Statistics and Application 统计学与应用, 2020, 9(4), 582-592
Published Online August 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.94062

Hans汉斯

User Credit Risk Prediction Model based on Big Data

Jingwen Hu, Xiao Liu, Zhe Feng

School of Mathematical Science, Tongji University, Shanghai
Email: 728669013@qq.com

Received: Jul. 16th, 2020; accepted: Jul. 28th, 2020; published: Aug. 5th, 2020

Abstract

Credit risk is the main risk of bank operation and affects the development of bank. It is necessary to establish credit risk prediction model to help banks avoid risks and reduce losses. In this paper, 80,000 pieces of thousand dimensional data of a commercial bank are taken as the research object, and the method of "group principal component" is used to preprocess the data of thousand dimensional variables. Then, the credit risk prediction model is established by using Logistic regression and random forest respectively. The analysis results of the two models show that the customer's credit card level, occupation, value level, basic information of personal business, deposits and foreign current holdings have great influence on predicting the probability of default. The area under the curve of logistic regression model is 0.847, and the prediction accuracy is 75%; the area under the curve of the random forest model is 0.848, and the prediction accuracy is 85%. Compared with previous studies, the prediction accuracy of the two models is significantly improved. In practical application, the two models can be combined with each other to give full play to their advantages.

Keywords

Credit Risk, Principal Component Analysis, Logistic Regression Model, Random Forest Model, High Dimensional Data

基于银行大数据的用户信用风险预测模型

胡竞文，刘　潇，冯　哲

同济大学，数学科学学院，上海
Email: 728669013@qq.com

收稿日期：2020年7月16日；录用日期：2020年7月28日；发布日期：2020年8月5日

摘　要

信用风险是银行经营的主要风险，影响银行的发展，有必要建立信用风险预测模型，帮助银行规避风险、

文章引用: 胡竞文, 刘潇, 冯哲. 基于银行大数据的用户信用风险预测模型[J]. 统计学与应用, 2020, 9(4): 582-592.
DOI: 10.12677/sa.2020.94062

减少损失。本文以某家商业银行的八万条千维数据作为研究对象，采用“分组主成分”的方法对千维变量进行降维的数据预处理，运用Logistic回归和随机森林建立信用风险预测模型。两种模型的分析结果显示，客户的信用卡级别、职业、价值等级、个人业务基本情况、存款及本外币持有额情况对违约风险预测的影响较大。Logistic回归曲线下面积为0.847，预测准确率为75%；随机森林曲线下面积为0.848，预测准确率为85%，相较于以往的研究，两个模型的预测准确率都有明显提高。实际应用时，两种模型可以相互结合，充分发挥二者的优越性。

关键词

信用风险，主成分分析，Logistic回归模型，随机森林模型，高维数据

1. 引言

银行经营的过程会面临许多风险，其中信用风险是主要风险。信用风险的涵义是，因为债务人未能按合约执行义务，或信用质量改变，给债权人带来损失的可能性。信用风险会给银行带来直接或间接的经济损耗、增加管理成本、降低资金利用率[1]。因此，建立风险预测模型，根据客户数据信息，预测是否可能违约，有助于银行控制风险、减少损失、保证收益。

信用风险预测模型一直被持续而广泛地研究，信用风险预测研究的主要模型如表 1。

Table 1. Main models of credit risk prediction
表 1. 信用风险预测研究的主要模型

广义线性模型	非线性模型	
	机器学习模型	非参数模型
线性判别分析模型	神经网络模型	样条回归模型
(Linear Discriminant Analysis, LDA)	(Neural Networks, NN)	(Spline Regression Models)
Probit 回归模型	支持向量机模型	核回归
(Probit Regression)	(Support Vector Machine, SVM)	(Kernel Regression)
Logistic 回归模型	决策树模型	局部多项式回归
(Logistic Regression)	(Decision Tree)	(Local Polynomial Regression)
	随机森林模型	
	(Random Forest)	

例如，庞素琳[2]等用线性判别分析方法建立企业信用评价模型，对我国 2000 年 106 家上市公司进行分析，选取 4 个指标，预测准确率达到 95.28%。迟国泰[3]等基于某商业银行 1231 笔小企业贷款数据，选取 81 个指标，用 Probit 模型建立债信评级模型，预测准确率达到 60%。Milad Malekipirbazari [4]等基于社交借贷平台的数据，选择 23 个变量，用随机森林、SVM、Logistic 等模型预测信用风险，预测准确率随机森林 88%，SVM47%，Logistic49 %。Sidney Tsang [5]等用神经网络、决策树等模型，针对 4000 个欺诈行为数据，选择 10 个变量进入模型，准确率分别达到 84.7%、96.7%。陈为民[6]根据客户消费行

为数据，使用多元自适应样条回归建立信用欺诈监测模型，用长沙某银行的2000条信用卡数据做实证研究，每个客户有15个属性，预测准确率达到83.91%。

这些模型较易理解，能较准确地预测违约状态。然而，已有的运用这些模型进行风险预测的研究，多基于数据量一万以内、变量不超过100的数据，对于数据结构复杂的高维数据，讨论并不充分。而在处理大批量的高维数据时，这些模型都存在各自的缺点。实际数据中，解释变量与违约状态可能并非线性或广义线性关系，广义线性模型无法处理非线性数据，容易欠拟合，一般准确率不高。机器学习模型虽然能较准确的预测违约状态，却不能得出相应的显示表达式，可解释性较差[7]。进一步，与传统的机器学习算法相比，神经网络、支持向量机在数据样本很多时运算效率较低；单个决策树模型相比于随机森林模型精确度较低。非参数模型虽然灵活且强大，但往往需要更多的数据、更长的训练时间，而结果是更容易过拟合，解释性更差。

本文基于一份我国2019年某家商业银行的客户信息数据，这份数据包含80,000条数据，986个变量，近千维的数据在已有的文献中是鲜有存在的，如果直接使用上述模型方法，不仅可能由于算法过于复杂无法得出模型结果，进入模型的变量过多还会导致模型的稳定性很差，可能引发“维度灾难”[8]。数据是建立模型的基础，对于高维复杂数据，数据预处理是提高模型稳定性、提高模型拟合精度的重要环节。因此本文首先对原始数据进行筛选、转换，通过主成分分析、重编码等方式，大大降低数据维数。结合数据样本的变量类型以及样本大小，决定采用Logistic回归和随机森林算法构建信用风险预测模型，为银行控制信用风险提供科学依据。本文的模型方法适应大样本，稳定性好、可推广、运算效率高。同时，模型对信用风险的预测效果好，结果显示两种模型预测准确率分别达到75%和85%，相较于张婷婷(2017)用Logistic回归模型评估个人信用评分，预测违约状态准确率为67.62% [9]，张亚琴(2019)基于集成学习的方法研究信用风险预测，随机森林模型预测准确率为77.1% [10]，本文建立的模型预测准确率提高。且Logistic回归模型与随机森林模型相结合，既发挥Logistic回归的可解释性优势，也发挥随机森林的高准确率优势。

2. 数据概述

本文所分析的数据是来自某家商业银行的客户信息[11]，共包含80,000条数据，986个变量，其中数值型变量944个，字符型变量42个。数值型变量经整理后，根据含义划分为17组，见表2，字符型变量见表3。

Table 2. Numerical explanatory variable
表2. 数值型解释变量

变量名称	变量个数	变量具体情况
个人业务基本情况	110	77个连续型，32个离散型，1个取值常数
存款及本外币持有额	91	55个连续型，21个离散型，15个取值常数
柜台业务	112	65个连续型，41个离散型，6个取值常数
网银业务	88	46个连续型，30个离散型，12个取值常数
电话业务	71	24个连续型，25个离散型，22个取值常数
手机银行业务	60	34个连续型，26个离散型
网络银行业务	20	14个连续型，6个离散型
自助设备业务	68	40个连续型，16个离散型，12个取值常数
乐收银POS机业务	48	30个连续型，14个离散型，4个取值常数

Continued

本行 POS 机业务	40	25 个连续型，10 个离散型，5 个取值常数
它行 POS 机业务	40	27 个连续型，10 个离散型，3 个取值常数
其他业务	22	14 个连续型，6 个离散型
大额业务	32	4 个连续型，4 个离散型，24 个取值常数
信用卡业务	26	10 个连续型，10 个离散型，6 个取值常数
定期存款业务	20	13 个连续型，7 个离散型
理财产品业务	33	26 个连续型，7 个离散型
基金业务	63	49 个连续型，12 个离散型，2 个取值常数

Table 3. Character explanatory variable
表 3. 字符型解释变量

变量名称	变量个数	变量具体情况
客户号	1	8 位字符
开户机构	1	4 位编码
证件类型	1	4 位编码
性别	1	取值 1、2
客户价值等级	1	取值 A、B、C
职业	1	取值 1-5
是否型变量	21	包括是否有欧元账户、是否有澳元账户、是否薪资理财等
持有标志型变量	15	包括持有活期产品标志、持有定期存款标志、个贷标识等

3. 建模方法

本批数据变量数量庞大，不仅大大增加了计算的负担，而且信息重复导致变量间存在共线性，对后续建立模型分析会造成严重后果。因此，首先对数据预处理，在较完整保留原始变量所含信息的基础上进行降维，同时尽量消除变量间的共线性。

使用 SAS 软件进行统计分析。把违约状态作为响应变量，取值为 1 表示可能违约，取值为 0 表示不会违约，以经过预处理的变量作为指标，划分数据集为训练集和测试集，分别建立 Logistic 回归模型和随机森林模型。最后，在测试集应用两种模型预测违约可能，用受试者工作特征(receiver operating characteristic curve, ROC)曲线下面积(area under the curve, AUC)和预测准确率比较两种模型的预测效果。

3.1. 数据预处理

3.1.1. 对数值型变量进行分组主成分分析

运用主成分分析法对数值型变量降维，该方法在保证一定方差贡献率的情况下，既能实现降维，还能消除变量间的相关性。但是如果对所有 944 个变量做主成分分析，会存在两个问题：一、计算损耗巨大，需要对 8 万条数据求 944 × 944 的相关系数矩阵，并进行特征根分解，计算量过于庞大；二、可解释性差，每一个主成分都是 944 个原始变量的线性组合，变量混杂严重，主成分缺乏可解释性。

因此，考虑牺牲一定的变量不相关性，增强主成分的可解释度和计算效率，采用“分组主成分分析”

的方法进行变量降维处理。具体来说，由于之前已将 944 个数值型解释变量根据变量名进行了分组，那么有理由认为不同组的变量之间的相关性较微小，而组内变量的相关性较明显，因此考虑在每个变量分组内进行组内变量的主成分分析，最后选取主成分替代该类别的原始变量，不仅大大降低了计算量和混杂性，而且也容易解释主成分的含义。

3.1.2. 对字符型变量进行筛选

根据变量含义，以及对部分变量使用 proc freq 做相关性分析，通过删除、整合重编码的方法，对字符型变量进行筛选。

3.2. Logistic 回归预测模型

① 划分数据集、抽样：从 80,000 个数据中，按照正例数据(可能违约)、负例数据(不会违约) 2:1 的比例抽取 20,000 个数据作为训练集，其余数据作为测试集。② 在训练集上构建 Logistic 函数：用 proc logistic，用 class 语句将字符型分类变量按照其取值个数编码，设置哑变量，将设置好的哑变量以及主成分作为模型的输入自变量。利用逐步回归的方法选择变量，进出模型的显著性水平分别设置为 0.05 和 0.15。③输出结果：输出模型系数的估计值，以及测试集中每个数据违约概率的估计值。

3.3. 随机森林预测模型

使用 proc split 以及 proc surveyselect 构建随机森林。① proc split 构造决策树：设置特征选择准则为基尼系数(Gini)，每个内部节点的最大分支树为 5，最大深度 20，每个叶节点的最小规模为 50，目标变量为违约状态，将生成的评分规则输出为外部 txt 文件。② 划分数据集：按照 3:1 的比例划分训练集和测试集。③ 抽取变量和样本：使用 proc suveryselect 为每一棵树随机抽取 30 个数值型变量、6 个字符型变量、约 10%负例数据和 80%正例数据，利用宏循环反复调用打分文件。④ 构造森林，打分预测：最终构建出随机森林，对测试集进行打分和预测，将所有树输出的预测概率值的平均值，作为最后的预测结果输出。

3.4. 验证模型

对测试集数据，使用 SAS 官网上提供的 rocplot 宏包，绘制 ROC 曲线[12]，得到 AUC 面积；同时可选择最优截断点，由最优截断点预测用户的违约状态，用 proc freq 得混淆矩阵，进而得到预测准确率。通过 AUC 面积和预测准确率评价模型预测的优劣。

$$\text{预测准确率} = \frac{\text{正确预测到的正例数}}{\text{实际正例总数}}$$

4. 数据结果分析

4.1. 变量预处理

对数值型和字符型变量分别进行预处理，986 个原始变量(944 个数值型+42 个字符型)最终保留 142 个(123 个数值型+19 个字符型)，大幅降低维数，为后续建立模型做准备。

4.1.1. 数值型变量

944 个数值型变量，删除所有取值为常数的变量后，还剩 832 个，按照变量含义可分为 17 组，使用 proc princomp 对每组变量分别进行分组主成分分析，保证累积方差贡献率在 70%左右，最终总共从原始变量中挑选出 123 个主成分，作为后续预测模型的指标。表 4 展示变量处理前后的情况。

Table 4. Before and after variable processing
表 4. 变量处理前后示意

变量类别	原始变量个数 (已删除常数取值变量)	选取主成分个数	累计方差贡献率
个人业务基本情况	109	17	70.55%
存款及本外币持有额	76	7	66.21%
柜台业务	106	17	68.77%
网银业务	76	8	67.82%
电话业务	49	5	70.38%
手机银行业务	60	6	70.00%
网络银行业务	20	4	66.39%
自助设备业务	56	7	68.27%
乐收银 POS 机业务	44	5	68.91%
本行 POS 机业务	35	4	68.31%
它行 POS 机业务	37	6	66.37%
其他业务	22	4	66.71%
大额业务	8	2	94.27%
信用卡业务	20	4	65.10%
定期存款业务	20	8	67.08%
理财产品业务	33	6	71.97%
基金业务	61	13	68.24%

4.1.2. 字符型变量

42 个字符型变量，经过以下处理，最终整合选取 19 个。

1) 删除 6 个：取值唯一的 4 个，包括 2 个是否型变量、2 个持有标志型变量；人为判断影响不大的 2 个，包括开户机构、证件类型。

2) 选择性保留 20 个：部分重要变量直接保留，包括客户号、性别等；取值相同、含义包含的变量，持有国债标志和持有凭证式国债标志，仅保留一个；使用 prop freq 做卡方检验，含义相近、不相互独立的变量，3 个持有钱生钱理财产品标志的变量、4 个持有基金标志的变量，分别仅保留一个。

3) 整合重编码 16 个：把 16 个原始的持有卡种信息变量转变为 5 个自定义变量。原始变量包括表 2 中的 15 种卡，以及“最高卡级别”变量，自定义变量分别为是否借记卡、是否信用卡、是否国际卡、是否国内卡、信用卡等级。前四个自定义变量取值为 0、1，即 1 为是，0 为不是，信用卡等级是按照卡种额度赋予的排序值。原变量重编码后的取值见表 5，信用卡等级标准见表 6。

Table 5. User defined variable value after the original variable is recoded
表 5. 原始变量重编码后的自定义变量取值

原始变量	自定义变量				
卡种	借记	信用	国际	国内	信用卡等级
欧元卡	1	0	1	0	0
澳元卡	1	0	1	0	0
美元卡	1	0	1	0	0
薪资理财卡	1	0	0	1	0
商务卡	1	0	0	1	0
国际普卡	0	1	1	0	1
国际银卡	0	1	1	0	2
国际金卡	0	1	1	0	4
国际钻石卡	0	1	1	0	5
金普卡	0	1	0	1	2
标准白金卡	0	1	0	1	3
豪华白金卡	0	1	0	1	4
白金理财卡	0	1	0	1	3
钻石卡	0	1	0	1	5
无限卡	0	1	0	1	6

Table 6. Credit card type and grade and quota
表 6. 信用卡卡种–等级–额度对照表

信用卡	等级	额度
借记卡	0	0
普卡/国际普卡	1	5 千~1 万
金卡/国际银卡	2	1~5 万
普通白金卡	3	5~10 万
豪华白金卡/国际金卡	4	10~30 万
钻石卡/国际钻石卡	5	30~100 万
无限卡	6	100 万以上

4.2. 两种模型的预测结果及效果比较

Logistic 回归预测模型最终选择了 69 个变量，部分系数估计值见表 7，从表中可以看出，当一名客户近期从柜台转出金额、或信用卡消费增加、或购买更多理财产品，他的违约概率会增加，而客户价值等级为 A 以及职业为 1 的客户，违约概率会减少。随机森林预测模型最终构建了 100 棵决策树，挑选其中一棵树，模型中部分变量重要性见表 8。表中“节点个数”表示在这棵树的所有判断节点中，以该变量的取值作为判断准则的节点个数，“基尼重要度”显示变量对预测结果的重要程度，从表中可以看出，近期 POS 机交易金额、近期资产总计、近期资产增加值，对预测违约状态影响较大。使用 SAS EM 还可

以绘制该树划分客户的具体流程图，相比于 Logistic 回归，树方法能提供更直观易懂的判别流程，该树部分判断流程见图 1。

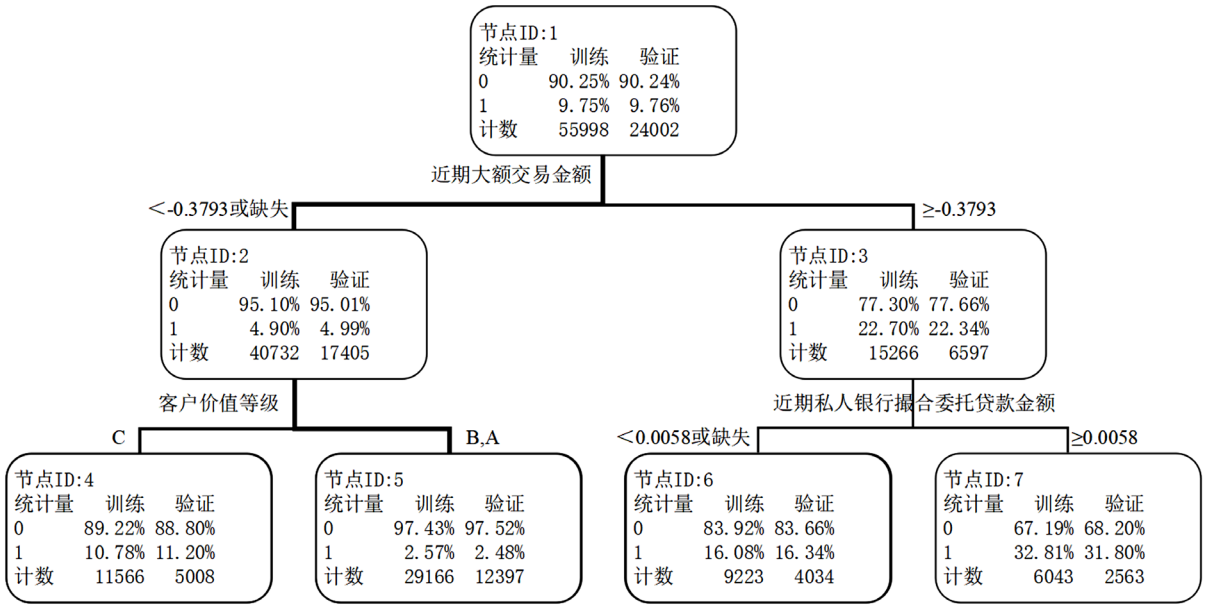

Figure 1. Part of the specific process of decision tree in random forest
图 1. 随机森林中某一棵决策树的部分具体判别流程

Table 7. Analysis results of some variables in logistic regression model
表 7. Logistic 回归模型部分变量的分析结果

变量	估计	标准误差	Wald 卡方	Pr≥卡方
近期柜台转出交易金额	0.01029	0.0182	32.1088	≤0.0001
近期信用卡人民币消费增加值	0.3418	0.0774	19.5105	≤0.0001
近期理财产品交易金额增加值	0.325	0.0277	137.3748	≤0.0001
客户价值等级(A)	−2.4561	0.1868	172.9554	≤0.0001
职业(1)	−0.4217	0.1549	7.4068	≤0.0001

注：形如“职业(1)”的变量名称表示该分类变量的取值为括号里的值。

Table 8. Analysis results of some variables in random forest model
表 8. 随机森林模型部分变量的分析结果

变量	节点个数	基尼重要度
近期 POS 机交易金额	4	1
近期资产总计	7	0.435939617
近期资产增加值	12	0.434289408
近期银保通金额	7	0.347330495
近期本币新增余额	7	0.324201205

将预测模型应用于测试集，使用 rocplot 宏包，绘制 ROC 曲线，由 Youden 指数选择最优截断点分别为 0.37283、0.49551，即 Logistic 回归预测违约的概率超过 0.37283，随机森林预测违约的概率超过 0.49551，就认为该客户可能违约。作为对比，又建立决策树预测模型，三种模型的混淆矩阵见表 9。从表中可以看出，对于 Logistic 回归模型，测试集 60,000 个数据中，预测和实际都可能违约的数据有 2514 个，占全体实际可能违约数据的 75.39%，即为本文定义的预测准确率。对于随机森林模型，测试集 20,000 个数据中，预测和实际都可能违约的数据有 1507 个，预测准确率 84.95%。对于决策树模型，测试集 20,000 个数据中，预测和实际都可能违约的数据有 1380 个，预测准确率为 77.79%，略高于 Logistic 回归预测模型，但其解释性较差，而随机森林预测模型对比决策树模型在精确性上占据明显优势。两种预测模型的 ROC 曲线比较见图 2。最终得到，Logistic 回归预测模型，AUC = 0.847，预测准确率为 75%；随机森林预测模型，AUC = 0.848，预测准确率为 85%。二者 AUC 面积差别不大，但随机森林模型的预测准确率较 Logistic 回归模型有显著提升。

Table 9. Confusion matrix of three models
表 9. 三种模型的混淆矩阵

			训练集		测试集	
			真实值		真实值	
			不会违约	可能违约	不会违约	可能违约
Logistic 回归模型	预测值	不会违约	9980	1734	43060	820
		可能违约	3003	5283(75.29%)	13606	2514(75.39%)
随机森林模型	预测值	不会违约	38628	813	12844	267
		可能违约	15342	5217(86.52%)	5382	1507(84.95%)
决策树模型	预测值	不会违约	36742	1239	13453	394
		可能违约	17228	4791(79.45%)	4773	1380(77.79%)

注：百分比为预测准确率。

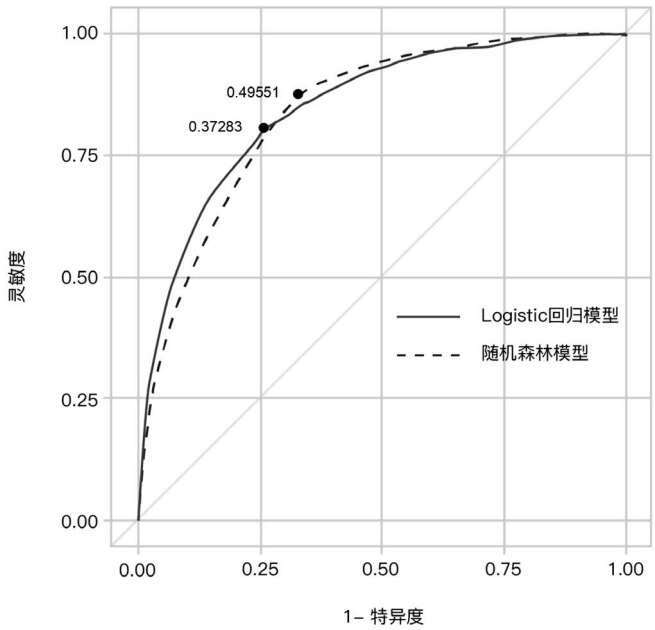

Figure 2. ROC curve of two models
图 2. 两种模型 ROC 曲线

5. 讨论

目前商业银行的利润仍主要来自信贷业务，准确及时地预测信用风险，是有效规避损失、保证银行稳健发展的重要前提。建立预测模型时，在采集信用资料阶段，应尽可能全面地考虑可能影响违约状态的因素，如基本信息姓名、性别、证件号码等，交易信息存转款、信用卡、理财产品等。在以往的研究中，往往根据经验选取若干重要指标，这是受数据、技术的限制所致。而在大数据时代，可以广泛地考虑所有可能的指标变量，为预测提供多方面的信息。想要从大数据中挖掘内在规律进行预测，首先需要对纷繁复杂的数据进行预处理。主成分分析法是一种考察多个变量间相关性的多元统计方法[13]，从大量原始变量中导出保留原始变量重要信息、又彼此互不相关的少量主成分。但是新生成的主成分往往不易解释，含义模糊。本文首先对原始变量分组，再在组内分别用主成分降维，对于各主成分，从组含义出发解释，保留了主成分的可解释性。在数据预处理的基础上，开展后续工作，建立预测模型。

Logistic 回归模型方法简单、训练高效、预测能力较强，由变量系数可知变量的影响程度。但不可否认，Logistic 回归也存在以下不足：预测结果对变量间的共线性敏感；需要设置很多哑变量，增加了变量维数和计算复杂度；逐步回归选择变量时，数据规模需有所控制，否则在有限的计算机内存下可能不收敛；对于银行的信用评估业务，如果拒绝贷款申请，需要给出一个准确的理由，让客户知道被拒绝的具体原因，虽然 Logistic 回归的系数有一定意义，但被拒绝或被接受的理由仍然不明确。而随机森林作为一种集成思想机器学习模型，不受变量间共线性的影响，无需设置哑变量，在特征选择标准方法为 GINI 系数时，对字符型变量和数值型变量都可以分析。最重要的是，树方法流程图的形式，可以为客户提供直观易懂的理由说明。

本文建立了 Logistic 回归预测模型和随机森林预测模型，比较了两种模型对银行信用风险的预测效果。结果显示，Logistic 回归模型 AUC = 0.847，预测准确率为 75%；随机森林模型 AUC = 0.848，预测准确率为 85%。随机森林预测模型的曲线下面积略优于 Logistic 回归预测模型，二者区别不大，而其预测准确率明显更优。

Logistic 回归输出结果显示，对预测概率正向影响最大的变量依次为高信用卡级别、职业取值 3/5、客户价值等级 B、个人业务基本情况、存款及本外币持有额，对预测概率负向影响最大的变量依次为低信用卡级别、客户价值等级 AC、职业取值 1/2/4、信用卡业务。随机森林模型变量重要性分析中，排名前五的变量依次为个人业务基本情况、客户价值等级、存款及本外币持有额、大额业务、柜台业务。两种模型的分析结果相似，说明模型较稳定，预测结果可靠。结果提示，客户的信用卡级别、职业、客户价值等级、个人业务基本情况、存款及本外币持有额对违约风险影响较大。

6. 结论

综上所述，随机森林模型对信用风险的预测效果较好，实际应用中，可以作为 Logistic 回归预测模型的有益补充，充分发挥两种模型的优越性。银行在预测信用风险时，在众多基本信息及交易信息中，应该着重注意信用卡级别、职业、客户价值等级、个人业务基本情况、存款及本外币持有额的状况，如发现这些指标异常，应尽早采取相应的干预措施，规避信用违约可能带来的损失。

参考文献

[1] 赵晓菊. 信用风险管理[M]. 上海: 上海财经大学出版社, 2008(5): 10-22.

[2] 庞素琳, 王燕鸣. 判别分析模型在信用评价中的应用[J]. 南方经济, 2006(3): 113-119.

[3] 迟国泰, 张亚京, 石宝峰. 基于 Probit 回归的小企业债信评级模型及实证[J]. 管理科学学报, 2016, 19(6): 136-156.

[4] Milad, M. and Vural, A. (2015) Risk Assessment in Social Lending via Random Forest. *Expert Systems with Applications*, **42**, 4621-4631. https://doi.org/10.1016/j.eswa.2015.02.001

[5] Tsang, S., Koh, Y.S., Dobbie, G., *et al.* (2014) Detecting Online Auction Shilling Frauds Using Supervised Learning. *Expert Systems with Applications*, **41**, 3027-3040. https://doi.org/10.1016/j.eswa.2013.10.033

[6] 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[D]: [博士学位论文]. 长沙: 湖南大学, 2009.

[7] 任晓萌. 基于逻辑样条回归的信用风险预测模型[D]: [硕士学位论文]. 大连: 大连理工大学, 2019.

[8] 王海雷. 面向高维数据的特征学习算法研究[D]: [博士学位论文]. 合肥: 中国科学技术大学, 2019.

[9] 张婷婷. Logistic 回归及其相关方法在个人信用评分中的应用[D]: [硕士学位论文]. 太原: 太原理工大学, 2017.

[10] 张亚琴. 基于集成学习的信用风险预测研究[D]: [硕士学位论文]. 兰州: 兰州大学, 2019.

[11] 2019SAS 大赛官方. “扬子江新金融杯”2019 年 SAS(中国)高校数据分析大赛暨首届国际邀请赛[Z], 2019.

[12] 李太顺, 刘沛. ROC 曲线绘制和曲线下面积比较的 SAS 宏包[J]. 中国卫生统计, 2018, 35(2): 302-304+309.

[13] 盖曦, 乔龙威. 基于主成分分析法的我国商业银行系统性风险的度量[J]. 长沙大学学报, 2013, 27(5): 100-103.