论文编号：A0958

基于多源数据融合的个体商户信用评估研究

目 录

图目录

表目录

摘要

近几年来，随着我国政府对个体工商户发展的支持，个体工商户已逐渐成为我国民营经济的重要组成部分，同时其融资需求也进一步扩大，银行所面临的债务人信用方面的信贷风险也不可避免地有所增大。因此对个体工商户信用进行准确评估具有极其重要的研究意义和价值。

首先，本文基于来自地区商业银行的数据集展开对本问题的研究，该数据集涵盖了 2157 条包括商户违约与否以及商户相关财务类指标、非财务类指标以及宏观环境指标等共计 57 个指标。为剔除无关变量、解决变量间多重共线性等问题，对数据预处理，采用**Lasso 回归算法**对其进行筛选，并利用**交叉验证法**选取参数，最终确立了由 26 个指标组成的个体商户信用评估指标体系。

其次，为找寻 26 个指标与商户信用（即是否会违约）间的内在联系，基于数据建立模型，本文选取了**逻辑回归模型**、**随机森林模型**、**KNN 分类**以及**支持向量机**四类算法在训练集上进行训练，对数据不平衡问题作了相关处理，分别预测商户的信用状况。

最后，为选择最优模型，本文选取 AUC 值、F1 值以及准确率 ACC 值以及分析其混淆矩阵，对个体工商户信用评估分类效果作出评价。其中支持向量机模型由于其对数据依赖性较低处理小样本数据效果较好，效果优于其他模型；随机森林模型对未违约商户的预测性能最好。因此本文选定**支持向量机与随机森林模型双重检验**作为最终个体工商户信用评估预测模型，模型准确度和精准度均可达 90%以上。

根据本文分析，得出以下结论：

①在个体商户信用评估问题上，商户的盈利能力和保证联保所涵盖的指标对个体商户信用影响较大，因此在对商户信用的评估作人为判断时，应着重查看这两项内容所对应的指标信息。

②利用模型对个体工商户作出信用评估时，违约商户可信度：支持向量机>随机森林；未违约商户可信度：随机森林>支持向量机。

关键词：个体工商户信用评估；Lasso；机器学习；支持向量机和随机森林双重检验

Abstract

In recent years, with the support of the Chinese government for the development of individual industrial and commercial households, individual industrial and commercial households have gradually become an important part of our private economy. At the same time, their financing needs have further expanded. The credit risk of debtors faced by banks is also not acceptable. Avoid an increase in the ground. Therefore, the accurate evaluation of the credit of individual industrial and commercial households has extremely important research significance and value.

First of all, conducted research on this issue based on a data set from a regional commercial bank. The data set covers a total of 57 indicators including whether merchants are in default or not, and merchant-related financial indicators, non-financial indicators, and macro environmental indicators. In order to eliminate irrelevant variables and solve the problems of multicollinearity among variables, the data was preprocessed, filtered by Lasso regression algorithm, and parameters were selected by cross-validation method, and finally established the individual merchant credit evaluation index composed of 26 indicators system.

Secondly, in order to find the internal connection between 26 indicators and merchant credit, based on the data to establish a model, this paper selects four types of algorithms: logistic regression model, random forest model, KNN classification, and support vector machine on the training set Conducted training and dealt with the problem of data imbalance, respectively predicting the credit status of the merchants.

Finally, in order to select the optimal model, this paper selects the AUC value, F1 value and accuracy rate ACC value and analyzes the confusion matrix to evaluate the effect of individual industrial and commercial households' credit evaluation classification. Among them, the support vector machine model has a better effect on processing small sample data due to its low dependence on data, and the effect is better than other models; the random forest model has the best predictive performance for non-default merchants. Therefore, this paper selects the dual test of support vector machine and random forest model as the final credit evaluation prediction model of individual industrial and commercial households, and the accuracy and precision of the model can reach more than 90%.

According to the analysis of this article, the following conclusions are drawn:

①In terms of the credit evaluation of individual merchants, the profitability of the merchant and the indicators covered by the guarantee joint guarantee have a greater impact on the credit of the individual merchant. Therefore, when making

artificial judgments on the evaluation of the merchant's credit, you should focus on the correspondence between these two contents Indicator information.

②When the model is used to evaluate the credit of individual industrial and commercial households, the credibility of the defaulting merchants: Support Vector Machine>Random Forest; the credibility of the non-defaulting merchants: Random Forest>Support Vector Machine.

Keywords: credit evaluation of individual industrial and commercial households; Lasso; machine learning; double test of support vector machine and random forest

一、绪论

（一）选题背景及意义

近几年来，随着国家对个体工商户政策的支持，个体工商户数量不断增加，据2019年国家统计局数据显示，新登记小型个人企业日均达到了1.89万户，个体工商户现已成为我国民营经济的重要组成部分。国务院总理李克强在会议上指出，加大支持小微企业、个体工商户等的普惠金融力度，引导扩大信用贷款、首贷、中长期贷款、无还本续贷业务规模，推广随借随还贷款[15]。为响应国家号召以及满足个体工商户经营的融资需求，银行进一步扩大个体工商户信贷业务，降低了个体商户的从业门槛，促进了社会的资金融通和经济发展，但同时也不可避免地产生了相应的信贷风险。

目前，银行的信贷风险主要集中于两方面。其一是债务人的信用风险，具体是指债务人由于经营状况不佳、无故违约等导致履约能力降低等造成对银行的损失，从而导致银行贷款的风险[2]；其二是银行内部人员或系统的操作风险，指银行内部人员由于存在流程不完善、操作不规范或系统本身具有缺陷等所造成的风险对银行产生不利影响。而银行的信用风险往往会直接导致其周转困难、停业甚至倒闭，在现代金融体系中，这种结果往往会引起连锁反应，这将涉及单个商业银行的生存安全及影响社会经济的发展稳定。此外信用风险还会使拥有现金的银行不敢对外放款，造成那些经营状况良好却资金难以运作的企业无法发展，市场经济低效萎靡，破坏和阻碍国民经济的正常发展。因此对个体商户信用进行正确评估，为银行规避潜在的债务人信用风险刻不容缓。

在如今的大数据时代，海量数据中蕴藏着大量丰富的信息。本文将基于多源数据融合对个体商户的信用评估进行研究并作出预判，通过对海量数据的处理及分析，建立恰当的统计模型并借助于目前先进的算法和软件，预测商户的信用状

况，从而为银行提供是否贷款的决策支持，规避银行对个体商户融资的风险，保障其利益，促进银行的正常运作，从而有助于国家经济的正常发展。

（二）国内外研究现状

目前国内外对信用风险管理话题都颇为关注，张维和李玉霜[4]对国内外商业银行信用风险的研究作出了详细的分析，其中国外信用风险管理的研究由于起步较早现在已较为成熟，传统的有比例分析法，在此基础上还有判别分析法，Logistic 回归分析，随机森林，主成分分析，聚类分析等多种统计方法也应用于风险管理，著名的有贝叶斯决策模型，Altman 的 Z-score 及在此基础上改进的 ZETA 模型。随着人工智能的发展，神经网络等模型也逐渐应用于银行行业并有较大发展。

国内对风险管理的研究起步尚晚，但也有所成就。传统的信用风险评估主要是由银行工作人员依据自身经验对个体商户作出分析和判断，该方法对相关人员的专业能力要求较高，业务流程比较复杂，且风险难以控制，具有很大的局限性。在 2010 年，白少步采用有序多分类的 logistics 回归模型，基于违约贡献度，建立了企业信用违约风险预警模型[9]。2016 年，辽宁大学刘艳春和崔永生采用探索性因子分析和结构方程模型的验证性分析建立了供应链下的中小型企业信用风险评价模型[10]，我国对贷款风险的研究从定性分析逐渐被定量分析所取代，运用灰色综合评价方法对中小企业信用风险进行评估，成为信用风险评估的主流。随着大数据时代的到来以及计算机技术的发展，统计机器学习模型、神经网络模型等成为热点并被大量应用，2014 年江训艳等采用 BP 神经网络对商业银行信用风险进行了研究[11]，2020 年耿成轩和李晓汨采用支持向量机模型对企业融资风险进行评估[12]等。我国国内研究人员对该课题的研究进程发展迅速，但对银行小额贷款以及个体工商户信用评估等的研究较少，本文将在前人研究所得结论的

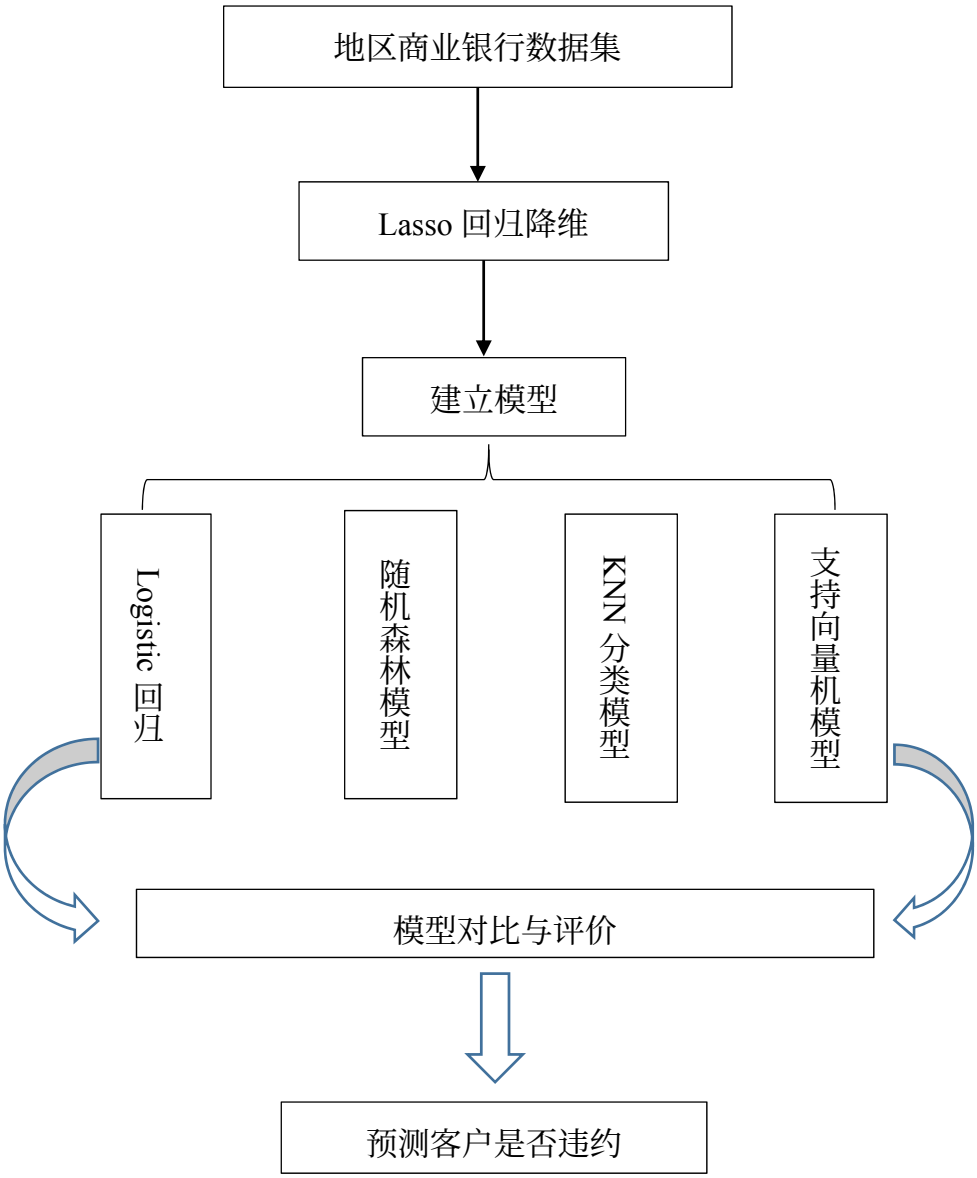
基础上，对该课题进行深入研究。

总结来说，对信用评估的研究方法可分为定性分析法和定量分析法。定性分析受人的主观影响较大，预测波动较大且准确性较差；定量分析法以数据为基础，现有研究方法、数学模型、统计模型等均较为成熟，预测结果更具可靠性。目前应用定量分析多为利用所搜集指标数据对信用风险先采用综合模糊评价、灰色评价等方法建立先验标记，进而利用不同模型进行预测。本文基于具有各类指标以及个体商户是否违约的数据集进行研究，减少了数据先验标记带来的误差，直接从大量数据中挖掘信息，建立模型对个体商户信用作出评估。

（三）研究内容及创新

本文以来自地区商业银行的数据集作为研究对象，数据集指标分布范围较广，有非财务指标如债务人以及其联保人的性别、学历、家庭成员及收入等信息，还有财务指标如商户经营收入、资产负债率、净资产报酬率等信息，共计 60 个指标。为对指标进行筛选，剔除无关指标，利用传统的 Lasso 算法对数据做降维处理，进而选取合适的指标。就目前的研究现状以及前人的研究经验看，logisitic 回归、随机森林、KNN 算法以及支持向量机模型被大量应用，均已被证明是较好的预测模型，因此本文采用以上模型应用于信用评估问题上。本文将采用 logisitic 回归、随机森林、K-近邻分类和支持向量机等统计模型建立个体工商户信用评估预测模型，并对不同模型比较分析其 AUC 值、F1 值和准确率，对模型评价选取最优模型，实现对商户信用的准确评估，并为银行提供可行性建议。

本文创新主要在于多源数据集的使用，数据更加真实客观，具有研究价值；



采用多种统计模型对信用评估问题做出二分类预测并进行对比评价。研究内容流程图如下表所示：

图 1　研究内容流程图

二、数据处理及分析

（一）数据来源及描述

1. 数据集的来源及说明

本文 data_1.csv 数据集来源于地区商业银行，为保护数据的隐私以及相关保密性，已删除相关隐私信息并对数据作标准化处理。数据集共包含两类数据，分类型变量和连续型变量，对不同类别数据所作具体处理如下，

① 分类型变量

建立不同类别与整数间的映射关系，并将不同整数映射于 0-1 之间使其标准化。

表 1　分类型指标的处理

分类型指标	分类标准	类别数
债务人及保证人学历	本科及以上，1；大专，0.8； 高中以及中专，0.6； 初中，0.4；小学，0.2； 其他，0	6
债务人及保证人 婚姻状况	已婚-初婚，1； 已婚-再婚或已婚-复婚，0.75； 未婚，0.5；丧偶，0.25；其他，0	5
居住状况	自有住房，1； 按揭贷款购买的住房，0.75； 共有住房，0.5；亲属住房，0.25； 其他，0	5
职务	单位主要负责人(处级以上领导)，1； 部门老总,处长,科长等领导，0.75； 一般员工，0.5；无工作单位，0.25	4
职称	高级，1；中级，0.75； 初级，0.5；其他，0.25；	4
是否有本地户口	0：否，1：是	2
公民身份联网核查结果	公民身份证号与其姓名一致,且有照片,1； 公民身份证号与姓名一致，但无照片，0.75； 公民身份证号存在，但与姓名不匹配，0.5； 联网核查公民身份信息系统故障或其他错误，0.25；	5

分类型指标	分类标准	类别数
与本行的关系	公民身份号码不存在，0 1 老个体商户，0 新个体商户	2
是否有营业执照	1 有；0.5 没有或者已过期；0 不清楚	3
营业执照年限	1 有营业执照且办照时间在一年以上 0.6 有营业执照但办照时间在一年以下 0.3 有过期或未年审营业执照 0 不清楚	4
债务人及保证人从事行业	批发零售业，1；餐饮业，0.75； 制造业，0.5；服务业，0.25； 其他行业，0	5
是否营业场所自有	0 否；1 是；0.5 不清楚	3
成员人数	4 人及以上，1；3 人，0.75； 2 人，0.5；1 人，0.25；	4
劳动力人数	4 人及以上，1；3 人，0.75； 2 人，0.5；1 人，0.25	4
供养人数	0 人，1；1 人，0.75；2 人，0.5 人，0.25；4 人，0；	5
负担人数	供养人数/负担人数	11
贷款用途	1:购买房产， 0.75:购买轿车、货车、卡车、设备， 0.5:营运资金,0.25:投资项目	4
保证人与借款人关系	非常好，1；良好，0.75； 一般，0.25；数据缺失，0	4
联保人信用状况	非常好，1；良好，0.75；一般，0.5；中等偏差，0.25；差，0	5
联保小组成员关系	1 朋友、老乡、亲戚、熟悉；0.75 同事、同学、合伙人、上下级、了解；0.25 个体商户的亲戚，个体商户的同事；0 数据缺失	4

② 数值型变量

采用 min-max 归一化对数值型变量处理，使其标准化，映射到[0，1]区间。

$$X' = \frac{X-min}{max-min}$$ 公式（1）

表 2　数值型指标的处理

连续型指标	指标说明	连续型指标	指标说明
债务人及联保人年龄	岁	家庭支出	（元）
保证人月收入	（元）	流动比率	$\frac{流动资产}{流动负债}$

连续型指标	指标说明	连续型指标	指标说明
资产负债率	$\frac{负债总额}{资产总额}$	产权比率	$\frac{负债总额}{所有者权益总额}$
所有者权益平均	$\frac{期初+期末}{2}$	每月偿还其他银行贷款数额	（元）
私人借款	元	每月偿还本行贷款数额占净收入比例	
净资产报酬率	$\frac{净资产}{平均净资产}$	净利润	（元）
平均营业收入	（元）	销售净利率	$\frac{净利润}{销售收入}$
总资产	（元）	总资产报酬率	$\frac{报酬总额}{资产平均总额}$
月净收入	（元）	月税收	（元）
应收账款周转率	$\frac{主营业务收入}{应收账款平均}$	存货账款周转率	$\frac{营业成本}{平均存货余额}$
总资产周转率	$\frac{销售收入}{总资产}$	固定资产	（元）
固定资产周转率	$\frac{销售收入}{固定资产}$	经营费用	（元）
经营年限	（年）	经营面积	（m^2）
雇员人数	（人）	人均储蓄余额	（元）
地区 GDP 增长率		居民消费价格指数	上年=100
人均国内生产总值	（元）		

2．数据指标说明与描述

①数据来源

本数据集来源于地区商业银行，数据真实有效，且经银行授权使用。由于数据涉及个体工商户以及银行隐私，已去除部分隐私数据并对数据作标准化处理，达到隐私保护的目的。

②数据描述

本数据集具有共计 57 个指标，包括财务类指标、非财务类指标、宏观环境类指标等，覆盖范围广，分类齐全；共计 2156 条数据，每条数据包含个体商户违约与否的信息以及 57 个指标。

（二）Lasso 回归处理

本文考虑到Lasso回归在变量选择上能够高效地从高维变量中提取出重要变量，解决高维数据所具有的普遍问题——稀疏性，并能提高模型的解释精度、解决变量之间的多重共线性问题等, 首先对数据降维处理, 在指标选取上采用 Lasso 回归的方法选取最优指标建立个体工商户信用评估指标体系。

1. 建立 Lasso 回归模型

Lasso 回归本质上是通过构造一个惩罚函数, 对所有变量系数进行回归惩罚, 使得相对不重要的独立变量或具有多重共线性的变量系数变为 0，排除在模型之外，从而实现变量筛选，得到一个较为精炼准确的模型。其具体原理如下,

设自变量矩阵为$\mathrm{X}=(\mathrm{x}_1,\mathrm{x}_2,\ldots,\mathrm{x}_\mathrm{n})$，其中$\mathrm{X}_j=(\mathrm{x}_{1\mathrm{j}},\mathrm{x}_{2\mathrm{j}},\ldots,\mathrm{x}_{\mathrm{nj}})\ j=1,2,\ldots,n$，因变量$Y=(y_1,y_2,\ldots,y_m)^T$，假定数据已进行标准化处理，则建立 Y 与 X 之间的线性模型：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n+\varepsilon \qquad \text{公式（2）}$$

其中，β_0为常量，$\beta_i,i=1,2,\ldots,n$为每个变量的系数，ε为随机扰动。

记$\beta=(\beta_1,\beta_2,\ldots,\beta_n)^T$,则$\beta$的 Lasso 估计表达式如下：

$$\hat{\beta}^{lasso}=argmin(Y-X\beta)^T(Y-X\beta)+\lambda\parallel\beta\parallel_1 \qquad \text{公式（3）}$$

其中，Lasso 回归模型的复杂程度由 λ 来控制，λ 值越大，对变量较多的线性模型的惩罚力度就越大，最终筛选所得变量也越少。随着 λ 的逐渐增大，某些

β_j的 Lasso 估计值也会随之变小甚至为 0，此时，与等于 0 相应的变量代表它与因变量 Y 的关系很小，将被剔除，进而实现变量选择的作用。本文通过利用以上原理借助 R 语言软件结合交叉验证的方法选取λ值，筛选得到最优指标。

借助 R 语言软件，以个体工商户违约与否作为被解释变量，记为 Y，其余 57 个变量作为解释变量，分别记为 X_1、X_2、…、X_{57}，调取 glmnet 包中的 glmnet 函数建立模型，并做出模型的系数的变化曲线，运行结果如下图所示：

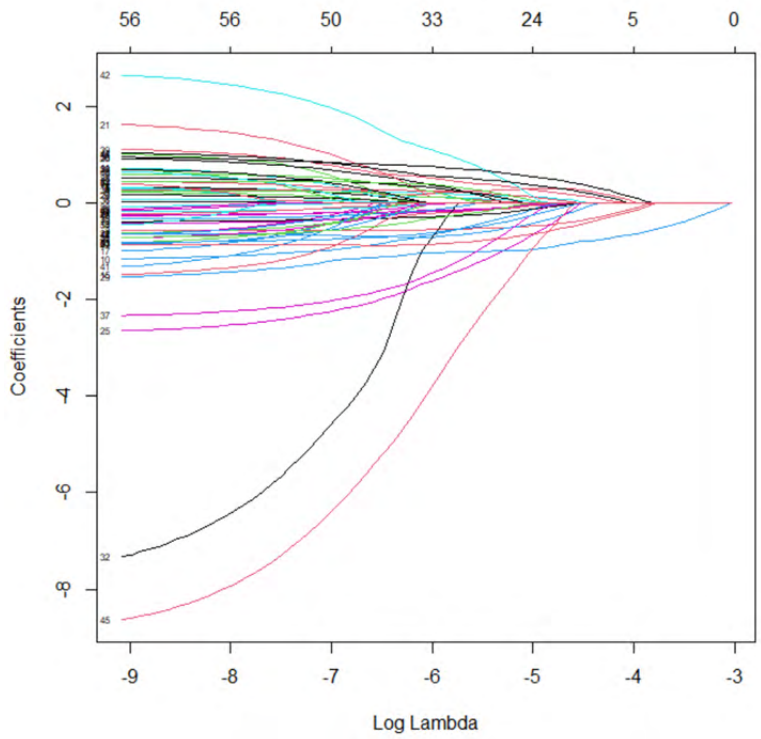


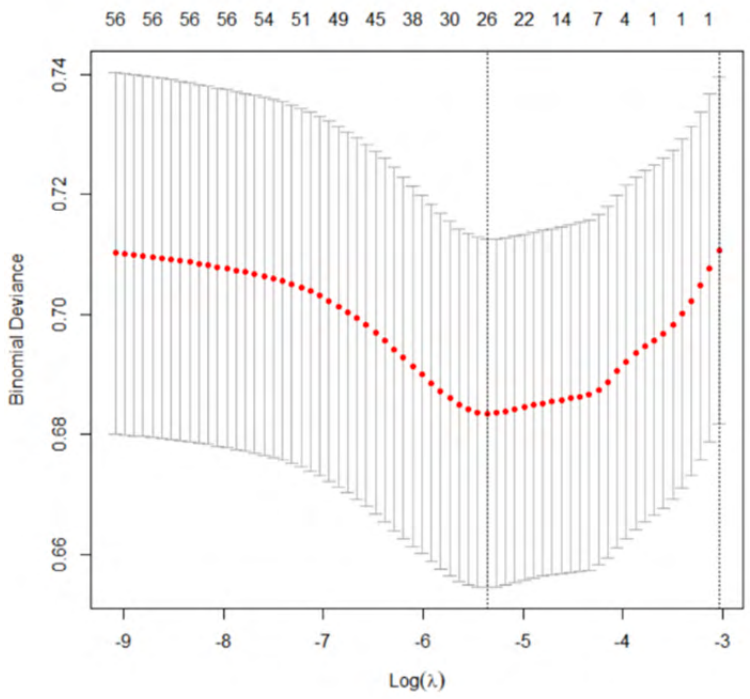
图 2　Lasso 回归的变量选择路径图

图 2 中的每一条曲线代表一个自变量系数的变化轨迹，以系数值作为纵坐标，此时模型中非零系数的个数为上横坐标，下横坐标表示此时$Log\ \lambda$的大小。图中随着$Log\ (\lambda)$的不断增大，越来越多的自变量的系数逐渐趋于 0，其中系数被压缩为 0 的变量，说明比较重要。此模型得出总资产、贷款用途、月税收、资产负债率、保证人学历、每月偿还本行贷款数额占净收入比例、联保人信用状况等因素对是否违约的影响较为重要。

2. 运用交叉验证选取参数

调用 R 语言中 glmnet 包中的 cv.glmnet 函数的交叉验证功能，采用交叉验证法拟合选取优化的模型，决定模型参数λ的选取，同时交叉验证还可对模型的性能有一个更准确的估计。

交叉验证的运行结果如下图：




```
Call:  cv.glmnet(x = x, y = y, type.measure = "deviance", alpha = 1,      family = "binomial")

Measure: Binomial Deviance

     Lambda Index Measure      SE Nonzero
min 0.00469    26  0.6835 0.02909      26
1se 0.04805     1  0.7106 0.02896       0
```

图 3　Lasso 交叉验证结果图

由交叉验证结果可以得出，当 lambda λ值取 0.00469 时，MSE 最小，此时模型含有 26 个自变量；当λ值取 0.04805 时，得到的模型最简单，此时没有自变量。

根据交叉验证的结果绘制 CV 图，如下图所示：

图 4　Lasso 回归交叉验证 CV 图

R 语言中 cv.glmnet 函数利用交叉验证，分别用不同的 lambda 值来观察模型误差。上图中的$Log\ (\lambda)$的值作为横坐标，模型误差作为纵坐标，图中红点表示每个 lambda 对应的目标参量, 两条虚线表示特殊的 lambda 值。从上图可以看出，最佳的 lambda 取值在红色曲线的最低点处，此时对应变量的个数为 26 个，因此我们选择此时的λ值为 0.00469。

基本情况	婚姻状况 年龄 居住状况 职务 是否有本地户口 与本行的关系 是否有营业执照 从事行业 成员人数 家庭支出 贷款用途
保证联保	保证人性别 保证人实力（月收入） 联保小组成员关系
盈利能力	营业收入 月平均税收 总资产报酬率
营运能力	应收账款周转率 存货周转率 经营年限

	经营面积 雇员人数
偿债能力	所有者权益
宏观环境	居民消费价格指数

3. 信用评估指标体系的确立

由上述交叉验证可得，最优λ的取值为 0.00469，运用 coef 函数进行变相筛选，运用 R 语言运行筛选得到的 26 个最优指标及其系数，并将其概括为个体商户信用风险评价指标体系，如下表所示，（具体指标系数见附录一）

表 3　个体商户信用风险评价指标

三、信用评估模型的建立

（一）逻辑斯蒂回归（Logistic）

Logistic 回归模型是常用的适用于二分类问题的模型，其本质上是通过分析因变量属于某一类别的概率与自变量的关系，建立数学模型并达到预测的目的。

在本研究问题中，因变量Y的取值为 0 或 1，假定$Y=1$的概率为P，即个体工商户违约的概率为P，则个体工商户没有违约的概率 P(Y=0)=$1-P$，X为通过 Lasso 回归算法筛选出来的 26 个自变量。因此，个体商户违约（Y=1）的概率 P 的公式如下，

$$P(X)=P(Y=1|X) \qquad \text{公式（4）}$$

其中 X 为 26 维向量，则 Logistic 回归模型可具体表示为：

$$ln\frac{P(X)}{1-P(X)}=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n \qquad \text{公式（5）}$$

对上式做 Logit 转换即得：

$$\frac{P(X)}{1-P(X)}=e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n} \qquad \text{公式（6）}$$

通过上式，可以看出等式两边的取值范围均为$(0,+\infty)$，越趋于 0 表示违约概率越小，越趋于$+\infty$表示违约的概率越大。上述公式可以转化为关于$P(X)$的更直接的公式：

$$P(X)=\frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n}} \qquad \text{公式（7）}$$

由该式可知，个体工商户违约的概率$P(X)$与各个自变量的系数β_i有关且呈正相关，β_i越大，$P(X)$的值就越大，即当前个体工商户违约的可能性越大。

（二）随机森林模型（RandomForest）

随机森林，顾名思义由多个分类回归树随机组合在一起，构成了随机森林模型。

分类回归树基于数据自动构建树的结构，对于多个变量的交叉相进行解释，具有容易处理缺失值、容易处理离散型变量等的优点，其表达式可表示为如下：

$$f(x)=\sum_{m=1}^{M}c_m*1\ (x\in R_m) \qquad \text{公式（8）}$$

该公式为分类回归树表达式，其中，“ 1 ”为指示函数(indicater function)，意为当 x 值落入R_m区域则取 1，否则取 0。

分类回归树除了具有其他模型无可比拟的优势，但其仍然有自身缺点，比如其阈值不稳定，同一个总体中不同样本的选取会影响模型阈值的选取导致其预测精度不稳定。预测精度不稳定性本质上指测试集样本的抽取具有随机性，可以用方差进行刻画。为克服分类回归树以上的缺点，采取“ 利用均值降低方差 ”这一基本思想解决预测精度不稳定这一缺点，思想如下：

$X_1,X_2,\ldots,X_n$ n 个均值为 0，方差为σ^2的随机变量，

$$\mathrm{Var}(\overline{X})=\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{\sigma^2}{n} \qquad \text{公式（9）}$$

利用以上基本思想，构建随机变量的均值可降低方差大小，减小样本抽取的不稳定性。基于此思想，机器学习提出 Bagging 思想独立同分布抽取 n 个数据集

分别训练，将所得预测函数求和取均值，该项操作可使原先不稳定的预测效果有所稳定并大大提高。

随机森林就是基于 bagging 的思想解决分类回归树存在的缺陷。从训练样本 X 数据集中有放回地抽样 n 个数据集，再对每个数据集从 P 个指标抽取不同特征对应于数据集，构建四棵分类回归树，最后由投票原则输出预测值，极大地提高了预测的稳定性和精度。本文利用随机森林处理二分类问题对个体商户信用作出评估，具体示意图如下：

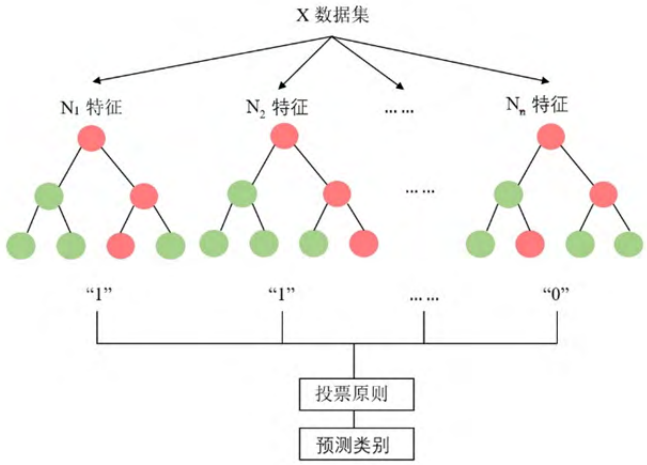


图 5　随机森林模型示意图

基于此模型建模时，由于数据原始类别不平衡，以 7：3 划分训练集和测试集，在 70%的训练集中共计 1509 条，其中违约数据 184 条，未违约占到 1325 条，容易导致模型准确率降低，因此需对数据做平衡处理。分别采用随机欠采样和随机过采样处理数据，并对两种方法的准确率比较分析，选取最优模型。

（三）KNN 分类模型（K- Nearest Neighbor）

KNN 又名 K-近邻模型，其本质为寻找距离未知类别 X 最近的 K 个数据，根据分类决策规则（如多数决策）推测该未知类别的分类。本文选取欧几里得距离计算不同数据间的距离，公式如下:

$$L_P(x_i, x_j) = \sum_{l=1}^{n} |x_i^{(l)} - x_j^{(l)}|^2 \qquad \text{公式（10）}$$

该算法原理简单易懂，且据以往研究结论来看，应用较广且准确率较高。其

原理示意图如下:

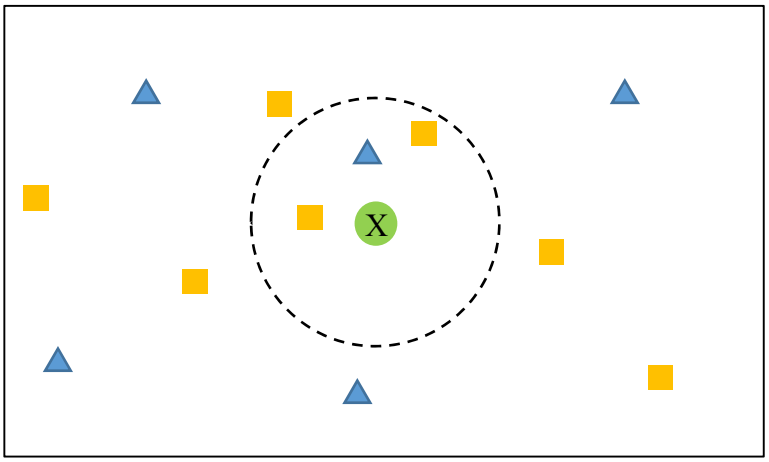


图 6　KNN 模型示意图

本文采用该算法建模的步骤为 :首先计算所求数据与其他数据的欧几里得距离，并对其作升序排列；其次选取前 K 个数据对其进行加权平均；最终根据分类决策预测 X 的类别。在建立此模型时，有两个难点，一是由于数据分布比例不平衡，因此简单的计算距离选取 K 值无法对数据作出有效预测，因此本文将类别为 0 和 1 的数据分别赋予权重，解决数据不平衡的问题；二是在对 K 值的选取上，该模型对 K 值的要求较高，K 值过高会导致分类模糊，而 K 值过低受个例影响较大，预测结果受干扰较大，本文在对 K 值选取上基于经验的判断，并选取不同 K 值运行最终得出最优模型，最终 K 值的取值为 30。

（四）支持向量机模型（Support Vector Machine）

支持向量机（SVM）为针对小样本数据进行高维分类效果较好的分类器，因此考虑信用风险评估的二分类问题时，对该模型加以运用。其原理为通过将不同变量所对应不同空间即个体商户是否违约以一条线进行划分，未知样本落入哪一区间则判定该样本属于什么类别，划分空间的线即为支持向量机，支持这条线成立的向量为支持向量。该模型示意图如下：

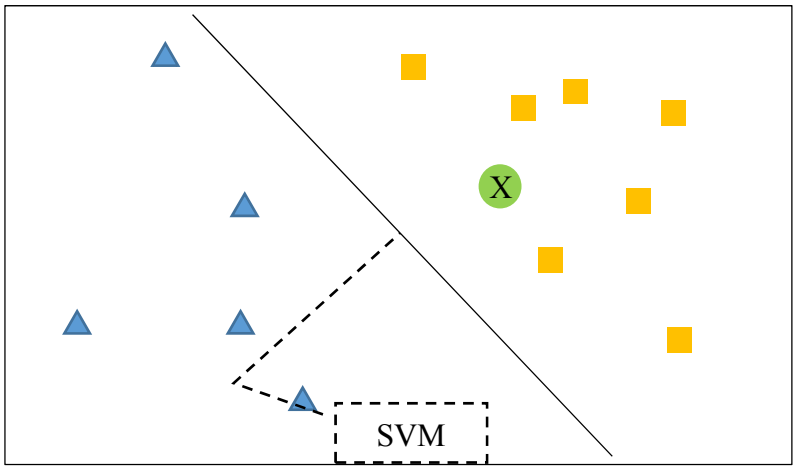


图 7　支持向量机模型示意图

针对信用评估问题，利用此原理建立模型存在两个难点：一是样本无法简单地以一条直线对其划分；二是划分区域直线的选取。因此，建立该模型时采用的基本原理为利用 python 首先将数据通过变换使其可映射到可用直线划分的空间，再使得距离直线最近的样本距离最大化。

四、模型在测试集上的检验

在本文将以混淆矩阵、ROC 曲线下的 AUC 值、F1 值对模型进行检验，并对模型预测效果做出评估。

其中 AUC 值为仅适用于二分类模型的评价指标，为 ROC 曲线下方的面积。ROC 曲线以假阳率 FPR（False Positive Rate）指标为横坐标，真阳率 TPR（True Postive Rate）为纵坐标，其中 FPR 指所有违约中预测为未违约的概率，TPR 指未违约正确预测的概率。计算公式如下，

$$FPR = \frac{FP}{FP+TN} \qquad \text{公式（11）}$$

$$TPR = \frac{TP}{TP+FN} \qquad \text{公式（12）}$$

其中 FP，TP，FN,TN 的定义以混淆矩阵的形式给出，如下所示：

表 4　混淆矩阵

	预测为未违约	预测值违约

真实未违约	TP	FN
真实违约	FP	TN

AUC 值也可用公式的形式表示出来, 具体含义为 :假设共有（x+y）个样本, x 个未违约样本，y 个违约样本，若未违约样本预测为未违约的概率值大于违约样本预测为未违约，则记为 1 并累加，以该值除以（x*y）则得到 AUC 值。具体公式如下,

$$AUC = \frac{sum(P(0\ 预测为\ 0) > P(1\ 预测为\ 0))}{x*y} \quad 公式（13）$$

其中 x 值为未违约样本个数，y 值为违约样本个数，sum 求和指对其计数累加。

F1 值综合了精准度（Precision）和召回率（Recall），精准度(precision)表示在预测出的违约个体商户中实际违约的个体商户占比，召回率（recall）表示在实际违约的个体商户中，成功预测出的违约个体商户的占比。其计算公式如下,

$$F1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \quad 公式（14）$$

$$precision = \frac{TP}{TP+FP} \quad 公式（15）$$

$$recall = \frac{TP}{TP+FN} \quad 公式（16）$$

准确率（accuracy）为预测正确数与预测总数量的比值，刻画了模型的整体准确率，但同时忽略了内在的精准度。

$$accuracy = \frac{TP+TN}{TP+FN+FN+TN} \quad 公式（17）$$

（一）逻辑斯蒂模型的验证及评价

借助 SPSS 软件对数据作 Logistic 回归分析，得到模系数的综合检验如下表 5 所示,

表 5　模型系数的综合检验

		卡方	df	Sig.
步骤 1	模型	133.335	26	.000

“模型”一行输出了 Logistic 模型中所有参数是否为 0 的似然比检验结果，这是总体评价的关键检验，也是判别模型是否有意义的标准。由上表可知，模型的显著性水平小于 0.05，说明本次拟合的模型总体有意义。

Hosmer 和 Lemeshow 检验可以评价模型是否充分利用了现有的信息、最大化地拟合了模型。利用 SPSS 分析可得下表：

表 6　Hosmer 和 Lemeshow 检验

步骤	卡方	df	Sig.
1	9.234	8	0.323

由上表可得，P 值（sig）=0.323。若 P>0.05，则可视作模型拟合优度效果较好，本模型中 P 值=0.323>0.05，所以该模型拟合优度效果较好。

对 SPSS 分析所得模型不同指标系数的结果观察可得，居民消费价格指数、月平均税收、贷款用途以及总资产报酬率等指标系数较大，由于该模型对分类型数据的处理上有一定缺陷，所以模型性能也因此有所局限。（各指标系数值见附录二）

观察模型准确度，利用其 F1 值、准确度进行验证，通过 Logistic 回归分析得到如下的混淆矩阵，计算得其 F1 值=93.99%，准确率=88.7%，但其对违约商户的预测值仅有 2%。虽然总体准确率较高，但预测性能较差。

表 7　逻辑回归混淆矩阵

	预测类别 0	预测类别 1	百分比校正
真实类别 0	1908	3	99.8%
真实类别 1	241	5	2.0%
总计百分比			88.7%

对该混淆矩阵的分析可得，该模型对预测未违约商户的准确率较高，难以判别违约商户，造成该结果的部分原因可能来自于数据的不均衡以及数据样本不足等因素，因此在模型预测上结果并不足以为银行提供决策支持。

（二）随机森林模型的验证及评价

本文采取随机森林的原理利用 python 建模，选取了 26 个最优指标，并得出 26 个指标的重要性以热力图形式呈现，如下所示，

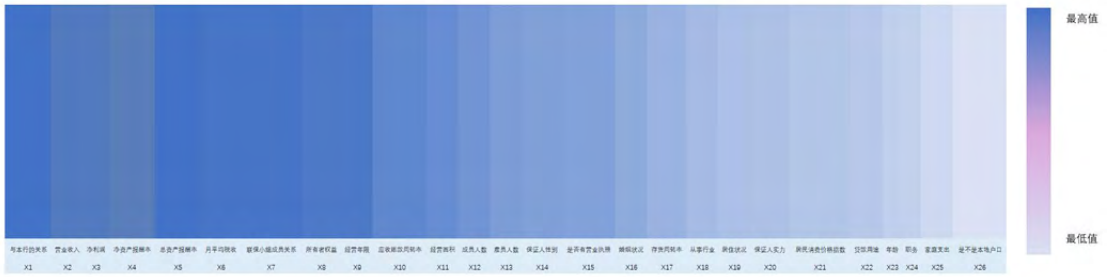


图 8　指标重要性热力图

观察图 8 可得，与本行的关系、营业收入、净利润、净资产报酬率、总资产报酬率等对个体商户信用评估的影响力较大，是否有本地户口、家庭支出影响极小，结果与人的主观感受相符。

采用随机欠采样和随机过采样处理数据后，并对两种方法比较分析，选取最优模型

表 8　随机欠采样和随机过采样比较

	随机欠采样	随机过采样
训练集预测准确率	78.26%	82.15%
测试集预测准确率	62.03%	72.06%

由表 8 分析可得，随机过采样效果较好，对模型预测效果绘制 ROC 曲线以及混淆矩阵如下，并得出其 AUC 值=0.5932，F1 值=82.9%，准确率=72.06%

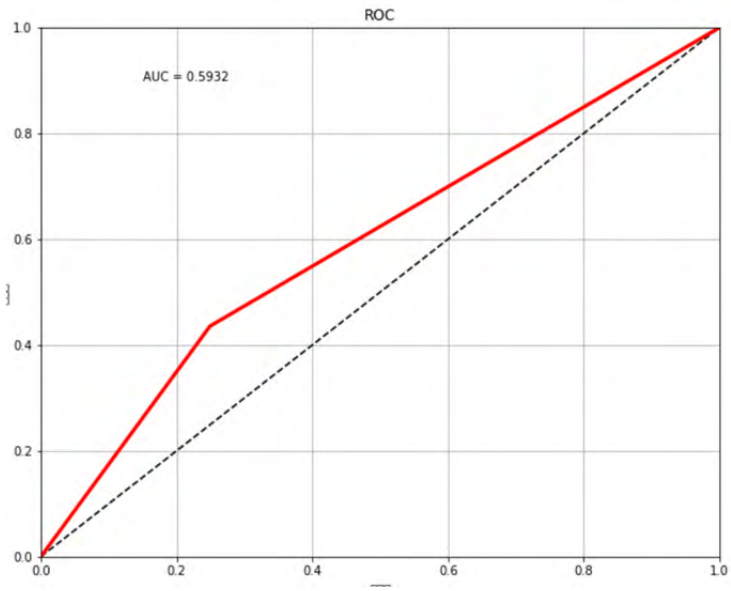


图 9　随机森林模型的 roc 曲线

表 9　随机森林混淆矩阵

	预测未违约 0	预测违约 1
真实未违约 0	440	146
真实违约 1	35	27

根据结果分析，随机森林虽然准确率总体较低，对商户信用的识别率较低，但其预测性能较好。随机森林在对商户信用评估问题的应用上，对未违约商户预测的可靠度较高，仅有约 7%概率误判，因此利用该模型预测为未违约的商户则可较为信任，而预测为违约客户极大可能不会违约，信用良好，因此对预测为违约的商户应了解更多信息判断其是否真的会违约。

（三）KNN 模型的验证及评价

基于 KNN 模型对信用评估作出的预测中，由于数据样本比例相对不平衡，违约与未违约数据比例 1：9，对于违约的预测性能有待提高。计算该模型预测的 F1 值=92.2%，准确率=85.9%，AUC 值=55.93%，其混淆矩阵如下：

表 10　KNN 模型混淆矩阵

	预测未违约 0	预测违约 1

真实未违约 0	226	14
真实违约 1	24	5

根据结果分析，其缺点与随机森林模型类似，均为在预测违约商户中存在较大误差，有约 75%的概率误判，因此需对预测为违约的客户进一步了解判断是否误判。

（四）支持向量机模型的验证及评价

支持向量机是适用于小样本数据的高效分类器，对数据依赖性较小，因此使用该数据可大大避免 KNN 模型的弊端，计算其 F1 值=94.0%，准确率=88.7%。AUC 值=64.82%，使用该模型对个体商户违约与否的预测混淆矩阵如下，

表 11　支持向量机混淆矩阵

	预测未违约 0	预测违约 1
真实未违约 0	1911	0
真实违约 1	244	2

由结果分析可得，该模型 AUC 值相对较高，所以其预测性能较好，使用该模型若预测值为违约，则该客户很大概率几乎一定会违约，若预测为不违约则仍需对顾客采取一定筛选，有大约 12%的概率会误判，因此该模型对违约商户的宽容度比较大，仍需采取其他人为措施更加精确的识别。

五、模型间对比与评价

（一）模型对比与分析

本文采用语 R 言、python 建立了逻辑回归、随机森林、KNN 和支持向量机四种模型对个体商户的信用作出评估，由于本文数据违约与否比例较为不平衡，对违约商户的预测性能较差，其模型评估比对效果如下，

表 12　模型评估对比表

	逻辑回归	随机森林	KNN	支持向量机
AUC 值		0.5932	0.5493	0.6481
F1 值	93.99%	82.9%	92.2%	94.0%
准确率	88.7%	72.06%	85.9%	88.7%

由表 12 中数据可得，在处理该个体商户信用评估问题上，支持向量机模型更为准确，验证了其处理小样本数据的优越性，同时其对违约商户的预测可信度极高，若其预测某个体工商户为违约商户，则可 100%认为该商户将存在违约行为，但也对违约商户具有较大的包容度；随机森林在处理该问题的总体准确度较低，但其具有指标重要性可视化的优势，且由于其对数据做了平衡处理，对未违约客户具有较准确的预测，若预测商户未违约，则仅有 7%的概率误判；逻辑回归模型虽然整体 F1 值以及准确率较高，且对未违约商户具有较高的判别性能，但其预测性能上若预测为违约商户有 12.1%的概率认为有误判可能，因此该性能不如随机森林，而且其对违约商户的预测性能较低，仅有 2%概率可预测违约商户，因此该模型不适合作为预测模型；KNN 模型在预测性能上较逻辑回归模型更为有效，但仍难以判别以及预测违约商户，仅有 20%的概率可判别商户信用是否可信。

因此，最终选取模型时，将支持向量机和随机森林模型均考虑在内，二者侧重点不同，利用两个模型双重验证从而提高预测精准度。

（二）个体商户信用评估模型的确立

结合以上分析，为对个体工商户作出更合理严谨的信用评估，本文建议使用随机森林和支持向量机双重检验。若支持向量机预测为违约商户，则极大可能该商户将存在违约行为，若支持向量机预测为未违约商户，则采用随机森林模型检验，若预测为未违约，则极大程度可信任其将不会存在违约状况，否则则需对该

个体商户进一步了解情况，人为判定其是否可靠，或对其降低贷款额度、签订相关保护协议等采取措施保障银行自身的利益。

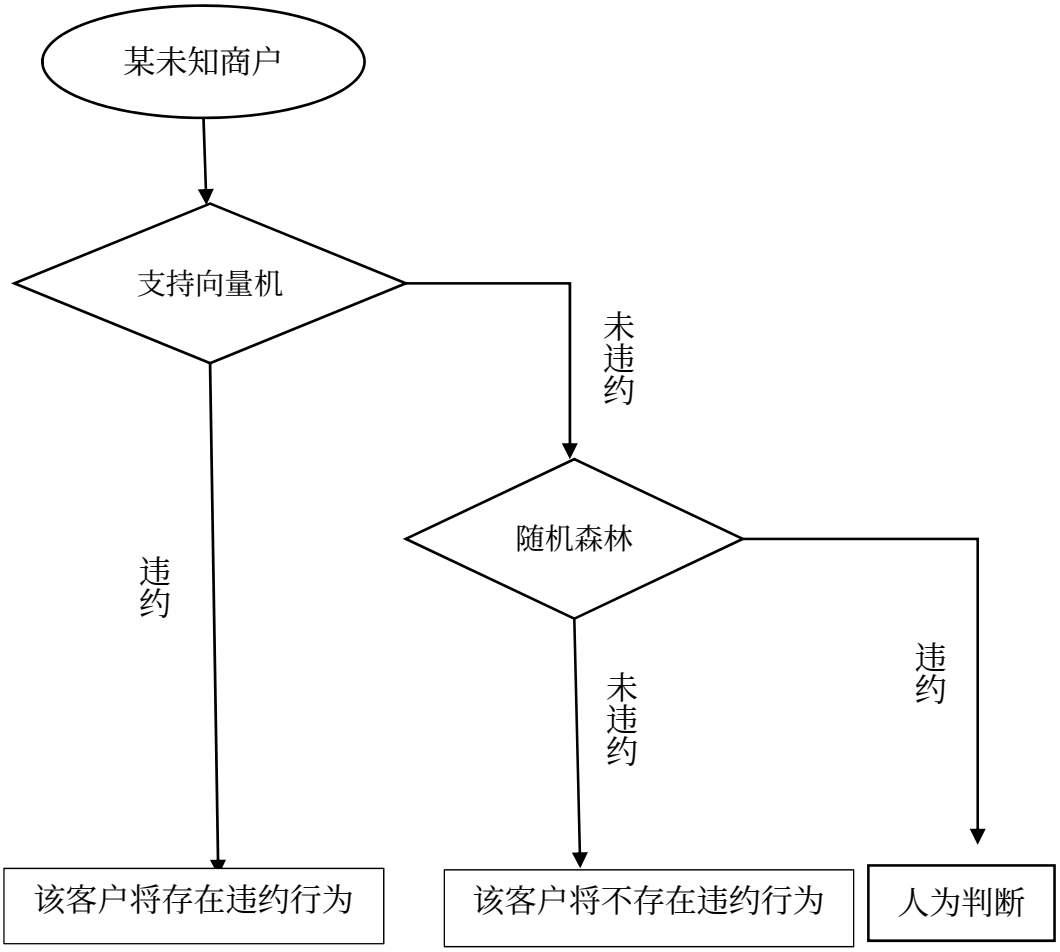


图 10　最优模型

六、总结与展望

（一）结论与分析

本文总体思路为先通过 Lasso 回归对数据降维处理，采用交叉验证法选取 lambda λ值，最终选取了 26 个指标作为最优指标建立信用评估模型指标体系；为探寻各项指标与个体商户信用之间的关系，采取机器学习领域的逻辑回归、随机森林、KNN 分类以及支持向量机的原理，选取合适的参数分别建立四个模型；比较分析四个模型的预测性能以及优缺点，分别计算其 AUC 值，并利用混淆矩阵计算 F1 值以及准确率（ACC），根据分析可得结论：

①逻辑回归（Logistic）模型总体准确率较高，但由于数据的不平衡性，根据混淆矩阵分析可得其预测性能仍然较差，不足以达到对信用评估的预测。

②随机森林（ForestRandom）模型虽然其 auc 值、F1 值以及准确率均偏低，但由于其采取随机欠采样和过采样两种方法对数据进行平衡处理，过采样效果明显优于欠采样，且对未违约商户的预测性能较好，精准度最高，且该模型解释性能也较好。

③KNN（K- Nearest Neighbor）分类模型对数据依赖性较强，由于该数据具有不平衡性且数据量较少，因此其模型准确度上略低于逻辑斯蒂回归，在对商户的预测性能上精准度也较低，不作为商户评估模型。

④支持向量机（Support Vector Machine）算法对数据依赖性较小，在对小样本数据的处理上具有先天的优势，因此该模型准确度和精准度均高于其他模型，其中对违约商户的预测精准度尤为准确。

综合以上分析，本文在选取模型上考虑到随机森林模型对未违约商户预测的良好性能以及支持向量机对违约商户预测的良好性能，最终选取支持向量机与随机森林模型双重检验作为模型评估。

本文利用 2157 条数据找寻数据间所存在的相对复杂的非线性关系，从数据中获取信息，并建立模型使其具有利用价值，可对个体工商户的信用作出可信度高于 90%的预测，为银行提供是否予以贷款的决策支持，规避银行对个体商户融资的风险等。除此之外基于本文所给模型，可对银行提以下建议：

①在个体商户信用评估问题上，商户的盈利能力和保证联保所涵盖的指标对个体商户信用影响较大，因此在对商户信用的评估作人为判断时，应着重查看这两项内容所对应的指标信息。

②利用模型对个体工商户作出信用评估时，违约商户可靠度：支持向量机>随机森林；未违约商户可靠度：随机森林>支持向量机。

（二）可进一步提升的方向

本文中所采用数据集共计 2157 条数据，违约与否大致比例为 1：9，存在极端不平衡问题，建立随机森林模型时为解决此问题采用了欠拟合和过拟合两种方法，从预测精度来看采取过拟合要更为准确，而其他模型未对数据平衡化处理，

对模型预测效果干扰较大，因此数据可进一步扩充使样本较为均衡，模型预测精度将有所提升。本文模型准确率很大程度上受制于此数据的局限性。

其次，本文在对指标的选取上，将 57 个指标利用 Lasso 回归筛选出 26 个指标作为后续的数据，该方法得到的可能并不是最优指标，因此在指标选择上可能有进一步优化。例如，可利用 KPCA 原理，即非线性主成分分析法对数据降维处理，更加全面、准确地寻找最优指标等

在模型选择上，本文主要基于机器学习准确率较高且比较传统的的逻辑回归、随机森林、KNN 分类以及支持向量机模型对数据作出分析，并未尝试神经网络模型，该原因主要还是局限于数据量的不足。因此在数据量大的前提下，可选择合适的神经网络模型，并结合相关原理调参优化，得到效果更好准确率更高的个体工商户信用评估模型，这也是进一步优化此模型的方向之一。

参考文献

[1] 夏冰.基于SLS-SVM的供应链视角下中小企业信用风险评估[J].工业技术经济,2021,40(06):77-82.

[2] 马铭泽.从贷款角度浅析商业银行信贷风险类型和防范措施[J].财经界,2016.

[3] 彭肖肖.农村信用社个体商户信用贷款风险及防范策略——以四川省眉山市农村信用社为例[J].农村经济与科技,2013,24(09):147-148.

[4] 张维,李玉霜 Zhangwei, Li,等.商业银行信用风险分析综述[J]. 管理科学学报, 1998.

[5] Daniel E O 'Lear y. On bankruptcy informatio n sy stems. Eur opean Jo urnal o f Opera tio na l Reasea rch . 1992; 56: 67～79

[6] Altman E, Eisenbeis R A, Sinkey J. Applica tio ns o f classification techniques in business. Banking a nd Finance . JAI Press , Greenwich. C T, 1981

[7] Colloms E Gho sh, Sco field C. An applicatio n o f a multiple neural-netw o rks lear ning system to emulatio n o f morg ag e underw riting judg ements . Pro ceeding s o f the IEEE Inter na tio na l Co nfer ence o n Neural Netw or ks 1988, 2: 459～466

[8] Yurt Alici. Neural netwo r ks in co rpora te failure predictio n: the UK ex perience. Pr oceeding s o f the Third Internatio nal Co nference o n Neura l Ne two rks in the Capita l Ma rket, 1995

[9] 白少布. 基于有序 logistic 模型的企业供应链融资风险预警研究[J]. 经济经纬, 2010, 000(006):66-71.

[10]刘艳春,崔永生.供应链金融下中小企业信用风险评价——基于 SEM 和灰色关联度模型[J].技术经济与管理研究,2016(12):14-19.

[11]江训艳.基于 BP 神经网络的商业银行信用风险预警研究[J].财经问题研究,2014(S1):46-48.

[12]耿成轩,李晓汨.基于样本加权 SVM 的科技型企业融资风险预警研究[J].工业技术经济,2020,39(07):56-64.

[13]谢梦龙,叶新宇,张升,盛岱超.LASSO 算法及其在边坡稳定性分析中的应用[J/OL].岩土工程学报:1-7

[14]农秋红,韦程东,罗文婷.Lasso 变量选择法在广西区域经济发展影响因素选取中的应用[J].中国商论,2021(10):162-164.

[15]期刊《李克强主持召开国务院常务会议 进一步支持微小企业》

附录一：Lasso 回归变量系数

最优变量	变量系数
婚姻状况	-0.52026595
年龄	-0.56485728
居住状况	0.15477463
职务	-0.11128449
是不是本地户口	0.17944537
与本行的关系	-0.35299691
是否有营业执照	0.06887163
从事行业	-0.21728043
成员人数	-0.13391928
家庭支出	-0.11905520
贷款用途	0.05522964
保证人性别	0.04059768
保证人实力	-1.07331520
联保小组成员关系	-1.00300987
所有者权益	0.25742123
净资产报酬率	-0.75191516
净利润	0.13006235
营业收入	0.35409730
总资产报酬率	0.57076551
月平均税收	0.45512160
应收账款周转率	-1.89767711
存货周转率	-0.13516968
经营年限	-0.75568901
经营面积	0.20358739
雇员人数	-0.43629672
居民消费价格指数	0.63552169

附录二：Logistic 回归分析结果

模型系数的综合检验

		卡方	df	Sig.
步骤 1	步骤	133.335	26	.000
	块	133.335	26	.000
	模型	133.335	26	.000

分类表[a]

已观测		已预测		
		违约与否		
		.0	1.0	百分比校正
步骤 1	违约与否 .0	1908	3	99.8
	1.0	241	5	2.0
	总计百分比			88.7

a. 切割值为 .500

模型汇总

步骤	-2 对数似然值	Cox & Snell R 方	Nagelkerke R 方
1	1397.680[a]	.060	.118

a. 因为参数估计的更改范围小于 .001，所以估计在迭代次数 7 处终止。

= Hosmer 和 Lemeshow 检验 =

步骤	卡方	df	Sig.
1	9.234	8	.323

方程中的变量

		B	S.E,	Wals	df	Sig.	Exp (B)	EXP(B) 的 95% C.I.	
								下限	上限
步骤 1[a]	婚姻状况	-.616	.278	4.904	1	.027	.540	.313	.932
	年龄	-.906	.453	4.009	1	.045	.404	.166	.981
	居住状况	.228	.215	1.122	1	.289	1.256	.824	1.915
	职务	-.214	.244	.770	1	.380	.807	.501	1.302
	是不是本地户口	.329	.237	1.927	1	.165	1.390	.873	2.212
	与本行的关系	-1.223	1.023	1.429	1	.232	.294	.040	2.185
	是否有营业执照	.175	.149	1.382	1	.240	1.192	.890	1.596
	从事行业	-.388	.278	1.945	1	.163	.678	.393	1.170
	成员人数	-.305	.403	.570	1	.450	.737	.334	1.626
	家庭支出	-.267	.571	.219	1	.640	.766	.250	2.343
	贷款用途	.665	1.135	.343	1	.558	1.945	.210	18.008
	保证人性别	.214	.400	.287	1	.592	1.239	.566	2.711
	保证人实力	-2.316	.966	5.744	1	.017	.099	.015	.656
	联保小组成员关系	-1.049	.419	6.279	1	.012	.350	.154	.796
	所有者权益	.113	.591	.037	1	.848	1.120	.352	3.564
	净资产报酬率	-2.689	1.087	6.123	1	.013	.068	.008	.572
	净利润	.502	.432	1.352	1	.245	1.652	.709	3.852
	营业收入	.581	.350	2.744	1	.098	1.787	.899	3.551
	总资产报酬率	1.968	.775	6.444	1	.011	7.160	1.566	32.733
	月平均税收	.684	.433	2.501	1	.114	1.982	.849	4.629
	应收账款周转率	-9.093	6.084	2.233	1	.135	.000	.000	16.982
	存货周转率	-.533	.556	.920	1	.337	.587	.197	1.744
	经营年限	-.990	.297	11.134	1	.001	.372	.208	.665
	经营面积	.550	.510	1.164	1	.281	1.733	.638	4.706
	雇员人数	-.866	.469	3.413	1	.065	.421	.168	1.054
	居民消费价格指数	.876	.301	8.481	1	.004	2.400	1.331	4.327
	常量	-.065	1.100	.003	1	.953	.937		

a. 在步骤 1 中输入的变量: 婚姻状况, 年龄, 居住状况, 职务, 是不是本地户口, 与本行的关系, 是否有营业执照, 从事行业, 成员人数, 家庭支出, 贷款用途, 保证人性别, 保证人实力, 联保小组成员关系, 所有者权益, 净资产报酬率, 净利润, 营业收入, 总资产报酬率, 月平均税收, 应收账款周转率, 存货周转率, 经营年限, 经营面积, 雇员人数, 居民消费价格指数.

致谢

在此首先要感谢全程指导我们、提供参考资料、指引模型方向的两位指导老师，为我们的模型建立、论文撰写等提供了宝贵的意见；其次要感谢院校联系人对此次比赛的负责认真，对论文进行查重、打包邮寄等；再次还要感谢帮助我们解决代码问题的同学；最后要感谢学校以及承办此次比赛的承办方，给了我们此次锻炼自己、展示自己能力的机会。

再一次对帮助、关心过我们的老师、学校以及同学等给予最衷心的感谢！