Modeling and Simulation 建模与仿真, 2021, 10(4), 991-1002
Published Online November 2021 in Hans. http://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2021.104099

Hans汉斯

基于数据挖掘的P2P网贷个人信用评价模型研究

何嘉欣，张　涛*，陈旭岚，关　悦

广西科技大学，广西　柳州

收稿日期：2021年8月24日；录用日期：2021年10月26日；发布日期：2021年11月2日

摘　要

近年来，随着互联网迅猛发展，P2P网络借贷平台凭借自身门槛低、收益高、操作便捷的特点，进入大众的视野并吸引了较多的借款者与投资者。如何提高、完善P2P网贷平台的风险监控能力，进一步加强对P2P网贷平台的管理，特别是对借款者的信用评价，降低投资者的投资风险是对P2P网贷行业未来发展十分重要的问题。本文针对国外Lending Club官网中给出的数据，对数据先进行预处理，再进行特征选择，最后选择评价指标在研究方法上分别使用传统的分类方法：Logistic回归模型、支持向量机模型和决策树模型。分析结果表明，支持向量机模型对结果预测准确度最高，相比之下，逻辑回归和决策树模型的预测能力相比较差些。在支持向量机模型的基础上，对模型的参数进行调优改进，改进后的模型预测准确度更高，因此该模型适合应用到P2P网贷个人信用评价中。

关键词

P2P网贷，个人信用评价，逻辑回归，支持向量机，决策树

Research on the Personal Credit Evaluation Model of P2P Net Loan Based on Data Mining

Jiaxin He, Tao Zhang*, Xulan Chen, Yue Guan

Guangxi University of Science and Technology, Liuzhou Guangxi

Received: Aug. 24th, 2021; accepted: Oct. 26th, 2021; published: Nov. 2nd, 2021

Abstract

As the Internet develops rapidly in recent years, the P2P network lending platform, with its low threshold, high income and convenient operation, has entered the public view and attracted more

*通讯作者。

文章引用: 何嘉欣, 张涛, 陈旭岚, 关悦. 基于数据挖掘的 P2P 网贷个人信用评价模型研究[J]. 建模与仿真, 2021, 10(4): 991-1002. DOI: 10.12677/mos.2021.104099

borrowers and investors. How to improve and perfect the risk monitoring ability of P2P network loan platform, further strengthen the management of P2P network loan platform, especially the credit evaluation of borrowers, and reduce the investment risk of investors is very important to the future development of P2P network loan industry. This paper aims at the data given in the official website of foreign Lending Club, carries on the characteristic selection to the data advanced, finally selects the evaluation index to use the traditional classification method respectively in the research method: the Logistic regression model, the support vector machine model and the decision tree model. Analysis results show that support vector machine model has the highest prediction accuracy and is suitable for P2P network loan personal credit evaluation model. Based on the SVM model, the parameters of the model are tuning and improved, and the improved model prediction accuracy is higher, so the model is suitable for application to the P2P online loan personal credit evaluation.

Keywords

P2P Net Loan, Personal Credit Evaluation, Logistic Regression, Support Vector Machine, Decision Tree

1. 引言

互联网行业的发展，使得互联网金融应运而生。借助大数据平台，将互联网技术与金融结合，则称之为互联网金融。而 P2P 就是互联网金融中的一种商业模式，主要针对于民间中的小额贷款，把资金提供给需要贷款的人。该过程为贷款人在平台上发放借款数目，投资者根据实际数目进行投资，双方交易自由。整个借贷过程主要依靠网络平台实现，成本相对较低，同时由于互联网的便利性，可随时通过网络管理资金，节省了大量的时间，其操作简单又便利。但由于 P2P 在国内发展正处于初级阶段，相关的法律以及监管机制并没有完善，极有可能出现平台跑路、用户逾期还款等现象，致使平台造成亏损。针对以上情况，各大借贷平台都会有借贷审批流程，主要对用户所提供的基本资料和还款能力进行分析，判断其是否有能力还款，最终确定是否给其提供贷款。在整个判断逾期的过程中，关键是依靠客户提供的信息，例如用户的历史贷款信息、历史还款信息、具备的存款金额等信息，找到影响客户逾期的关键因素，以此来建立客户逾期预测模型，最终判断用户是否逾期。本文对用户贷款信息数据进行分析，利用影响用户是否逾期的关键因素，建立多个用户逾期预测模型，并对各模型进行比较，找到对用户逾期判断较为有效的模型，更好地提高用户逾期预测的准确性。

目前，该领域在我国的发展相比外国较为滞后，因此国内学者在这方面的研究是从之前的仅停留在理论层面的研究，到近几年开始进行实证分析。随着该行业在我国的发展不断完善，对该领域的研究也日益不断深入，取得了较大的成果。王薇[1]主要以某贷款机构的历史贷款数据为例，建立信贷逾期行为的预测模型。首先将所得数据清洗和处理，然后通过 WOE 分箱和 IV 值选取包含信息量较大的特征，进行相关系数的计算，确定强相关的变量并去除，以免影响实验结果；并采用随机欠采样与 SMOTE 过采样相结合的方法去平衡训练集，以避免仅仅使用欠采样造成数据过度损失或仅使用过采样引入太多噪声。在模型选择方面，采用 Logistic 回归、支持向量机以及基于决策树的集成算法随机森林和 Light GBM，在平衡过的训练集上分别建模。最后在原测试集上预测，并根据预测的准确率和 AUC 值对模型进行评价，

综合各项指标选出最优模型进行信贷逾期行为预测。王洋琼[2]构建了以随机森林模型为主的数据挖掘模型，再研究数据集进行分析和预处理；接着运用随机森林模型及其他对比模型如 Logistic 回归模型、支持向量机模型、朴素贝叶斯模型等，对平台调整前后的借款人数据进行训练和预测；最后就各模型对借款人信用风险综合预测效果做出相关的评价和总结。王冬一等[3]从社会资本角度构建动态的个人信用评估体系，并利用人工智能算法进行指标筛选和性能评估对比。根据数据分析结果发现，极限梯度提升算法在预测准确性、稳定性以及灵敏度上性能较佳。因此，本文在前人的研究方法基础上使用 Logistic 回归模型、支持向量机模型和决策树模型对数据进行预测。同时，选取三个模型中预测性能最佳的一个模型进行优化，从而进一步提升预测水平，能更精准地挖掘出潜在逾期客户和优质客户。

2. 数据的搜集与处理

2.1. 数据来源

本文所用数据来自 Lending Club 公司官网，该公司是全球最大的 P2P 借贷平台公司。借助网络平台，连接投资者和借款者两端的端口，缩短资金流动的过程，为两方都提供了很大的方便与快捷。

为便于统计，只考虑原始数据中借款状况为“Fully Paid”、“Charged Off”和“Default”的贷款，将“Fully Paid”视为好客户，将“Charged Off”和“Default”视为坏客户。整理后得到的数据集共有 145 个变量，7828 条记录。内容有大致四个方面：一是基本信息，贷款金额等；二是还款能力，借款人的年收入等；三是不良信用记录，如逾期月数等；四是外部授信状况，如授信账户数等。

2.2. 数据预处理

在对构建模型前，原始数据需要进行预处理。如果省略了预处理阶段，则会使得所建模型不准确，错误的模型会得到错误的分析结果，严重误导决策。一个完整的建模过程至少有一半的时间和精力是用来进行数据的清洗、转换、抽样等预处理工作的，本文主要针对数据集的冗余信息、缺失值、不平衡问题、特征选择等方面的问题进行处理。

2.2.1. 剔除部分数据

由于原始数据经常存在异常值、缺失值、无效值等等，我们需要在建模之前先进行数据清洗，以提高数据质量，并且利于后续模型分类性能的提高。经过观察比较，本文剔除了以下几类变量。

1) 与本文研究对象“客户是否会逾期”无关的变量，如“title”，“hardship_flag”，“verification_status”。

2) 贷后信息，如“last_pymnt_d”，“last_credit_pull_d”，“initial_list_status”。

3) 数据分布极度不平衡且对于了解数据集帮助不大的变量，如“disbursement_method”，“debt_settlement_flag”。

4) 种类过多的变量，如“earliest_cr_line”，“zip_code”。

2.2.2. 缺失值处理

数据缺失的原因是多方面的，可能是由于数据采集过程中因工作疏忽导致的数据遗漏，也可能是由于数据特征本来就不存在，例如没有工作的人收入状况是无法填写的，也可能由于涉及到一些敏感信息导致人们不愿意透露，例如高收入人群不愿意透露收入情况等。数据的缺失会对数据挖掘造成一定的影响，从而增加模型预测的不确定性，可能会导致一些难以解释的结果。

对于缺失值的处理，一般采取以下几种方法：一是当特征包含大量缺失值，只有少量有效数据时，直接将该特征删除；二是当特征存在少量缺失值，采用相似的数据来替代；三是对缺失值不做处理。

本文选取的数据集中针对不同的缺失情况和缺失程度，使用不同的数据缺失处理方法。

1) 对于缺失程度大于或等于50%的特征，选择将该特征删除。

2) 对于缺失程度小于或等于5%的特征，选择将包含该特征缺失值的记录删除。

3) 对于缺失程度没有超过设定阈值的特征，则根据已有的数据进行填补。填补的方法主要包括以下几个方面：① 对于定量数据，若数据服从正态分布，采用均值填补的方法；若数据呈现偏态分布，则用中位数填补；② 对于定类数据，使用众数填补；③ 缺失值作为新的变量进行填补，如表1所示。

Table 1. County level planning schedule
表 1. 县域等级规划一览表

序号	变量名	缺失值个数	填补方法
1	il_util	1108	均值
2	mths_since_recent_inq	569	补0
3	emp_length	495	补0

2.2.3. 特征编码

通过观察可以发现，本文数据集中的这些变量并不都是连续的，存在一些特征是分类型变量。对于这些特征，许多机器学习的模型都无法识别，需要我们将这些特征用数字表示。对于一些自然有序的变量，在编码的过程中，为了尽量保持次序关系，可以按照取值水平将特征取值用数字依次表示。如果变量本身取值没有顺序，若直接采用数字编码，计算机会自动根据取值的大小进行判断，由此会产生错误信息。因此对于取值无顺序的分类变量，采用One-Hot编码，如表2所示。

Table 2. Characteristic encoding of the categorical variables
表 2. 分类变量的特征编码

变量名	特征编码
loan_status	“Fully Paid”: 1, “Charged Off”: 0, “Default”: 0
emp_length	“< 1 year”: 0.5, “1 year”: 1, “2 years”: 2, “3 years”: 3, “4 years”: 4, “5 years”: 5, “6 years”: 6, “7 years”: 7, “8 years”: 8, “9 years”: 9, “10+ years”: 10, np.nan:0
home_ownership	“Own”: 2, ”Mortgage”: 1, ”Rent”: 0
purpose	“credit_card”: 0, “home_improvement”: 1, “debt_consolidation”: 2, “other”: 3, “major_purchase”: 4, “medical”: 5, “small_business”: 6, “car”: 7, “vacation”: 8, “moving”: 9, “house”: 10, “renewable_energy”: 11

2.2.4. 数据规范化

本文选用Z-score规范化方法对本文的数据进行规范化处理。设变量 X 的取值为 $x_1, x_2, \cdots, x_n$，计算公式如下：

$$x_i^* = \frac{x_i - \overline{x}}{s}, i = 1, 2, \cdots, n \tag{2.1}$$

其中，$\overline{x}$ 为变量 X 的均值，s 为变量 X 的标准差。

2.2.5. 非平衡数据处理

本文采用SMOTE算法来解决数据不平衡的问题。具体步骤如下：

① 采取最近邻算法，常用欧式距离计算每个少数类样本的 K 个近邻。

② 在计算出的 K 个近邻样本中，随机抽取一个样本 $y_j\left(j = 1, 2, \cdots, k\right)$。

③ 构造新的少数类样本

$$NEW = x_i + rand(0,1) * (y_j - x_i) \tag{2.2}$$

其中，x_i 表示第 i 个少数类样本，$rand(0,1)$ 表示(0,1)区间中的任意随机数。

④ 最后合成新的数据集。

2.3. 特征选择

经过数据预处理，本文的数据集中还有 69 个变量，7370 条记录。本文因变量为用户是否为优质客户，在构建分类模型时，需要对特征进行选择，而不会选择所有变量输入模型，否则会增加模型的复杂度，容易造成过拟合。这就涉及到特征选择的问题。

2.3.1. GBDT 梯度提升树

在特征选择的过程中，一般需要注意以下两个方面：首先，特征是否包含足够信息。如果一个特征取值变化比较小，则该特征能反映的信息很少，对最后的分类结果贡献不大，将其视为无效变量。其次，要选取对建模有意义的变量，充分考虑到变量的预测能力、变量间的相关性等等。

本文利用 Python 软件的 Gradient Boosting Classifier()函数来确定各个特征的重要程度，之后建模所选取的特征即为从该方法中选取的排名前 10 的特征。

2.3.2. 相关性分析

本文采用 Person 相关系数表示变量间的相关性。

$$Corr(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} \tag{2.5}$$

一般而言，相关系数绝对值小于 0.4，认为相关性不强；相关系数绝对值大于 0.6，认为存在较强相关性。本文以 0.6 为界，找出存在较强相关的变量，再根据变量的 GBDT 值进行筛选，将 GBDT 值较小的变量剔除，如图 1 所示。

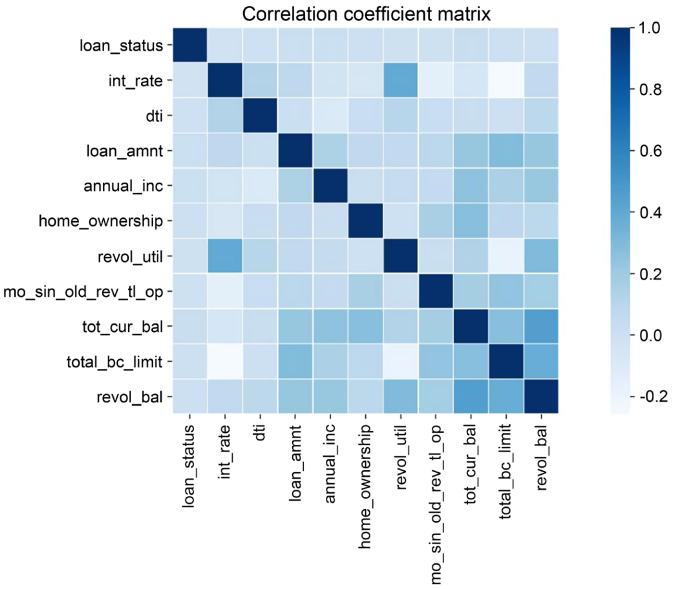

Figure 1. Person's correlation coefficient among some variables
图 1. 部分变量间的 Person 相关系数

经过计算发现，筛选出的变量间的相关系数都没有超过设定界限，因此不需要剔除变量。

综上所述，本文模型最终选用的特征变量如表 3 所示。

Table 3. The feature variables selected for the model
表 3. 模型所选特征变量

序号	变量名	变量解释
y	loan_status	贷款状况
x_1	int_rate	贷款利率
x_2	dti	贷款占收入比例
x_3	loan_amnt	贷款金额
x_4	annual_inc	贷款人年收入
x_5	home_ownership	住房性质：自有/按揭/租住
x_6	revol_util	透支额度占信用比例
x_7	mo_sin_old_rev_tl_op	自最早的周转账户开立以来的月份
x_8	tot_cur_bal	所有账户的总现金余额
x_9	total_bc_limit	银行卡的总信用额度
x_{10}	revol_bal	尚未结清的总贷款金额

2.4. 评价指标

针对分类问题，模型评估标准主要依据混淆矩阵进行。下面是二分类问题的混淆矩阵，用 T 和 F 表示预测是否准确，P 和 N 表示预测的分类是正类还是负类，具体形式如表 4 所示。

Table 4. Confusion matrix of the secondary classification problem
表 4. 二分类问题的混淆矩阵

混淆矩阵		预测类型	
		0	1
真实类型	0	TN	FP
	1	FN	TP

根据混淆矩阵可得如下指标：

识别准确率 $\text{Accuracy} = \dfrac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}}$；

识别精确率 $\text{Precision} = \dfrac{\text{TP}}{\text{TP}+\text{FP}}$；

召回率 $\text{TPR} = \dfrac{\text{TP}}{\text{TP}+\text{FN}}$；

假正率 $\text{FPR} = \dfrac{\text{FP}}{\text{FP}+\text{TN}}$。

Accuracy 称为识别准确率，预测准确样本占总体的比例。Precision 称为识别精确率，表示预测结果为正类的样本中实际为正类的比例。TPR 称为召回率，表示实际为正类的样本中预测为正类的比例。FPR

称为假正率，表示实际为负类的样本中被预测为正类的比例。在实际问题中，根据不同的需求，对于查准率和查全率各有侧重，但通常情况下查准率和查全率是相互矛盾的指标，提高其中一个指标会引起另一个指标的下降，任何一个指标过低都是不被希望的。

另外，ROC 曲线以及 AUC 值也常常被用来作为二值分类模型性能的度量指标。ROC 曲线将假正率 FPR 定义为 x 轴，真正率 TPR 定义为 Y 轴，虚线上的点表示 TPR = FPR，是随机预测的结果。当 FPR = 0，TPR = 1 时，表示所有的样本都分类正确，这是最理想的情况。实际中分类器基本不可能达到这么完美的结果。我们只能尽可能让曲线靠近左上角，此时表示分类正确的样本多于分类错误的样本。

3. 模型建立与预测结果

从海量的数据中搜寻、挖掘出有价值的信息，这一过程称之为数据挖掘。数据挖掘可以为人们在做决策的时候提供建议与支持，并在社会的各个行业中广泛应用。

在建立模型之前，先将数据分为训练集和测试集。本文抽取数据的 80%作为训练集，剩下的 20%用来测试。

3.1. 逻辑回归模型

3.1.1. 原理

Logistic 回归属于概率型非线性回归，是以因变量作为分类变量的回归分析方法。通过在线性回归的基础上套用函数，从而引进非线性因素来解决分类问题。

一般的线性回归方程如下：

$$f(x)=\omega_1 x_1+\omega_2 x_2+\cdots+\omega_d x_d+b=\omega^{\mathrm{T}}x+b$$

若函数是单调可微的，可得到广义线性模型：

$$y=g^{-1}\left(\omega^{\mathrm{T}}x+b\right)$$

对数几率函数为：

$$y=\frac{1}{1+\mathrm{e}^{-z}}$$

于是似然函数的似然项为：

$$\ln\ p\left(y_i\mid x_i;\omega,b\right)=y_i p_1\left(\hat{x}_i;\beta\right)+\left(1-y_i\right)p_0\left(\hat{x}_i;\beta\right)$$

目标函数为最大化似然函数，根据似然项可以看出，最大化 L 相应需要最小化 $l(\beta)$：

$$l(\beta)=\sum_{i=1}^{m}\left(-y_i\beta^{\mathrm{T}}\hat{x}_i+\ln\left(1+\mathrm{e}^{\beta^{\mathrm{T}}\hat{x}_i}\right)\right)$$

于是，目标函数转化为高阶连续可导凸函数，求解时可采用梯度下降法、牛顿法等数值优化算法，从而得到参数 ω 和 b 的估计值。

3.1.2. 预测结果

用逻辑回归模型对测试集进行预测，得到混淆矩阵如表 5 所示。根据表，整体准确率 Accuracy 为 0.6078，其中未逾期用户正确分类的样例为 730，逾期用户正确分类样例为 1000。精准率 Precision 为 0.6101，

召回率为 0.6061。

Table 5. The logistic regression model predicts the confusion matrix
表 5. 逻辑回归模型预测混淆矩阵

混淆矩阵		预测类型	
		0	1
真实类型	0	1000	440
	1	700	730

逻辑回归的 ROC 曲线如图 2 所示，AUC 值为 0.65。

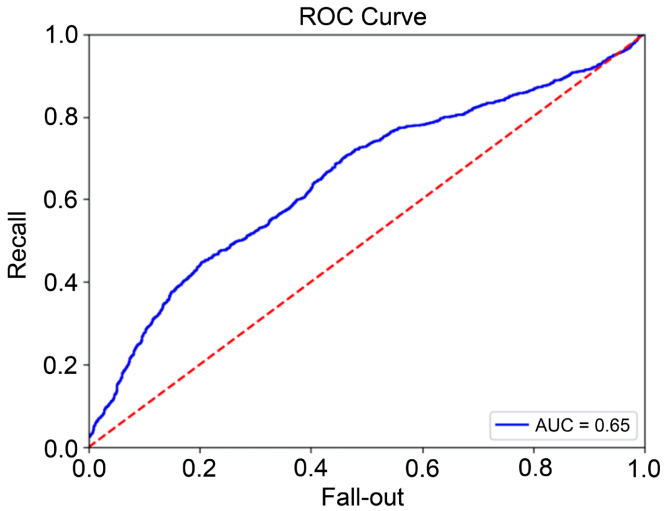

Figure 2. The ROC curve of the logistic regression model
图 2. 逻辑回归模型 ROC 曲线

3.2. 支持向量机模型

3.2.1. 原理

支持向量机的原理基于感知机。感知机的目标是寻找一个分离超平面，将数据中的不同类别分开，感知机寻找到的分离超平面有无数个。在此基础上，支持向量机的目的是寻找一个能够使间隔最大的最优分离超平面。间隔对应着分类的确信度。对于一个样本点而言，其距离分类超平面的距离越大，对于其分类的结果就越值得相信。于是，求解支持向量机转化为求解以下的凸二次优化问题：

$$\arg\min_{\omega,b}\frac{1}{2}\|\omega\|^2$$

$$\text{s.t. } y_i\left(\omega^{\mathrm{T}}x+b\right)\geq 1, i=1,2,\cdots,n.$$

3.2.2. 预测结果

用高斯核函数建立支持向量机模型，用测试集进行预测，得到混淆矩阵如表 6 所示。整体准确率 Accuracy 为 0.9048，其中未逾期用户正确分类的样例为 1200，逾期用户正确分类样例为 1400。精准率 Precision 为 0.9148，召回率为 0.9062。

支持向量机的 ROC 曲线如图 3 所示，AUC 值为 0.65。

Table 6. The logistic regression model predicts the confusion matrix
表 6. 逻辑回归模型预测混淆矩阵

混淆矩阵		预测类型	
		0	1
真实类型	0	1400	27
	1	250	1200

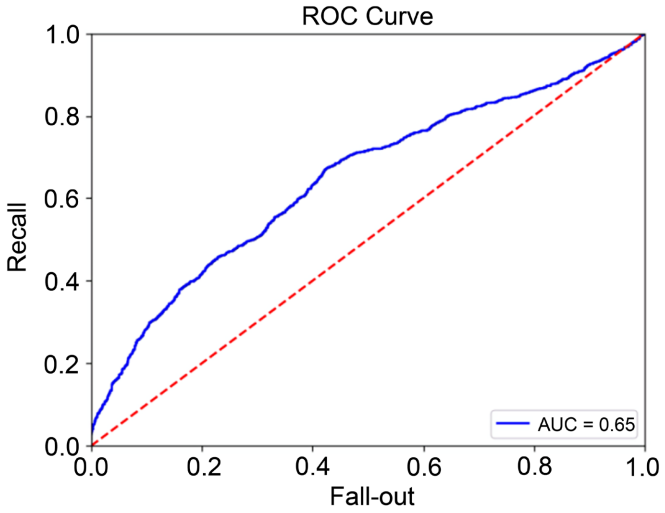

Figure 3. Support vector machine model ROC curve
图 3. 支持向量机模型 ROC 曲线

3.3. 决策树模型

3.3.1. 原理

决策树作为机器学习中的一类经典算法，主要用于分类和预测。决策树算法的目标是根据给定的训练数据集构建一个决策树模型，使它能够对实例进行正确的分类，其本质是从训练集中归纳出分类规则，或者说是由训练数据集估计条件概率模型。决策树算法的损失函数一般为正则化的极大似然函数，各节点的测试采用最小化损失函数，使最优决策树路径的综合损失函数最小。

3.3.2. 预测结果

用决策树模型对测试集进行预测，得到混淆矩阵如表 7 所示。整体准确率 Accuracy 为 0.7368，其中未逾期用户正确分类的样例为 870，逾期用户正确分类样例为 1300。精准率 Precision 为 0.7659，召回率为 0.7397。

Table 7. Decision tree model confusion matrix
表 7. 决策树模型混淆矩阵

混淆矩阵		预测类型	
		0	1
真实类型	0	1300	150
	1	620	870

决策树的 ROC 曲线如图 4 所示，AUC 值为 0.66。

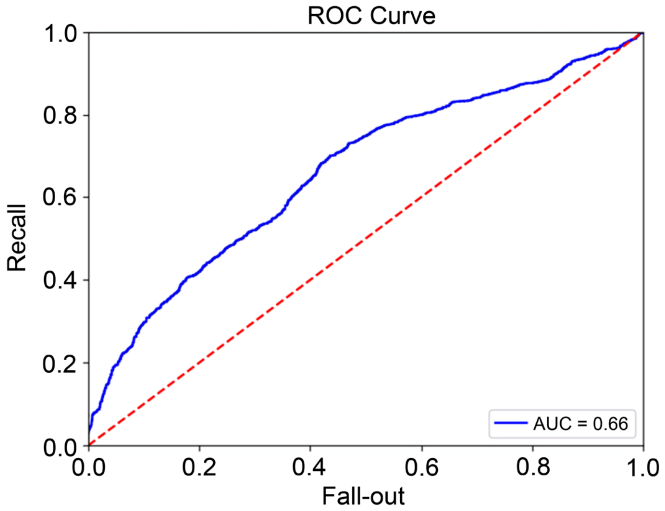

Figure 4. Decision tree model ROC curve
图 4. 决策树模型 ROC 曲线

4. 模型对比与评估

表 8 对各个模型的预测结果进行了整理，以便于对比和分析各模型的分类性能，如表 8 所示。

Table 8. Predictive comparison
表 8. 预测对比

	整体分类准确率(Accuracy)	精准率(Precision)	召回率(Recall)	AUC 值
逻辑回归	0.6078	0.6101	0.6061	0.65
支持向量机	0.9048	0.9148	0.9062	0.65
决策树	0.7368	0.7659	0.7397	0.66

针对上述的预测结果，可以明显看出各模型的预测能力，支持向量机模型的预测效果比逻辑回归模型和决策树模型的效果要好。针对这一结果，再结合各个模型在处理数据时的优劣势，来推断本实验中数据的特点以及从模型原理上分析各个模型的结果。

各个模型进行预测时的优劣势总结如表 9 所示。

Table 9. Comparison of the three models
表 9. 三个模型的比较

评估算法	优点	不足
逻辑回归(Logistic 回归模型)	既能处理分类问题也能处理概率问题	只能用于线性问题的解决；容易欠拟合，且在特征空间比较大时性能比较差
支持向量机(SVM 模型)	处理小样本能力强；能够解决非线性问题并具有泛化能力，找出的解是全局中的最优解	对异常值不敏感；训练数据过大时会占用大量计算资源
决策树(C5.0 模型)	处理速度更加快速；能够自动归并自变量中的类别，使其显著性达到最大	无法处理缺失值；忽略集中属性的相互关联；易出现过拟合问题

在本实验的预测结果中，支持向量机的预测结果最佳，可以推断出本次实验所使用的数据中异常值较少，并且数据集是小样本的数据所得到的结果是全局中的最优解；支持向量机可以抓住关键样本，避免大量冗余样本的影响，所以未逾期用户与逾期用户的分类正确率有所提高；决策树的预测效果次之，决策树算法的损失函数一般为正则化的极大似然函数，各节点的测试采用最小化损失函数，使最优决策树路径的综合损失函数最小，可以推断出数据集中的缺失值较多，集中属性的关联度较差甚至出现了过拟合现象；逻辑回归模型的预测效果最差，可以推断出该模型在预测时出现了欠拟合的现象，以及数据集的特征空间比较大，对于该模型来说处理较为困难。

针对以上的预测结果，为了进一步使模型的预测能力有所提升，因此选用最佳的支持向量机模型进行改进，以得到更好的预测结果。

5. SVM 模型改进

使用 scikit learn 库中的 GridSearchC V 进行参数调优。首先设置调参的参数列表，将参数列表变量设置为 C 在 0.1 到 10 之间取值，gamma 在 0.01 到 1 之间取值，这样网格搜索中就会有多种参数组合进行训练，最终得到一个训练效果最优的系数组合。因为本文训练样本较大，所以设交叉验证的折数为 10，把训练集分成 10 份进行交叉验证。最后得到输出结果最佳参数组合，C 取值 1.7，gamma 取值 0.0156 时分数最高为 0.84。使用优化后的参数分别对校验集与测试集进行预测，得到混淆矩阵如表 10 所示。

Table 10. The modified SVM model predicted the confusion matrix
表 10. 改进后支持向量机模型预测混淆矩阵

混淆矩阵		预测类型	
		0	1
真实类型	0	1450	40
	1	100	1300

由表中可以看出，模型预测结果的识别准确率为 0.9516，精准率为 0.9701，召回率为 0.9286。相较于改进前模型各个指标都有显著提高，说明改进后的模型对结果的预测更为精准。以该模型建立信贷逾期行为预测模型可以更为精准的识别出机构的优质用户和逾期用户。

6. 总结

6.1. 结论

本文对数据先进行预处理，再进行特征选择，最后选择评价指标在研究方法上分别使用传统的分类方法：Logistic 回归模型、支持向量机模型和决策树模型。分析结果表明，支持向量机模型对结果预测准确度最高，相比之下，逻辑回归和决策树模型的预测能力相比较差些。在支持向量机模型的基础上，对模型的参数进行调优改进，改进后的模型预测准确度更高，更适合建立个人信用评估模型。

6.2. 建议

在大数据背景下，根据个人信用体系以及相关的实证研究结果，尝试从以下三个方面提出相应建议：

1) 完善个人信用信息。个人信用与社会的各个相关指标间有着密切的关系，因此我们需要尽可能地完善个人的信息，也要重视社会资本对其的注重。

2) 完善大数据个人信用评价指标体系。建议在未来的大数据个人信用评估指标体系中加入一些能够评价社会资本的指标，从个人的家庭结构、社会关系网络规模、社会关系稳定程度、社交频率等方面对

个人的信用状态进行个人信用的画像分析。

3) 选取大数据个人信用评估算法。在大数据个人信用评估算法的选取上，预测的精确程度与预测算法的选择以及数据处理的方式以及数据的特征之间有着极大的关联。因此针对每个样本都应选取与其相适应的模型从而可以得到更为精准的预测效果。

6.3. 研究局限

尽管本文采用了较新的算法并严格控制实证研究和对比分析的过程，但仍存在一些缺陷和不足。首先，数据维度还不够丰富，有待进一步地扩展和丰富；其次，在实证研究中仅比较了 Logistic 回归模型、支持向量机模型和决策树模型 3 种算法的性能，对算法的分析和比较还不够全面。

因此，未来的研究应更加侧重于在实现多维度数据采集的基础上进行算法的优化整合和应用对比，从而更大程度地提升大数据个人信用评估的精准度。

参考文献

[1] 马薇. 基于数据挖掘算法的信贷逾期行为预测[D]: [硕士学位论文]. 太原: 山西大学, 2020.

[2] 王洋琼. 基于数据挖掘技术的 P2P 网贷借款人信用风险预测研究[D]: [硕士学位论文]. 重庆: 重庆理工大学, 2020.

[3] 王冬一, 华迎, 朱峻萱. 基于大数据技术的个人信用动态评价指标体系研究——基于社会资本视角[J]. 国际商务(对外经济贸易大学学报), 2020(1): 115-127.