个人信用风险评估方法的研究[①]

--基于 lending club 数据

四川大学　　刘晶、谭峰、柴容倩

摘　　要

信用风险的评价方法不断推陈出新，管理技术正日臻完善，许多定量技术、支持工具和软件已付诸商业应用。由于我国商业银行和金融市场尚处转轨和新兴发展阶段，缺乏对个人信用风险评估的基础理论的深究，个人征信系统的发展十分滞后和缓慢，严重阻碍了中国社会经济的健康持续发展。因此，个人信用风险的评估方法的探讨成为了热点话题。

鉴此，　本文以 lending club 公司 2014 年 1 月 1 日到 2015 年 3 月 31 日的借款人数据为源数据，首先根据 spearman 秩相关性检验和主成分分析法从众多变量中筛选出解释能力比较强的变量，然后依据筛选出的变量建立随机森林模型、判别分析模型和 logistic 回归模型，通过这三类模型判断准确性的比较，力图构建更为有效的个人信用风险评估方法，从而为该平台及出借人决策提供科学依据。

实证结论如下：

第一，主成分筛选变量后可以减少模型变量个数起到降维的作用，但是在对随机森林模型来说，由于其并非线性模型，而且筛选变量造成了信息的损失，经过线性组合的特征并不一定能给模型带来更好的效果。

第二，通过训练集 10000 个数据样本测试出三个模型的预测能力。其中，随机森林模型和 logistic 模型的预测准确度比较高，分别为：81.87%和 70.89%。而判别分析的结果相对较低为：65.92%。

第三，通过测试集 2325 个样本检验三个模型的预测能力。总体来看随机森林模型和 logistic 模型对测试集的预测准确度相近，远远高于判别分析的结果。

最后，通过预测的结果，和对模型分析之后得出三个模型的预测准确率分别为：随机森林模型的判断准确率最高，而判别分析模型的准确率最低。 即：随机森林的准确率>logistic 模型的准确率>判别分析模型的准确率。其中，随机森林模型判别的准确率又和其特征的选取有一定的关系。

综上所述，本文在建立 P2P 信贷风险评价模型相比较中认为，随机森林模型

[①] 注:该论文获得由中国统计教育学会举办的“ 2015 年（第四届）全国大学生统计建模大赛 ”大数据统计建模类研究生组二等奖。

会更加准确和可靠。

关键词：信用风险　　随机森林　　判别分析　　logistic 回归

一．问题的提出及研究概述

（一）问题的提出

在市场经济的社会，产权，法制，信用和风险是市场经济的四个关键。信用不论对于人，社会，还是国家都尤为重要，是现代社会的基石[②]。然而，我国 1999 年建立征信体系至今，我国一直缺乏对个人信用风险评估的基础理论的深究，个人征信系统的发展十分滞后和缓慢，严重阻碍了我国信用借款和中国社会经济的健康持续发展。我国信用风险评估的问题主要体现在：第一，信息不对称。个人信用活动的随机性和非连续性不利于银行和企业对消费者作出正确的判断；第二，目前我国个人信用评估标准不统一，缺少一整套科学，严密，可推广使用的个人信用评分模型；第三，目前我国相关的信用评估机制体制不健全，缺少必要的法律保障。

基于以上情况，本文立足征信的基本理论和实践，运用数据挖掘和计量经济学的相关方法，以国外最大 p2p 公司的借款人数据为依据，借鉴国外的经验建设，对我国个人信用风险的评估方法进行研究和分析。

（二）国外研究概述

1.国外机构对个人信用评估的方法研究

（1）美国 FICO 信用评分模型

FICO 信用模型是一种被广泛使用的信用评估模型。FICO 评分系统将信用分数设为 325 – 900 分之间，得分越高则借款人的违约风险越小。根据大量实践数据证明，当借款人的 FICO 得分大于 800 分时，违约率为 0.0773%， 700-800 分之间时，违约率为 0.81%，低于 600 分以下时，违约率为 12.5%。[③]

[②] 王富全．个人信用评估与声誉机制研究[M]．山东大学:王富全，2010.

[③]刘峙廷.我国 P2P 网络信贷风险评估研究.广西大学，2013.5：44

下表 1-1 为 FICO 个人信用评分法评分指标。由评估项目，评估占比和评估指标组成。

表 1-1： FICO 个人信用评分法评分指标

评估项目	评分占比	评估指标
偿还历史	35%	各信用账户还款记录
		公开记录
		逾期偿还情况
信用账户数	30%	需要偿还的信用账户数
		信用账户余额
		总信用额度使用率
		账户偿还率
信用历史	15%	使用信用的年限
新开信用账户	10%	新开信用账户数
		新开信用账户账龄
		当前的信用申请数量
		最近的信用状况
正使用的信用类型	10%	正使用的信用账户类型
		每种类型账户数

（2）德国 IPC 微贷技术评估法

德国国际项目咨询公司（IPC 公司）多年来在小微借款领域形成了一套独特的技术，称作：IPC 微贷技术[④]。全球有十多个国家引进了该技术，运作的平均不良借款率小于 3%，在技术输出上取得了良好的效果。IPC 微贷技术在评价客户信用风险方面有其独到之处，将评价标准分为了软信息和硬信息两部分（祥见表 1-2）。

表 1-2：德国 IPC 微贷技术评估法指标体系

评估类型	评估项目	评估指标
软信息	基本信息	年龄

[④]龙新庭,王晓华.德国 IPC 公司微贷技术在我国经济欠发达地区的运用.区域金融研究,2013（10）:69

		教育水平
		他人对客户的评价
		婚姻状况
		性格特征
		是否有不良嗜好和犯罪记录
		是否为本地人
		是否有其他收入或支出
		社会地位
	经营信息	经营经验
		经营记录
		借款用途
硬信息	财务信息	损益表
		资产负债表
		现金流量表

2. 国内典型个人信用风险评估方法

我国典型的个人信用风险评估方法大都来源于各商业银行，但我国各大商业银行的个人信用风险评估体系都相互独立，没有一套统一的评估体系，在某些具体的评估指标上有一些差异。本文整理了国内各大商业银行债务人信用风险评价指标如表 1-3 所示。图示中打勾表示该银行存在此类债务人信用风险评级指标。

表 1-3：我国商业银行债务人信用风险评级指标

		建设银行	交通银行	工商银行	民生银行	光大银行
基本信息	年龄	√	√	√	√	√
	性别	√	√	√		√
	健康	√				√
	婚否	√	√	√		√
	户口	√	√	√	√	√
	教育背景	√	√	√	√	√
	单位情况	√	√	√	√	√
	职务	√	√	√		√
	职称	√	√	√		√
	工作年限	√	√	√	√	√
经济信息	收入	√	√	√	√	√
	金融资产			√		
	其他资产			√		

	家庭平均收入	√	√	√	√	
	储蓄账户余额	√			√	
信用情况	是否银行职员	√	√	√		
	在本行账户	√	√	√		√
	是否有其他贷款	√	√			
	本行业务往来	√	√	√		√
	不良信用记录	√	√	√	√	√

资料来源：我国各商业银行网站信息整理（建设银行：http://www.ccb.com/；工商银行：http://www.icbc.com.cn/；交通银行：http://www.bankcomm.com/；光大银行：http://www.cebbank.com/；民生银行：http://www.cmbc.com.cn/）

二. 数据预处理

（一）基础数据的情况

本文采用 lending club 公司 2014 年 1 月到 2015 年 3 月 31 日的借款人数据为源数据。该数据包含了在 lending club 借款的所有人的基本个人信息以及产品信息，包括借款金额，借款利率，工作年限，违约情况等。

总数据一共存在 20 多万条，本文选取 2014 年 1 月至 2015 年 3 月借贷状态已经完成的数据 30719 条，其他正在进行当中的记录不予考虑。

（二）研究方法选择

本文基于大数据分析，采用数据挖掘和计量经济学的方法，选用随机森林，判别分析法和 logistic 回归的方法分别进行研究，并比较三种模型的预测准确率及重要指标。

本文采用 spss 和 R 软件结合使用。其中，SPSS 使用于数据的描述性统计；R 软件使用于具体的建模过程，即：随机森林，判别分析和 logistic 回归。

（三） 定义变量

本文选取了借款状况为被解释变量，选取了借款金额，借款利率，借款期限等 20 个指标。见表 2-1

表 2-1： 变量的定义及解释

变量定义			
变量类型	变量	变量名称	备注
被解释变量	y	借款状况	0：违约；1：履约
解释变量	x1	借款金额（美元）	度量变量
	x2	借款期限（年）	度量变量
	x3	借款利率（%）	度量变量
	x4	分期付款（美元）	度量变量
	x5	工作年限（年）	度量变量
	x6	住房所有权（按揭）	0：租用和自有 1：按揭
	x7	住房所有权（租用）	0：按揭和自有 1：租用
	x8	年收入（美元）	度量变量
	x9	收入认证	0：未认证 1：认证
	x10	借款目的（购买固定资产）	0：偿还债务和其他 1：购买固定资产
	x11	借款目的（偿还债务）	0：购买固定资产和其他 1：偿还债务
	x12	收入负债比（%）	度量变量
	x13	逾期次数（次）	度量变量
	x14	公开信用账户（个）	度量变量
	x15	毁誉记录（次）	度量变量
	x16	循环额度利用率（%）	度量变量
	x17	信用账户（个）	度量变量
	x18	清单初始状态	0：f 1：w
	x19	月还收入比（%）	度量变量
	x20	建立信用年限（年）	度量变量

基于模型分析，本文设置被解释变量为“是否履约”，履约取“1”、违约取“0”；解释变量分别为借款金额、借款期限、借款原因、就业时长、住房所有权、月还

款收入比、负债收入比、身份认证等 20 个解释变量。根据表 2-1，名义指标对应的属性为“真”时取“1”，为“假”时取“0”。以“住房所有权”为例，当样本数据中“住房所有权”指标为按揭时，虚拟变量“住房所有权 1”取 1，而其他虚拟变量都取 0；“住房所有权”指标为租用时，虚拟变量“住房所有权 2”取 1，而其他虚拟变量都取 0:；“住房所有权”指标为自有时，虚拟变量“住房所有权 1”，“住房所有权 2”都取 0。

（四）样本处理

1.随机抽取样本。

本文从 lending club 公司 2014 年 1 月到 2015 年 3 月 31 日的借款人样本中，筛选出交易状态已经完成的 30719 条样本做为基础样本。然后从中随机随机筛选 10000 个交易完成的样本作为训练集，其中 7000 个样本为履约样本，3000 违约样本；另外，基于真实数据履约和违约的比例，随机筛选 2325 个样本作为测试集，其中 2000 个为履约样本，325 个为违约样本。

2.补充空白值和缺失值。

筛选出总共 12325 个样本，这些样本的数据完整性比较好。在选取的 20 个变量中，只存在工作年限指标和毁誉记录指标需要定义缺失值和填补空白值。本文采用的方法如下：

从属度量性指标的工作年限，小于一年的和大于十年的，分别赋值 0.5 和 15。

从属名义变量的毁誉记录，筛选有完整数据的样本替代出现空白值的样本。

三．数据的探索性分析

（一）变量基本情况分析

表 3-1 为本文选取变量的描述性统计。表中显示：年收入，借款目的（购买固定资产），逾期次数，公开账户个数，毁誉记录的峰度大于 3，呈现尖峰的特征。其次，偏度大于 0 为右偏，小于 0 为左偏，从表中统计得 5 个变量是左偏的，16 个变量是右偏的。

表 3-1　变量的描述性统计

变量的描述性统计							
	N	极小值	极大值	均值	标准差	偏度	峰度
借款状况	10000	0.00	1.0000	0.7000	0.4580	-0.8730	-1.2380
借款金额（美元）	10000	1000.00	35000.0000	13956.1800	8490.9320	0.8080	-0.0180
借款期限（年）	10000	36.00	60.0000	42.9400	10.8810	0.9310	-1.1340
借款利率	10000	0.06	0.2606	0.1473	0.0469	0.2960	-0.3410
分期付款（美元）	10000	30.40	1370.8000	424.3670	250.1780	1.0160	0.8610
工作年限（年）	10000	1.00	15.0000	7.8100	5.6620	0.2310	-1.5680
住房所有权(按揭)	10000	0.00	1.0000	0.4900	0.5000	0.0320	-1.9990
住房所有权(租用)	10000	0.00	1.0000	0.4100	0.4920	0.3620	-1.8690
年收入（美元）	10000	6000.00	4900000.0000	76257.5500	71904.1660	32.2900	2040.0870
收入认证	10000	0.00	1.0000	0.6700	0.4710	-0.7130	-1.4920
借款目的（购买固定资产）	10000	0.00	1.0000	0.0200	0.1310	7.3380	51.8630
借款目的（偿还债务）	10000	0.00	1.0000	0.8000	0.4010	-1.4850	0.2070
收入负债比（%）	10000	0.00	39.9900	17.5088	8.1044	0.1960	-0.5390
逾期次数（次）	10000	0.00	15.0000	0.3300	0.8860	5.0200	41.0820
公开信用账户(个)	10000	1.00	58.0000	11.5300	5.2020	1.3480	3.8930
毁誉记录（次）	10000	0.00	9.0000	0.2200	0.5350	3.6400	22.9490
循环额度利用率	10000	0.00	1.0910	0.5238	0.2447	-0.0920	-0.7850
信用账户（个）	10000	2.00	118.0000	26.5700	12.4420	0.9250	1.3750
清单初始状态	10000	0.00	1.0000	0.5100	0.5000	-0.0550	-1.9970
月还收入比	10000	0.00	0.2260	0.0770	0.0411	0.5130	-0.3910
建立信用年限(年)	10000	1126.00	21823.0000	5659.4400	2479.1320	1.0730	1.7620

（二）描述性统计分析

1.从总体样本中获取的 30719 条数据中，可以看出违约的条数为：4306 条，约占：14%；履约的条数为：26413 条，约占 86%。如图 3-1 所示

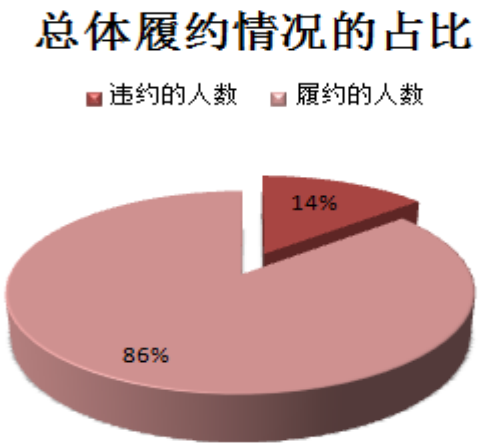

图 3-1 总体履约情况的占比

2．从贷款期限来看： 贷款 36 个月的数量为 22476 条；贷款期限为 60 个月的数量为 8244 条。说明的大部分人选择了时间相对较短一点贷款。其中贷款为 36 个月的违约条数为 2616 条， 占比为： 11.639% ；贷款为 60 个月的违约条数为：1690 条， 占比为 20.5%。

可以初步看出：相对来说贷款的期限越长，其违约的概率会越大。同时也可以想象得到人们为什么更加偏好 3 年的贷款。具体情况见图 3-2。

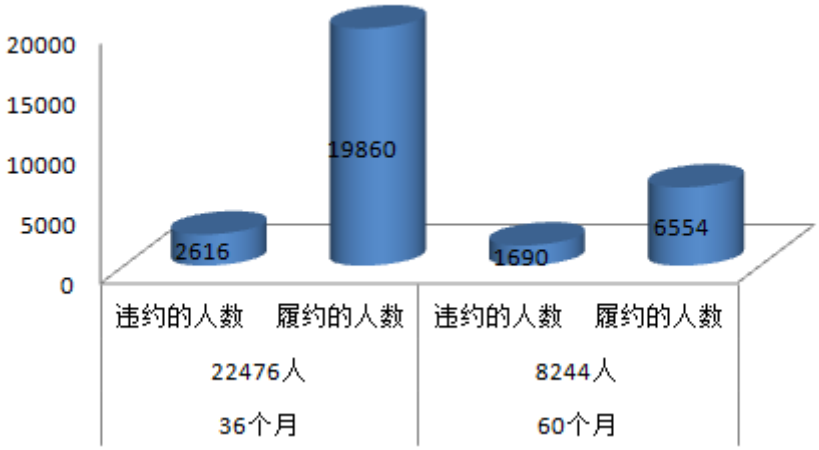

图 3-2 不同期限下的履约人数情况比较

四．解释变量筛选

（一） spearman 相关性分析

1．解释变量与被解释变量的相关性

为了研究哪些变量会影响到借款人是否履约，由于本文选取变量比较多，如果均加以考虑，必然会带进许多糅杂信息，从而导致数据冗余。于是本文首先通过 spearman 相关性检验来初步探究解解释变量和被解释变量之间的相关关系。

本文最终决定根据解释变量与被解释变量之间的相关关系的强弱来进行变量的选取，即根据检验结果的相关系数由大到小选取相关性较强的变量。

其结果如下表 4-1 所示:

表 4-1　spearman 相关系数表

	y		y
x1	-.081**	x11	-0.012
x2	-.140**	x12	-.106**
x3	-.307**	x13	-.042**
x4	-.086**	x14	.039**
x5	.056**	x15	.023*
x6	.077**	x16	-.148**
x7	-.073**	x17	.095**
x8	.127**	x18	.065**
x9	-.131**	x19	-.201**
x10	0.013	x20	.059**

根据相关性检验结果, 除了 x10,x11 没有通过显著性检验其它变量均通过显著性检验。相关关系比较强的变量依次分别为 x3,x19,x16,x2,x9,x8,x12 。综上本文主要选取指标变量为：x3,x19,x16,x2,x9,x8,x12。

2．解释变量之间的相关性

如果解释变量之间的相关性较高，会影响到本文后面的建模效果，如出现多重共线性或者伪回归因等现象，因此在此之前本文对变量之间的相关做了检验。其结果如下表 4-2 所示。其结果显示变量之间的相关性不高。

表 4-2　变量之间的相关性

	y	x2	x3	x8	x9	x12	x16	x19
y	1.000	-.140**	-.307**	.127**	-.131**	-.106**	-.148**	-.201**
x2	-.140**	1.000	.462**	.112**	.274**	.091**	.096**	.190**
x3	-.307**	.462**	1.000	-.155**	.281**	.177**	.254**	.294**
x8	.127**	.112**	-.155**	1.000	.120**	-.252**	.066**	-.377**
x9	-.131**	.274**	.281**	.120**	1.000	.074**	.066**	.258**
x12	-.106**	.091**	.177**	-.252**	.074**	1.000	.192**	.254**
x16	-.148**	.096**	.254**	.066**	.066**	.192**	1.000	.101**
x19	-.201**	.190**	.294**	-.377**	.258**	.254**	.101**	1.000

3. 多重共线性分析

为确保模型结果的准确性，还需对解释变量是否存在多重共线性进行检验，结果见表 4-3 ，可以看出 6 个变量的膨胀因子 VIF 介于 0-10 之间，容忍度大于 0.1，可以判断 6 个变量之间不存在严重的多重共线性，不会对回归模型的参数估计结果的精确性产生较大影响。

表 4-3 变量多重共线性检验结果

模型	非标准化系数		t	Sig.	共线性统计量	
	B	标准 误差			容差	VIF
(常量)	1.232	.021	58.930	.000		
x2	.000	.000	.389	.697	.759	1.318
x3	-2.327	.112	-20.793	.000	.680	1.471
x9	-.033	.010	-3.343	.001	.867	1.153
x12	-.001	.001	- 1.21	.263	.906	1.104
x16	-.134	.019	-7.215	.000	.907	1.102
x19	-1.209	.114	-10.568	.000	.850	1.177

（二） 主成分法筛选变量

1．主成分分析的基本思想

主成分分析师通过最原变量进行线性组合来达到降维的目的，从而通过较少的综合指标来反映数据的信息。如在本文中，最初本文选取 20 变量进行观察和分析。但是变量过多会增加分析问题的复杂程度，而变量之间可能存在相关性带来信息的重叠。于是我们试图寻找到一系列的线性组合，将他们作为综合指标来代替原来的变量，来达到降维的目的，又不会损失太多的信息。实现主成分分析，我们一般采用下面的步骤。

（1） 计算原始数据的协方差矩阵Σ。

（2） 求出Σ的特征值$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$，对应的特征向量为$a_1, a_2, \cdots, a_p$，

则$Y_i = a_i' X$为第 i 个主成分，λ_i为第 i 个主成分的方差，$i = 1, 2, \cdots p$。

（3） 计算每个主成分的方差贡献率。

（4） 计算观测样本在m个主成分上的得分。

2. 主成分筛选变量过程

本文首先根据相关性分析结果，按相关性大小顺序，从 20 个变量中筛选出相关性较强的 10 个变量加以分析。分别为：x3,x19,x16,x2,x9,x8,x12,x17,x4,x1。然后通过降维，寻找到维数更低的综合变量代替原有变量。运行结果如下表 4-4 所示。

表 4-4　主成分分析重要结果

	comp.1	comp.2	comp.3	comp.4	comp.5	comp.6	comp.7	comp.8	comp.9	comp.10
主成分的标准差	1.7324	1.2848	1.0856	1.067	0.9932	0.8609	0.7345	0.7125	0.4988	0.088
方差贡献率	0.3001	0.165	0.1179	0.1138	0.0986	0.0741	0.05395	0.0508	0.0249	0.00078
累计贡献率	0.3001	0.4651	0.583	0.6969	0.7955	0.8696	0.9236	0.9743	0.9991	1

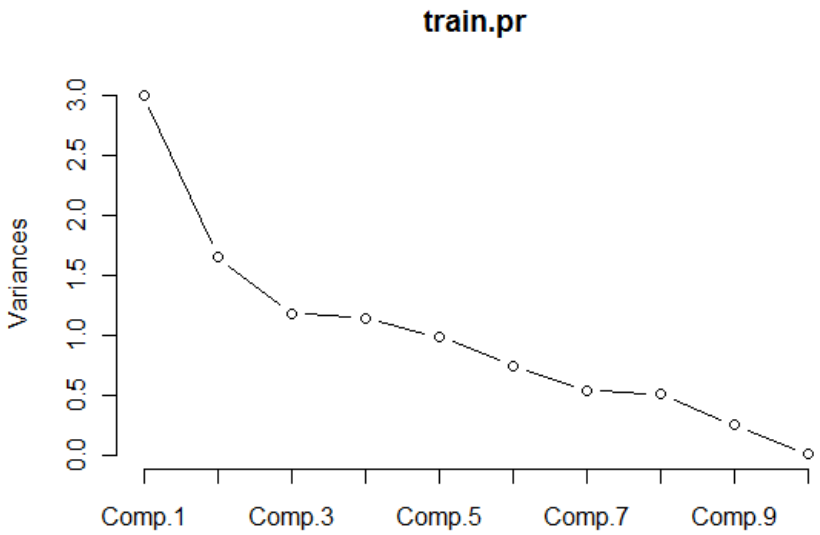

图 4-1：各主成分方差的贡献率

从累计贡献率结果图 4-1 可以看出：前 6 个主成分的累计贡献率已经达到 87%，后面主成分的贡献率均很低。所以本文选取 6 个主成分作为新的综合变量，重新命名为 Z1 到 Z6。套用前面分析的所用模型，以探究通过主成分方法筛选的变量是否可以使模型更加优化。同时达到降维的目的。

五．个人信用风险评估方法的建模

（一）各类模型建模

1.随机森林模型

（1）基本思想

随机森林，指的是利用多棵树对样本进行训练并预测的一种分类器。简单来说，随机森林就是由多棵 CART（Classification And Regression Tree）构成的。对于每棵树，它们使用的训练集是从总的训练集中有放回采样出来的，这意味着，总的训练集中的有些样本可能多次出现在一棵树的训练集中，也可能从未出现在一棵树的训练集中。在训练每棵树的节点时，使用的特征是从所有特征中按照一定比例随机地无放回的抽取的。

随机森林的训练过程可以总结如下：

首先,从给定的训练集通过多次随机的可重复的采样得到多个数据集。然后,对每个数据集构造一棵决策树,构造是通过迭代的将数据点分到左右两个子集中实现的,这个分割过程是一个搜索分割函数的参数空间以寻求最大信息增量意义下最佳参数的过程。然后，在每个叶节点处通过统计训练集中达到此叶节点的分类标签的直方图经验的估计此叶节点上的类分布。这样的迭代训练过程一直执行到用户设定的最大树深度（树的棵树的设定）或者直到不能通过继续分割获取更大的信息增益为止。

（2） 建模过程

① 基于 spearman 秩检验筛选的变量

本文随机选取10000条信贷完成记录，其中违约3000条记录，履约7000条记录作为训练样本集,以各个变量作为训练的特征。通过调用 R 软件随机森林软件包,实现建模过程。

a 特征的选取：

随机森林模型在训练的时候，特征的选取决定了模型的好坏，虽然模型对特征的依赖程度不像线性模型对解释变量的条件要求那么苛刻,但是选择适当的变量个数，不仅可以提高准确度还可以使模型复杂的计算过程，提高模型的效率。当初次筛选变量为20个的时候，见表5-1履约的被判成违约的错误率为7.37%，而违约的被判成履约的错误率比较高为65.87%，总的错误达到24.92%。 考虑到变量过多，其中相关性较小的变量可能会影响到模型的训练效果，造成预测的准确度下降。本文考虑逐步剔除一些不相关的变量（根据随机森林的特征重要性大小从小到大依次筛选），以达到最优的预测效果。当变量个数为20个的时候，通过随机森林训练结果可以查看其变量的重要程度。如下图5-1所示：

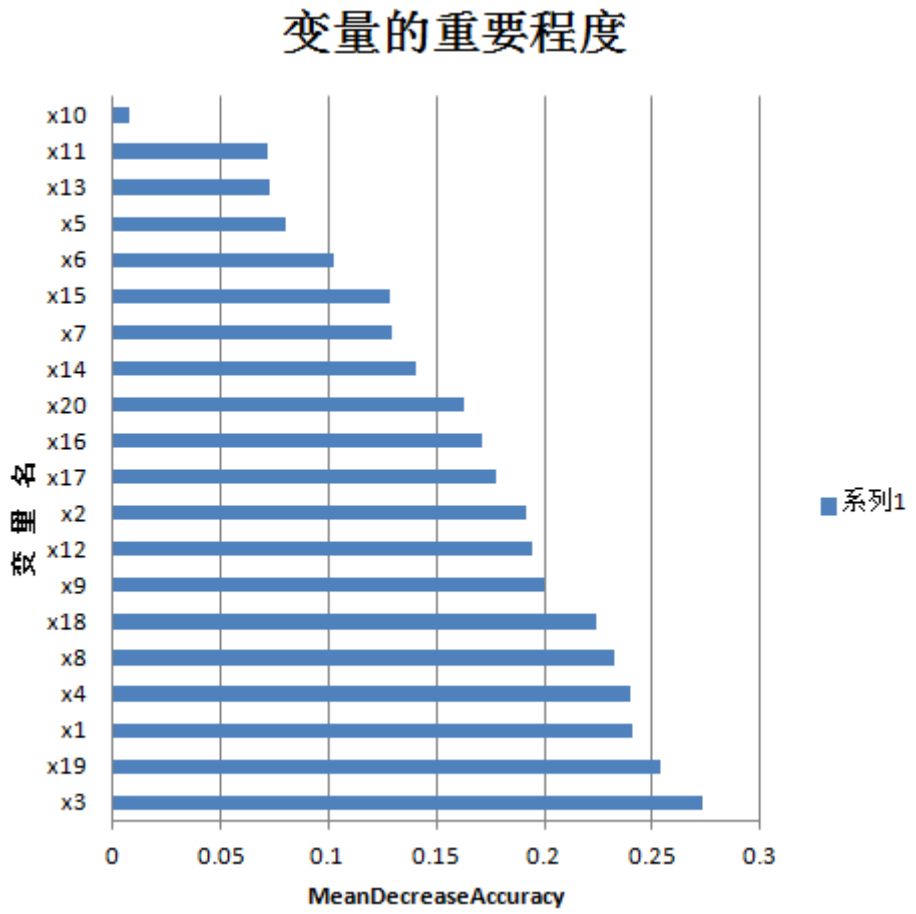

图5-1 各变量的重要程度

从上图5-1可以看出，对借款人是否违约的分类结果贡献最大的几个特征分别是：x3:借款利率；x19:月还收入比；x1：借款金额；x4：分期付款额；x8:年收入；x18：清单的初始状态；x9：收入认证；x12：收入负债比。另一方面，对比模型中所判断出来的特征重要程度与前文中变量相关系数的判断差异不大。进一步说明上述变量是建模过程中应该选取的重要特征变量。

依据重要程度的高低值逐步剔除表格后的准确率（由于多数变量与被解释变量相关性较小，故直接筛选到8个变量）。 其结果如下图5-2和表5-1，显示 ：

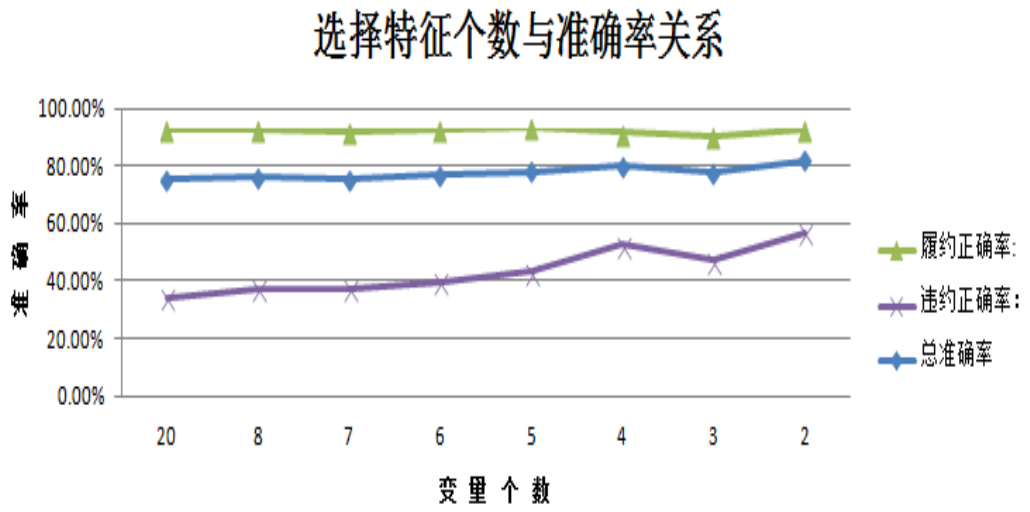

图5-2 选择特征个数与准确率的关系

表5-1 逐步剔除变量后的准确率

变量个数	20	8	7	6	5	4	3	2
履约正确率	92.63%	92.30%	91.76%	92.62%	93.07%	91.50%	90.07%	92.63%
违约正确率	34.14%	37.60%	37.50%	39.73%	43.17%	52.73%	47.17%	56.77%
总准确率	75.08%	75.89%	75.48%	76.75%	78.10%	79.87%	77.64%	81.87%

由上图 5-2 和上表 5-1 可得当变量为四个，五个和两个的时候，准确率较高。但是，变量为两个又不足以判断一个人信用状况，因为变量过少会导致信息的获取不充分，导致真实情况判别结果的失真，因此在实际应用中应该考虑选取较多变量为好。

但是本文从下表 5-2 可以看出，当只存在最显著的两个变量建模的时候，犯两类错的概率都有所下降即准确率较明显提升。

因此在此种情况下，本文最终选取为此模型的最优。

b 树的颗树的选择：

树的颗树的选择直接影响到随机森林训练结果的准确度。树选择过少，则预测的结果会不理想。选择的树过多，在提高准确率上没有显著效果，反而会

直接影响计算的速度。本文选取了 300 棵树以探究树的颗数的选取对判断准确性的影响，其结果如下图 5-3 所示：

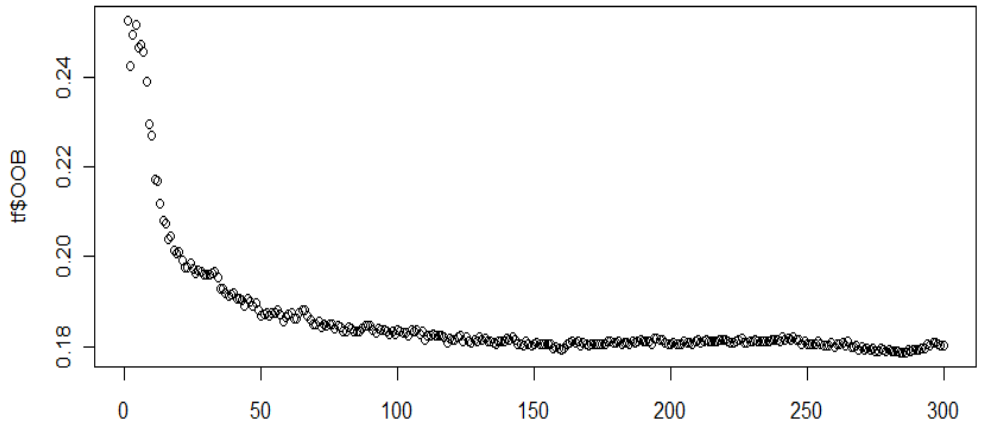

图 5-3 数的棵树与准确率的关系

如上图 5-3 所示 :图形横坐标表示树的棵树, 纵坐标表示模型的判断错误率。从上图中可以看出，当树的棵树选择小于 100 的时候，随着树选择的增多，其模型判断的错误从大约 26%下降到 18%左右。而继续随着棵树的增加，模型的错误率下降的速率变得缓慢，基本趋于平稳。由此可以看出模型的判断准确率大约为 82%。

随机森林模型预测最终结果：

根据对特征的选取和模型参数的设定，本文最终选取 x3 和 x19 变量，参数 ntree 选择为 500 棵。通过 R 软件调用随机森林程序包，运用训练集的数据得到随机森林模型的训练结果如下表所示：

表 5-2　预测准确率

原始结果		预测结果		预测准确率
		履约	违约	
履约	7000	6484	516	92.63%
违约	3000	1297	1703	56.77%
总体准确率：81.87%				

从结果可以看出，模型总的准确率可达 81.87%。其中对履约人群的判断更为准确，其概率达到 92.63%。而对违约人群的判断并不理想。仅仅只有 56.77%。其原因可能与模型特征的选取有很大关系。

（2）基于主成分分析筛选的变量

筛选出的主成分作为随机森林的特征变量。运用前文所述方法，选取树的棵树为 500 棵。下图展示了判断错误的概率与树的棵树选择的关系。从中可以看出：随机森林的预测准确率最终恒定在 70%左右，比较之前的准确率（82%左右）有明显的下降，可以得出如下结论：主成分筛选变量的方法在随机森林特征提取中没有较好的效果，反而因为其引起的信息损失，造成模型预测的效果下降。

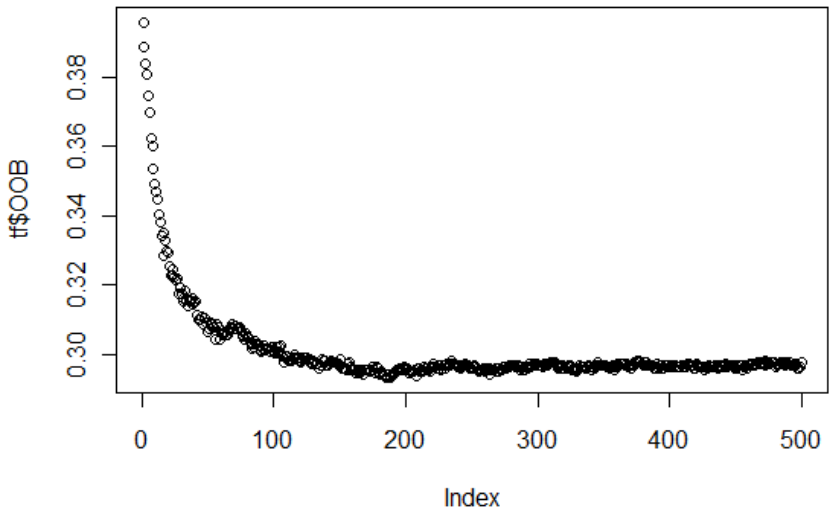

图 5-4　数的棵树与准确率的关系

2.判别分析法

（1）判别分析基本思想

判别分析是用以判别个体所属群体的一种统计方法，近年来，在许多现代自然科学的各个分支和技术部门中，得到了广泛的应用。本文运用判别分析的方法，选取一批参与 P2P 信贷的人中已知履约的人，观测其若干指标的数据。然后再选取一批已知违约的人，同样也测得相同的指标和数据。利用这些数据建立一个判别函数，并求出相应的临界值。这是对于需要进行甄别的人，也同样利用这些指

标的数据，将其带入判别函数，求得判别得分，在依据判别临界值，就可以判断出此人是否属于违约（或者履约）的群体。

判别分析有多种方法，本文采用的是距离判别方法，它是最简单、最直观的方法，该方法适用于连续性随机变量的判别类，对变量的概率分布没有限制。(处于论文篇幅限制本文对距离判别法的判别准则和具体判别函数的建立过程从略)。

（2）判别分析建模过程

①基于 spearman 秩检验筛选的变量

本文根据前文 spearman 秩相关性分析的检验结果，选取了与被解释变量(是否违约）相关关系最高的 5 个变量作为其观测指标。其分别是贷款期限 x2；贷款利率 x3；收入认证 x9；循环额度利用率 x16；月还收入比 x19。

选取 10000 个数据作为训练集数据（train.csv），其中 7000 条为履约记录，3000 条为违约记录。通过 R 软件编程，运行判别分析算法，输出判别分析的结果。程序见（判别分析程序.R）

最终由判别分析的计算方法得出结果如下：

①当方差取为相等时，其预测准确率判断见下表 5-3 所示：

表 5-3 方差取为相等时的模型预测率

原始结果		预测结果		预测准确率
		履约	违约	
履约	7000	4662	2338	66. 6%
违约	3000	1070	1930	64.33%
总体准确率：65.92%				

②当方差取为不等时，其预测准确率判断间下表 5-4 所示：

表 5-4 方差取为不等时的模型预测率

原始结果		预测结果		预测准确率
		履约	违约	
履约	7000	4457	2543	63.67%
违约	3000	982	2019	67.3%
总体准确率：64.76%				

本文代入 10000 组借款人数据，由距离判别模型的计算结果可以看出：当选取方差相等的时候，平均预测准确率为 65.92%，其中履约的被误判为违约的有 2338 条数据，履约客户的预测准确率为 66.6%，违约客户的预测准确率为 64.3%，总的预测准确率为 65.92%，略微高于方差选取为不等的时候。但是方差取为不等的时候，模型对违约的人的判断会更加准确为 67.3%；而履约的被误判为违约的有 2543 条数据，履约客户的准确率为 63.67%，总的预测准确率为 64.76%。

②基于主成分分析筛选的变量

方法如同前文，将筛选出的主成分作为判别变量。重新计算模型在新的变量下的准确效果。从下表对比前文可以看出，主成分筛选变量对判别分析影响较小，没有引起其准确性显著的变化。

表 5-5　预测准确率

原始结果		预测结果		预测准确率
		履约	违约	
履约	7000	4500	2500	64.29%
违约	3000	1039	1961	65.37%
总体准确率：64.61%				

3. Logistic 回归

（1）Logistic 回归基本思想

根据国内外大量研究结果，Logistic 回归分析方法常用于个人信用评分，该模型的优点在于模型稳健，能有效判断客户还款概率，结果较为精确，并且容易理解和运用。

对于响应变量有 p 个自变量（或称解释变量），

记为 $X_1, X_2, \cdots, X_p$ 在 p 个自变量的作用下出现　成功的条件概率记为　$P = \{Y = 1 \mid X_1, X_2, \cdots, X_p\}$

则 logistic 回归模型为：

$$p = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots\ldots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots\ldots + \beta_k x_k}}$$

其中：

β_0为常数或截距项， 称$\beta_1, \beta_2, \ldots, \beta_p$为logistic模型回归函数。

（2）建模

① 基于 spearman 秩检验筛选的变量

a 显著性检验

运用 R 软件计算 logistic 回归, 将 10000 组数据纳入 logistic 二分类回归方程 ，算出的显著性检验结果如表 5-6 ：

表 5-6 显著性检验结果

Estimate	Std. Error	z	value	Pr(>|z|)	
(Intercept)	3.19E+00	1.34E-01	23.788	<2e-16	***
x2（借款期限）	-1.06E-03	2.36E-03	-0.448	0.6538	
x3（借款利率）	-1.14E+01	6.11E-01	-18.611	<2e-16	***
x8（年收入）	2.87E-06	6.65E-07	4.321	1.56E-05	***
x9（收入认证）	-2.82E-01	5.62E-02	-5.013	5.37E-07	***
x12（收入负债率）	-6.05E-03	3.23E-03	-1.875	0.0608	.
x16（循环额度利用率）	-7.60E-01	1.03E-01	-7.402	1.35E-13	***
x17（总共的信用账户）	1.23E-02	2.12E-03	5.774	7.74E-09	***
x19（月还收入比）	-4.60E+00	6.30E-01	-7.309	2.69E-13	***
Signif. codes: 0；‘***’ 0.001；‘**’ 0.01；‘*’ 0.05；‘.’ 0.1；‘ ’ 1					

由上表各变量系数的估计值可得：借款利率 x3，年收入 x8，收入认证 x9，循环额度利用率 x16，总共的信用账户 x17，月还收入比 x19 比较显著，与我们预想的对这些指标的定性判断是正确的, 认定上述这 6 个解释变量与被解释变量的线性关系是统计上显著的。而变量 x2，x12 则不显著，没有显著解释作用，故我们考虑将其从模型中去掉。

在剔除不显著变量借款期限 x2 和收入负债率 x12 后，模型中还剩 6 个解释变量，重新进行 Logistic 回归，检验结果如表 5-7:

表 5-7 剔除不显著变量后的显著性检验结果

Estimate	Std. Error	z	value	Pr(>|z|)	
(Intercept)	3.12E+00	1.22E-01	25.61	<2e-16	***
x3	-1.15E+01	5.54E-01	-20.775	<2e-16	***
x8	3.19E-06	6.38E-07	5.003	5.64E-07	***
x9	-2.91E-01	5.56E-02	-5.228	1.71E-07	***
x16	-8.02E-01	9.99E-02	-8.031	9.69E-16	***
x17	1.09E-02	1.99E-03	5.453	4.96E-08	***
x19	-4.17E+01	6.27E-01	-7.544	4.57E-14	***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

回归结果可以明显看出，这六个解释变量均显著，对被解释变量有较强的解释作用，因此可以保留在模型之中。由之前的多重共线性也可知，这六个变量不存在多重共线性，变量之间的容忍度比较高，不会对回归模型的参数估计结果的精确性产生较大影响。

b 个人风险评估模型的创建

根据（表 4-6）的回归结果， Logistic 回归模型可表示为：

$$P=\frac{\exp(3.115-11.152X3+0.000003191X8-0.2906X9-0.8023X16+0.01087X17-4.728X19)}{1+\exp(3.115-11.152X3+0.000003191X8-0.2906X9-0.8023X16+0.01087X17-4.728X19)}$$

根据本次 logistic 回归检验时 R 软件得出结果（表 5-8）显示：本文代入 10000 组借款人数据所建立的 Logistic 回归模型，平均预测准确率为 70.89%，其中履约客户的预测准确率为 91.66%，违约客户的预测准确率为 22.43%。该模型在预测客户是否履约方面具有较高准确率，而判断违约客户的准确率则很低。故还需进一步测试模型判断履约和违约方面的准确率。

表 5-8　logistic 回归预测准确率

原始结果		预测结果		预测准确率
		履约	违约	
履约	7000	6416	584	91.66%
违约	3000	2327	673	22.43%
总体准确率：70.89%				

② 基于主成分分析筛选的变量

用提取出的前六个主成分作为新的变量，重新拟合 logistic 模型。在建模，由于后面三个变量未能通过显著性检验，故予以剔除。最终得到的模型为：

表 5-9 剔除不显著变量后的显著性检验结果

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.96448	0.02416	39.925	<2e-16 ***
Z1	0.25523	0.01360	18.762	<2e-16 ***
Z2	0.50353	0.02157	23.345	<2e-16 ***
Z3	-0.19825	0.02190	-9.051	<2e-16 ***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

根据上表的回归结果， Logistic 回归模型可表示为：

$$P=\frac{\exp(0.9645+0.2552*Z1+0.5035*Z2-0.1983*Z3)}{1+\exp(0.9645+0.2552*Z1+0.5035*Z2-0.1983*Z3)}$$

其中 Z_i 为原始变量的线性组合。

本文将 10000 组借款人数据回带入所建立的 Logistic 回归模型，得到平均预测准确率为 70.98%（具体情况如下表 5-10 所示），与之前模型 logistic 模型相比较略有提升。说明主成分筛选变量后对 logistic 模型有一定的优化效果。

表 5-10：预测准确率

原始结果		预测结果		预测准确率
		履约	违约	
履约	7000	6461	539	92.3 %
违约	3000	2363	637	21.23%
总体准确率：70.98%				

（二） 总结与预测

1. 变量的筛选对模型的影响

本文从 spearman 秩检验和主成分分析两种方法筛选变量，再分析各个模型的判断准确率。

主成分筛选变量后可以减少模型变量个数起到降维的作用，但是在对随机森林模型来说，由于其并非线性模型，而且筛选变量造成了信息的损失，经过线性组合的特征并不一定能给模型带来更好的效果。而对于线性的 logistic 模型，则可以运用主成分分析的办法提取变量以降低模型变量的个数。因此，本文采用 spearman 秩检验筛选变量，从而进行模型预测。

2.各模型的比较

通过上述三种建模，得到各个模型的预测准确率汇总如下表 4-8：

表：5-11 各模型比较分析

模型名称	履约正确率: 违约正确率：	总准确率
随机森林	92.63%	81.87%
	56.77%	
判别分析	66.60%	65.92%
	64.33%	
logistic	91.66%	70.89%
	21.43%	

由上表可总结得：

第一，相对于履约正确率：随机森林模型和 logistic 模型的预测准确度比较高，都超过了 90%。而判别分析的结果相对较低。而相对于违约正确率来讲，判别分析正确率最高，而 logistic 模型的正确率远远低于其他两种模型。

第二，总体来看随机森林模型预测最为准确，其次是 logistic 模型，最后是判别分析的结果。但是如果贷款者属于风险厌恶性，要追求违约的人的判断准确性，则很可能舍弃对违约人的判断准确率低的模型。

3. 各模型的预测

为了更加贴近原始数据的真实情况，即违约的概率为 14%左右，本文预测模型选取了 2000 个履约样本和 325 个违约样本作为测试集的样本。根据以上建模的训练结果对测试集进行检验。检验程序如（表 5-12）所示：

表 5-12 模型准确率预测

建模方法	判断正确个数	预测准确率	训练准确率
随机森林	1932	83.10%	81.87%
判别分析	1534	65.98%	65.92%
logistic	1933	83.14%	70.89%

由上表可以看出，三个模型的预测准确率分别为：随机森林模型（83.10%），判别分析模型（65.92%），logistic 模型（83.14%）。

比较出乎意料的是 logistic 模型的预测集准确率远远高于训练集准确率。但是，通过分析我们可以看到，在建立 logistic 模型并且用训练集进行回代的时候 logistic 模型将履约的人判断正确的概率为：91.66% 。而将违约的人判断正确的概率却只有：22.43%。可以看出本文中指标和数据建立的 logistic 模型更加偏向于将所有人判断成为履约的人。而本文的选取预测集中履约的比重比训练集更大。从某种程度上讲迎合了 logistic 模型的判断，使其准确率变高。而并非 logistic 模型在解决此类问题上的优越性。

综上所述，本文在建立 P2P 信贷风险评价模型相比较中认为，随机森林模型更加准确和可靠。

参考文献

[1]刘峙廷.我国 P2P 网络信贷风险评估研究.广西大学, 2013.5：44

[2]王富全．个人信用评估与声誉机制研究[M]．山东大学:王富全，2010．96-106

[3]龙新庭,王晓华.德国 IPC 公司微贷技术在我国经济欠发达地区的运用.区域金融研究,2013（10）:69

[4]薛毅．R 统计建模与 R 软件[M]．清华大学出版社:王静仪，2015．307-313

[5][5]何晓群．应用回归分析[M]．中国人民大学出版社:刘文卿，2000．159-162

[6]何晓群．多元统计分析[M]．中国人民大学:何晓群，2012．88-94

[7]Holmstrom, B.,1982:“Moral hazard in term”, Journal of economics, 13, pp.324-340

[8]William, k.,2000:”Online recuriting:a powerful tool ,”Strategic Finance , Montval.

[9]Milgom,P.and Roberts, J., 1982:“Predation,reputation and entiy deterence”,journal of economic Theory,27,pp.280-312.