Finance 金融, 2022, 12(4), 363-370
Published Online July 2022 in Hans. http://www.hanspub.org/journal/fin
https://doi.org/10.12677/fin.2022.124036

Hans汉斯

P2P网络借贷违约风险预测模型实证研究

陈思年[1]，邓家奇[2]，王　玭[2]

[1]云南财经大学统计与数学学院，云南　昆明
[2]西华师范大学数学与信息学院，四川　南充

收稿日期：2022年5月7日；录用日期：2022年5月19日；发布日期：2022年7月13日

摘　要

本文的实证研究数据来自Kaggle网站的比赛数据，该数据集爬取自某上市公司的用户信息，主要包含了6万多个借贷人的情况及贷款状态(是否违约)。本文依次基于传统的logistic回归模型、贝叶斯决策树、支持向量机和随机森林算法构建违约预测模型。按照文章中构建的评估指标来进行比较，得到随机森林模型的预测效果最佳。结果表明，本文选出的影响客户信用好坏的特征和风险预测模型有解释性，可以通过随机森林模型来预测客户的违约风险，有利于P2P网贷的发展同时也极大地为借贷公司减小损失。

关键词

P2P网络借贷，风险预测，个人信用，机器学习，随机森林

Empirical Research on Default Risk Prediction Model of P2P Network Lending

Sinian Chen[1], Jiaqi Deng[2], Pin Wang[2]

[1]School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
[2]College of Mathematics and Information Science, China West Normal University, Nanchong Sichuan

Received: May 7th, 2022; accepted: May 19th, 2022; published: Jul. 13th, 2022

Abstract

The empirical research data of this paper is from the contest data of Kaggle website. The data set is extracted from the user information of a listed company, mainly including the situation and loan status (default or not) of more than 60,000 borrowers. This paper constructs default prediction

文章引用: 陈思年, 邓家奇, 王玭. P2P 网络借贷违约风险预测模型实证研究[J]. 金融, 2022, 12(4): 363-370.
DOI: 10.12677/fin.2022.124036

model based on traditional Logistic regression model, Bayesian decision tree, support vector machine and random forest algorithm successively. According to the evaluation indexes constructed in this paper, the prediction effect of random forest model is the best. The results show that the characteristics and risk prediction models selected in this paper are explanatory, and the default risk of customers can be predicted through the random forest model, which is conducive to the development of P2P lending network and greatly reduces the loss of lending companies.

Keywords

P2P Network Lending, Risk Prediction, Personal Credit, Machine Learning, Random Forest

1. 引言

P2P (Peer To Peer)网贷模式[1] [2] [3]自出现在人们的视野后就成为了金融行业的焦点，围绕此模式衍生出的平台也渐渐浮出市场，加之我国科学技术和经济社会的高速发展，使得 P2P 网贷的发展更为安稳。本文将某公司作为研究对象，构建适用的 P2P 平台风险评估模型，主要选取了传统的 logistic 回归模型、贝叶斯决策树、支持向量机和随机森林算法，以便于为平台提供具有参考价值高的数据，更好的判断客户的借贷行为是否可靠，将公司的损失降到最低。

2. 文献综述

在美国等西方国家，个人信用评价体系已比较成熟。尤其以美国为例，它作为世界上信贷消费最高的国家，开创了信用评价的先河。正是因为美国倡导先消费、后还款的生活理念，美国人的一生都将伴随着信用交易，这些信用记录贯穿在他们生活的方方面面，包括申请贷款、就业、购置保险、交通出行等。

David Durand 是个人信用体系构建史上第一个认识到能将所学的数学公式等方法应用在信用评价上的学者，他首次使用线性判别来判断消费贷款以及个人信用的好坏，他成功地拉开数学方法应用在个人信用评价的序幕[4]。紧接着 1941 年较为完善的信用评价模型 Z-score 模型被 Altman 建立，同时也奠定了他在企业信用研究方面的地位[5]。1977 年，Altman、Haldeman 和 Narayanan 在 Z-score 模型的基础上，加入二次判别模型思想，形成了第二代信用评价模型，ZETA 信用风险模型(ZETA Credit Risk Model) [6]。ZETA 模型内含多种数理分析方法：Logistic 分析与判别方法、神经网络分析法、聚类分析法、判别分析法(Discriminant Analysis，简称 DA)、多元判别分析法(Multivariate Discriminant analysis)、层次分析法等等[7]。

第一个应用且至今仍在使用的信用体系评价模型是基于线性判别分析，在众多学者的研究过程中，不断地引入了新的统计模型和算法模型。目前我国对于个人信用体系模型构建的研究主要分为两个方面，一方面是评价指标体系的构建，另一方面则是评价体系信用分计算和风险预测。何建奎和岳慧霞结合我国现实情况与西方国家经验总结出适合中国个人信用体系构建的模式应是[8]：政府推动与市场运作相结合的模式，即采取以政府和中央银行为主导，以股份制资信公司为支撑，以现有信用中介公司为主体，以地区会员制为框架的全国个人信用体系[9]。

3. 模型简述

3.1. Logistic 回归模型

本文研究的因变量是贷款状态，违约者赋值为 1，未违约者赋值为 0。对于这种二元离散现象的数量分析，首先使用 Logistic 模型进行回归分析，模型的被解释变量 Y 是一个 0~1 变量，事件的发生概率是依赖于解释变量，即 $P(Y=1)=f(X)$，依赖于其影响因子解释变量的研究需要。

$$\text{Logit}(p)=\beta_0+\beta_1x_1+\cdots+\beta_px_p \tag{1}$$

根据 Logit 变换的定义：$\text{Logit}(p)=\ln\left[p/(1-p)\right]$，$p/(1-p)$称为发生比(odds)，得到最终的 Logistic 回归模型：

$$P=\exp\left(\beta_0+\beta_1x_1+\cdots+\beta_px_p\right)/\left(1+\beta_0+\beta_1x_1+\cdots+\beta_px_p\right) \tag{2}$$

3.2. 贝叶斯决策树

贝叶斯决策树是一种利用先验信息处理数据间非同质关系的树型分类法。该模型不需要分布的假定，它的求解采用非参数技术；贝叶斯(策树算法的关键是选择节(的分裂属性，常将((Entropy)、卡方(χ^2)以及基尼系数(Gini Index)作为计算信息增益的算法。熵是表示随机变量不确定性的度量，将 p_i 定义为分类变量 U 取值为 i 时的发生概率，若事件类型共有 s 类，则随机变量的熵定义为：

$$H(p_1,p_2,L,p_3)=-\sum_{i=1}^{s}(p_i\log_2 p_i)=H(U) \tag{3}$$

在本文中，s 取值为 2；又假设自变量为自变量 $X_1,X_2,\cdots,X_s$，则自变量 i 对应的 2，因子水平 k 记为 X_{ik}；将信息增益定义为：

$$I=H(U\mid X_i)=\sum_{k=1}^{n}P(X_{ik})H(U\mid X_{ik})=\sum_{k=1}^{n}P(X_{ik})\left(-\sum_{i=1}^{s}P(U_i\mid X_{ik})\log_2\left(P(U_i\mid X_{ik})\right)\right) \tag{4}$$

因此，对于自变量 $X_1,X_2,\cdots,X_s$，计算其对应的 $I(U,X_i)$，$I(U,X_i)$取值越大，则表示自变量 X_i；对于贝叶斯决策树分类具有更多的信息，则优先将 X_i；作为识别。

3.3. 支持向量机

支持向量机(Support Vector Machine, SVM)是基于统计学习 VC 维理论和结构风险最小原理共同建立的，它区别于经典统计学模型的最大特点是只需要将小样本作为特定的训练样本，通过训练得到学习精度和学习能力的最佳选择，便于获得最好的推广。上世纪九十年代，Cortes 和 Vapnik 提出了线性支持向量机，而后 Boser、Guyon 与 Vapnik 引入核技巧，提出了非线性支持向量机，功能较为完整的支持向量机应用于当今社会的诸多方面。总的来说，SVM 主要解决了两类问题，一是寻找到最优的超平面二是能够划分非线性可分的样本[10]。

由于非线性均在线性的基础上进行研究，如图 1 所展示的线性支持向量，其中小叉叉和空心圆点分别表示两类样本，w 表示该平面的法向量，H_1 和 H_2 为上边界和下边界，且令 $\text{margin}=\dfrac{2}{\|w\|}$ 为间隔距离，H 为最优分割线。

3.4. 随机森林

随机森林(Random Forest，简称 RF)是 Bagging 的一个扩展体，它在以决策树为基学习器构建 Bagging 集成的基础上，进一步在决策树的训练过程中引入了属性选择。随机森林相比于单纯的决策树算法消除

了许多局限性，减少了数据集的过拟合、提供了一种处理丢失数据的方法，因此提高了精度[10]。具体算法结构如图 2 所示：

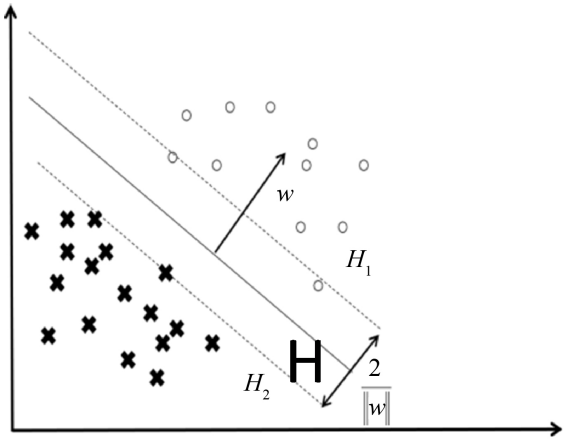

Figure 1. Schematic representation of the linear classification
图 1. 线性分类示意图

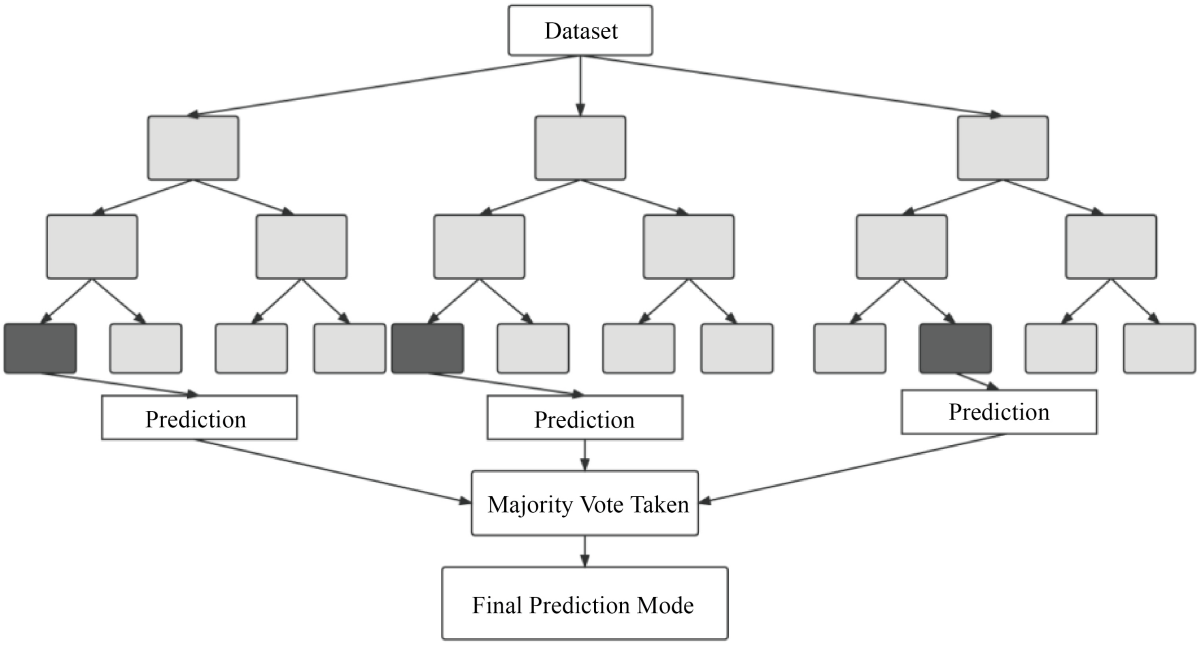

Figure 2. Random forest algorithm structure
图 2. 随机森林算法结构

3.5. 评价指标

评估指标是用来衡量模型优劣水平的算法，对于分类很多指标可以对其进行评价，如精确率、召回率(precision-recall)、错误率(error rate)、精度(accuracy)、误差(error)、训练误差、泛化误差等等。对于本文的分类模型，就是区分客户是否违约，分别对应正、负样本数，用 P、N 来表示。在统计分析当中，我们可以用一个混淆矩阵(见表 1)来表示分类正确与否的情况，并在此基础上计算：

Table 1. Classifies the resulting confusion matrix
表 1. 分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

正确率(precision/accuracy)：测试集上被模型正确分类的样本数所占的百分比。即：

$$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{P}+\mathrm{N}} \tag{5}$$

错误率(precision/accuracy)：测试集上被模型错误分类的样本数所占的百分比。即：

$$\frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{P}+\mathrm{N}} \tag{6}$$

召回率/真正率(sensitive/recall)：即在所有实际为正样本的情况下，成功预测为正样本的百分比。即：

$$\mathrm{TRP}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{7}$$

精度(sensitive/recall)：即即在所有预测为正样本的情况下，实际上为正样本的百分比。即：

$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{8}$$

假正率(False Positive Rate, FPR)即被预测为正的负样本结果数/负样本实际数。

$$\mathrm{FRP}=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{FN}} \tag{9}$$

F 度量即(F_1 和 F_P)度量的方法就是将精度与召回率的计量方法结合后组成新的计量方法。定义如下：F_1 是精度和召回率的调和平均数

$$F_1=\frac{2*\mathrm{precision}*\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}} \tag{10}$$

F_P 加的权系数是召回率和精度的 β 倍：

$$F_P=\frac{\left(1+\beta^2\right)*\mathrm{precision}*\mathrm{recall}}{\beta^2*\mathrm{precision}+\mathrm{recall}} \tag{11}$$

当 $\beta=1$ 时，模型对于准确率和召回率都一视同仁。

当 $\beta<1$ 时，模型主要重视准确率。

当 $\beta>1$ 时，模型主要重视召回率。

β 的取值通常是 2 或者 0.5，基于本文是研究客户的违约风险，故越精确越好，也就是尽可能地避免有违约风险的客户被预测为没有违约风险，这里选取 β 的取值为 2，即

$$F_P=\frac{5*\mathrm{precision}*\mathrm{recall}}{4*\mathrm{precision}+\mathrm{recall}} \tag{12}$$

ROC 全称是“受试者工作特征”(Receiver Operating Characteristic)，AUC (area under the curve)其实就是 ROC 曲线的面积，两者相结合还有 AUROC (area under the receiver operating characterstic curve)指标。

ROC 曲线和 AUC 曲线常常作为衡量二分类分类器的重要指标，其中 ROC 曲线越靠近左上角说明在 FPR 很小时 TRP 很大，AUC 在分析中常类似成绩一样地被分为五个区间：

$AUC \in (0.9,1)$ 意味着优秀(Excellent)；

$AUC \in (0.8,0.9)$ 意味着良好(Good)；

$AUC \in (0.7,0.8)$ 意味着尚可(Fair)；

$AUC \in (0.6,0.7)$ 意味着不好(Poor)；

$AUC \in (0.5,0.6)$ 意味着失败(Fail)；

也有学者对 AUC 的评价准则做出了不同的解释：

AUC = 1，最好的分类器，若使用此种分类器，可以得到不止一个阈值的最优预测结果。但是大部分的猜测状况下并不会出现最好的分类器。

0.5 < AUC < 1，优于随机猜测，这个分类器若是选择合适的阈值，会有预测价值；

AUC = 0.5，与随机猜测原理相同，模型是没有预测价值的；

AUC < 0.5，比随机猜测还差；

但不管是哪种评价准则，都有样的规律，即 AUC 值越大，分类结果越好。

4. 实证结果分析

数据说明

本研究使用的数据集主要包含各种属性，例如资金金额，位置，贷款，余额等，具体说明如表 2：

Table 2. Variables introduction table
表 2. 变量介绍表

变量名称(中文)	取值方式	变量名称(中文)	取值方式
ID	由数字构成唯一 ID	开户	数值
贷款金额	数值	公共记录	数值
资助金额	数值	循环余额	数值
投资者资助金额	数值	循环公用事业	数值
期限	数值	总账户	数值
批量注册	字符串数据	初始列表状态	2 = w、1 = f
利率	数值	收到利息总额	数值
等级	7 = G、6 = F、5 = E、4 = D、3 = C、2 = B、1 = A	收到滞纳金总额	数值
子等级	字符串数据	回收	数值
就业期限	3 = MORTGAGE、2 = RENT、1 = OWN	托收回收费	数值
房屋所有权	数值	收集 12 个月医疗	数值

Continued

验证状态	3 = Not Verified、2 = Source Verified	申请类型	2 = INDIVIDUAL、1 = JOINT
	1 = Verified	上周付款	数值
付款计划	2 = n、1 = y	拖欠的账户	数值
贷款所有权	字符串数据	总收款金额	数值
借方与收入之比	数值	总当前余额	数值
拖欠——两年	数值	总循环信用额度	数值
查询——6 个月	数值	贷款状态	1 = 违约者，0 = 非违约者

将数据集的前百分之八十作为训练集用以训练模型，将后百分之二十作为测试集，比较不同模型的预测结果(表 3)：

Table 3. Comparison table of the model prediction effect
表 3. 模型预测效果对比表

模型	正确率	精度	F_P
logistic 回归模型	90.75%	90.75%	98.00%
贝叶斯决策树	90.75%	90.75%	98.00%
支持向量机	91.90%	90.82%	98.10%
随机森林模型	93.36%	91.41%	98.16%

比较模型预测结果，以得到以下结论，logistic 回归模型和决策树模型在精度上都较低，支持向量机模型和随机森林模型的精度较高，但效果最好的是随机森林模型。为了更好的观察随机森林模型的预测效果，接着做出 ROC 曲线图 AUC 图进行观察，通过图 3 可以看出，该模型的 ROC 曲线远离纯随机分类器的 ROC 曲线(AUC 等于 0.5)，AUC 等于 0.889∈(0.8, 0.9)，故该模型的违约风险预测效果良好。

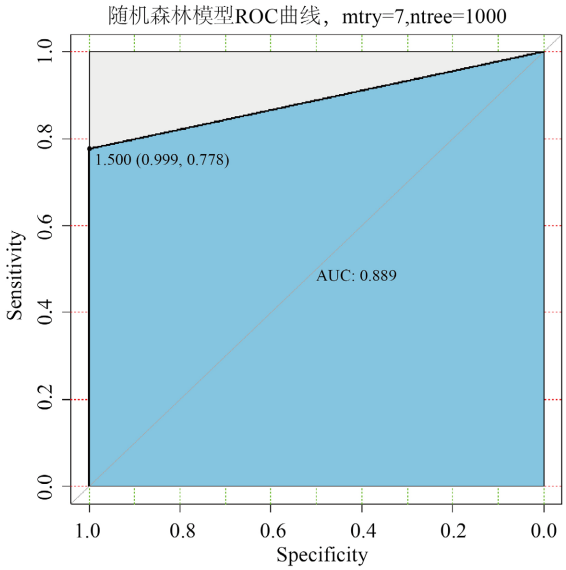

Figure 3. The ROC plot of the random forest model
图 3. 随机森林模型 ROC 图

5. 结论

本文基于某公司信贷数据集，通过比较 logistic 回归模型、贝叶斯决策树、支持向量机和随机森林算法构建违约预测能力，根据结果得知对于此数据集，随机森林的预测效果最佳。

随着市场经济的不断发展与成熟，信息时代的到来，金融信用发展到一个新的层次与水平，加之我国法制的进步和人民法院强制执行力的不断加强，我们越来越频繁地从媒体上听到“老赖”这个词，一旦被列入“老赖”行列，既坐不了高铁，也乘不了飞机，出行非常不方便，而且还会影响到孩子的教育。但是近年来，在个人消费的各个方面出现的失信行为，信用缺失已成为当前经济社会发展的“瓶颈”，不仅影响到人们的日常生活，也影响到我国市场经济的健康发展。

为了国家更好的建设和完善社会信用体系，也为了自身的金融信用、日常生活不受影响，客户要做到以下几点：

1) 按时归还欠款，防止贷款逾期或欠息；

2) 若无法一次性付清借款可分期；

3) 理性做好规划，加强个人金融管理，避免超前消费和过度消费。

同时我国现在处于建全违约风险相关法律法规的重要阶段，国家和政府的相关部门也需要付出相应的行动，以下是对相关部门提出的建议：

1) 加快个人征信体系的建设；

2) 加强对借贷人的教育；

3) 推动金融惠普发展。

模型考虑的影响因素较单一，基本为微观数据，后期尝试引入部分宏观影响因素如：国际环境、国家政策、社会舆论等，自 2019 年新冠疫情的爆发，不仅对我国经济发展造成了严重的影响，甚至对国际经济而言影响都很大，故在后续的研究当中，会收集资料将新冠疫情也作为影响因素之一构建借贷风险预测模型。

基金项目

国家级大学生创新创业项目(202110638006)。

参考文献

[1] 邓春生. 演化博弈视角下 P2P 网络借贷的信用风险及其法律规制研究[D]: [博士学位论文]. 成都: 西南财经大学, 2020. https://doi.org/10.27412/d.cnki.gxncu.2020.000320

[2] 傅一帆. 社会网络视角的 P2P 平台机制设计研究[D]: [硕士学位论文]. 杭州: 浙江大学, 2015.

[3] 戴宙松. P2P 网络借贷相关会计核算问题研究[D]: [硕士学位论文]. 西安: 长安大学, 2015.

[4] Grobben, M. (1943) Risk Elements in Consumer Instalment Financing. David Durand. *Journal of Political Economy*, **51**, 185-186. https://doi.org/10.1086/256026

[5] Zhang, X.Y., Xie, Q. and Song, M. (2021) Measuring the Impact of Novelty, Bibliometric, and Academic-Network Factors on Citation Count Using a Neural Network. *Journal of Informetrics*, **15**, Article ID: 101140. https://doi.org/10.1016/j.joi.2021.101140

[6] Ptak-Chmielewska, A. (2021) Bankruptcy Prediction of Small- and Medium-Sized Enterprises in Poland Based on the LDA and SVM Methods. *Statistics in Transition New Series*, **22**, 179-195. https://doi.org/10.21307/stattrans-2021-010

[7] Altman, E.I., Esentato, M. and Sabato, G. (2020) Assessing the Credit Worthiness of Italian SMEs and Mini-Bond Issuers. *Global Finance Journal*, **43**, Article ID: 100450. https://doi.org/10.1016/j.gfj.2018.09.003

[8] 鲁秀秀. P2P 网络借贷借款人的信用风险研究[D]: [硕士学位论文]. 济南: 山东大学, 2021.

[9] 何建奎, 岳慧霞. 中国个人信用体系模式选择[J]. 消费经济, 2004(3): 49-52.

[10] 史小伍. 基于支持向量机的组合预测模型及其个人信用评价方法[D]: [硕士学位论文]. 镇江: 江苏科技大学, 2012.