Statistics and Application 统计学与应用, 2016, 5(3), 263-275
Published Online September 2016 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2016.53026

Hans汉斯

Study on Analysis and Influence Factors of Credit Card Default Prediction Model

Ruiting Mei*, Yang Xu*, Guochang Wang#

College of Economics, Jinan University, Guangzhou Guangdong
Email: #wanggc023@amss.ac.cn

Received: Aug. 31st, 2016; accepted: Sep. 14th, 2016; published: Sep. 20th, 2016

Abstract

Credit cards are a bank business in which high income and heavy risk coexist. Along with the development of the credit card business, banks are using the Internet and mobile data to establish customer credit rating system. How to evaluate customer credit from the information that customers fill in, and how to identify the information true or false, and what type of information that customers are asked to fill are crucial for banks. Based on the credit card customer data of 2005 in Taiwan, this article established Lasso-Logistic model and random forest model to explore the key factors which effect customer credit, including individual characteristics and some objective characteristics. Through comparing the prediction accuracy of the model and F score index, we selected the model of better prediction effect to forecast the bank credit card defaults. The establishment of the credit card default prediction model and the exploration of the key factors influencing the customer credit not only have a important guidance value for banks to choose customers and design data, but also can provide certain theoretical support for the credit decisions. In addition, it has a strong theoretical and practical significance.

Keywords

Credit Card, Credit Risk, Random Forest, Lasso-Logistic Model

信用卡违约预测模型分析以及影响因素探究

梅瑞婷*，徐　扬*，王国长#

*并列第一作者。
#通讯作者。

文章引用: 梅瑞婷, 徐扬, 王国长. 信用卡违约预测模型分析以及影响因素探究[J]. 统计学与应用, 2016, 5(3): 263-275.
http://dx.doi.org/10.12677/sa.2016.53026

暨南大学经济学院，广东 广州
Email: #wanggc023@amss.ac.cn

收稿日期：2016年8月31日；录用日期：2016年9月14日；发布日期：2016年9月20日

摘　要

信用卡对于银行来说是高收益和高风险并存的业务，伴随信用卡业务发展的是各大银行都在利用网络和移动端的数据来建立客户的信用评分系统。如何从客户所填的资料里对客户进行信用评估、如何鉴别所填资料的真假性及应该要求客户填什么类型的资料等对银行来说是至关重要的。本文基于2005年台湾信用卡客户数据，建立Lasso-Logistic及随机森林模型来探索影响客户信用的关键因素，包括个体特征及某些客观特征，通过比较模型的预测准确度以及*F*得分等指标来选择预测效果更优的模型对银行信用卡违约进行预测分析。信用卡违约预测模型的建立以及影响客户信用的关键因素的探索，对于银行选择客户和设计资料填写具有重要的指导价值，并且能够为信贷决策提供一定的理论支持，具有很强的理论和现实意义。

关键词

信用卡，信用风险，随机森林，Lasso-Logistic模型

1. 引言

随着我国金融体制改革和发展，信用卡业务应运而生，信用卡业务的兴起，为我国的金融市场做出了巨大的贡献。信用卡作为一种无抵押、无担保的消费信贷工具，具有为满足客户的日常消费需求提供相关信贷的功能。然而在金融市场的客观规律下，风险和收益相对称，高收益的信用卡业务必定伴随着高风险。我国商业银行开展信用卡业务的时间不长，业务经验匮乏，经营模式是粗放式的，银行信用卡业务也面临着越来越大的挑战，因此也必定蕴藏着很大的风险。如图 1 所示[1]，2008~2011 年我国信用卡违约金额由 33.70 亿元增至 110.31 亿元，增加 76.61 亿元，增长率为 227.4%。由于信用卡违约金额逐渐增大，因此信用卡风险管理是重中之重，而风险管理体系是建立在评估个人信用风险的基础之上的。传统的信用评估方法是一种人工信用风险评估，就是信用分析人员通过对信用卡申请者所上交的资料的审核来进行的，一般包括客户个人资料(年龄、身份证等)、工作(收入)证明、名下资产、稳定的还款能力情况等。银行为了控制风险，看重办卡人的还款能力，还款能力越强，银行越容易发卡。如今申请信用卡人数越来越多，银行的信用卡业务逐渐体现出的发行量大、交易频繁、交易信息全面准确等的特点，让我们意识到传统的人工估算已越来越不能胜任这个工作了。而现代数据挖掘技术的出现，则为信用卡的信用风险评估提供了一个客观准确的控制机制[1]。随着金融行业信息化的快速发展，几乎每天都产生了大量的个人贷款及还款等信息，银行如何利用这些信息来判定申请者的信用能力，是很多学者关注的主要问题。

从早期传统的人工信用风险评估到之后的自动评估系统，也就是信用卡个人信用评分系统。评分系统利用大量的客户历史数据，给出一个分数，根据客户的信用分数，可以分析客户按时还款的可能性，以此给予客户相应的额度及利率。国内对于信用卡的研究以理论研究居多，崔萌(2013) [2]硕士论文主要研究信用风险与宏观经济因素之间的问题，采用 CPV 模型与压力测试的方法进行研究，使用居民消费价格指数、国民生产总值、固定资产投资、居民收入四个变量，将变量与不良贷款的违约率进行回归分析，

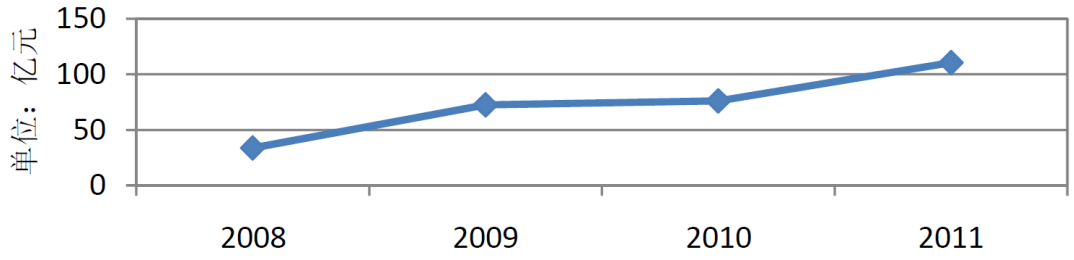

Figure 1. China credit card default amount of 2008-2011
图 1. 2008~2011 年中国信用卡违约金额

得到结果是宏观经济会对银行的不良贷款率产生影响。高嘉晔(2014) [3]考察了宏观经济以及微观两个层面的数据利用 Logistic 回归模型进行研究，发现两者对于违约率均有影响。许多文献结果表明宏观经济形势对于信用卡行业有着很大的影响，商业银行在一定程度上需要依靠有利的宏观经济优势，制定合适的信用卡业务的发展战略。

在微观经济层面对信用卡违约的研究中，朱醒亮(2012) [4]基于统计学中的 bi-probit 模型，同时考虑信用卡信用评分的授信过程以及申请信用卡成功之后的还贷过程，通过调查问卷的数据，分析持卡人的违约率为持卡人的信用情况进行评分。徐少峰、王延臣(2003) [5]在个人信用卡评估中建立了 Logistic 模型，实证表明，以此模型为依据决定放贷与否效果较好。石庆焱(2005) [6]进行了国内多种方法的比较研究，研究发现神经网络等非线性方法的精度往往要高于 Logistic 回归、线性规划等线性评分方法。但是在总体上看在预测精度范围内，这些线性评分模型还是有较强的区分“好”“差”客户的能力，因此可以用于信贷决策，而且这些线性模型的稳健性要强于神经网络等非线性方法。

本文就是基于 3 万的样本数据包括客户个体特征(如性别，年龄婚姻状况等)以及信用卡额度、还款情况等信息，通过对 Lasso-Logistic 回归、随机森林两种模型的建立以及预测效果的对比研究来进行的。信用卡风险评估的预测模型建立能够为信贷决策提供一定的理论支持，对于推动银行信用卡市场良性发展、保持国民经济持续稳定发展有着十分重要的意义。

2. 模型简介

2.1. Lasso-Logistic 模型介绍

Logistic 模型是 1938 年提出的用于人口预测的模型，现如今主要用来预测离散因变量与一组解释变量之间最常用的二值型模型。在本文中预测 10 月份客户信用卡违约情况的二值多元 logistic 模型为：

$$p=\frac{\exp\left(\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n\right)}{1+\exp\left(\alpha+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n\right)} \tag{1}$$

上式中，因变量 p 是违约发生的概率，$x_1,x_2,\cdots,x_n$ 是自变量，$\beta_1,\beta_2,\cdots,\beta_n$ 是回归待定系数。对上式进行 logit 变换，得到多元 logixtic 回归模型如下：

$$\ln=\left[p_i/(1-p_i)\right]=\alpha+\sum_{n=1}^{n}\beta_nx_{ni} \tag{2}$$

式中，$p_i=P\left(y_i=1\middle|x_{i1},x_{i2},\cdots,x_{in}\right)$ 是在自变量给定时违约发生的概率，α 为截距。

Lasso (The Least Absolute Shrinkage and Selectionator operator)算法是 Tibshirani (1996) [7]提出的一种降维方法，本质是通过惩罚函数将对因变量没有影响或影响较小的自变量的回归系数压缩到 0，可以有效消除多重共线性的影响。

为了消除指标量纲的影响，在使用 Lasso 方法时，需要对数据进行标准化处理，即满足：$\sum_{i=1}^{n} x_{ij} = 0$，$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$，其中 $j = 1, 2, \cdots, p$。Lasso 估计定义公式：

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \quad \text{s.t.} \quad \sum_{j=1}^{p} \left|\beta_j\right| \le t \tag{3}$$

t 是调和参数，当 t 充分小时将导致某些系数恰好为 0，这样就是实现了变量选择。

Lasso-Logistic 模型在普通的 Logistic 模型上，加上 Lasso 变量选择，来同时实现选择参数和变量估计。之所以选择这种方法是因为本文的 23 个变量存在很高的相关性，用 Lasso 在实现变量选择的同时可以解决多重共线性问题。还有本文信用卡违约预测时，其被解释变量是二元离散取值，所以用这种模型较适合。

假设信用卡观测值 $(x_i, y_i), i = 1, 2, \cdots, n$，其中 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ 和 y_i 分别是模型的解释变量和被解释变量，并且 y_i 是二元离散数据变量，即 $y_i \in \{0, 1\}$，则

Logistic 线性回归模型的条件概率为：

$$\log\left\{\frac{p(y_i = 1|x_i)}{1 - p(y_i = 1|x_i)}\right\} = \eta\beta(x_i) \tag{4}$$

其中，$\eta\beta(x_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$。Lasso-Logistic 回归模型中的系数估计值 $\hat{\beta}_\lambda$ 由下面公式凸函数的极小值给定：

$$S_\lambda(\beta) = -l(\beta) + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{5}$$

其中 $l(.)$ 是对数似然函数，则式(4)中的 $l(\beta)$ 可以写成如下式：

$$l(\beta) = \sum_{i=1}^{n}\left\{y_i\eta\beta(x_i) - \log\left\{1 + \exp\left[\eta\beta(x_i)\right]\right\}\right\} \tag{6}$$

Lasso-Logistic 回归模型中的系数估计值 $\hat{\beta}$ 可写成如式(6)的形式：

$$\hat{\beta} = \arg\min \sum_{i=1}^{n}\left\{y_i\eta\beta(x_i) - \log\left\{1 + \exp\left[\eta\beta(x_i)\right]\right\}\right\} + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{7}$$

其中调和参数 λ 可以用交叉验证的方法来确定。

2.2. 随机森林简介

随机森林方法通俗的讲，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类(对于分类算法)，然后看看哪一类被选择最多，就预测这个样本为那一类[8]。

在建立每一棵决策树的过程中有两点需要注意，采样与完全分裂。

1) 有放回的随机采样选择样本和无放回的随机选择特征属性[9]

首先是两个随机采样的过程，random forest 对输入的数据要进行行、列的采样。对于行采样，采用有放回的方式，也就是在采样得到的样本集合中，可能有重复的样本。假设输入样本为 N 个，那么采样的样本也为 N 个。这样使得在训练的时候，每一棵树的输入样本都不是全部的样本，使得相对不容易出现 over-fitting。然后进行列采样，随机森林的做法不同于决策树，随机森林中，是从 M 个 feature 中，选择 m 个($m \ll M$)，

一般来说，*m* 取值为根号 *M* 比较合理。由于是无放回的随机选取，因此，在构建单棵决策树的过程中，每次选取一个特征，然后不放回，接着选取其他特征，最终拿到合适数量的特征属性用来创建单棵的决策树。

2) 完全分裂建立决策树[9]

之后就是对采样之后的数据使用完全分裂的方式建立出决策树，这样决策树的某一个叶子节点要么是无法继续分裂的，要么里面的所有样本的都是指向的同一个分类。一般很多的决策树算法都有一个重要的步骤——剪枝，但是这里不这样做，由于之前的两个随机采样的过程保证了随机性，所以不剪枝，也不会出现 over-fitting。

按这种算法得到的随机森林中的每一棵都是很弱的，但是大家组合起来就很厉害了，我们一般会重复很多次得到很多棵决策树，树的数量不能太多也不能太少。太少就会过拟合，不能客观的进行预测，太多就会增加计算量。我觉得可以这样比喻随机森林算法：每一棵决策树就是一个精通于某一个窄领域的专家(因为我们从 *M* 个 feature 中选择 *m* 让每一棵决策树进行学习)，这样在随机森林中就有了很多个精通不同领域的专家，对一个新的问题(新的输入数据)，可以用不同的角度去看待它，最终由各个专家，投票得到结果。

3) OOB 错误评估[10]

在采样时，我们每次做有放回的随机采样选取训练样本集合的过程中，会有一部分样本采样不到。选中了就放到 bag 中，没有选中的就是 out of bag。平均而言，每次放回采样中，37% (即大概三分之一)的数据不会被选中。所以我们可以用没有抽取到的数据集来测试这棵决策树的分类的泛化能力也就是准确度，即可以作为模型效果评估的一个指标即 *OOB*，而且 *OOB* 错误估计被证明是无偏的。

3. 数据简介

3.1. 数据认识

本文采用的数据是来自 UCI 网站中台湾地区信用卡客户数据，研究目的是对信用卡客户是否违约做一个预测，因此响应变量是一个二分类变量，违约记为 1，未违约记为 0。并且本文使用 23 个变量作为解释变量：

X1：银行给予信用卡客户的信用额度包括个人信用额度和客户的家庭信用额度。

X2：信用卡客户的性别。男性记为 1，女性记为 2。

X3：信用卡客户的教育水平。研究生及以上记为 1，大学记为 2，高中记为 3，其它记为 4。

X4：信用卡客户的婚姻状况。已婚记为 1，未婚记为 2，其它记为 3。

X5：信用卡客户的年龄。

X6~X11：这六个变量是 2005 年 4 月到 9 月每月的还款记录。如：X6 为 2005 年 9 月的还款情况；X7 为 2005 年 8 月的还款情况；…；X11 为 2005 年 4 月的还款情况。还款情况的测量量表为：0 = 及时还；1 = 还款延迟一个月；2 = 还款延迟两个月；3 = 还款延迟三个月；…；9 = 还款延迟九个月及以上。

X12~X17：这六个变量是 2005 年 4 月到 9 月每月的账单记录，即每月用信用卡消费记录。如：X12 为 2005 年 9 月的账单金额；X13 为 2005 年 8 月的账单金额；…；X17 为 2005 年 4 月的账单金额。

X18~X23：这六个变量是 2005 年 4 月到 9 月每月的支付记录，包括还账单金额和存入信用卡的金额，其中还账单金额不能低于银行规定的最低还款额。如果支付金额大于上月账单金额则视为及时还，剩余金额存入信用卡留做下次消费；如果支付金额小于上月账单金额则视为延迟还款。如：X18 为 2005 年 9 月的支付金额；X19 为 2005 年 8 月的支付金额；…；X23 为 2005 年 4 月的支付金额。为了对数据有个整体性认识，我们对本文数据进行简单描述性统计分析如下：

从表 1 的简单描述性数据可以看出，在所有样本信用卡用户中，女性所占比例为 60.37%，是男性

Table 1. Descriptive statistics of data
表 1. 数据描述性统计

		Y		Y = 1	
		样本总数	比重(%)	样本量	比重(%)
X2	男性	11,888	0.396767	2873	0.241672
	女性	18,112	0.603733	3763	0.207763
X3	研究生	10,585	0.352833	2036	0.192348
	大学	14,030	0.467667	3330	0.237349
	高中	4917	0.163900	1237	0.251576
	其它	468	0.015600	33	0.070513
X4	已婚	13,659	0.455300	3206	0.234717
	未婚	15,964	0.532133	3341	0.209283
	其它	377	0.012567	89	0.236074
X6	及时还	23,182	0.772733	3207	0.138340
	延迟还	6818	0.227267	3429	0.502933
X7	及时还	25,562	0.852067	4160	0.162742
	延迟还	4438	0.147933	2476	0.557909
X8	及时还	25,787	0.859567	4434	0.171947
	延迟还	4213	0.140433	2202	0.522668
X9	及时还	26,490	0.883000	4757	0.179577
	延迟还	3510	0.117000	1879	0.535328
X10	及时还	27,032	0.901067	4987	0.184485
	延迟还	2968	0.098933	1649	0.555593
X11	及时还	26,921	0.897367	5025	0.186657
	延迟还	3079	0.102633	1611	0.523222

信用卡客户的 1.524 倍；从教育水平角度来看，大学生所占比重最大，研究生次之；从婚姻角度来看，已婚所占比重大于未婚所占比重；从 4 月~9 月的还款情况来看，及时还款人数占大多数。但对于 10 月份违约样本的记录中，男性信用卡客户违约的比例占总男性的比重为 24.17%，女性违约比例为 20.78%，说明男性客户较之女性更易违约；从教育水平方面来说，高中生违约比例最大，大学生次之，研究生第三，某种程度上说明教育水平越低越易发生违约情况；从婚姻方面来说，已婚客户较之未婚客户更易违约；从是否及时还款方面来说，超过半数的延迟还款客户会违约，而及时还款客户发生违约的比重比较小，说明延迟还款是衡量是否违约的重要变量。

3.2. Smote 数据平衡法及预测指标 *F* 得分

由于本文样本数据“0”(未违约)和“1”(违约)分布的不均匀，考虑到为了建立预测效果更好的模型，我们使用 Synthetic Minority Oversampling Technique (Smote)方法即合成少数类过采样技术。Smote 是一种基于过采样技术的经典采样算法，是针对稀有事件(少数类)进行过采样。算法的基本思想是对少数类样本进行分析并根据少数类样本人工合成新样本添加到数据集中[11]。合成的策略是对每个少数类样本 a，从

它的最近邻中随机选一个样本 b，然后在 a、b 之间的连线上随机选一点作为新合成的少数类样本[12]。

Smote 算法里主要参数有 prec.over：过采样时生成少数类的样本个数；k：过采样中使用 K 近邻算法生成少数类样本时的 K 值，默认为 5；prec.under：欠采样时，对应每个生成的少数类样本，选择原始数据多数类样本个数。如 perc.over = 500 表示对原始数据集中的每个少数样本，都将生成 5 个新的少数样本；perc.under = 80 表示从原始数据集中选择的多数类的样本是新生的数据集中少数样本的 80% [8]。本文正是采用了 Smote 算法平衡训练集来解决数据分类的不平衡化问题。

正是由于数据的严重不平衡，因此会使得更多的预测类偏向多数类即“0”(不违约)这一边，这里我们所采取的措施是尽量使得本应预测分类为“1”的样本数据，通过更好的模型建立让其实际上也预测为“1”，提高少数类样本“1”的分类准确率。这里具体使用的指标是分类“1”的精确度和召回率综合得到的调和平均数 F scores，将其作为一个模型预测效果提升的参考指标。F scores 的具体意义如表 2 所示。

4. 模型分析

4.1. 随机森林模型分析

由于本文中变量自个数有 $p = 23$ 个，因此随机森林模型选择 $m = \sqrt{p}$ 个数较为适合，因此选择变量节点数为 5 个。从表 3 中看出 ntree 个数 500~1000 内 OOB 误差以及 F 得分并无明显区别，在随机森林模型中决策树的树的个数一定要设置的比较合理，树的个数比较少会导致训练的不够充分，树之间的相关性会变得越来越小，反之比较大则会大大的增加模型的计算量，因此我们选择 ntree = 500 即可。

接下来为了建立一个预测效果好的随机森林模型，我们将 Smote 方法与随机森林结合起来。表 4 即为 smote 不同采样比例下的相关预测指标，我们主要考察 OOB 误差率以及 F 得分来选择 smote 参数。综合评价这两个指标，选择 Pr. Over = 600，Pr. Under = 120，此时 OOB 为 3.88%，F 得分 49.95%。相比较于表 3 中未平衡处理得出的 OOB (18.42%)，随机森林模型的预测效果明显增强了。

随机森林选择出的特征的重要性如图 2，基于其他文献资料大多使用 Mean Decrease Accuracy 指标，本文着重阐释 Mean Decrease Accuracy 指标。图 2 左图即为 Mean Decrease Accuracy 指标，是指平均精确度的降低，就是对每一个变量比如 X1 随机赋值，如果 X1 重要的话，预测的误差会增大，所以误差的增加就等同于准确性的减少，精确度减小也就是反映了这个变量越重要。在图中我们看到了这些变量的重要性排序，前几个重要变量依次是 X5 (年龄)、X6 (2005 年 9 月的还款状态)、X12~X17 (9 月~4 月的账单)、X1 (信用卡总的信用额度)。很容易理解，前几个月份的还款及账单情况对于 10 月份的违约与否有着比较直接的影响。

我们还可以继续分析特定某个变量对随机森林模型的影响，如图 3 和图 4，横坐标表示变量取值，纵坐标则为该变量对模型的影响程度，这里我们显示 X1 (信用卡总的信用额度)、X5 (年龄)对模型的影响情况。信用卡额度在 40 万的时候对模型的影响程度最大，其次是在 40 万额度上升阶段的情况下，而最后是在额度比较低的情况下影响最小即违约可能性最小。这可以理解成一般都是财富、资本多的少数人才会去申请 40 万以上的高额度，此时此类申请人有资本去还款，并且因为高额度违约风险太大，这样也导致了不敢轻易冒险违约。X5 (年龄)变量值在 35 岁左右对模型影响程度最大即预测的违约概率最大，其次是 20 多岁的年龄阶段，最后为老年龄。这个可以解释为客户 35 岁左右正是成家立业阶段，信用卡使用动机强而且额度也会相应比较高，因此风险也比较大。而 20 多岁的客户一般刚出校园进入社会，资本匮乏，大多数事业也还未起步，对于信用卡的使用不敢冒太大风险，相对于 35 岁左右的客户风险低违约率也就降低，最后最不判为违约的就是老年龄了，这个也是比较符合常理的。

Table 2. Confusion matrix
表 2. 混淆矩阵

		True class		
		Yes	No	
Predicted class	Yes	*TP*	*FP*	*p*
	No	*FN*	*TN*	*n*
		P	*N*	

$$\text{precision} = \frac{TP}{TP + FP} \quad \text{recall} = \frac{TP}{TP + FN}$$

$$F\text{-score} = \frac{2}{1/\text{precision} + 1/\text{recall}}$$

Table 3. Selection of trees number
表 3. 树的个数选择

ntree	500	600	700	800
OOB	18.42%	18.48%	18.32%	18.30%
F Scores	48.30%	48.13%	48.80%	48.61%

Table 4. Correlation prediction index of smote under different sampling proportions
表 4. Smote 不同采样比例下的相关预测指标

Prec.over	Prec.under	OOB	F scores	precise	recall
300	100	8.74%	49.50%	42.08%	60.08%
400	100	5.16%	50.36%	45.31%	56.68%
500	100	5.16%	50.49%	45.41%	56.83%
600	100	4.19%	49.78%	45.33%	55.21%
300	100	8.74%	49.50%	42.08%	60.08%
500	120	4.53%	49.63%	47.28%	52.45%
600	**120**	**3.88%**	**49.95%**	**47.51%**	**52.66%**
700	120	3.78%	49.61%	47.25%	52.21%
500	80	5.85%	49.87%	47.25%	52.21%

同时可发现这些变量特征按重要性大小的顺序大致可以划分为如下几个方面：第一组，X5 (年龄)，X6 (9 月的还款情况)，X12~X17 (9 月~4 月的账单金额)，第一方面的特征对于 10 月份还款情况的影响比较明显；第二组，X1，X2，X4，X3，对应的是信用卡的总额度、教育水平、婚姻状况以及性别，这几个即为客户的个体特征，在某些程度上是一种主观影响因素；第三组，X18~X23，为 4~9 月的支付金额；第四组，X7~X10，为 3~8 月(中间月份)的还款延期情况。其中，表现出来的一现象是，X6~X11，X12~X17，X18~X23 作为 4~9 月份的三组变量，9、8、7 月份的还款情况对 10 月的还款情况影响最大。即对于 10 月份信用卡的违约预测中，样本数据的最近三个月份的消费、还贷情况比其他月份的影响要大。

4.2. Lasso-Logistic 模型分析

接下来看看 Logistic 模型的建立以及模型结果分析，表 5 为 smote 不同采样比例下的相应指标，可以

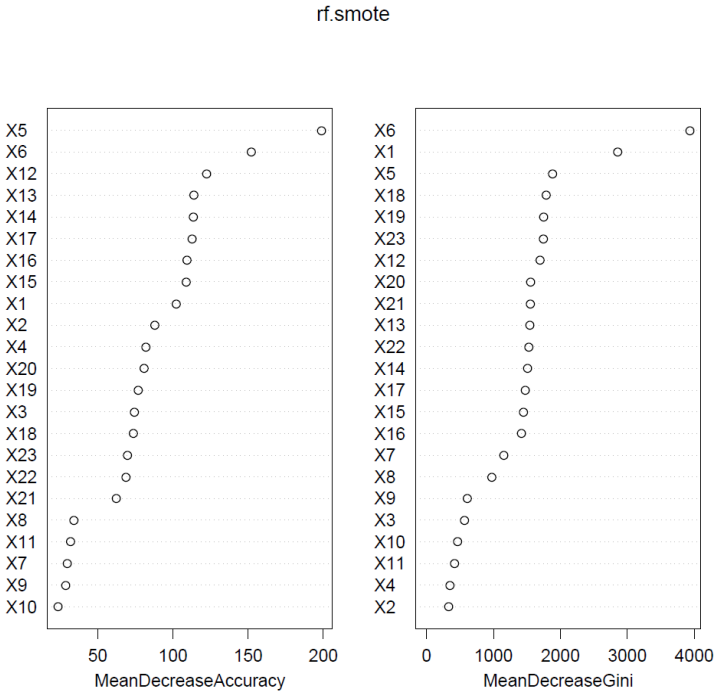

Figure 2. Variable feature importance evaluation
图 2. 变量特征重要性评估

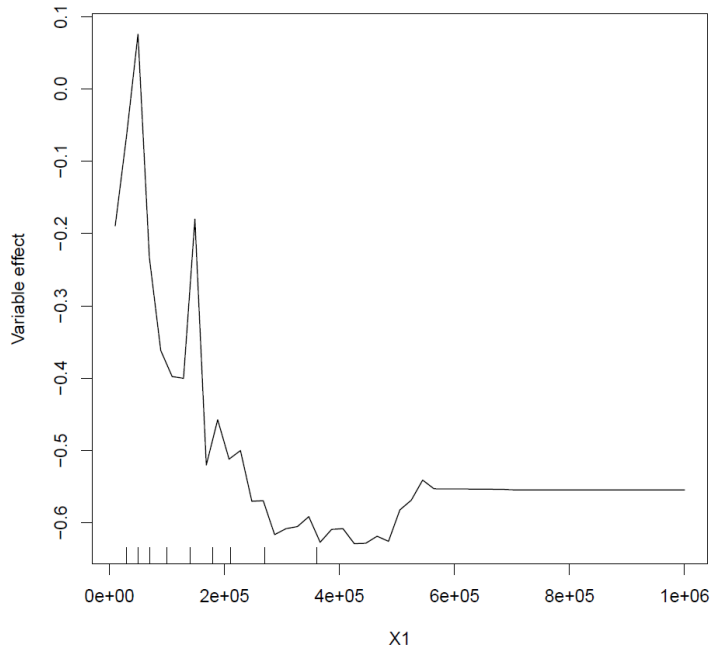

Figure 3. Effects of X1 variables on the model
图 3. X1 变量对模型的影响

看出预测效果 F 得分最高的是 Prec.over = 600，Prec.under = 120，F 得分为 53.47%，对应的 precise 为 51.15%，recall = 56.02%。

我们用此组参数来建立 Lasso-Logistic 模型，通过交叉验证方法选得的惩罚参数 λ = 0.01541 进行变量选择建立 Logistic 模型得到表 6 系数，其中 X1 (信用卡总额度)，X2 (性别)，X3-3 (教育水平－“高中生”

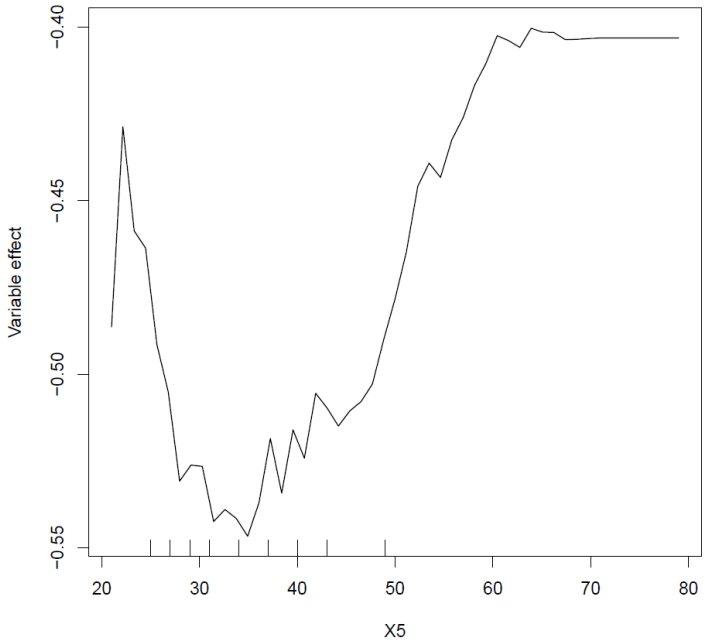

Figure 4. Effects of X5 variables on the model
图 4. X5 变量对模型的影响

Table 5. Correlation prediction index of smote under different sampling proportions
表 5. Smote 不同采样比例下的相关预测指标

Prec.over	Prec.under	F scores	precise	recall
300	100	51.32%	41.25%	67.90%
400	100	52.62%	44.41%	64.55%
500	100	34.83%	32.84%	37.07%
600	100	52.48%	46.35%	60.49%
600	**120**	**53.47%**	**51.15%**	**56.02%**
500	120	53.31%	50.56%	56.38%
700	120	53.45%	51.66%	55.37%
500	80	48.65%	36.62%	72.44%
300	100	51.32%	41.25%	67.90%
400	100	52.62%	44.41%	64.55%

Table 6. Coefficient of Lasso-Logistic model
表 6. Lasso-Logistic 模型系数

Intercept	X1	X2	X3-3	X4-1	X6	X7	X8	X9	X11	X18	X19	X20	X21	X22
−0.235	−0.173	−0.225	0.133	0.012	0.834	0.089	0.109	0.059	0.048	−0.080	−0.101	−0.004	−0.039	−0.040

哑变量)，X4-1 (“已婚”哑变量)，X6~X10 (5 月份~9 月份还款延期情况)，X18~X22 (5 月份~9 月份的支付金额)，这些经过筛选出来的变量在某种程度上来说对于违约预测的影响是比较大的，其中说明了高中生及已婚客户更易发生违约。

变量 X2 性别系数为−0.225，说明男性比女性更易发生违约。X18~X22 (9 月份~5 月份的支付金额)系

数都为负，说明支付额越多越不易发生违约，支付越少越易发生违约。其中 X6 影响系数较大，即客户在 9 月份的还款延期会很大程度的导致 10 月份发生违约，并且 X10~X6 即 5、6、7、8、9 月份的影响系数有逐月递增的趋势，说明越到信用卡的后期延期，则违约性质越严重，10 月份被判为违约的可能性越大。

4.3. 模型预测效果对比

本文建立了随机森林以及 Lasso-Logistic 两种模型作为本文信用卡违约的预测模型，两种模型的预测效果即预测准确率(Accuracy)以及 F scores (Precise, recall)指标如表 7 所示：发现随机森林预测准确度以及 F 得分都比 Lasso-Logistic 高。因此在本文信用卡违约预测模型分析中，随机森林预测效果要优于 Lasso-Logistic 模型。

随机森林在本文数据集的基础上表现良好，源于随机森林本身所具有的优势。随机森林能够处理高维度大数据，在处理过程中不需要做特征选择，但在训练完后，能够给出哪些特征变量比较重要，并且可以进行并行化的训练，训练速度快。因此数据挖掘技术在很多时候能解决经典统计方法在进行数据分析的时候所遇到的一些问题。

5. 总结与建议

5.1. 总结

本文主要针对信用卡客户的基本数据建立了两种客户违约预测模型，并进行了模型预测效果的对比，发现预测效果随机森林优于 Lasso-Logistic 模型。

综合以上分析，我们看到在预测信用卡客户违约中，客户的某些个体特征如教育水平、性别、年龄以及总信用额度、月份的消费及还贷情况等对信用违约与否有着明显的影响。而这些客户的资料属于很容易收集到的基本信息，我们可以建立一种信用评分机制。如表 8 所示，根据客户这些自然情况的

Table 7. Comparison of model prediction effect
表 7. 模型预测效果对比

Model	Accuracy	F scores	Precise	recall
Random Forest	78.25%	49.95%	47.51%	52.66%
Lasso-Logistic	75.60%	49.94%	45.95%	54.69%

Table 8. Credit card customer credit score
表 8. 信用卡客户的信用评分

项目	描述	得分	项目	描述	得分
性别	男	1	信用卡总额度	30~50 万	1
	女	2		50 万以上	2
教育水平	高中生	1	延期还款	30 万以下	3
	本科生	2		延期九个月	1
	研究生	3		延期八个月	2
婚姻状况	已婚	1		延期七个月	3
	未婚	2		.....	.....
年龄	30~50	1	支付金额	—	(支付越少得分越低)
	20~30	2	消费金额	—	(消费越高得分越低)
	50 以上	3	.....	.....	.....

信息，为其打分，表中得分是按照本文模型分析结果所设置的得分等级。而由于不同银行所设置的标准也可能不同，所以银行在面对具体客户进行信用评分时，可以按照银行自身所设置的具体标准进行评分。根据最终的信用评分结果，对于客户分数低的信用卡申请者不给予办卡或者降低信用额度，相反分数高的，即在一定程度上被看做是信用高的客户，银行可以适情增加额度，通过这样简单易操作的办法既可以降低风险同时又能增加银行收益。

5.2. 建议

目前国内申请信用卡填表资料与该数据所涉及到的变量相同，所以可以从该数据的研究结果适用于国内。基于以上研究，本文结合国内信用卡及公民信用相关知识从以下几个角度给出政策建议：

1) 严格审批、预防为主[13]，银行应设立信用卡申请的初审制度，规范相关申请批准流程，严格把关，特别是对于被评估为潜在高违约的人群，应设置较高的审批标准，比如设定更高的手续费率、违约金、滞纳金等，在源头上控制风险，可以有效降低风险。

2) 实时监督[13]，银行在信用卡发放后，可对持卡人的消费行为进行动态跟踪及评级，随时掌握客户的消费情况，建立动态机制，及时更新客户的信用特征信息。特别是高违约客户，一般消费额也高，对其账户实时监督，并且对于消费额高又长期逾期不还的客户可及时加入黑名单中，以此对风险进行很好的防范。

3) 征信系统完善，本文数据所涵盖的信息只局限于信用卡业务的信用信息(性别年龄、教育水平、婚姻状况及总额度、消费和还款情况)，还不能充分反映作为社会中的一个个体所有的信用。这些信用资料都散布在各个职能部门机关和相关单位，因此这些个人的信用数据征信困难，部门单位之间信用数据共享能力差，导致了银行只能依靠自身所具有的数据，建立的信用卡评分模型也就不能全面反映客户的信用情况。因此需要加强全国统一联网的个人征信体系的建设，完善征信系统平台，才能让我们的经济活动更安全，更有保障。

4) 信用卡业务创新，当信用卡违约率上升时，若只凭信用卡的某项业务的收益，是无法保证信用卡业务的持续发展。因此银行需要创新信用卡的业务，增加业务的多样性，比如与公交公司进行合作，开发新型公交－信用卡，可解决忘记充值人群一时的窘境，而且这种公交－信用卡额度也不高。再如和互联网企业合作，进行轻松网上购物等这些丰富多样的业务形式既能提高银行和合作方公司两方的收益，实现了合作共赢，又能将经济活动一体化，对于我国经济具有促进作用。

致　谢

本论文是在王国长老师的悉心指导下完成的，从论文的选题到中途的撰写、修改，他都给了我们很大的帮助。从老师渊博的专业知识，严谨的治学态度，精益求精的工作作风，严以律己、宽以待人的崇高品质，不论是在研究方法和个人品质方面，我们都收获良多。我们学到了高校阅读文献、搜集整理分析数据，用心对待才会有收获。这对我们将来的工作生活以及人生道路产产生了很大的帮助。在此向老师表示衷心的感谢！

还要感谢这篇论文里涉及到的学者和他们的研究，谢谢你们的研究！

最后，向在百忙中抽出时间对本文进行评审并提出宝贵意见的各位专家表示衷心地感谢！

参考文献 (References)

[1] 聂雨. 基于数据挖掘的信用卡个人客户信用评价研究[D]: [硕士学位论文]. 西安: 西安科技大学, 2012.

[2] 崔萌. 基于 CPV 模型和压力测试的我国商业银行信用风险研究[D]: [硕士学位论文]. 长春: 吉林大学, 2013.

[3] 高嘉晔. 信用卡违约风险影响因素实证研究[D]: [硕士学位论文]. 大连: 东北财经大学, 2014.

[4] 朱醒亮, 王佳, 葛姣菊. 基于 Probit 模型对消费者信用卡还贷影响因素的实证分析[J]. 消费经济, 2013(4): 48-51.

[5] 徐少锋, 王延臣. 个人信用评估中的 Logistic 模型[J]. 天津科技大学学报, 2003(12): 46-49.

[6] 石庆焱. 一个基于神经网络——Logistic 回归的混合两阶段个人信用评分模型研究[J]. 统计研究, 2005(5): 45-49.

[7] 方匡南, 章贵军, 张惠颖. 基于 Lasso-Logistic 模型的个人信用风险预警方法[J]. 数量经济技术经济研究, 2014(2): 125-136.

[8] 周丽峰. 基于非平衡数据分类的贷款违约预测研究[D]: [硕士学位论文]. 长沙: 中南大学, 2013.

[9] 佚文. 机器学习中的算法(1)——决策树模型组合之随机森林与 GBDT [DB/OL]. http://www.cnblogs.com/LeftNotEasy/archive/2011/03/07/random-forest-and-gbdt.html, 2011.

[10] 李伯韬. Spark 随机深林扩展——OOB 错误评估和变量权重[DB/OL]. http://www.cnblogs.com/bourneli/p/4536778.html, 2015.

[11] 佚文. 不平衡数据下的机器学习方法简介[DB/OL]. http://www.jianshu.com/p/3e8b9f2764c8, 2015.

[12] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002) SMOTE: Synthetic Minority Over-Sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321-357.

[13] 张晓蕾. 信用卡分期业务违约的影响因素及研究[D]: [硕士学位论文]. 广州: 暨南大学, 2014.