性能，得到一个 OOB 准确率（或基尼值）；然后随机改变 OOB 数据中某个特征的值（即人为地加入噪声干扰），再用加入噪声后的 OOB 数据测试 RF 的性能，得到一个新的 OOB 准确率（或基尼值）；则原始 OOB 准确率（基尼值）与加入噪声后的 OOB 准确率（基尼值）之差作为相应特征的重要性度量值。

四、变量选择与数据预处理

（一）变量选择

从影响因素角度来分析，可以把影响个人住房贷款的因素分为借款人特征、贷款特征、房产特征和经济文化特征等四大类。

1. 借款人特征。

借款人特征是指借款人的性别、职业、年龄、婚姻状况、个人收入、信用状况等特征。大量研究表明借款人特征与违约风险具有高度相关性。Lawrence 等（1995）研究发现借款人过去信用记录、年龄、每月偿还额占家庭收入比例等因素显著影响违约风险。此外，Cowan（2004）利用借款人特征预测违约风险，发现借款人特征对违约风险影响显著。

2. 贷款特征。

贷款特征是指贷款金额、贷款期限、贷款价值比（Loan to value，LTV）等特征。研究最多的是有关 LTV 和贷款违约关系的研究，如 Qi 等（2009）对抵押贷款的违约损失进行了实证研究，发现 LTV 是最重要的影响变量，而且在房市低迷时，违约损失显著高于正常情况。

3. 房产特征。

房产特征是指房屋类型、房屋面积、房龄、房屋地理位置、房屋评估价与购买价之比等特征。Gau（1978）利用 873 个正常贷款样本和 212 个违约贷款样本对美国个人住房贷款违约风险进行研究，发现房屋的类型、房龄和住房的地理位置对违约有显著影响。

4. 经济文化特征。

经济文化特征是指一个国家或地区宏观经济形势、货币政策、文化习惯等特征。不少学者对利率和违约的关系进行了研究，比如 Magri 等（2011）研究了使用

意大利的住房按揭贷款数据对违约率与利率的关系进行了研究,研究表明高违约率通常伴随着高利率,利率和违约率相互影响;Daglish(2009)对违约率进行计算,研究表明在低利率环境下,信用升级的潜在可能会减少违约,而且违约率对利率和房价非常敏感。

(二)数据来源与变量说明

本文数据来自于中国某大型商业银行数据库,共收集到 2004 – 2009 年 3364 个违约样本和 14503 个非违约样本,共 17867 笔原始数据。该数据集包含 18 个变量,包括借款人特征(年龄、年收入、职级、学历、行业、性别、婚姻、贷款收入比)、贷款特征(利率、核准额度、核准期限、核准层次、其他抵押品、担保人、已偿还比例)、经济文化特征(利率)、房产特征(地理位置)和信用记录(违约记录)。各变量及类型见表 1。根据相关研究,违约有多种定义,本文的违约定义是指存在催收的客户,也就是存在"转催收日"的客户。

表 1　　变量说明

变量类型	变量名	代码	变量赋值
因变量	违约	Default	违约 = 1,非违约 = 0
借款人特征	性别	Sex	女 = F,男 = M
	婚姻	Marr	未婚 = 1,已婚 = 2,离婚或丧偶 = 3
	学历	Educ	博士 = 1,硕士 = 2,大专或本科 = 3,高中或中专 = 4,初中及以下 = 5
	行业	Work	农业 = 1,矿业 = 2,制造业 = 3,运输业 = 4,商业 = 5,金融业 = 6,政府 = 7,水电煤气 = 8,餐饮 = 9,学生 = 10,其他 = 11
	职级	Occu	自营 = 1,高管 = 2,职员 = 3,一般主管 = 4,专业技术人员 = 5,公务员 = 6,军人 = 7,教师 = 8,其他 = 9
	年龄	Age	借款人年龄
	年收入	Income	家庭年收入额
	贷款收入比	LOI	月还款额占家庭平均月收入的比值
贷款特征	额度	Loan	核准的贷款总额度
	期限	Term	核准的贷款总期限
	核准层次	Admi	董事会 = 1,总经理 = 2,信贷部 = 3,作业中心 = 4,其他 = 5
	有无其它抵押品	Mort	无其它抵押品 = 0,有其他抵押品 = 1
	有无担保人	Spon	无担保人 = 0,有担保人 = 1
	已偿还比例	Repa	已偿还的本金占总贷款本金的比例
经济文化特征	利率	Interest	贷款的年利率
房产特征	地理位置	Locate	房屋所处地理位置

(三)数据预处理

由于数据可能存在缺失、异常值等,所以在分析前有必要先进行缺失值与异常值的检测,其中有 113 个缺失值,同时数据集中共有 960 个样本是收入为 0、但有正当职业的样本,这部分收入数据明显不符合实际,由于本文的样本量足够多,删除这部分缺失、异常值,对研究的影响不大,剩余有效样本有 16794 个。

统计模型(比如 CART、logistic 等)往往只能对数据分布比较对称的数据集具有较好的预测能力。然而在实际的运用中,数据集往往集中于非目标数据,使得建立的模式无法对数量较少的类别数据进行正确的预测,即样本的非对称分布问题。因此,本文采用"减少多数法"来平衡数据分布,该方法是从数据量比较多的类别中除去特征差异性较大的数据,如噪声资料(noise data)和边界资料(borderline data),再用抽样技术从样本较多的数据集中选取部分具有类别代表性的资料,用以降低类别间的不对称性。

本文利用原始数据集,生成训练集、测试集和预测集。首先,为了检验平衡后所建模型对原始数据的适应性,本文将原始数据集中的约 20% 数据(3682 个样本)预留为预测集,进行最后的外推预测检验。其次,用剩余的 13112 个数据进行平衡取样:使用 Clementine 的平衡(Balance)工具,减少样本中非违约的数目,以其中的 2635 笔违约数据为 1 单位,将违约和非违约的比例大约平衡为 1:1。平衡后的数据集中包含 2635 笔违约户和 2640 笔非违约户;最后,将平衡后的数据集按一定的比例划分为训练集和测试集。

五、实证分析

(一)指标体系的确定

在确定信用风险评估模型之前,必须选择合适的风险评估指标体系。由于原始自变量很多,但这些变量并不都有助于信用风险的预测,反而可能由于变量间的相关性等降低模型的有效性,因此从原始变量中选取少数几个合适的变量建立风险评估指标体系,有助于住房按揭贷款发放银行的审核和重点监控。如何选取合适的评估指标体系是个难点,本文利用 RF 法来筛选预测模型的自变量,RF 采用启发式算法,通过在变量被加入噪声前后的预测准

确性差异来判断变量的重要性，见表 2。

表 2 变量重要性

变量	Gini 值平均减少量	变量	Gini 值平均减少量
已偿还比例	764. 98	学历	69. 78
利率	294. 21	行业	38. 57
贷款收入比	187. 47	地理位置	34. 1
额度	178. 56	性别	33. 14
年龄	166. 01	婚姻	30. 97
年收入	127. 32	核准层次	9. 65
期限	117. 02	担保人	9. 42
职级	90. 07	其它担保品	0. 16

从表 2 可以看出，贷款特征变量重要性强于借款人特征变量，其中“已偿还比例”是影响贷款违约的最重要变量，这在一定程度上印证了相关研究中常用到的“违约机会曲线”的重要性。此外，利率是第二重要的影响变量，与 Magri 等(2011)和 Daglish (2009)的研究结果相吻合。实际上，利率相当于信贷市场中资本的价格，高的利率会产生逆向选择，使得那些收益率低、风险小的项目由于支付不起高利率而被挤出市场，剩下的往往是那些高风险高收益的项目。

为了确定最优的指标体系，本文根据变量重要性排序，分别选取前 6 个变量、前 8 个变量、前 10 个变量、前 12 个变量为输入变量建模。其中，RF 参数 mtry 的取值由系统默认设定，同时将训练集和测试集的划分比例设置为 8:2，分别计算训练集和测试集的违约样本预测准确率和非违约样本预测准确率，并利用训练所得的模型对预留的预测集进行预测，见表 3。

表 3 预测准确率

输入变量个数	训练集		测试集		预测集	
	违约	非违约	违约	非违约	违约	非违约
6	98. 70	96. 57	98. 41	96. 77	99. 04	96. 76
8	98. 64	97. 36	97. 80	98. 56	98. 56	97. 71
10	98. 46	97. 12	98. 60	96. 91	98. 92	97. 11
12	98. 54	96. 93	98. 26	97. 28	98. 92	97. 20

从表 3 可以看出，并不是模型所包含的自变量个数越多，预测准确率就越高，这也进一步验证了选取合适指标体系的必要性。相对来讲，前 6 个变量模型在训练集里对违约客户的预测准确率最高，准确率为 98. 70%，在测试集里第二高，为 98. 41%，而且在预测集中也最高，为 99. 04%；虽然对非违约客户的预测准确率不是最高的，但总体上精度较高，比如在预测集中准确率为 96. 76%。通常，银行更注重对违约客户的预警，故综合来看，本文认为 6 变量模型最为合适，而且该模型由于所需变量最少，故收集数据的成本最小，银行的审核和监控成本也最小。

(二)模型结果与解释

根据上文确定的指标体系，以前 6 个变量为输入变量，建立 RF 风险评估模型，同时为了比较 RF 方法与其他方法的优劣，本文分别利用支持向量机(SVM)、神经网络、bayes 分类和 Logistic 模型建立预测模型，各模型的预测结果如表 4 所示。

表 4 五种模型的比较 (%)

方法	训练集		测试集		预测集	
	违约	非违约	违约	非违约	违约	非违约
SVM	85. 32	88. 13	85. 85	84. 04	81. 51	77. 56
神经网络	91. 98	86. 43	94. 57	85. 33	89. 98	86. 03
Logistic	87. 33	86. 14	86. 17	84. 89	87. 64	85. 4
Bayes	78. 03	61. 36	82. 1	62. 34	74. 72	56. 21
RF	98. 70	96. 57	98. 41	96. 77	99. 04	96. 76

注：本研究中，上表的 SVM 模型已经采用归一化；SVM 模型的参数设置如下：迭代容忍度 = 0. 001，内核：RBF，gamma = 0. 1，规则化参数 = 10。

从表 4 可以看出，RF 模型的预测准确率最高，神经网络模型次之，接着是 SVM 模型，预测效果最差的是 Bayes 模型，其次是 Logistic 模型，而 Logistic 模型是商业银行以往使用最广泛的模型。此外，各个模型对违约客户的预测准确率都要高于对非违约客户的预测准确率。

此外，为了进一步比较不同模型的优劣，本文还分别画出各个模型的提升图，并进行了比较。在提升图中，横轴代表了总体数据的百分比，纵轴是各百分比对应的提升度，提升度指的是每个模型分位点中正例的命中率与随机排列所得的命中率之比，计算公式为 $lift = \frac{n}{r}$，其中，n 为分位点中的正例数占分位点中的样本数的比例，r 为总正例数占总样本数的比例。对于一个好的模型来说，提升图应该是恰好从左端高于 1. 0 处开始，当移动到右边时能够保持在一个高度稳定的水平上，然后到图像右端时突然急剧地减小到 1. 0。而对于一个没有提供任何信息的模型来说，整个图像中曲线将一直在 1. 0 附近围绕。图 6 是 SVM、神经网路、Logistic、Bayes 和 RF 在训练集与测试集 8:2 下的提升曲线，该图中基线就是一条取 1 的水平线，和图中的 x 轴重合。

从各个模型的提升图可以看出，RF 的提升曲线与理想模型(BEST)几乎重合，明显优于其他曲线，且提升率也是最大的，其次是神经网络，最差的是

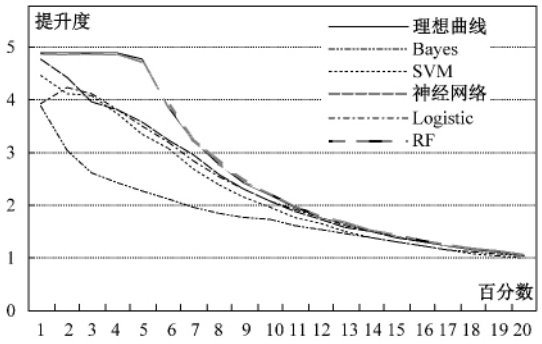

图 2　五个模型的提升曲线

Logistic 模型,这与准确性分析结果相一致。从图 2 中还可以发现,RF 不仅具有很高的准确率,而且训练集和测试集的提升曲线几乎没有差别,这充分说明了模型具有良好的外推性。

以上的研究表明,RF 方法的预测准确度明显高于其他方法,而且没有出现过度拟合的现象,其次为神经网络,而 logistic 表现一般,最差的是 bayes 模型。近年来也有许多将这些方法应用到信用风险评估的研究,通常情况下 CART、神经网络和 SVM 比 logistic 回归更准确(方匡南,吴见彬等,2010)。

六、利率政策模拟

前文研究发现利率是影响住房贷款违约的一个重要变量,本文利用所建立的 RF 模型,模拟利率调整对住房贷款违约的影响,模拟时仅改变利率值,其他变量保持不变,从而得到利率的边际效应。所利用的数据为平衡前预先保留的预测集数据,其中有 2953 个非违约的样本,729 个违约样本。倘若按原先的利率进行违约预测,预测为非违约的样本有 2888,违的有 794 个。若利率上调 0.25 时,模型预测为非违约样本减少为 2860 个,违约样本在原先的基础上增加至 822 个,即在原先利率下,预测为非违约样本有 28 个被划分到违约类。同理,对其他利率调整幅度进行了模拟,分别模拟了当利率上调和下调 0.25、0.5、1 和 2 个基点的情况,以及利率在初始值基础上分别上升和下降 1%、2.5%、5%、10% 和 20% 的情形,见表 5。表 5 中第 1 列是利率按不同的基点上调和按不同的百分比上升的情况,第 6 列表示利率按不同的基点下调和按不同的百分比下降的情况。"非违约"表示预测集中预测为非违约的客户数。"违约"表示预测集中预测为违约的客户数。"变化数"表示利率调整后的违约人数变化情况,正数表示违约人数增加,负数表示违约人数减少;"变化率"表示利率调整导致违约人数的变化率,计算公式为:变化率 = 变化数/原违约人数(794)。

表 5　　利率调整模拟

加息	非违约	违约	变化数	变化率(%)	降息	非违约	违约	变化数	变化率(%)
I + 0.25	2860	822	28	3.53	I - 0.25	2894	788	-6	-0.76
I + 0.5	2858	824	30	3.78	I - 0.5	2913	769	-25	-3.15
I + 1	2843	839	45	5.67	I - 1	2919	763	-31	-3.90
I + 2	2810	872	78	9.82	I - 2	2957	725	-69	-8.69
I* 1.01	2863	819	25	3.15	I/1.01	2888	794	-0	0.00
I* 1.025	2858	824	30	3.78	I/1.025	2895	787	-7	-0.88
I* 1.05	2862	820	26	3.27	I/1.05	2900	782	-12	-1.51
I* 1.1	2848	834	40	5.04	I/1.1	2914	768	-26	-3.27
I* 1.2	2827	835	61	7.68	I/1.2	2943	739	-55	-6.93

注:I + 0.25 表示加息 25 个基点,I - 0.25 表示降息 25 个基点,I* 1.01 表示在现有利率基础上加息 1%,I/1.01 表示在现有利率基础上降息 1%,其余类推。

从表 5 的模拟结果来看,当利率上调 25、50、100 和 200 个基点时,违约率分别增加 3.53%、3.78%、5.67% 和 9.82%;当利率下调 25、50、100 和 200 个基点时,违约率分别减少 0.76%、3.15%、3.90% 和 8.69%。当利率上升和下降 1%、2.5%、5%、10% 和 20% 时,违约率也有类似的变化。这说明利率对违约存在负影响,即加息导致违约率上升,而降息导致违约率下降,这与 Magri 等(2011)和 Daglish(2009)的研究结论一致。

同时,利率对违约率的影响呈现一定的不对称性。加息导致违约率的增加值要大于降息相同基点或者相同比例导致违约率的减少值,也就是说,加息造成的违约增加效应比降息造成的违约降低效应更加显著。例如加息 25 个基点时,使得违约率增加 3.53%,但是降息 25 个基点,违约率仅减少 0.76%;又例如加息 200 个基点,违约率增加 9.82%,而降息 200 个基点,违约率仅减少 8.69%。

此外,利率调整对违约率的影响是非线性的,这说明研究利率对违约率的影响不能简单使用线性模型分析。例如,利率上调 25 个基点,违约率增加 3.53%,而利率从上调 100 个基点变化到上调 200 个基点的时候,违约率增加 4.15%(4.15% = 9.82% - 3.53%)。

七、小结

本文首次构建了基于非参数 RF 方法的个人住房贷款违约风险评估模型,并应用于我国商业银行个人住房贷款违约预测,通过对 17469 个个人住房贷款样本的研究,发现 RF 的预测准确率明显高于

SVM、神经网络、bayes 分类和 Logistic 模型等；发现已偿还比例、年龄、年收入、利率、职级、核准期限等是影响个人住房贷款违约的最重要的 6 个变量，基于这 6 个变量建立的违约预测模型能比较准确地判断出借款人是否违约。此外，本文还对利率调整对个人住房贷款违约进行了模拟，模拟结果发现利率对违约率是负的影响，并存在不对称性和非线性特征。本文的研究结论可为银行的风险压力测试提供一定的参考，进一步而言，为国家的宏观调控提供参考。本文构建的基于 RF 的个人住房贷款违约模型方法不仅准确性高，而且适合用属性数据处理，十分适用于银行信用风险预警模型，可以为银行建立起有效的基于信用风险甄别的风险控制机制。本文的不足之处在于，RF 模型只能判断出是否违约，虽然在 RF 模型也会给出概率度，但是该指标与传统违约概率的概念有差别，还不能直接得到违约概率，这方面有待进一步研究。

参考文献

[1] Chen T. H. and Chen C. W. Application of data mining to the spatial heterogeneity of foreclosed mortgages [J]. Expert Systems with Applications, 2010, 37(2):993 – 997.

[2] Medema L., Koning R. H., and Lensink R.. A practical approach to validating a PD model [J]. Journal of banking and finance, 2009, 33(4):701 – 708.

[3] Bellotti, T. and Crook, J. Support vector machines for credit scoring and discovery of significant features [J]. Expert Systems with Applications, 2009, 36(2): 3302 – 3308.

[4] Lee Y. C. Application of support vector machines to corporate credit rating prediction [J]. Expert Systems with Applications, 2007, 33:67 – 74.

[5] Twala, B. Multiple classifier application to credit risk assessment [J]. Expert Systems with Applications, 2010, 37(4): 3326 – 3336.

[6] Bhekisipho T. Multiple classifier application to credit risk assessment [J]. Expert Systems with Applications, 2010, 37(4): 3326 – 3336.

[7] 马宇. 我国个人住房抵押贷款违约风险影响因素的实证研究 [J]. 统计研究, 2009 (5):100 – 107.

[8] 颜新秀. 个贷违约率与宏观经济指标相关性研究 [J]. 国际金融研究, 2009 (10):59 – 67.

[9] 平新乔，杨慕云. 信贷市场信息不对称的实证研究——来自中国国有商业银行的证据 [J]. 金融研究, 2009 (3):1 – 18.

[10] Lawrenc E. C. and Arshadi N. A multinomial logit analysis of problem loan resolution choices in banking [J]. Journal of Money, Credit and Banking. 1995, 27(1):202 – 216.

[11] Cowan, A. M. and Cowan, C. D. Default correlation: An empirical investigation of a subprime lender [J]. Journal of Banking and Finance, 2004, 28(4): 753 – 771.

[12] Qi M. and Yang X. L. Loss given default of high loan-to-value residential mortgages [J]. Journal of Banking and Finance, 2009, 33(5): 788 – 799.

[13] Gau G.. W. A taxonomic model for the risk-rating of residential mortgages [J]. The journal of business. 1978, 51(4):687 – 706.

[14] Magri, S. and Pico, R. The rise of risk-based pricing of mortgage interest rates in Italy [J]. Journal of banking and finance, 2011, 35(5):1277 – 1290.

[15] Daglish T. What motivates a subprime borrower to default? [J]. Journal of Banking and Finance, 2009, 33(4):681 – 693.

[16] 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——基于非参数随机森林模型的实证分析 [J]. 经济研究, 2010 (1):97 – 107.

作者简介

方匡南，男，浙江台州人，2010 年毕业于厦门大学经济学院，获经济学博士学位，厦门大学经济学院统计系副教授，硕士生导师。研究方向为数据挖掘与计量经济学。

吴见彬，女，福建宁德人，比利时鲁汶大学统计学博士研究生。研究方向为金融时间序列与数据挖掘。

（责任编辑：程 晞）