“东证期货杯”全国大学生统计建模大赛

**基于Lasso-Logistic的个人信用**

**违约预测模型**

参赛单位 中南财经政法大学

指导老师 徐映梅教授

参赛队员 李坤 武山 张果 李明堋 康振英

队伍编号 343

所选题目 选题2

目录

[表格和插图清单 III](#_Toc512378321)

[表目录 III](#_Toc512378322)

[图目录 III](#_Toc512378323)

[摘要 1](#_Toc512378324)

[一、引言 2](#_Toc512378325)

[（一）问题提出 2](#_Toc512378326)

[（二）研究内容及章节安排 3](#_Toc512378327)

[（三）本文的创新之处 4](#_Toc512378328)

[二、Lasso-Logistic模型介绍 4](#_Toc512378329)

[（一）Lasso理论介绍 5](#_Toc512378330)

[（二）Lasso-Logistic模型 6](#_Toc512378331)

[（三）调和参数λ的选择 7](#_Toc512378332)

[三、个人信用违约预测模型 7](#_Toc512378333)

[（一）数据来源 7](#_Toc512378334)

[（二）指标体系建立 8](#_Toc512378335)

[（三）数据处理和变量设置 11](#_Toc512378336)

[（四）模型建立与求解 15](#_Toc512378337)

[（五）模型预测与分析 20](#_Toc512378338)

[（六）原始数据测试集预测 21](#_Toc512378339)

[四、结论和建议 21](#_Toc512378340)

[（一）结论 21](#_Toc512378341)

[（二）建议 22](#_Toc512378342)

[参考文献 23](#_Toc512378343)

[附录 24](#_Toc512378344)

表格和插图清单

表目录

[表3-1部分变量优势比分析 8](#_Toc512376692)

[表3-2指标体系设置 9](#_Toc512376693)

[表3-3地区和婚姻状态样本量分布 12](#_Toc512376694)

[表3-4全部数据缺失情况 12](#_Toc512376695)

[表3-5人口特征缺失数据插补结果 13](#_Toc512376696)

[表3-6变量调整说明 15](#_Toc512376697)

[表3-7模型参数估计（含欺诈记录变量） 16](#_Toc512376698)

[表3-8模型参数估计（剔除欺诈记录变量） 18](#_Toc512376699)

[表3-9调整前后模型预测效果 18](#_Toc512376700)

[表3-10模型预测效果对比 20](#_Toc512376701)

[附录：模型指标体系 24](#_Toc512376702)

图目录

[图3-1部分指标数据分布直方图 14](#_Toc512376720)

[图3-2模型误差变化情况（含欺诈记录变量） 16](#_Toc512376721)

[图3-3模型误差变化情况（剔除欺诈记录变量） 17](#_Toc512376722)

[图3-4神经网络拓扑图 19](#_Toc512376723)

[图3-5 ROC曲线 20](#_Toc512376724)

基于Lasso-Logistic的个人信用违约预测模型

摘要

互联网金融在加快业务处理速度、提高资金配置效率的同时也存在着信用风险和用户欺诈等问题，通过构建合理的个人信用评估指标体系对借款人违约情况进行精准预测具有重要意义。

本文基于组委会提供的人口特征和信用记录数据，从人口特征、资产负债状况、信用历史和信用行为四个方面选取了27个指标，在对变量缺失值和异常值进行处理后，将30000个样本中的1875个违约客户样本全部抽取,在非违约客户样本中使用聚类分析和分层抽样方法抽取1875个样本，得到样本量为3750的样本集，然后从中随机抽取80%的样本组成训练集，剩余样本作为测试集。使用Lasso-Logistic和BP神经网络算法建立个人信用违约预测模型，预测测试集中750个样本的违约情况，并使用借款人欺诈记录变量对Lasso-Logistic模型的预测结果进行调整。

通过对比两个模型的预测效果，可以发现，使用欺诈记录变量调整的Lasso-Logistic模型兼具了较高的预测准确率和较好的外推性，能够对样本中违约客户进行有效识别，该模型对训练集和测试集样本总体预测准确率分别为79.47%和75.33%，对应于BP神经网络模型的86.17%和73.37%，可以发现该模型具有更高的样本外预测准确率和更好的外推性；特别地，使用欺诈记录变量对Lasso-Logistic模型预测结果进行调整后，模型对违约客户的预测准确率提高到81.32%，比BP神经网络模型高13个百分点。

关键词：指标体系；众数插补法；聚类分析；Lasso-Logistic模型

一、引言

（一）问题提出

以大数据、深度学习、人工智能等为代表的互联网技术对金融领域的业务和商业模式产生了巨大影响，互联网金融机构能够突破时间和地域的约束，为有融资需求的客户提供更快捷的金融服务，使互联网金融发展成为我国主流商业模式之一，并呈现出多种多样的业务模式和运行机制。

但是，互联网金融在加快业务处理速度、提高资金配置效率的同时也存在着信用风险和用户欺诈等问题，需要互联网金融机构对个人信用进行评估，提高风险控制水平。但是现阶段的制度对风险的控制能力相对较弱，同时我国还未建立起完善的社会信用体系，导致国民违约成本极低，国民信用意识淡薄，网络借贷违约事件层出不穷。因此完善制度和提高风险控制水平成为我国互联网金融关注的核心，尤其是对信用违约风险的控制更是学界和金融行业关注的重点。

随着网民数量的逐年增加以及互联网金融的蓬勃发展，个人信用风险表现出的多面性和复杂性更加突出，对个人信用风险评估理论与技术的要求也越来越高，评估难度越来越大。互联网金融信用风险较高的一个重要原因是存在信息非对称，金融机构缺少客户的完整征信信息，使得其对客户的信用风险评估存在偏差，导致客户评级结果不准确。因此，完善信用风险管理制度的核心：一方面是从互联网现有复杂信息中提取出与借款人信用水平密切相关的信息，构建简洁有效的信用评估指标体系；另一方面是通过建立科学合理的信用违约预测模型，更加准确地预测个人未来的信用表现，降低授信成本。

现有学者对以上两方面均有研究，在指标体系构建方面，除了参考已有研究设置指标体系外，比较流行的指标变量选取方法有交叉列联表、t检验、逐步回归和lasso等方法。段昊（2012）[1] 结合美国FICO个人信用评价体系和我国商业银行个人信用评估方法构建具有P2P网贷平台特点的评价指标体系；孙燕（2012）[3]认为以逐步 （Stepwise）回归为代表的子集变量选择法，需要进行多次重复计算，算法复杂度是np，当数据变量众多时，该方法往往就不适用了；方匡南（2014）[4]提出在构建指标体系中使用lasso算法同时进行变量选择和参数估计，筛选出的指标体系效果良好；张国政等（2015）[2]根据统计学中的交叉列联表初步分析了自变量和因变量的关联性。

个人信用风险评估模型包含判别分析、logistic等为代表的经典统计方法和以支持向量机、神经网络、决策树等机器学习理论为代表的新型信用风险评估模型。其中曲秋实（2010）[5]认为logistic模型在预测商业银行个人信用风险评估中具有很好的有效性，能够为商业银行判定借款人的信用风险程度、降低不良贷款率提供准确性较好的客观依据；罗方科等（2017）[6]运用光大银行某分行样本数据，构建二分类logistic信用风险评估模型，评估商业银行互联网金融个人小额贷款信用风险。他认为使用logistic模型衡量中小企业信用风险可以取得最佳效果，且该模型限制条件较少，操作便利，且具有较高的预测性。Lee和Sung(2000) [7]比较logistic模型和神经网络模型后认为，logistic模型稳定性更强，更适合分析消费信用违约问题。

本文通过对组委会提供的人口特征和信用记录数据进行分析，通过对原始变量的重组、拆分、汇总和信息提取（如从身份证中提取地区和性别信息），从人口特征、资产负债状况、信用历史和信用行为四个方面选取了27个变量构成个人信用评估指标体系，通过比较Lasso-Logistic算法和BP神经网络算法在个人信用评估模型的优劣，考虑将Lasso-Logistic引入到我国互联网金融个人信用评估模型中。

（二）研究内容及章节安排

本文从组委会提供的原始数据出发，通过对原始变量进行组合、拆分和汇总，从人口特征、资产负债状况、信用历史、信用行为四个方面构造了27个变量组成个人信用评估指标体系，在对缺失值和异常值进行处理后，使用平衡后的样本集建立Lasso-Logistic模型和BP神经网络模型，对测试集违约状况进行预测，并使用欺诈记录变量对Lasso-Logistic模型的预测结果进行调整（后文简称为“调整的Lasso-Logistic模型”），比较两模型的预测效果，最终选择调整的Lasso-Logistic模型对原始数据中10000个借款人的违约状况进行预测。

本文的章节安排如下：

第一部分介绍了本文要研究的问题和相关文献综述，说明了本文的创新之处。

第二部分介绍了本文所使用的Lasso-Logistic模型的基本原理和模型参数的选择方法。

第三部分构建个人信用违约预测模型。首先设置指标体系，本文通过对组委会提供的原始数据进行重新组合，选取和构建了27个变量作为指标体系；然后处理缺失值和异常值，对样本进行平衡处理得到训练集和测试集数据，之后分别使用Lasso-Logistic方法和BP神经网络方法建立模型，对测试集样本进行预测并与真实值比较；选出预测效果较好的模型，对10000个违约状况未知的样本进行预测。

第四部分根据模型结果得出结论并提出相关建议。

（三）本文的创新之处

为了对借款人信用违约状况作出尽可能准确地预测，本文运用数据挖掘思想并结合前人研究方法，在数据清洗、指标体系设置、模型建立和优化等方面进行了针对性的处理，我们认为这些处理对优化指标体系、提高预测准确率具有积极作用，也是本文的亮点和可取之处。本文的创新之处主要有以下几点。

1. 构建了新的个人信用评估指标体系。本文从已有文献、变量特征、经济含义三个层次出发，除了直接选取已有变量外，还通过组合、拆分、汇总等方法，从四个方面构建了个人信用评估指标体系，其中一些变量平时不被关注，却对违约状况有着显著影响，如借款人所属地区、信贷审批查询次数等变量，也被纳入指标体系。

2. 采用众数插补法对人口特征缺失数据进行插补。借款人人口特征信息中，除收入变量缺失值较多外，其余变量缺失值所占比例约为4.21%，因此本文参考K近邻分类算法的思想，使用众数插补法对借款人人口特征信息缺失值进行插补。

3. 使用聚类分析和分层抽样方法平衡样本数据。在处理训练样本不均衡问题时，首先对非违约样本进行聚类分析，然后使用分层抽样方法抽取样本，使违约样本量和非违约样本量比例为1：1，保证训练集中能够包含不同类型的借款人特征。

4. 使用Lasso-Logistic模型同时进行变量降维和参数估计。由于本文构建的指标体系包含较多变量，相对于BP神经网络模型，我们选择使用Lasso-Logistic模型同时进行变量降维和参数估计，在使用CPU进行计算的情况下，Lasso-Logistic模型能更好地兼顾预测准确度和运算效率。

5. 对特殊变量使用特殊方法进行分析。我们发现指标体系中欺诈记录变量和违约状况存在“确定性单向因果关系”，即当借款人存在欺诈记录时，其违约状况确定为违约，反之则不一定成立。使用此变量直接建立模型时，所得到的模型并不理想，因此我们将该变量从模型中剔除，直接使用该变量对预测结果进行调整，结果表明此方法不仅有利于模型解释，而且能够有效提高预测准确率。

二、Lasso-Logistic模型介绍

（一）Lasso理论介绍

Tibshirani（1996）[8]提出Lasso方法的动机来源于Breiman（1995）[9]的非负绞除法（Non-negative Garrote），该方法的目标函数可以概括为式（1）的形式。

(1)

式（1）中非负因子的加和受常数t控制，该算法思想是从最小二乘估计法出发，通过对部分非负因子进行控制从而对估计的回归系数进行压缩。Breiman（1995）基于大量数据模拟后得出结论：相对于子集选择方法，非负绞除法的预测误差相对较小，并且由于非负绞除法去除了模型中很多接近0但非0的变量，从而增强了模型的解释性；与岭回归方法比较，在高维数据分析中，非负绞除法对高维变量进行了压缩，简化了计算过程并且增强了重要变量的解释性。Tibshirani（1996）[8]认为非负绞除法的缺点是其运算结果依赖于最小二乘估计的符号和数值大小，并且在过度拟合和变量存在高度相关的情况下，由于最小二乘估计效果不好从而影响预测准确性。相比之下，Lasso避免了非负绞除法的缺陷。

假设有数据变量,其中分别是解释变量和被解释变量的观察值。在一般的回归模型中，常常认为观察值彼此独立或者被解释变量在给定解释变量的条件下相互独立。同时，假设是标准化的，即。令，Lasso方法的估计量定义为

(2)

式中是调和参数，此时对一切的t有。为不失一般性，假定，这样就可以将式（2）整理为式（3）

(3)

对调和参数t的控制将会使回归系数总体变小。若令是回归参数的最小二乘估计值，，就会使一些回归系数缩小并趋于0，有些甚至会等于0。例如，当，计算的结果就会使不为0的回归系数的数目减少到大约为p/2个。上述式（3）的表达还可以用式（4）的惩罚函数的形式表达,即

(4)

式（4）的第一部分表示模型拟合的优度，第二部分表示参数的惩罚。调和系数越小，模型的惩罚力度就越小，保留的变量就越多；λ越大，模型的惩罚力度就越大，保留的变量就越少。AIC、BIC等子集选择方法是一个离散、无序的过程，变量被保留或者被删除，常常表现为方差较大，不能降低整个模型的预测误差，而Lasso方法是一个连续的、有序的过程，方差较小。Tibshirani 提出的Lasso 算法在模型变量选择时需要用二次规划方法求解。 Efrom等（2004）[10]认为，Tibshirani的求解方法比较复杂，他们提出了计算速度更快的最小角回归算法（Least Angle Regression，LARS），并利用该算法计算 Lasso参数路径

（二）Lasso-Logistic模型

Lasso方法主要应用于线性模型，其本质是在残差平方和上添加惩罚函数，在估计参数时，系数被压缩，部分系数甚至被压缩到0来实现模型选择。但是对于违约情况预测时，其因变量是二元离散取值，此时不能再利用线性回归模型，而应该使用Lasso-Logistic回归。

假设有独立同分布的观测值,其中分别是模型的解释变量和被解释变量，并且是二元离散数据变量，即，则Logistic线性回归模型的条件概率为：

(5)

其中，。Lasso-Logistic回归模型中的系数估计值由式（6）凸函数的极小值给定：

(6)

其中是对数似然函数，则式（6）中的可以写成式（7）：

(7)

Lasso-Logistic回归模型中的系数估计值可写成如式（8）的形式。

(8)

（三）调和参数λ的选择

Lasso-Logistic模型的变量选择，关键在于调和参数λ的选取，常用方法有Bootstrap、交叉验证、广义交叉验证法等，本文采用广义交叉验证方法确定罚参数λ的值，其具体算法如下：

若令，由此可定义广义交叉验证值GCV统计量为式（9）所示形式

(9)

显然，使交叉验证值GCV达到最小的罚函数为最优的罚参数λ（Lambda），则最优λ的估计值为式（10）所示的表达式。

(10)

三、个人信用违约预测模型

（一）数据来源

本文数据来源于组委会提供的40000个借款人的人口特征和信用记录数据，其中个人特征信息包括教育水平、婚姻状况、收入水平、本地籍和违约情况等信息，信用记录数据包括借款人的贷款相关信息、贷记卡和准贷记卡相关信息、借款人逾期情况和欺诈行为等信息。其中对于借款人违约情况，只有30000个样本违约情况已知，10000个样本违约情况未知，本文的目的是在充分挖掘现有信息的基础上，构建合适的指标体系，建立个人信用违约预测模型，并对这10000个借款人的违约情况进行预测。

（二）指标体系建立

1.基于文献研究的指标选择。本文在充分考虑互联网金融的特点和风险特征的基础上，结合前人研究的成果，通过对所提供数据进行分析，初步选取了性别、本地籍、教育水平、婚姻状态、收入水平、公积金、本月应还款金额、本月实还款金额8个变量。

2.根据变量特征选择指标。通过对原始数据信息进行整理，分析部分变量不同取值状态与违约状况之间的关系，部分优势比分析结果如表3-1所示。从表中我们可以发现投资（经营）性贷款笔数、消费性贷款笔数、住房性贷款笔数等变量不同取值对应的违约风险有所差异，其中欺诈记录变量与违约状况之间还可能存在“确定性单向因果关系”，即当借款人存在欺诈记录时，其违约状况确定为违约，反之则不一定成立。基于此，本文选取投资（经营）性贷款笔数、消费性贷款笔数、住房性贷款笔数、异常贷款账户数、异常贷记卡账户数、当前贷款逾期期数、当前贷记卡逾期期数、最近6个月平均贷款使用额度、最近6个月平均贷记卡使用额度、地区和欺诈记录共11个变量进入指标体系。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 表3-1部分变量优势比分析 | | | | | | | | | |
|  | 指标项 | 履约 | 违约 | 优势比 |  | 指标项 | 履约 | 违约 | 优势比 |
| 投资经营性贷款笔数 | 0-2 | 22167 | 1375 | 16.12 | 消费性贷款笔数 | 0-4 | 15799 | 997 | 15.85 |
| 3-5 | 568 | 16 | 35.5 | 5-9 | 5528 | 314 | 17.61 |
| 6-8 | 228 | 1 | 228 | 10-14 | 1427 | 59 | 24.19 |
| ≥9 | 339 | 2 | 169.5 | ≥15 | 548 | 24 | 22.83 |
| 住房性贷款笔数 | 0 | 20768 | 1282 | 16.20 | 异常贷款账户数 | 0-4 | 24690 | 1535 | 16.08 |
| 1 | 2466 | 110 | 22.42 | 5-9 | 6 | 0 | Inf |
| 2 | 67 | 1 | 67 | 10-14 | 2 | 0 | Inf |
| ≥3 | 1 | 1 | 1 | ≥14 | 1 | 0 | Inf |
| 地区 | 地区1 | 3822 | 767 | 4.98 | 欺诈记录 |  |  |  |  |
| 地区2 | 3396 | 619 | 5.49 | 0 | 25594 | 1104 | 23.1830 |
| 地区3 | 15130 | 489 | 30.94 | 1 | 0 | 510 | 0 |
| 地区4 | 5777 | 0 | Inf |  |  |  |  |
| 注：表中数据由作者对原始数据处理所得。 | | | | | | | | | |

3.根据数据经济含义选择指标。从变量的经济含义出发，本文选取了非信用担保贷记卡笔数、信用担保贷记卡笔数、贷记卡共享额度、还款进度、贷款审批查询次数、担保资格查询次数、信用卡审批查询次数、贷记卡透支度8个变量。

其中贷款审批查询、担保资格查询、信用卡审批查询都是银行进行的信用报告査询工作，如果同一借款人存在多条查询记录，说明银行对其较为关注，此人信用水平可能出现异常，违约风险相应提高。无论是信用担保还是非信用担保，贷记卡笔数越多，一方面体现了其较大的负债压力，另一方面也反映了银行对其信用的认可程度。贷记卡共享额度反映了金融机构对该借款人的信用评估水平，一般来说信用越好，共享额度越高。还款进度是借款人已还贷款金额和合同金额的比值，表示借款人当前的还款比例，还款进度越高，表明借款人未偿还的贷款比例越小，违约率可能越低。贷记卡透支度是借款人所有贷记卡已用额度加总与对应的账户总限额的比值，透支度越大表明当前资金缺口越大，违约风险可能越高。

基于以上三个角度，本文从人口基本特征、资产负债状况、信用历史、信用行为四方面选取了27个变量构成个人信用评估指标体系，完整的指标体系如表3-2所示。

|  |  |  |
| --- | --- | --- |
| 表3-2指标体系设置 | | |
|  | 变量名称 | 符号 |
| 人口特征 | 地区 | area |
| 性别 | sex |
| 本地籍 | local |
| 教育水平 | EDU\_LEVEL |
| 婚姻状态 | MARRY\_STATUS |
| 收入水平 | SALARY |
| 公积金 | HAS\_FUND |
| 资产负债状况 | 投资（经营）性贷款笔数 | loantype1 |
| 消费性贷款笔数 | loantype2 |
| 住房性贷款笔数 | loantype3 |
| 非信用担保贷记卡账户数 | guarantee\_C1 |
| 信用担保贷记卡账户数 | guarantee\_C2 |
| 本月应还款金额 | should\_money |

（续表3-2）

|  |  |  |
| --- | --- | --- |
| 信用历史 | 贷记卡共享额度 | sharelimit |
| 还款进度 | payschedule\_money |
| 贷款审批查询次数 | time1 |
| 担保资格审查次数 | time2 |
| 信用卡审批查询次数 | time3 |
| 欺诈记录 | Y\_FRAUD |
| 信用行为 | 异常贷款账户数 | errorloan |
| 异常贷记卡账户数 | errorcard |
| 本月实还款金额 | actual\_money |
| 当前贷款逾期期数 | overcyc\_loan |
| 当前贷记卡逾期期数 | overcyc\_card |
| 最近6个月平均贷款使用额度 | ave6\_loan |
| 最近6个月平均贷记卡使用额度 | ave6\_card |
| 贷记卡透支度 | used\_card |

其中，人口特征指标中，本地籍、婚姻状态、收入水平、公积金4个变量来自原始数据;教育水平是将原始数据中不同学历重新分组合并为专科及以下学历、本科学历、硕士及以上学历得到的;地区、性别数据来源于原始数据中的身份证序列，根据我国身份证号码编排规则，身份证号码第一位表示其所在的地理区域，倒数第二位表示性别，奇数表示男性，偶数表示女性。

资产负债状况指标中，投资（经营）性贷款笔数是每个借款人当前未结清的个人经营性贷款、个人助学贷款和农户贷款账户数目统计；消费性贷款笔数是对借款人当前未结清的个人汽车贷款、个人消费贷款和其他贷款账户数目统计；住房性贷款笔数是对借款人当前未结清的个人商用房（包括商住两用）贷款、个人住房贷款和个人住房公积金贷款账户数目统计；非信用担保贷记卡账户数表示持有人当期未结清的采用非信用担保方式（保证、抵押担保、其他担保、质押（含保证金）担保、组合（不含保证）担保）的贷记卡数。信用担保贷记卡账户数表示持有人当期未结清的采用信用/免担保方式的贷记卡数；本月应还款金额是对借款人所有账户（包括未结清贷款账户、未注销贷记卡和准贷记卡账户）的本月应还款进行加总得到。

信用历史指标中，贷款审批查询次数、担保资格审查次数、信用卡审批查询次数和欺诈记录来源于原始数据。还款进度数据来源于未销户贷记卡或未结清贷款中的已用额度和合同金额。共享额度是持卡人在所有借款机构的共享额度中位数，其中每个借款机构的共享额度是持卡人在该机构所有币种共享额度经汇率转化为人民币后的总额度。

信用行为指标中，异常贷款账户数和异常贷记卡账户数是每个借款人当前未结清贷款账户和未注销贷记卡账户的异常账户数统计；本月实还款金额由借款人所有账户（包括未结清贷款账户、未注销贷记卡和准贷记卡账户）的本月实还款加总得到，反映借款人的短期偿债能力；当前贷款逾期期数、当前贷记卡逾期期数、最近6个月平均贷款使用额度和最近6个月平均贷记卡使用额度由借款人当前所有未结清贷款账户和未注销贷记卡、准贷记卡账户对应原始数据加总得到。贷记卡透支度是借款人所有贷记卡已用额度加总与对应的账户额度总限额的比值，由于贷记卡可以预存额度，因此对于使用额度大于对应账户限额的贷记卡账户，其使用额度按照账户限额计算。

（三）数据处理和变量设置

1.变量类型

本文所用模型为Lasso-Logistic模型，被解释变量为客户的违约状况，该变量为二元离散数据变量，解释变量为上文构建的个人信用评价指标体系中所包含的变量。

反映借款人人口特征信息的7个变量中，性别、本地籍和公积金变量为0-1变量；教育水平和收入水平变量为顺序变量；地区和婚姻状态变量为分类变量。

反映借款人资产负债状况的6个变量中，本月应还款金额变量为连续型数值变量；其余变量均为离散型数值变量。

反映借款人信用历史信息的6个变量中，已用额度和贷记卡共享额度变量为连续型数值变量；欺诈记录为0-1变量；其余变量均为离散型数值变量。

反映借款人信用行为信息的8个变量中，异常贷款账户数、异常贷记卡账户数、当前贷款逾期期数和当前贷记卡逾期期数为离散型数值变量；其余变量均为连续型数值变量。

由于解释变量中地区变量和婚姻状态变量为分类变量，变量取值大小无实际意义，因此需要将其转换为虚拟变量放入模型。我们在已知违约状况的30000个样本数据中得到不同地区和婚姻状态的样本量分布，如表3-3所示。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 表3-3地区和婚姻状态样本量分布 | | | | | | | | | |
|  | 地区 | | | | 婚姻状态 | | | | |
| 变量取值 | 1 | 2 | **3** | 4 | 离婚 | 离异 | 丧偶 | 未婚 | **已婚** |
| 样本量 | 4589 | 4015 | **15619** | 5777 | 314 | 719 | 41 | 9566 | **19199** |

由上图可以看出，样本数据中处于3地区的人最多，大多数人的婚姻状态为已婚，由Lasso-Logistic模型参数估计表达式

我们可以发现，解释变量观测值中观测值为0的个数越多，参数估计的计算量越小，模型运行效率越高，因此我们以地区取值3和已婚状态为基准设置虚拟变量，设置area1、area2、area4为地区虚拟变量，marry1、marry2、marry3、marry4为婚姻状态虚拟变量。

2.缺失值与异常值处理

①缺失值处理

由于原始数据存在缺失值，因此需要对缺失值进行插补。本文所用数据有两类缺失值，一类是借款人相关人口特征信息，另一类是借款人相关信用记录数据。全部数据缺失情况如表3-4所示。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 表3-4全部数据缺失情况 | | | | | |
| 变量名称 | 个数 | 变量名称 | 个数 | 变量名称 | 个数 |
| area | 1667 | loantype3 | 6261 | sharelimit | 160 |
| sex | 0 | payschedule\_money | 6269 | errorcard | 150 |
| local | 276 | should\_loan | 4406 | ave6\_loan | 6267 |
| EDU\_LEVEL | 5593 | actual\_loan | 4406 | ave6\_card | 163 |
| MARRY\_STATUS | 233 | should\_card | 150 | used\_card | 163 |
| SALARY | 26666 | actual\_card | 150 | time1 | 1022 |
| HAS\_FUND | 1079 | overcyc\_loan | 4406 | time2 | 31786 |
| errorloan | 4406 | overcyc\_card | 150 | time3 | 1794 |
| loantype1 | 6261 | guarantee\_C1 | 204 | Y\_FRAUD | 9857 |
| loantype2 | 6261 | guarantee\_C2 | 204 |  |  |
| 注: 由于本月应还款和本月实还款由贷款数据和贷记卡数据组合得到，贷款数据和贷记卡数据缺失状况并不完全相同，因此两类数据需要分别处理之后再组合。 | | | | | |

需要说明的是，地区变量本身不存在缺失值，但是由于已知违约状况的30000个样本数据中地区变量只有4个取值，而未知违约状况的10000个样本数据中地区变量有6个取值，无法使用已知违约状况的30000个样本建立的模型对后者进行预测，因此将未知违约状况的10000个样本数据中独有的1667个地区变量值视为缺失值。

对于第一类数据缺失，由于反映人口特征信息的各变量取值都是有限且离散的，因此可以参考K近邻分类算法的思想，采用众数插补法对借款人人口特征信息进行插补。

K近邻分类算法的基本思想是，如果一个样本在特征空间中的k个最相似样本中的大多数属于某一个类别，则将该样本划分到这个类别中。基于此，本文所用众数插补法的基本过程为：在对样本a的某一缺失变量值进行插补时，首先需要在全样本中找到该变量值为非缺失的样本集A，然后在样本集A中寻找人口特征信息其余6个变量值与样本a尽可能相同的样本集B，使用样本集B中相应变量值的众数对样本a的缺失值进行插补。

由于数据集中样本是有限的，与样本a的人口特征信息完全相同的样本集很可能不存在，因此需要放宽匹配条件，寻找与样本a的部分人口特征信息相同的样本集对缺失值进行插补。为了使样本集B中的样本量尽可能大，我们根据人口特征信息缺失量升序，对性别、本地籍、婚姻状态、公积金、地区、学历和收入水平依次进行匹配，按照该匹配顺序得到的样本集B中的样本与样本a具有较高的相似度。人口特征信息插补结果如表3-5所示。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 表3-5人口特征缺失数据插补结果 | | | | | | |
| 变量 | 取值 | 个数 | 变量 | 取值 | | 个数 |
| local | 0 | 72 | HAS\_FUND | 0 | 645 | |
| 1 | 204 | 1 | 434 | |
| MARRY\_STATUS | 4 | 6 | SALARY | 1 | 23 | |
| 5 | 227 | 2 | 2362 | |
| EDU\_LEVEL | 1 | 5365 | 3 | 20843 | |
| 2 | 228 | 4 | 2622 | |
| area | 1 | 15 | 5 | 362 | |
| 2 | 28 | 6 | 5 | |
| 3 | 1594 | 7 | 449 | |
| 4 | 30 |  |  | |

对于第二类借款人信用记录数据缺失，通过分析发现此类数据缺失是由于借款人历史上没有相关借款还款记录，并非数据采集过程中产生的缺失，因此直接根据变量的经济含义进行赋值。

②异常值处理

在信贷业务中，大部分客户都能够按时还款履约，只有部分客户会出现违约情况，而客户违约的特征和程度各不相同，从图3-1所示的部分指标数据分布直方图可以发现，多数变量具有严重的右偏分布特征，这会对模型的参数估计产生影响，因此需要对异常数据进行处理。

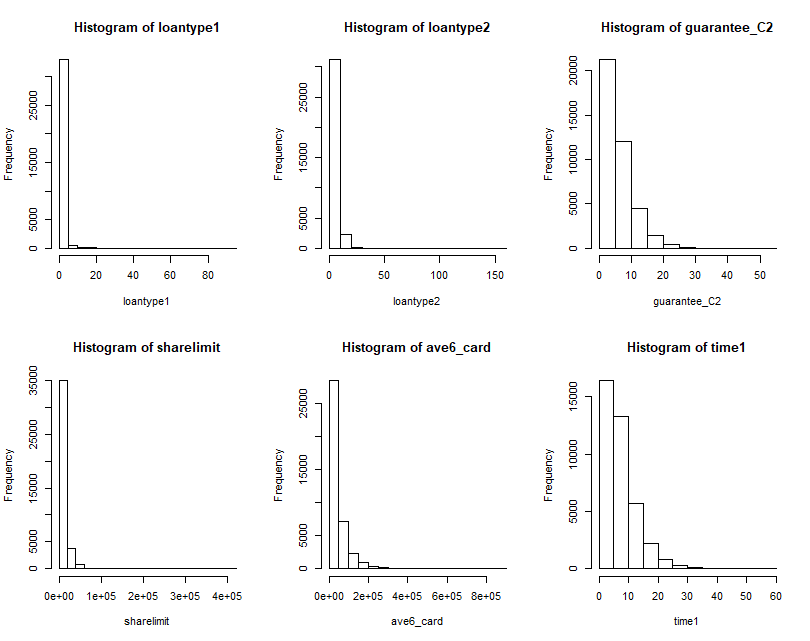


图3-1部分指标数据分布直方图

本文使用缩尾方法对异常值进行处理。缩尾是将超出变量特定百分位范围的数值替换为特定百分位数值的方法，一般选择异常值水平为5%，即将所有小于2.5%分位数或大于97.5%分位数的观察值分别替换为2.5%分位数或97.5%分位数。

训练集中包含30000个样本，其中违约样本有1875个，违约率为6.25%。因为将违约客户判定为非违约客户的损失要比将非违约客户判定为违约客户的损失大，按照谨慎的原则，本文将样本的异常值水平定为6.25%。结合本文变量数据右偏分布特征，使用变量取值的0.9375分位数上侧的最小变量值为阈值，替代高于该阈值的变量值。

由于某些变量数据分布过于集中，使用该方法进行异常值处理之后，部分变量由原来的离散型数值变量变成0-1变量，因此需要根据变量取值含义对变量名称进行调整，调整的变量及含义说明如表3-6。

|  |  |  |
| --- | --- | --- |
| 表3-6变量调整说明 | | |
| 初始变量名称 | 调整变量名称 | 变量含义 |
| 异常贷款账户数 | 异常贷款记录 | 借款人当前未结清贷款中是否有“逾期、转出、呆账”状态的贷款 |
| 异常贷记卡账户数 | 异常贷记卡记录 | 借款人当前未注销贷记卡中是否有“冻结、止付、呆账”状态的贷记卡 |
| 当前贷款逾期期数 | 贷款逾期 | 借款人当前所有未结清贷款是否存在逾期行为 |
| 当前贷记卡逾期期数 | 贷记卡逾期 | 借款人当前未注销贷记卡是否存在逾期行为 |
| 非信用担保贷记卡账户数 | 非信用担保贷记卡账户 | 借款人当前所有未注销贷记卡中是否有采用非信用担保方式申请的贷记卡 |
| 注：完整的调整后模型指标体系设置及说明在附录中呈现。 | | |

3.样本平衡

由于在金融借贷中违约客户所占比例较低，样本数据存在非对称分布问题，直接使用已知违约状况的30000个样本数据进行建模会大大降低模型的预测精度，因此本文采用欠抽样方法对样本数据进行平衡处理，即首先将30000个样本中的1875个违约客户样本全部抽取后，对非违约客户样本进行聚类分析，将非违约客户分为4类，其次使用分层抽样方法抽取1875个非违约客户样本，共得到3750个样本，然后随机抽取其中80%的样本，将3000个样本作为训练集进行建模，用剩下的750个样本作为测试集进行样本外预测，对模型进行评价。为了避免各变量取值水平不同带来的量纲影响，还要对数据进行归一化处理。

（四）模型建立与求解

1.Lasso-Logistic模型

本文使用R软件中Glmnet程序包建立Lasso-Logistic模型。

首先使用被解释变量Y和包含虚拟变量在内的33个解释变量建立Lasso-Logistic模型，通过广义交叉验证，得到模型误差对应调和参数λ值的变化情况，并在图最上方给出模型筛选出来的对应变量数，如图3-2所示。

随着λ的取值不同，模型变量的压缩程度也会随之变化，即模型选择出的变量数目受λ估计值大小的影响。图中可以看出，随着调和参数λ逐渐增大，模型压缩程度增大，模型中包含的自变量个数减少，由于冗余变量被剔除，模型误差逐渐增大，这是因为越来越多的自变量系数被压缩到0，导致原始变量信息损失。

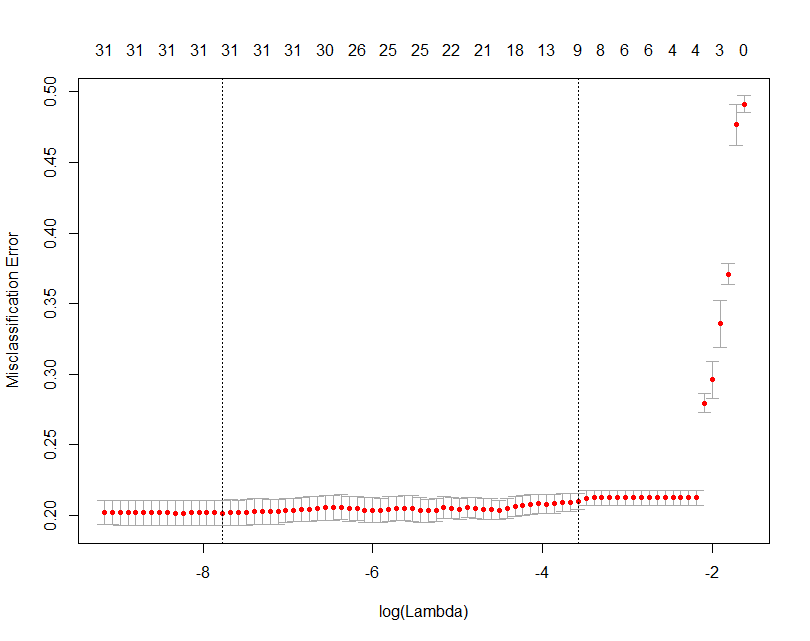


图3-2模型误差变化情况（含欺诈记录变量）

图中左边虚线对应的λ取值为使模型误差最小时的调和参数λ值，可称为最佳λ值；右边虚线对应的λ取值为在最佳λ值一个标准差范围内得到的最简单模型所对应的λ值。因为λ值达到一定大小之后，继续增加模型自变量个数及缩小λ值并不能显著提高模型性能，而右边虚线对应的λ值确定了一个具备优良性能同时自变量个数最少的模型，因此我们选取图中贴近右边虚线的log(Lambda)值，即λ=0.0279，模型中保留的变量个数为9个，此时，基于lasso变量选择的logistic模型参数估计结果如表3-7所示。

|  |  |  |  |
| --- | --- | --- | --- |
| 表3-7模型参数估计（含欺诈记录变量） | | | |
| 变量 | 参数估计 | 变量 | 参数估计 |
| 截距项 | 0.325 | should | 0 |
| area1 | 1.394 | actual | 0 |
| area2 | 1.326 | overcyc\_loan | 0 |
| area4 | -1.656 | errorcard | 0 |
| sex | 0 | guarantee\_C1 | 0 |
| local | 0.014 | guarantee\_C2 | -0.434 |
| EDU\_LEVEL | 0 | sharelimit | 0 |
| marry1 | 0 | overcyc\_card | 0 |
| marry2 | 0 | payschedule\_money | 0 |
| marry3 | 0 | ave6\_loan | 0 |
| marry4 | 0 | ave6\_card | -0.506 |
| SALARY | 0 | used\_card | 0 |
| HAS\_FUND | 0 | time1 | 0.359 |
| errorloan | 0 | time2 | 0 |
| loantype1 | -0.698 | time3 | 0 |
| loantype2 | 0 | Y\_FRAUD | 2.379 |
| loantype3 | 0 |  |  |
| 注：系数为0表示在变量选择中，被剔除出去的变量 | | | |

从表中可以看出，Lasso-Logistic模型只保留了地区虚拟变量、local、loantype1、guarantee\_C2、ave6\_card、time1和Y\_FRAUD变量，模型中大部分解释变量系数都被压缩到0，导致模型中留下的变量较少，与现有研究有较大差异，模型回归结果解释能力较差。通过上文的分析我们可以推测，欺诈记录变量和违约状况之间存在“确定性单向因果关系”，因此，把借款人欺诈记录变量放入模型时，由于该“确定性单向因果关系”的存在，导致部分变量系数被压缩到0，从而被模型剔除。

为了解决这个问题，我们分为两步建立模型，首先将欺诈记录变量从解释变量中剔除，使用剩下的解释变量和违约状况Y建立Lasso-Logistic模型并对测试集进行预测，然后根据欺诈记录变量和违约状况之间的“确定性单向因果关系”对预测结果进行调整，如果借款人存在欺诈记录，则直接将其违约状况预测为违约。

在剔除欺诈记录变量后，通过广义交叉验证，得到模型误差对应调和参数λ值的变化情况如图3-3。

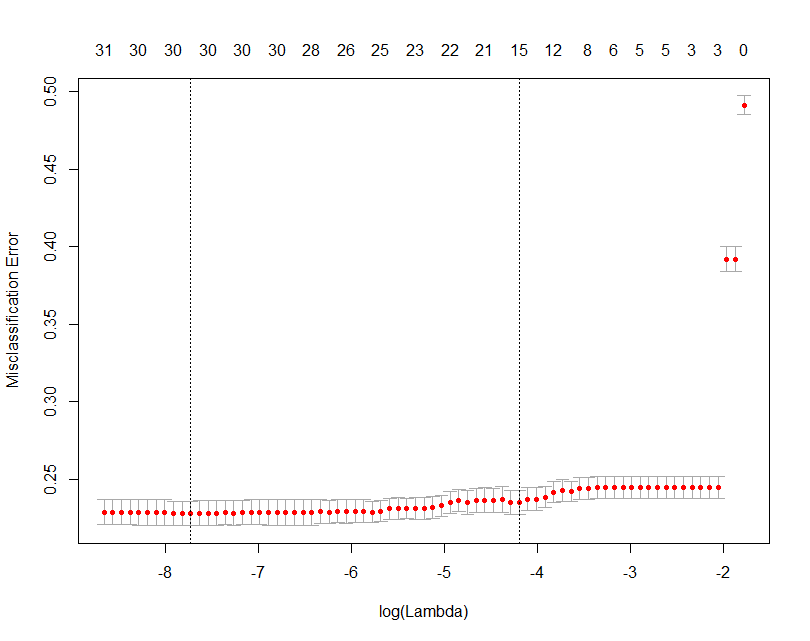


图3-3模型误差变化情况（剔除欺诈记录变量）

图中右边虚线对应的λ值确定了一个具备优良性能同时自变量个数最少的模型，因此我们选取图中贴近右边虚线的log(Lambda)值，即λ= 0.015，模型中保留的变量个数为15个，此时，基于lasso变量选择的logistic模型参数估计结果如表3-8所示。

|  |  |  |  |
| --- | --- | --- | --- |
| 表3-8模型参数估计（剔除欺诈记录变量） | | | |
| 变量 | 参数估计 | 变量 | 参数估计 |
| 截距项 | -0.634 | loantype3 | 0 |
| area1 | 1.661 | should | 0 |
| area2 | 1.685 | actual | 0 |
| area4 | -2.422 | overcyc\_loan | 0 |
| sex | 0.160 | errorcard | 0 |
| local | 0.184 | guarantee\_C1 | -0.173 |
| EDU\_LEVEL | -0.195 | guarantee\_C2 | -0.732 |
| marry1 | 0.326 | sharelimit | 0 |
| marry2 | 0 | overcyc\_card | 0 |
| marry3 | 0 | payschedule\_money | 0 |
| marry4 | 0 | ave6\_loan | 0 |
| SALARY | -0.018 | ave6\_card | -0.849 |
| HAS\_FUND | 0 | used\_card | 0 |
| errorloan | 0.148 | time1 | 0.744 |
| loantype1 | -1.721 | time2 | 0 |
| loantype2 | 0 | time3 | 0.153 |
| 注：系数为0表示在变量选择中，被剔除出去的变量 | | | |

在剔除了欺诈记录变量后，模型中有更多变量被保留下来，解释变量系数也具有更强的可解释性。从表中可以看出，违约风险较高的借款人主要为1地区和2地区的低收入、低学历的本地籍离婚男性。借款人贷款账户存在异常状态时，相应的违约风险也会提高；借款人未结清贷款中投资（经营）性贷款笔数增多时，显示其会有较低的违约风险；借款人持有的贷记卡越多、使用额度越高，说明其信用状况越好，违约风险越低，但是申请过多的贷款和贷记卡会让借款人信贷审批查询记录和信用卡审批查询记录增加，借款人的违约风险可能增大，金融机构对该借款人的关注度也会有所提高。

得到Lasso-Logistic模型后，将测试集中的750个样本带入模型，预测借款人的违约状况，并与真实值进行比较，得到模型的预测效果。

在此基础上，利用欺诈记录变量和违约状况之间的“确定性单向因果关系”对预测结果进行调整，得到调整后训练集和测试集预测结果准确率，如表3-9所示。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 表3-9调整前后模型预测效果 | | | | | | |
| 模型 | 训练集准确率（%） | | | 测试集准确率（%） | | |
| 违约 | 非违约 | 总体 | 违约 | 非违约 | 总体 |
| 调整前 | 75.97 | 77.07 | 76.51 | 74.14 | 70.15 | 72 |
| 调整后 | 81.79 | 77.07 | 79.47 | 81.32 | 70.15 | 75.33 |

从表中可以发现，在对Lasso-Logistic模型预测结果进行调整后，并没有降低模型对非违约样本的预测效果，而是使违约样本的预测准确率进一步提升，使得总体预测准确率显著提升，说明本文对欺诈变量的处理方法是有效的。

2.BP神经网络模型

本文R软件中neuralnet程序包建立BP神经网络模型，该模型使用与上文Lasso-Logistic模型相同的训练集和测试集进行建模和预测，同时将欺诈记录变量包含在指标体系中，保证两个模型使用相同的信息量，便于与Lasso-Logistic模型进行对比。

神经网络中的关键问题是网络结构的确定，即隐藏层的层数和每层的隐节点的个数，它们决定了网络的复杂程度，层数和隐节点个数越多，网络的复杂程度就越高，预测准确度也会越高，但同时也会使得问题的解决需要更多的时间，并且可能会导致过拟合问题。这需要我们在模型训练中逐步调整，寻求最优解。本文选取的隐藏层数为1，隐藏层有6个隐藏节点，在经过111323次迭代后得到模型，神经网络拓扑图如图3-4所示。

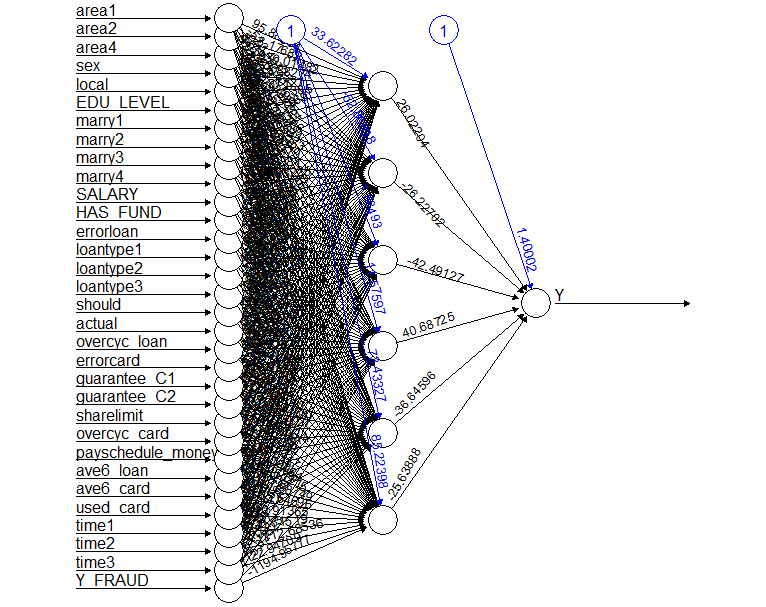


图3-4神经网络拓扑图

图中的黑线表示每一层与其相关权重直接的关系，而蓝色线表示拟合过程中，每一步被添加到蓝色线上的误差项，而这些误差可以表示一个线性模型的误差区间。

在二分类问题中，BP神经网络输出节点给出的是预测类别为1的概率。通常概率大于分割值τ是预测类别为1，小于τ时预测类别为0。尽管一般情况下分割值τ取0.5，但这并非适合于所有情况。所以我们绘制ROC曲线来寻找分割值τ，如图3-5。

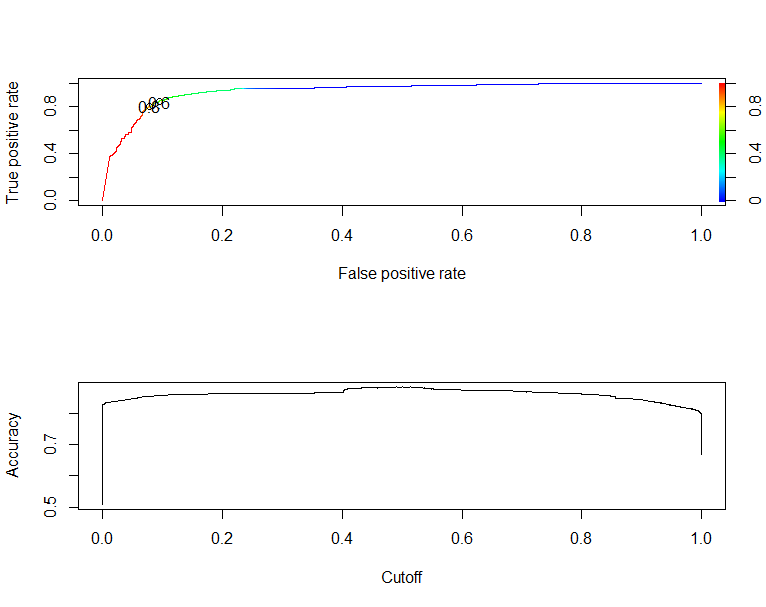


图3-5 ROC曲线

图3-5显示，当以0.8作为分割值时，TPR较大而且FPR处在一个相对较小的水平，因此选择τ=0.8作为分割值。使用该模型对测试集进行预测，模型的预测效果如表3-10。

（五）模型预测与分析

Lasso-Logistic模型和BP神经网络模型在训练集和测试集中对违约客户和非违约客户预测准确率的对比情况如表3-10所示。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 表3-10模型预测效果对比 | | | | | | |
| 模型 | 训练集准确率（%） | | | 测试集准确率（%） | | |
| 违约 | 非违约 | 总体 | 违约 | 非违约 | 总体 |
| Lasso-Logistic | 81.79 | 77.07 | 79.47 | 81.32 | 70.15 | 75.33 |
| BP神经网络 | 80.41 | 92.13 | 86.17 | 68.19 | 77.86 | 73.37 |

从表中可以看出，Lasso-Logistic模型对训练集的总体预测准确率要低于BP神经网络模型，但是Lasso-Logistic模型对测试集的总体预测准确率比BP神经网络高，而且相对于训练集中的预测准确率并没有降低太多，说明该模型比BP神经网络模型更稳定，具有较好的外推性，更适合做样本外预测。

在实际的信用风险评价中，将违约客户误判为非违约客户对授信人或社会而言造成的潜在经济损失会更大，因此对违约客户的准确判断更为重要，其次考虑的才是对非违约客户的判断，Lasso-Logistic模型在测试集中对违约样本的预测准确率比BP神经网络高13个百分点，对训练集中违约样本的预测效果也优于BP神经网络模型。虽然Lasso-Logistic模型对非违约客户的预测准确率比BP神经网络低，但是综合来看本文所构建的Lasso-Logistic模型要优于BP神经网络模型，更适合用该模型对个人信用违约状况进行预测。

（六）原始数据测试集预测

使用本文所建立的Lasso-Logistic模型对测试样本集中没有给出是否违约信息的10000个样本进行预测，然后使用欺诈记录变量对预测结果进行调整，预测结果放在附件中。

四、结论和建议

（一）结论

本文从组委会提供的原始数据出发，通过组合、拆分、汇总等方法，从人口特征、资产负债状况、信用历史、信用行为四个方面构造了27个变量组成个人信用评估指标体系，对数据进行缺失值插补和异常值处理，分别建立了调整的Lasso-Logistic模型和BP神经网络模型对借款人的违约情况进行预测。分析发现调整的Lasso-Logistic模型预测准确率较高，具有更好的外推性。

本文建立的调整的Lasso-Logistic模型结果显示除性别、本地籍、教育水平、婚姻状况、收入等常见变量对个人信用风险有着显著影响外，地区类型、投资（经营）性贷款笔数、贷记卡担保类型数量、最近六个月平均贷记卡使用额度、信贷审批查询记录等变量对个人信用违约风险有着显著而重要的影响。同时欺诈变量与借款人违约状况存在“确定性单向因果关系”，直接放入回归模型将影响模型的结果，因此我们将该变量从模型中剔除，直接使用该变量对预测结果进行调整，模型的可解释性和准确率得到提升。

通过分析可以发现本文所构建的指标体系和调整的Lasso-Logistic模型中保留下来的指标变量与前人的研究在某些方面存在差异，而且之前的研究结果之间也并不是完全相同。除了研究所用数据质量差异带来的不同，可以推测，在不同经济和社会发展阶段中，影响个人信用的指标也会发生变化，而且违约客户的特征表现也会有所不同，因此在实际应用时，不能照搬本文建立的指标体系和预测模型，需要结合实际情况进行适当调整，但是本文在指标体系设定和模型建立时所使用的方法在解决类似问题时是具有参考意义的。

（二）建议

信用风险问题是全社会需要高度重视的问题，提高风险控制水平不只是金融机构的任务，也需要政府和个人的努力。基于上文研究和结论，本文从金融机构、政府和个人三个角度提出如下建议。

1.金融机构需要从借款人信息完整度和评估体系方面进行改进。

金融机构需要完善借款人的各类基础信息和信用记录，提高借款人信息完整度。高质量的信用数据可以提高金融机构对借款人进行评估的准确度，由于本文所用原始数据存在不同程度的缺失，使用众数插补法进行插补后的数据不可避免与真实值存在误差，影响模型的预测准确度。因此，金融机构可以建立规范的客户信息填写标准，将对借款人信用状况有重要影响的指标列为借款人申请借款的必填信息，同时各金融机构可以借助大数据平台，建立金融机构借款人信息共享制度。

金融机构也需要对借款人信用评估体系进行阶段性调整。由于在不同经济和社会发展阶段中，影响个人信用的指标会发生变化，而且违约客户的特征表现也会有所不同，这就需要金融机构在构建个人信用评估体系时，要适当关注到之前关注较少甚至没被关注的信息，对评估体系进行阶段性调整，保证所建立的个人信用评估指标体系能够与时俱进，一直对真实情况进行反映。

2.政府需要加强社会诚信建设，提高全民信用意识。

我国网络借贷违约事件层出不穷，原因之一是我国还未建立起完善的社会信用体系，现阶段的制度对风险的控制能力相对较弱，国民信用意识淡薄。因此政府需要加强社会诚信建设，健全公民和组织守法信用记录，完善守法诚信褒奖机制和违法失信行为惩戒机制，同时也要加强征信宣传教育，营造诚信社会环境，提高全民信用意识。

3.个人应提高自身财务管理水平，保持良好征信记录。

人们在日常的金融和社会活动中，要树立信用意识，合理安排财务收支计划，定期关注自己的债务情况并按时偿还，避免出现延期、违约、欺诈等各种不良行为，保持良好征信记录。

参考文献

[1]段昊.基于P2P网贷平台特点的信用体系实证研究[D].北京邮电大学, 2015.

[2]张国政,陈维煌,刘呈辉.基于Logistic模型的商业银行个人消费信贷风险评估研究[J]. 金融理论与实践, 2015(3):53-57.

[3]孙燕.随机效应Logit计量模型的自适应Lasso变量选择方法研究——基于Gauss—Hermite积分的EM算法[J].数量经济技术经济研究, 2012(12):147-157.

[4]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.

[5]曲秋实,李莉.基于logit模型的商业银行个人信用风险评估[J]. 商业经济, 2010(12):72-73.

[6]罗方科,陈晓红.基于Logistic回归模型的个人小额贷款信用风险评估及应用[J]. 财经理论与实践, 2017,38(1):30-35.

[7]Lee T H, Jung S. Forecasting creditworthiness: Logistic vs. artificial neural net[J]. Journal of Business Forecasting Methods & Systems, 1999(4):28.

[8]Tibshirani R J. Regression shrinkage and selection via the LASSO. J R Stat Soc B[J]. Journal of the Royal Statistical Society, 1996, 58:267-288.

[9]Leo Breiman. Better Subset Regression Using the Nonnegative Garrote[M]. American Society for Quality Control and American Statistical Association, 1995.

[10]Efron B, Hastie T, Johnstone I, et al. Least angle regression[J]. Annals of Statistics, 2004, 32(2):407-451.

附录

|  |  |  |  |
| --- | --- | --- | --- |
| 附录：模型指标体系 | | | |
|  | 变量名称 | 符号 | 变量说明 |
| 人口特征 | 地区 | area | 以虚拟变量形式引入模型，area1为1地区，area2为2地区，area4为4地区 |
| 性别 | sex | sex=1表示男性，sex=0表示女性 |
| 本地籍 | local | local=1表示本地籍，local=0表示非本地籍 |
| 教育水平 | EDU\_LEVEL | EDU\_LEVEL=1表示教育水平为专科及以下，EDU\_LEVEL=2表示教育水平为本科，EDU\_LEVEL=3表示教育水平为硕士及以上， |
| 婚姻状态 | MARRY\_STATUS | 以虚拟变量形式引入模型，marry1为离婚，marry2为离异，marry3为丧偶，marry4为未婚 |
| 收入水平 | SALARY | 取值范围为1-7，数值越大表示收入水平越高 |
| 公积金 | HAS\_FUND | HAS\_FUND=1表示有公积金，HAS\_FUND=0表示无公积金 |
| 资产负债状况 | 投资（经营）性贷款笔数 | loantype1 | 贷款类型中每个贷款人当期未结清的个人经营性贷款、个人助学贷款、农户贷款总笔数，变量取值为0-10，其中loantype1=10表示投资/经营性贷款笔数大于9笔 |
| 消费性贷款笔数 | loantype2 | 贷款类型中每个贷款人当期未结清的个人汽车贷款、个人消费贷款、其他贷款总笔数，变量取值为0-19，其中loantype2=19表示投资/经营性贷款笔数大于18笔 |
| 住房性贷款笔数 | loantype3 | 贷款类型中每个贷款人当期未结清的个人商用房（包括商住两用）贷款、个人住房贷款、个人住房公积金贷款笔数之和，变量取值为0-2，其中loantype3=2表示住房性贷款笔数大于等于2笔 |
| 非信用担保贷记卡账户 | guarantee\_C1 | guarantee\_C1=1表示借款人当前所有未注销贷记卡中有采用非信用担保方式（保证、抵押担保、其他担保、质押（含保证金）担保、组合（不含保证）担保）申请的贷记卡，guarantee\_C1=0表示借款人无非信用担保贷记卡账户 |
| 信用担保贷记卡笔数 | guarantee\_C2 | 每个贷记卡持有人当期未结清的采用信用/免担保方式的贷记卡笔数，取值为0-11，其中guarantee\_LC2=11表示信用/免担保贷记卡笔数大于10笔 |
| 本月应还款金额 | scheduled | 每个贷款人本月应还贷款和应还贷记卡总金额 |

|  |  |  |  |
| --- | --- | --- | --- |
| 信用历史 | 贷记卡共享额度 | limitamount | 每个贷记卡持有人所有贷记卡/准贷记卡的共享额度中位数，反映其他金融机构对借款人的信用评估状况 |
| 还款进度 | payschedule\_money | 每个借款人还款金额比例，取值范围未0-1 |
| 贷款审批查询次数 | time1 | 每个贷款人贷款审批查询次数，取值范围为0-17，其中time1=17表示贷款审批查询次数大于等于17次 |
| 担保资格审查次数 | time2 | 表示每个贷款人担保资格审查次数，取值范围为0-3，其中time2=3表示担保资格审查次数大于等于3次 |
| 信用卡审批查询次数 | time3 | 每个贷款人信用卡审批查询次数，取值范围为0-20，其中time3=20表示信用卡审批查询次数大于等于20次 |
| 欺诈记录 | Y\_FRAUD | Y\_FRAUD=1表示有欺诈记录，Y\_FRAUD=0表示无欺诈记录 |
| 信用行为 | 异常贷款记录 | errorloan | errorloan=1表示借款人当前未结清贷款中有“逾期、转出、呆账”状态的贷款，errorloan=0表示无异常账户 |
| 异常贷记卡记录 | errorcard | errorcard=1表示借款人当前未注销贷记卡中有“冻结、止付、呆账”状态的贷记卡，errorcard=0表示无异常账户 |
| 本月实还款金额 | actual | 借款人本月实还贷款和贷记卡总金额 |
| 贷款逾期 | overcyc\_loan | overcyc\_loan=1表示借款人当前所有未结清贷款存在逾期行为，overcyc\_loan=0表示借款人当前不存在贷款逾期行为 |
| 贷记卡逾期 | overcyc\_card | overcyc\_card=1表示借款人当前未注销贷记卡存在逾期行为，overcyc\_card=0表示借款人当前不存在贷记卡逾期行为 |
| 最近6个月平均贷款使用额度 | ave6\_loan | 每个贷款人最近6个月所有贷款的平均使用额度加总 |
| 最近6个月平均贷记卡使用额度 | ave6\_card | 每个贷记卡持有人最近6个月贷记卡平均使用额度加总 |
| 贷记卡透支度 | used\_card | 贷记卡持有人所有未注销贷记卡/准贷记卡额度使用程度，数值越大说明贷记卡额度使用越多，取值范围为0-1 |