

个人住房按揭贷款违约风险研究^{*}

钱争鸣 李海波 于艳萍

内容提要:近年来以个人住房按揭贷款为主要形式的居民个人消费金融在国内发展迅速,为全面深入研究个人住房按揭贷款违约风险的主要影响因素,本文构建了四维度整体分析框架,提出用非参数随机森林法进行异常点度量和特征变量选取,设计个人住房按揭贷款违约风险的 Logistic 回归模型,对影响方向和影响程度进行深入的定量分析。结果表明,在所选择的 17 个分析变量中影响显著的有 8 个因素。研究结论为商业银行和其他金融机构在开展个人住房按揭贷款等消费金融业务时如何防范和管理风险,提供了新的分析方法与实际指导作用。

关键词:个人住房按揭贷款 违约风险 非参数随机森林法 Logistic 回归模型

一、引言

近年来,随着我国金融行业对外开放的水平不断提升,各商业银行纷纷将业务拓展的重点转向个人金融业务。金融产品的竞争日益激烈,以个人住房按揭贷款为主要形式的居民个人消费金融在国内发展迅速,逐渐成为各家银行业务拓展的重点,在信贷规模和效益贡献上的占比得到了很大的提高。但随着信贷业务的增长,不良贷款和不良贷款率等不良风险指标也日益突显,借款人违约风险有继续增加的趋势。个人住房按揭贷款违约会给贷款银行等金融机构带来预期利润损失,这种利润损失可能来自于贷款无法按时收回,预期收益无法实现;也可能来自于收回抵押住房并将其拍卖出让时出现的困难,即实现抵押物兑现时的交易成本增大。但更值得关注的是,一旦出现个人住房按揭贷款大规模违约,势必会给整个金融系统带来巨大的冲击,而这种资产价格变化所带来的冲击在金融加速器的作用下,会对宏观经济产生巨大影响。因此,需要研究个人住房按揭贷款违约风险的主要影响因素及其对贷款违约的影响方向和影响程度。特别是根据当前国内居民消费金融中住房按揭贷款市场的风险现状,揭示个人住房贷款背后隐藏的风险,分析主要影响因素,为商业银行在发展消费金融中防控个人住房按揭贷款违约风险提供新的分析方法与实际指导作用,具有积极和重要的现实意义。

国外学者在消费金融中个人住房按揭贷款及其违约风险管理方面的研究,主要集中在个人住房按揭贷款的信用风险方面,通过梳理相关文献,可以归结为以下三个方面:20 世纪 70 年代,以 George W. Gau 为代表采用特征变量对个人住房按揭贷款违约风险的相关特征进行综合分析;20 世纪 70—80 年代对个人住房按揭贷款违约行为研究,而 Foster & Van Order 最早应用期权理论对违约行为及其后果进行了探讨;20 世纪 80 年代以后,以 Green & Shoven 为主要代表,提出了提前支付决定经验模型,对个人住房按揭贷款集群的违约问题进行研究。相比之下,国内学者在个人住房按揭贷款违约风险方面的研究还十分欠缺,主要是对国外一些研究成果的借鉴。形成的主要研究成果有:一是刘萍(2002)对我国个人住房按揭贷款违约行为和所面临风险形式的探析;二是汪利娜

^{*} 钱争鸣、李海波、于艳萍,厦门大学经济学院,邮政编码:361005,电子信箱:zmqianxmu@sina.com。本研究得到国家社会科学基金项目(08BJY076)、教育部人文社会科学重点研究基地重大课题项目基金(07JJD790145)的研究资助。感谢匿名审稿人的意见。

(2001)从保险视角对个人住房按揭贷款保险品种和运作方式的研究,强调可以借鉴美国个人住房抵押贷款保险制度,采用住房抵押保险的混合模式来防范违约风险;三是施海松(2007)从美国次贷危机的启示出发,对当前我国住房按揭贷款市场形势与美国市场的同质性和差异性进行分析,并就其风险管理进行研究。

而在个人住房按揭贷款违约风险的建模分析与实际应用方面,Jung Allen 成为早期通过对数据的统计分析寻找影响违约风险因素的代表。其后国内外学者从不同角度进行研究,归纳起来,研究方法 与模型选择、数据资料来源、研究的重要结论见表 1 所示。

表 1 国内外个人住房按揭贷款违约风险影响分析相关文献汇总

作者	模型选择	数据来源	重要结论
Von Furstenberg	多元回归模型	联邦住宅委员会	贷款价值比、贷款期限、收入、贷款期限 为重要影响因素
Lawrence et al.	Logistic 模型	国家财务公司	信用状况、年龄、贷款额度等
Canner et al.	Logistic 模型	消费者财务状况调查 及人口普查资料	贷款价值比与违约率、损失率均显著相 关
Quigley and Order	比例风险模型	联邦住宅抵押贷款公 司	30 年期固定利率、低贷款价值比或自用 住宅有较低的违约率
Epperson et al.	Logistic 模型、Probit 模型和 判别分析	住宅融资年度调查报 告	三种统计方法在分类正确性与影响因素 均有不同结果
Campbell and Dietrich	Logistic 模型	抵押贷款保险公司	贷款价值比与偿还能力对违约有显著影 响
Deng et al.	期权模型	联邦住宅委员会	贷款价值比、住房价值、失业率及离婚率 均为重要影响因素
George W. Gau	因子分析、判别分析、聚类 分析	两家私营抵押保险公 司	影响各个类别的违约特征是不同的,对 不同类别的研究以达到风险评级的目的
李馨苹	聚类分析、因子分析、 Logistic 模型、判别分析	台湾台中地区某银行	影响不同类别借款人的因素存在差异 性,应分类研究

资料来源:详见参考文献。

从以上的研究综述中看出,传统和已有的研究中仍存在两个方面的主要缺陷:一是对个人住房按揭贷款违约风险因素来源的全面系统分析不够,二是较少开展对个人住房按揭贷款违约风险的影响方向和影响程度分析的数量分析,特别是如何解决分析对象的复杂数据集中异常点检测和特征变量的选取,是传统研究与实际应用分析中仍未涉及和没有解决的关键问题。为此,我们提出构建个人住房按揭贷款违约风险的四维度整体分析框架,对当前国内居民消费金融中个人住房按揭贷款风险现状及来源进行全面分析,探讨影响违约风险的各类因素。通过非参数随机森林法和建立 Logistic 回归模型,揭示出背后隐藏风险及其影响因素,并对其特征值、影响方向与影响程度做深入的定量分析。

本文第二部分提出并构建个人住房按揭贷款违约风险的四维度整体分析框架;第三部分从方法论角度构建用于定量分析与解决问题的非参数随机森林法和 Logistic 回归模型;第四部分是变量设计与模型实证,第五部分是结果分析。

二、个人住房按揭贷款违约风险的四维度整体分析框架

笔者认为个人住房按揭贷款违约风险的发生,主要是与权益状况的变动和支付能力的变化有

关。因此,从理论上就可以用支付能力理论和权益理论来分析违约发生的成因:支付能力理论与借款人个人财务状况有关,支付能力的变化往往发生在合同签订后借款人家庭收入发生重大变动,或由于借款人经济财务状况变动而引起的支付能力变化,它受借款人月收入和住房持有成本因素的影响。可能会导致借款人无法按时偿还贷款,或继续还款的行为,从而导致被动违约;权益理论着重分析个人的房产权益数额,拥有极大正权益的贷款人会在市场上出售房产来提前偿还住房贷款而主动违约。相反,若权益为负值,即使借款人有能力支付每月的还款额,主动违约仍会发生。另外,在借款人主观上认为放弃继续还款可能带来更大收益,或是在房价下降和市场利率降低,借款人将住房转让后所获的价款不足以归还抵押贷款余额时也会发生这类违约行为。

住房贷款具有金额大、期限长的特点,在信息不对称的市场中,个人住房按揭贷款违约风险来源于许多方面。为全面、深入研究个人住房按揭贷款违约风险,我们提出和构建以借款人特征、贷款特征、房产特征和外部经济特征四个维度的系统整体分析框架,如图 1 所示。下面对各个维度的主要影响因素进行分析。

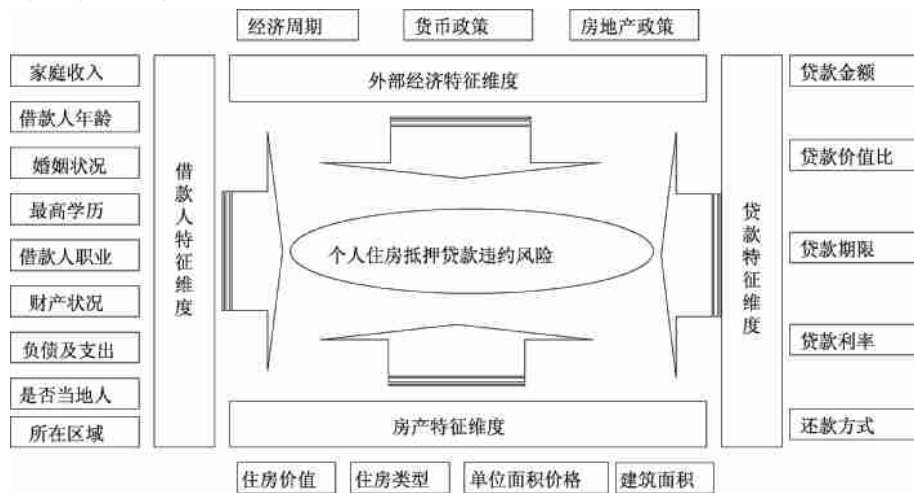


图 1 影响个人住房按揭贷款违约的四维度分析框架

(一)借款人特征维度

借款人特征维度的主要影响因素包括家庭收入、年龄、婚姻状况、最高学历、职业稳定性、财产状况、负债和支出状况、所在区域等。其中各主要变量选取的作用、测度对象及其经济含义是：(1)家庭收入的变动作为衡量借款人偿还抵押贷款能力的重要指标,能反映借款人的财务负担能力。(2)年龄的变化说明不同年龄借款人的收入与风险倾向不同,从而成为影响住房抵押贷款违约风险的一个重要因素。(3)婚姻状况不同反映着家庭收入来源的不同,表明家庭收入来自单亲还是双亲,来自单薪还是来自双薪。(4)学历的变化反映在相同的行业背景下借款人学历越高收入可能越高且收入稳定性更强。(5)借款人职业差别决定其能够享受的住房、医疗、劳保、退休等方面的待遇。因此,职业状况成为评估的一个重要指标。(6)财产状况的变化说明借款人的个人财产(包括住房、金融资产等)状况,它可以作为债务担保而成为清偿债务的来源。如果拥有充足的流动性和保值性强的财产作抵押,其履约还款的可能性会大大提高,资信程度也较高。(7)负债及支出变动状况不仅是评估个人资信状况的重要依据,也是决定借款人举债能力的关键。负债状况主要调查借款人在银行是否有其他负债或担保,而支出状况则主要调查其抚养和赡养情况。(8)所在区域的差别表明不同区域可能有不同的产业结构、经济发展状况和不同的失业和个人收入波动。而且不同地区银行对个人住房抵押贷款的要求存在一定的差异。

(二)贷款特征维度

贷款特征维度的主要影响因素包括还款方式、贷款金额、贷款期限、贷款利率、贷款价值比等。各主要变量选取的作用、测度对象及其经济含义是：(1)贷款金额反映借款人债务负担量情况。负担越大，借款人财务恶化时贷款发生违约的可能性就越大。而贷款金额越大，银行在单笔贷款上遭受损失的严重性可能也越大。(2)贷款价值比反映借款人初始权益投入，它与违约风险有正相关关系，即个人住房抵押贷款违约风险水平是贷款价值比的增函数。(3)贷款期限越短，月还款额越大，财务负担就越重，使违约造成的潜在权益损失也越大。(4)贷款利率越高，利息负担就越大，一般认为违约风险水平是贷款利率的增函数。由于当前我国实行管制下的浮动利率，利率波动风险主要由借款人承担，因此它可能对违约风险产生较大影响。(5)还款方式有等额还款(等额本息)和等本还款(等额本金)。前者指在贷款偿还期内借款人每期还款的数量固定不变(不考虑利率调整)。后者指借款人每期偿还贷款本金额固定不变，偿还利息随贷款本金的减少而减少，而这种类型的借款人大多属于风险回避型，所以其违约的概率相对较低。

(三)外部经济特征维度

外部经济特征维度包含的影响因素很多，对住房贷款的违约构成直接影响的主要有三个方面：(1)经济周期的影响。一般而言，当经济处于复苏与繁荣阶段房价也处在上升通道和高位运行，则按揭贷款市场繁荣且贷款违约率较低。而当经济处于衰退和萧条期间房价下跌，容易出现大量违约的风险。(2)货币政策主要指政府或中央银行为影响房市等经济活动所采取的控制货币供给及调控利率的各项措施。一般而言，在宽松政策下货币供应量大，虽然违约风险较小，但各银行可能会在审批时放松标准，从而为违约埋下伏笔，一旦经济恶化或个人遭遇困难就可能无力按期偿还贷款，从而出现违约，这是美国次级贷款危机中的一个典型现象。而紧缩政策会导致贷款成本增加，使以浮动利率计算的每期还款额相应上升。(3)房地产相关政策的影响表现在如果提高首付比例和收紧二套房政策，都将影响住房贷款市场。而贷款利率是否优惠也对借款人每期还款额产生直接影响，优惠一旦取消，利率上浮，会增加借款人负担从而可能导致贷款违约。

(四)房产特征维度

房产特征维度变量包括住房价值、住房类型、建筑面积、单位面积价格等。(1)充当贷款抵押物的住房价值是抵押贷款的最终保证。如果购买价格大于市场评估价值，借款人出售房产不能弥补成本，在房地产市场下滑时就会加大贷款违约风险。(2)住房类型主要指用于抵押的房产是期房还是现房。前者在抵押贷款发生违约时可能会导致购房权的丧失。后者是借款人已获得房屋产权，从而会助长违约风险发生。(3)住房单位面积购买价格综合反映了住房户型、建筑质量、环境等，如果与实情发生较大背离，则发生违约的可能性会增加。(4)建筑面积主要反映人们对居住需求的满足程度。在房价波动时会增加借款人违约的概率。

如果从建房和卖房的视角进行分析，房产特征维度的风险还可能来自于经销商和开发商对住房的“假按揭”。经销商风险主要有：一是经销商不具备销售资格或在住房销售合同上违规，导致购房行为与销售合同无效。二是经销商高度负债经营时，存在经销商卷款外逃而使购房者产生损失的风险。而“假按揭”指开发商以本单位职工或其他关系人冒充客户作为购房人，通过虚假销售方式套取银行贷款的行为。从而产生开发商利用积压房产套取银行信用，欺诈信贷资金的风险。

上述的四个维度中第一维和第二维是违约风险的主要来源。特别是由于借款人的经济财务状况变动而导致借款人无法按期偿还贷款而成为违约风险的重要来源，也是商业银行在放贷前关注的重点。违约风险很大程度上是源于信息不对称性，而商业银行很难或完全无法掌握借款人个人经济财务状况的全部信息。在国家个人社会信用系统、信用环境和法制环境不够完善的条件下，借款人可能隐瞒自身真实收入和风险状况，特别是那些有问题的借款客户的信息得不到充分披露，致使商业银行对其缺乏足够的认识与识别，则使违约风险难以防范。而且随着信息不对称面不断扩

大,会产生高风险问题客户的“挤入”和低风险潜力客户被“挤出”的效应。因此,在信息不对称条件下为提高违约风险的管理水平,一方面商业银行需尽可能获得真实完善的客户信息,但获得完备可靠的信息不仅成本高而且几乎不可能。另一方面就需要着重研究对个人住房按揭贷款违约风险的有效管理方法。我们认为可以从对借款客户的历史数据分析中得到一些情况,来弥补由于信息不对称而产生的一些缺陷。商业银行通过构建由借款人特征维度和贷款特征维度变量的历史和当前状态数据组成的数据库或由四个维度构成的数据仓库,利用非参数随机森林法等现代数据挖掘和统计方法对数据库中的数据集合进行计算分析,就能获得一些有价值的信息,从而大幅提高对个人住房按揭贷款违约风险的识别和管理能力。而这些恰恰是用传统方法难以实现的。

三、非参数随机森林法和 Logistic 回归模型

由于个人住房按揭贷款违约风险分析的数据分布复杂,特征较多,且噪声很大,同时还含有连续属性和离散属性,用传统方法难以达到满意的评估性能与效果。为解决问题,本文引入非参数随机森林法和 Logistic 回归模型对个人住房按揭贷款违约风险进行系统建模与定量分析。

(一)非参数随机森林法

非参数随机森林法是 Breiman 2001 年在 CART 算法基础上,结合 Bagging 和随机特征选取而提出的一种新组合分类器算法,具备了可用于分类、回归、变量重要性评估、异常点检测分析等多种功能,能够很好地容忍噪声,并具有许多优良的特性。它作为一种自然的非线性建模工具,只需要对样本信息利用多个决策树进行学习训练,就有很好的自适应功能。而且能同时处理连续属性和离散属性,具有对随机变量特征选择的组合算法,能克服单个模型的过拟合现象,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全的应用问题。利用非参数随机森林法进行变量的分类与选择的基本思路是:首先,利用 bootstrap 重抽样方法从研究对象的原始样本集 D 中抽取 k 个 bootstrap 样本集,然后生成随机向量序列 $\Theta_i(i, L, k)$,记为 $D_i(i=1, 2, L, k)$;其次,对每个 bootstrap 样本集 $D_i(i=1, 2, L, k)$ 分别建立个人住房按揭贷款风险评价决策树模型 $\{h(X, \Theta_i), i=1, K, k\}$,其中 X 是筛选出来的用来研究个人住房按揭贷款风险的自变量,其中参数 $\{\Theta_k\}$ 集是独立同分布的随机向量;其三,通过 k 轮训练,得到分类模型序列 $\{h_1(X), h_2(X), h_k(X)\}$,再用它们构成一个多分类模型系统,系统的最终分类结果由最终分类决策模型来决定,其公式表示为: $H(x) = \arg \max_{Y \in \{1, \dots, K\}} \sum_{i=1}^k I(h_i(x) = Y)$, 其中, h_i 表示单个决策树分类模型, $H(x)$ 是组合分类模型, Y 为输出变量(或称目标变量), $I(\cdot)$ 表示示性函数。分析的目标是以多数投票决策的模式来确定最终的分类。

理论和大量实证研究都证明非参数随机森林法具有良好的泛化性、稳健性和噪声容忍度,而且不容易出现过拟合(Breiman, 2001)。能同时处理连续属性和离散属性,特别是具有较为突出的三大功能:一是很强的数据集异常点识别、比较和删除功能。在对所有样本的异常点进行测度与比较过程中,如果发现某样本的异常点度量较大,或是某个样本与其他样本的相似程度较小,则判别为可能存在异常点。若事先预知置信区间,可把超过某个阈值的样本视为异常点;二是特征值与变量选择的功能,使用 Wrapper 算法,从许多特征中找出那些最有效的特征作为特征变量。而且随机森林法通过构造不同的训练集增加分类模型间的差异,具有很强的组合分类和预测与外推能力,并具有很高的准确率;三是可以进行参数调整的功能,目的是使模型具有较好的分析性能,达到较好的分析效果。随机森林法中两个比较重要的可调整参数是(1)ntree:调整随机森林模型中树的数目;(2)mtry:调整每个节点处候选特征的个数。其中 ntree 的设置比较简单,只要让森林的整体误差率趋于稳定就行。而 mtry 相对更重要一些,因为它对模型性能的影响更大。随着森林中树数目的增加,森林的总体误差率会趋向一个有限上界,只要让 ntree 的值足够大,就可以保证 RF 收敛。而随

机森林法的分析效果主要表现在它对变量选择和分类与预测的准确率。

(二)Logistic 回归模型

Logistic 模型最大优点在于它解决了因变量不连续回归问题,特别是因变量为分类变量时非常适合使用该模型进行研究。应用 Logistic 回归可识别自变量对个人住房抵押贷款违约风险的影响方向和程度。Logistic 模型的具体形式为:

$$\ln(\frac{P_i}{1-P_i}) = Z_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + + \beta_n X_n + \mu_i$$

其中: P_i 是个人住房抵押贷款违约事件发生的概率, $1-P_i$ 是不发生违约的概率, β_i 是待估计系数, X_i 是自变量, μ_i 是随机误差项。Logistic 模型主要采用极大似然法进行模型的估计与拟合,回归系数检验采用 Wald 统计量检验,检验值越大表明该自变量的作用越显著。

Logistic 回归模型在个人住房抵押贷款违约事件中应用可以表示为,如果假设: y 表示发放一笔房贷这一事件,用 $y=1$ 表示到期后借款人违约(通常称为一个“坏”的贷款), $y=0$ 表示借款人不违约(称为一个“好”的贷款)。我们的目的是利用收集的样本数据和资料建立模型,对借款人违约(即用 p 表示 $y=1$)的概率 p 进行预测。 x_i 是描述与违约概率有关的自变量影响因素的一些指标变量,这些指标变量影响了解释变量 $\ln(\frac{p}{1-p})$ 的方向与程度,其中 $(\frac{p}{1-p})$ 称为发生比(odds)。利用已有的样本指标数据对模型中的参数 β_i 进行估计,并对模型进行相关的统计检验。因此 Logistic 回归是处理这一类问题较好的方法。待得到一个较为稳定的、预测准确性较高的模型后,模型即可投入使用。

四、变量设计与模型实证

为便于对新的处理方法、模型设计和应用分析作深入探讨,本文仅对借款人特征维度和贷款特征维度下的变量进行分析。主要是运用非参数随机森林法,以较高的准确率完成对复杂变量和多类型、大容量的数据进行的筛选和分类与预测,然后利用 Logistic 回归模型对经过 RF 所遴选出来的特征变量进行建模,分析各因素对个人住房按揭贷款违约风险的影响方向和影响程度。下面是系统建模与实证分析过程。

(一) 样本数据与变量设计

本文使用的数据是中国某地区 2008 年的个人住房按揭贷款数据,该数据集共包含 25060 条贷款记录,剔除数据不完整的记录 1563 条和 571 条重复记录之后,获得有效样本 22989 个,其中违约样本共 4670 个,非违约样本 18294 个。同时为降低无关和冗余数据对后续分析造成的不必要负担和导致的结果偏差,在正式分析之前,需要对原始数据集进行三个步骤的预处理:第一步数据清理是清除与建模主题无关的数据或无使用价值的数据;第二步数据整合是根据分析的需要把来自多个数据源的数据整合成一个或几个指标变量;第三步是变量选择与量化,并且以平均值替代连续变量的缺失值,以众数替代分类变量的缺失值。为了删除异常值,在变量量化时应尽可能保持其原来的数据特征。本文研究所选变量及量化方法用表 2 进行说明。

其中,因变量为贷款是否违约,用虚拟变量表示,违约贷款赋值为 0,正常贷款赋值为 1。

(二)建模与实证分析

个人抵押类贷款违约风险的实证分析,主要是通过对借款人自身特征与贷款特征等变量进行统计分析,寻找、测量和评价影响违约风险的因素,分析各变量对违约风险的影响方向与影响程度。因此,首先应用非参数随机森林法进行异常点的识别与删除,特征变量的选择。我们运用 R-plus 软件中 randomForest 程序包中的 Outlier 程序,计算出各个样本点与数据集的异常点度量的阈值,对异

常样本进行删除(删除异常样本 98 个,占全体样本总数的 1.96%),删除后数据集剩下 4912 个样本。然后,利用 ntree 和 mtry 进行参数调整,我们把 ntree 设为 1000,即森林中共有 1000 棵树。根据随机森林法的要求 $mtry \ll M$,其中 M 为数据集中特征的总个数。实验中取其缺省值, $mtry = \text{sqr}(M)$,因为数据集中共有 17 个特征,故令 mtry 为 4,即在每棵树的分裂节点处从所有属性中随机抽取 4 个属性作为候选特征。采用 Rplus 程序包 varSelRF 进行计算,就可得到分类效果最优且包含特征个数相对较少的特征子集。通过分析,从 17 个特征变量中剔除泛化性影响后,遴选出 8 个特征变量,而对个人住房抵押贷款违约风险影响最为显著的是每期还款额 (AmtPT) 和还款能力 (Capacity)。其中还包括一些定性因素,如性别 (Gender)、地区 (District)、教育背景 (EduBg) 和职位级别 (Rank)。此前,这 4 个变量中性别是采取定性方法对男性女性分别赋值,而地区、教育背景和职位级别则采用定序方法赋值。用随机森林算法分析时,这样赋值对结果无影响。但在 Logistic 模型进行回归时,要求引入的虚拟变量应尽量少。因为地区变量有 7 个属性类型,所以只要引入 D1—D6 个地区虚拟变量。

表 2 个人住房按揭贷款违约风险分析中各个变量及其量化方法

变量名称	英文名	变量解释及量化方法说明
编号	Number	贷款记录的编号
贷款特征维度:		
贷款期限	CreditLength	用贷款的具体月数表示
贷款额	Volume	用贷款的具体数额表示
审批层次	AppLevel	1:董事会;2:总经理;3:销售总处;4:授信管理部;5:作业中心
适用利率	InterestRate	具体利率表示
可否提前清偿	PrePay	0:不可以;1:可以
每期还款金额	AmtPT	具体金额表示
有无担保品	Gage	0:无担保品;1:有担保品
有无担保人	Garantor	0:无保证人;1:有保证人
借款人特征维度:		
性别	Gender	0:女 1:男
年龄	Age	申请贷款时借款人的年龄,实际年龄表示
婚姻状态	MaritalStatus	1. 未婚;2. 已婚;3. 离婚、丧偶
学历	EduBg	1. 博士;2. 硕士;3. 大专;4. 高中;5 初中及以下
地区	District	1.A 区 2.B 区 3.C 区 4.D 区 5.E 区 6.F 区 7. G 区
职位级别	Rank	贷款人的职位等级,01—自营;02—高级主管;03—一般主管;04—职员;05—专业技术人员;06—公务员;07—教师;08—军人;09—其它
行业	Industry	1. 农林牧渔;2. 矿业、土石业;3. 制造业;4. 水电煤气;5. 营造业;6. 商业;7. 运输、仓储;8. 通讯;9. 金融保险、工商服务;10. 社会团体、个人服务; 11. 军公教、公营事业;12. 学生;13. 餐旅;14. 新闻广告; 15. 卫生保健;16. 娱乐、职业运动;17. 家馆;18. 资讯;19. 其他
企业规模	CorpScale	贷款人所在单位的规模,1:大企业;2:中小企业;3:非营利事业
还贷能力	Capacity	还贷能力=年收入/(12* 每期还款额)
因变量:		
违约登记	Default	0:未违约;1:违约

利用 Logistic 回归模型进一步分析 8 个变量中各因素对个人住房按揭贷款违约风险的影响方向和影响程度,需要对 8 个变量进入分析的顺序进行选择,选择的常用方法有:(1)向前逐步选择法(Forward Stepwise), (2) 向后逐步选择法 (Backward Stepwise), (3) 混合逐步选择法 (Combined Stepwise),以揭示和发现各因素的影响大小与方向。我们逐一采用各方法并利用 SAS 软件进行估计、检验和比较,采用极大似然法对 Logistic 模型的回归系数进行 Wald 统计量检验的结果(见表 3)。

表 3 个人住房按揭贷款违约风险影响因素的 Logistic 模型估计和检验

	B	S.E.	Wald	df	Sig.	Exp(B)
Volume**	1.987	1.350	216.548	1	0.000	4.257E8
InterestRate**	-0.650	0.023	8.142	1	0.004	0.937
AmtPT**	-0.348	0.054	41.251	1	0.000	0.706
Gender**	0.706	0.156	20.567	1	0.000	2.026
EduBg**	-0.653	0.123	28.353	1	0.000	1.921
D1	0.308	0.235	1.725	1	0.189	1.361
D2	-0.379	0.308	1.509	1	0.219	0.685
D3**	0.982	0.313	9.816	1	0.002	2.669
D4	0.790	0.420	3.539	1	0.060	2.204
D5	0.020	0.306	0.004	1	0.948	1.020
D6**	2.528	0.503	25.233	1	0.000	12.524
Rank	-0.033	0.032	1.067	1	0.302	0.967
Capacity*	3.282	1.302	6.349	1	0.012	0.038
常量**	-4.608	0.697	43.744	1	0.000	0.010

注：* 表示在 95% 的置信水平下显著，** 表示在 99% 的置信水平下显著。

从表 3 模型估计和检验的结果可以看出：变量贷款总额 (Volume)、贷款利率 (InterestRate)、每期还款额 (AmtPT)、性别 (Gender)、教育背景 (EduBg)、还款能力 (Capacity) 以及地区虚拟变量中的 D3 和 D6 都非常显著。模型的函数表达形式可以表示为：

$$Z = -4.608 + 1.987 \text{Volume} - 0.650 \text{InterestRate} - 0.348 \text{AmtPT} + 0.706 \text{Gender} \\ - 0.653 \text{EduBg} + 0.982 \text{D3} + 2.528 \text{D6} + 3.282 \text{Capacity}$$

当 $P_i = \frac{1}{1 + e^{z_i}} > 0.5$ 时,这笔个人住房按揭贷款很可能会发生违约;反之,即为正常贷款。

模型投入使用,把一个新的借款人的相关指标数据输入模型,就对其的违约发生比(或违约概率)进行分析与预测。在实际使用时,通常将违约发生比或违约概率通过某种线性变换转换成分数,银行就可以根据申请人的信用得分情况决定是否发放贷款及发放的额度。

五、结果分析

实证分析结果表明,在所选择的 17 个分析变量中,对个人住房按揭贷款违约影响显著的 8 个因素,其具体的分析结果为:(1)贷款总额 (Volume) 的系数估计值为 1.987,说明贷款总额越大,个人住房按揭贷款违约发生的概率越小,即贷款总额每增加 1% 便会降低 1.987 个百分点的违约概率。产生这种结果的原因是贷款额较大的基本上都是大型高档住宅,或者是投资性住房,其贷款人一般经济状况较好,抗风险能力强,并且这类贷款人整体的综合素质较强,对信用记录及声誉较为在意,违约的意愿不强烈。(2)贷款利率 (InterestRate) 的系数估计值为 -0.650,说明贷款利率越高,个人住房按揭贷款违约发生额概率越大,贷款利率每上升 1%,将会导致违约概率上升 0.65%,这种关系在 99% 的置信水平上显著。(3)每期还款额 (AmtPT) 的系数估计值为 -0.348,表明每期还款额越高,贷款发生违约的概率越大。每期还款额每上升 1%,将会增加 0.348 个百分点的违约概率,且这

种关系在 99% 的置信水平上显著。(4)性别(Gender)的系数估计值为 0.706,表明性别对贷款违约有所影响,其中男性的违约概率较小。(5)教育背景(EduBg)的系数估计值为-0.653,表明教育程度越高,贷款发生违约的概率越小(数据处理时,令博士取值为 1,初中及以下取 5,学历越低取值越大),这种关系在 99% 的置信水平上显著,说明教育程度较高的人群,整体而言,其经济条件、综合素质、抗风险能力及对信用记录的珍惜程度都较高,发生违约的意愿比教育程度低的人群较弱。(6)还款能力(Capacity)的系数估计值为 3.282,表明还款能力越强,个人住房按揭贷款发生违约的概率越小。还款能力(月收入/月还款额)强 1% 的借款人,其发生违约的概率小 3.282 个百分点,且这种关系在 99% 的置信水平上显著。(7)反映地区(District)影响因素的六个虚拟变量其系数值差异很大,其中 D6 最大,为 2.528。D2 最小,为-0.379;而且除了 D2 以外其他各变量均为正值。表明来自 F 地区的借款人发生违约的概率最小。可见来自不同地区的借款人发生违约的概率有较大差别,地区因素对个人住房按揭贷款违约风险有所影响。但这 6 个变量中,除了 D3 和 D6 在 99% 的置信水平上显著之外,其余几个不显著。

从表 4 可以看出, 表 4 用 Logistic 模型估计个人住房按揭贷款违约
Logistic 回归模型分析有
较好的分类性能,模型对
选定案例的分类和预测结果
是:训练集当中非违约
记录的分类准确率为
72.1%,而违约记录的分
类准确率高达 91.2%,模
型整体的分类准确率达到

已观测		已预测					
		选定案例			未选定的案例		
		Default		百分比 校正	Default		百分比 校正
		0	1		0	1	
Default	0	483	187	72.1	197	71	73.5
	1	59	611	91.2	33	234	87.6
总计百分比				81.6			80.6

81.6%;模型对未选定案例的分类和预测结果是:对违约贷款预测的准确率达到 87.6%,对正常贷款预测准确率达到 73.5%,模型整体的预测准确率为 80.6%。说明模型整体的预测效果很好,进一步支持了以上结论的可靠性。需要指出的是:利用非参数随机森林法中的分类决策与回归分析功能也可以完成 Logistic 模型的一些分析和预测功能,从而用一种模型方法就可完成分析任务,并且还可以与其它方法(比如支持向量机 SVM 和人工神经网络 ANNs)的效率与效果进行对比,有关的研究将另文探析。

总之,随着我国房地产金融市场的扩大,银行等金融机构对个人住房抵押贷款的业务必将获得更大发展。各商业银行在积极应对机遇与竞争的同时,如何有效地防控贷款违约风险,将其损失控制在合理的范围内,已成为需要直面和重视的问题。本文提出的研究架构、实证分析过程的模型与方法,以及研究的结论不仅为商业银行和其它金融机构为分析个人住房按揭贷款影响因素提供更加全面系统的四大维度整体分析框架,也为利用非参数随机森林法和 Logistic 模型解决个人贷款违约风险的复杂数据集中异常点检测和特征变量选取,以及对违约风险的影响方向和影响程度分析,提供了新的分析方法与实际操作和指导。而且这些方法具有可移植性和适应性,同样也可用于汽车贷款和信用卡等消费金融业务的分析。这对于在我国消费金融的研究与应用中,如何防范与管理风险,如何进行深入的定量分析,提供了理论分析与实际应用的启示与借鉴作用。

参考文献

刘萍, 2002,《个人住房按揭贷款风险探析》,《金融研究》第 8 期。
汪利娜, 2001,《北美、西欧的信用风险防范》,《中外房地产导报》第 11 期。
施海松, 2007,《中美住房按揭贷款市场比较: 基于同质性与差异性的分析》,《上海金融》第 10 期。

李馨苹, 1997,《住宅抵押贷款违约风险之实证分析》, 国立台湾工业技术学院博士学位论文。

Jung Allen, 1962, “Terms of Conventional Mortgage Loans on Existing Houses”, *Journal of Finance*, Vol. 17, 43—244.

George W. Gau, 1978, “A Taxonomic Model for the Risk-Rating of Residential Mortgages”, *Journal of Business*, Vol. 51 No. 4, 687—706.

Foster, Chester and Robert Van Order, 1984, “An Option-Based Model of Mortgage Default”, *Housing Finance Review*, Vol. 3, No. 4, 351—372.

Green, Jerry and Shoven John B., 1986, “The Effects of Interest Rates on Mortgage Prepayments”, *Journal of Money, Credit, and Banking*, Vol. 1, No. 1, 41—59.

Von Furstenberg, George M. and R. Jeffery Green, 1989, “Mortgage Delinquencies: A Cohort Analysis”, *Journal of Finance*, Vol. 29, 1545—1548.

Lawrence, Edward C. and Nasser Arcade, 2002, “A Multinomial Logit Analysis of Problem Loan Resolution Choices in Banking”, *Journal of Money, Credit and Banking*, Vol. 27, No. 1, 202—216.

Canner, Glenn B. Stuart A. Gabriel and J. Michael Woolley, 2001, “Default Risk and Mortgage Lending: A Study of the FHA and Conventional Loan Markets”, *Southern Economic Journal*, 249—262.

Quigley, John M. and Robert Van Order, 1979, “Defaults on Mortgage Obligations and Capital Requirements for U.S. Savings Institutions—a Policy Perspective”, *Journal of Public Economics*, Vol. 44, 353—369.

Epperson, James F., James B. Kau, Donald McKenna and Walter J. Muller, 1984, “Pricing Default Risk in Mortgage”, *AREUEA Journal*, Vol. 13, No. 2, 261—272.

Campbell, Tim S. and J. Kimball Dietrich, 1983, “The Determinants of Default on Insured Conventional Residential Mortgage Loans”, *Journal of Finance*, Vol. 38, and No. 5, 1569—1581.

Deng Yongheng, Quigley John M, Van order Robert, Mae Freddie, 1996, “Mortgage Default and Lowdown Payment Loans: The Cost of Public Subsidy”, *Regional Science and Urban Economics Amsterdam*, vol. 26, 63.

Quercia, Roberto G and Michael A. Stegman, 2006, “Residential Mortgage Default: A Review of the Literature”, *Journal of Housing Research*, Vol. 3, 341—379.

Breiman L., 2001, “Random Forests”, *Machine Learning*, 45(1): 5—32.

Pall Oskar Gislason, Jon Atli Benediktsson, and Johannes R. Sveinsson, 2004, “Random Forest Classification of Multisource Remote Sensing and Geographic Data”, *IEEE*, 2, 1049—1052.

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.

A Study of Default Risks in Personal Housing Mortgage Loan of China’s Consumer Finance

Qian Zhengming, Li Haiibo and Yu Yanping
(School of Economics, Xiamen University)

Abstract:To meet a rapid growth in personal household consumer finance (PHML), we examine the major factors that influence default risks in PHML. We build a four-dimensional analytical framework, measure abnormal points and select characteristic values using nonparametric random forest algorithm, and construct logistic regression model on default risks in PHML. The results indicate that the eight out of 17 variables are significant. The findings offer new analytical methods and guide financial institutions how to guard against the risks in conducting PHML businesses.

Key Words: Consumer Finance; Personal Housing Mortgage Loan; Default Risk; Nonparametric Random Forest Algorithm; Logistic Regression Model

JEL Classification:F29, C45

(责任编辑:晓 喻)(校对:梅 子)