

Logistic模型在银行贷款中的应用

杨巾铤

云南财经大学统计与数学学院, 云南 昆明

收稿日期: 2023年4月15日; 录用日期: 2023年5月18日; 发布日期: 2023年5月25日

摘要

本文使用Logistic模型对台湾客户是否违约支付建立预测模型, 通过这个模型可以在银行给客户贷款时判断客户是否会违约。首先, 由于数据中有23个变量, 其中有些变量并不显著, 遂采用最优子集的方法判断出模型最优的变量个数为8。再通过Forward Stepwise Selection方法选择出8个变量并对此建立Logistic模型。通过将数据分为训练集和测试集来得到模型的精准度: 模型整体预测准确率为80.2%, 总体精度还算可以, 模型对客户不违约的预测还是非常准确, 但对客户违约的预测非常不理想。同时, 采用另一种可视化的方法衡量模型的优劣, 即ROC曲线, 计算出AUC的值为0.66。模型的结果优于我们随机猜测, 具有预测价值。

关键词

Logistic模型, 次级贷危机, 分类预测

The Application of Logistic Model in Bank Loan

Jinyi Yang

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan

Received: Apr. 15th, 2023; accepted: May 18th, 2023; published: May 25th, 2023

Abstract

In this paper, the Logistic model is used to establish a prediction model for the default payment of Taiwan customers. Through this model, the bank can judge whether the customer will default when lending to the customer. First of all, since there are 23 variables in the data, some of which are not significant, the optimal number of variables in the model is judged to be 8 by the optimal subset method. Then the Forward Stepwise Selection method selects 8 variables and establishes the Logistic model. The accuracy of the model was obtained by dividing the data into training set

and test set: the overall prediction accuracy of the model was 80.2%, and the overall accuracy was reasonable. The prediction of non-default by the model was still very accurate, but the prediction of default by the customer was very unsatisfactory. At the same time, another visual method was used to measure the merits of the model, namely the ROC curve, and the value of AUC was calculated as 0.66. The results of the model are better than our random guesses and have predictive value.

Keywords

Logistic Model, The Subprime Crisis, Classification Prediction

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

美东时间 3 月 8 日, 美国 Silicon Valley Bank (硅谷银行) 因资不抵债遭金融监管部门关闭。事实上, 在硅谷银行遭到接管的 48 小时以后, 位于纽约的 Signature Bank 同样遭到了关闭与政府接管。硅谷银行的倒闭, 是美国历史上第二大银行倒闭案, 仅次于在 2008 年金融危机中倒闭的 Washington Mutual (华盛顿互助银行)。3 月 10 日, 美国银行业动荡, 风波也在向美国以外的国家传导, 英国和韩国正在密切关注风波影响[1]。

如此大范围的影响不禁让人联想到 2008 年的由雷曼兄弟破产而引发的全球金融危机。银行为了消化过剩产能, 向人们发放大量贷款拉动内需, 刺激消费。此次房贷包括了巨额次级贷[2]。次级贷, 指的是次级抵押贷款(subprime mortgage loan)。次级抵押贷款的受贷人一般信用等级比较低, 信贷记录差, 因而违约风险大[3]。随着楼市降温, 短期利率上升, 次贷还款利率也大幅上升, 这部分购房者的还贷压力随之陡增。2006 年夏, 美国媒体上关于次贷购房者违约的报道多了起来, 越来越多的人还不起房贷, 被银行扫地出门。银行虽然收回房子, 但却卖不出高价, 因而面临大面积亏损, 次贷危机爆发。

胡冰在次贷危机对我国商业银行的启发中阐述了次级贷危机爆发的根源, 其中就有宽松的贷款条件积累了大量的信用风险[4]。任重在美国次级贷危机的成因分析中提出要对个人住房贷款坚持严格的审查标准, 防止无偿还能力的低收入者任意进入房屋抵押贷款市场[3]。

如今, 银行会对贷款客户的财务指标进行量化分析, 计算出该客户的初始违约概率并对其划分一个初始信用等级, 然后根据信用等级将客户归于某一类债务人, 并根据一组具有相同风险特征的债务人的历史违约记录确定客户的最终违约概率。这样得到的最终违约概率既考虑了客户的财务数据又考虑了客户的历史违约情况, 提高了银行测算公司类客户贷款是否违约的准确度[5]。

2. 研究设计

数据来源于 UCI Machine Learning Repository: Data Sets 数据库网页中 default of credit card clients Data Set。此数据集的内容是针对台湾客户的违约支付, 共有 3000 个数据, 其中有 23 个变量为解释变量, Y 是二分类变量: 违约支付(是 = 1, 否 = 0)作为响应变量。本文采用 R 软件对数据集进行分析。

我们想要了解银行贷款的客户是否可以信赖, 能否按时还上贷款, 数据中已有的二分类变量作为响应变量, 故采用二元 Logistics 回归模型进行分类。以便银行在给客户贷款时可以将客户数据输入模型中

可以得到客户是否会违约，在一定程度上可以有效避免次级贷危机的出现。

Logistics 模型的基本形式如为：

$$\ln\left(\frac{p(y=1)}{1-p(y=1)}\right) = \beta_0 + \sum_{i=1}^k \beta_i x_i$$

模型中， x_i 表示判断客户是否违约的第 i 个变量， k 为变量个数， β_0 为截距项， β_i 为 x_i 的系数，反应该变量对客户违约的程度，通常用极大似然估计法求得。客户违约的概率与客户不违约的概率比值为事件发生比 $p(y=1)/[1-p(y=1)]$ [6]。我们选择的变量内容见表 1。

Table 1. Variable selection

表 1. 变量选取

变量	变量显示	变量赋值及含义
给定信贷的金额(新台币)	X_1	包括个人消费信贷和 TA 的家庭(补充)信贷
性别	X_2	1 = 男性, 2 = 女性
教育	X_3	1 = 研究生, 2 = 大学, 3 = 高中, 4 = 其他
婚姻状况	X_4	1 = 已婚, 2 = 单身, 3 = 其他
年龄	X_5	
过去付款的历史	$X_6 \sim X_{11}$	X_6 = 2005 年 9 月的还款情况; X_7 = 2005 年 8 月还款情况; …… X_{11} = 2005 年 4 月的还款情况。 还款状况的测量尺度: -1 = 按时还款; 1 = 延迟付款一个月; 2 = 付款延迟两个月; …… 8 = 延迟支付 8 个月; 9 = 延迟付款 9 个月或以上。
账单金额(新台币)	$X_{12} \sim X_{17}$	X_{12} = 2005 年 9 月账单金额; X_{13} = 2005 年 8 月账单金额; …… X_{17} = 2005 年 4 月账单金额。
前期付款金额(新台币)	$X_{18} \sim X_{23}$	X_{18} = 2005 年 9 月支付金额; X_{19} = 2005 年 8 月支付金额; …… X_{23} = 2005 年 4 月支付的金额。

3. 研究结论

3.1. 模型应用

本研究首先将运用 R 软件对台湾客户的违约支付的 23 个变量进行二元 logistic 回归分析，其中我们发现有些变量并不显著，希望能够减少变量个数，选择出较为显著的变量进行二元 logistic 回归分析。

最优子集法能使模型更简洁、更有效地筛选出影响客户违约的因素，根据这些选择出的变量个数能为银行提供更有指导[7]。于是采用最优子集方法进行变量最优个数的选择，根据调整后 R^2 越大越好，Cp 和 BIC 越小越好的原则选择个数。

最优子集选择变量的个数与调整后 R^2 、Cp 和 BIC 的关系见图 1。

在 R 软件中，我们画出上图，图中的红点是当变量数为 13 时，调整后 R^2 、Cp 和 BIC 的值。绿点是变量数为 8 时，红点是变量数为 4 时。结合图中点的位置综合考虑，我们将变量数选择为 8 个变量。再通过 Forward Stepwise Selection 方法选择出变量 X_4 、 X_6 、 X_8 、 X_{10} 、 X_{11} 、 X_{16} 、 X_{18} 、 X_{19} 。再通过这 8 个变量进行二元 logistic 回归方程，得到表 2。

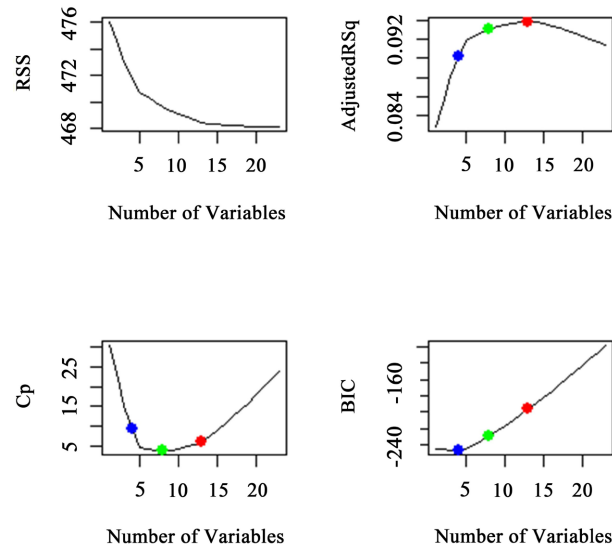


Figure 1. Optimal subset result
图 1. 最优子集结果

Table 2. Logistic regression equation
表 2. Logistic 回归方程

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-7.576e-01	1.541e-01	-4.918	8.76e-07***
X_4	-2.524e-01	8.806e-02	-2.866	0.00416**
X_6	5.406e-01	5.082e-02	10.638	<2e-16***
X_8	8.206e-02	5.450e-02	1.506	0.13215
X_{10}	1.988e-01	7.185e-02	2.767	0.00565**
X_{11}	-1.173e-01	6.736e-02	-1.742	0.08158
X_{16}	-6.308e-07	9.352e-07	-0.674	0.50003
X_{18}	-2.939e-05	9.634e-06	-3.050	0.00229**
X_{19}	-1.012e-05	6.203e-06	-1.632	0.10274

从表 2 我们可以得到各变量的系数和常数项, 同时发现婚姻情况和常数项非常显著, 过去付款历史是否按时还款、延期几个月和最近两期的支付金额在模型中是非常重要的变量, 这有助于银行在贷款时着重考虑这几个变量去衡量这个客户能否每月按时还款。

3.2. Logistic 回归模型的预测精度

本文通过将数据集划分为训练集(2500 个数据)和测试集(500 个数据)。通过训练集中的 8 个变量对模型重新建模, 将得到的模型在测试集中进行预测, 就得到表 3 Confusing matrix。不用原模型预测是因为用该数据建模再对同样的数据进行预测会使模型的预测精度高于模型的真实预测精度。

我们发现:

- 1) 模型对客户不违约的预测是非常准的($384/(384 + 7) = 0.982$);
- 2) 模型对客户违约的预测非常不理想($17/(92 + 17) = 0.156$);

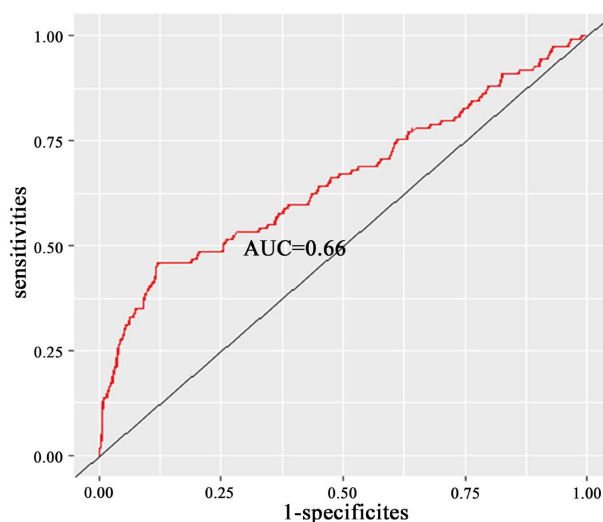
Table 3. Confusing matrix**表 3.** 混淆矩阵

Y.test	模型预测结果	
	0	1
0	384	7
1	92	17

3) 模型的整体预测准确率为 $0.802((384 + 17)/(384 + 7 + 92 + 17))$ ，预测精度还可以。

模型对客户不违约的预测精准，而对客户违约预测非常差，我认为可能原因是模型对非平衡数据非常敏感，即构建模型的训练集中客户不违约的个数为 2332，违约个数为 668，两者相差非常大。

通过对模型的预测精度进行验证，也说明了模型的整体预测能力比较理想。接下来我们通过另一种可视化的方法衡量模型的优劣，即 ROC 曲线(图 2)，该曲线可以在出现类不平衡现象中基本保持原貌。这里的 AUC 为 ROC 曲线和 $y = 0$ 直线之间的面积，得到 $AUC = 0.66$ ，模型的结果优于我们随机猜测，具有预测价值。

**Figure 2.** ROC curve**图 2.** ROC 曲线

4. 总结

近日，硅谷银行的倒闭，是美国历史上第二大银行倒闭案，仅次于在 2008 年金融危机中倒闭的 Washington Mutual (华盛顿互助银行)，这不禁让我们想起 08 年的全球金融危机。正是因为银行没有考虑贷款客户的信贷记录以及还款情况，大量发放次级贷拉动内需。最后次贷违约的人越来越多，面临大面积亏损，导致全球金融危机爆发。如今，银行需要对个人住房贷款坚持严格的审查标准，防止次级贷危机的产生。

本文通过对 default of credit card clients Data Set 数据集进行分析并建立二元 Logistics 回归模型，银行可以通过模型得到客户是否违约的情况，能更好的帮助银行对个人住房贷款进行审查，从一定程度上可以有效避免次级贷危机的发生。由于数据中有 23 个变量，我们发现有一些变量并不显著。于是，想要通过最优子集方法综合考虑，最终选择建立模型的变量个数为 8 个。再采用 Forward Stepwise Selection 方

法选择出变量 X_4 、 X_6 、 X_8 、 X_{10} 、 X_{11} 、 X_{16} 、 X_{18} 、 X_{19} ，使用这 8 个变量建立二元 Logistics 回归模型，得到各变量的系数和常数项，同时注意到婚姻情况和常数项非常显著，过去付款历史是否按时还款、延期几个月和最近两期的支付金额在模型中是非常重要的变量。

由于数据集中有 3000 个数据，本文再将数据分为训练集(2500 个)和测试集(500 个)来得到模型的预测精准度：模型对客户不违约的预测还是非常准确，精度可以达到 98.2%，但对客户违约的预测非常不理想，精度仅有 15.6%。模型整体预测准确率为 0.802，总体来说还算不错。通过另一种可视化的方法衡量模型的优劣，即 ROC 曲线，计算出 AUC 的值为 0.66。模型的结果优于我们随机猜测，具有预测价值。

参考文献

- [1] 宫宏宇, 谢艺观. 硅谷银行倒闭会引发新一轮金融危机吗? [J]. 宁波经济(财经视点), 2023, 589(4): 45-47.
- [2] 石伊凡. 1929 年经济危机、2008 年金融危机的成因以及相似度分析[J]. 经济师, 2023, 407(1): 27-28.
- [3] 任重. 美国次级贷危机的成因分析[J]. 杭州电子科技大学学报(社科版), 2009, 5(S1): 25-28.
<https://doi.org/10.13954/j.cnki.hduss.2009.s1.009>
- [4] 胡冰. 次贷危机对我国商业银行的启示[J]. 合作经济与科技, 2009, 376(17): 76-77.
- [5] 彭建刚, 张丽寒, 刘波, 等. 聚合信用风险模型在我国商业银行应用的方法论探讨[J]. 金融研究, 2008, 338(8): 72-85.
- [6] 彭佳琪, 徐璐. 武汉市在校生垃圾分类认知和行为调查研究——基于二元 Logistic 回归模型[J]. 科教导刊, 2022, 475(7): 151-154. <https://doi.org/10.16400/j.cnki.kjdk.2022.07.050>
- [7] 李健民, 刘添文, 符思远, 等. 基于最优子集法建立肠道准备预测模型的研究[J]. 中国实用内科杂志, 2020, 40(3): 231-236. <https://doi.org/10.19538/j.nk2020030111>