Operations Research and Fuzziology 运筹与模糊学, 2019, 9(3), 222-231
Published Online August 2019 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2019.93025

Hans汉斯

A Credit Scoring Technology Based on Bayes Discriminant Analysis

Shenghua Zhou

Rizhao Marine Engineering Vocational College, Rizhao Shandong
Email: 66838031@qq.com

Received: Jul. 19th, 2019; accepted: Aug. 2nd, 2019; published: Aug. 9th, 2019

Abstract

According to FICO Score theory and Bayes Discrimination, the credit scoring model is derived, which ends up as an optimization model with linear objective function and quadratic equality constraints. Finally, compared with Logistic Regression through an example, the result shows that the credit scoring model is effective and can support application scenarios of practical business better.

Keywords

Credit Scoring, Bayes Discriminant Analysis, Divergence Model, Binning Weight, Score Scaling

一种基于贝叶斯判别的信用评分方法

周声华

日照航海工程职业学院，山东 日照
Email: 66838031@qq.com

收稿日期：2019年7月19日；录用日期：2019年8月2日；发布日期：2019年8月9日

摘　要

本文借鉴了FICO评分的思想，基于贝叶斯判别定理推导出一套评分模型，评分模型最终为一个目标函数是线性函数，约束条件含有二次等式约束的最优化问题。最后，通过一个实例与Logistic回归做了对比，实例结果表明模型是有效的，且模型能够更好的支持实际业务应用场景。

文章引用: 周声华. 一种基于贝叶斯判别的信用评分方法[J]. 运筹与模糊学, 2019, 9(3): 222-231.
DOI: 10.12677/orf.2019.93025

关键词

信用评分，贝叶斯判别，区分度模型，分箱权重，尺度化

1. 概述

在信用评分方面，美国 Fair Isaac Corp 公司的评分(以下简称 FICO 评分[1])是应用最为成熟的一种评分模型，该评分在 300~850 之间，信用评分越高，说明客户的信用风险越小。除此之外，Logistic 回归[2][3]、决策树、神经网络等也是常用的评分方法。

本文借鉴了 FICO 评分的思想，并基于贝叶斯判别定理推导出了一套信用评分模型(以下简称评分模型)，该模型是一个目标函数为线性函数，约束条件为二次型的最优化模型。本文评分模型所具有的优点是：

① 本文评分模型得到的是非常直观的整数权重，这对不懂评分技术的业务人员来讲，能够很方便的对评分结果进行解读和应用。

② 当业务人员拒绝客户的信用业务申请时，可以依据评分结果给予合理的拒绝原因。

③ 利用本文评分模型得到的多张评分卡，可以方便的比较、混合使用。

④ 鉴于以往的项目经验：本文评分模型的稳健性是非常好的，利用本文方法建立好的评分卡应用3年后仍然有很好的预测性，而Logistic回归、决策树要逊色很多。

2. 基于贝叶斯判别的评分模型

评分模型的建立分为 5 个步骤：输入变量的筛选、输入变量的分箱、评分模型的求解与评估、评分结果的拟合与尺度化、评分模型的部署。

本文研究的内容主要是：评分模型的建立、评分结果的拟合与尺度化。

假设有一批信用良好的客户样本(以下简称好客户样本)和信用不良的客户样本(以下简称坏客户样本)，我们要通过这两组样本数据建立评分模型。基于以往业务经验对评分模型做如下要求和假设：

① 客户信用评分 S 越大，代表该客户是好客户的概率越大；反之，代表是坏客户的概率越大。

② 好客户样本评分 S_g 服从正态分布 $N\left(\overline{S}_g,\sigma_g^2\right)$，$\overline{S}_g$ 为好客户样本均值，σ_g^2 为好客户样本方差；S_g 的密度函数为 $f_g(S)$。

③ 坏客户样本评分 S_b 服从正态分布 $N\left(\overline{S}_b,\sigma_b^2\right)$，$\overline{S}_b$ 为坏客户样本均值，σ_b^2 为坏客户样本方差；S_b 的密度函数为 $f_b(S)$。

④ $p(g|S)$：信用评分为 S 的客户是好客户的概率；$p(b|S)$：信用评分为 S 的客户是坏客户的概率；

⑤ $odds(S)=p(g|S)/p(b|S)$：信用评分 S 对应的好、坏客户的概率比。

⑥ 评分模型有 p 个输入变量，各分箱组数分别是 $q_1,q_2,\cdots,q_p$ 个，各分箱权重分别如下：

第 1 个输入变量的分箱权重为：$w_{11},\cdots,w_{1q_1}$

.....

第 p 个输入变量的分箱权重为：$w_{p1},\cdots,w_{pq_p}$

记：$\boldsymbol{w}=\left(w_{11},\cdots,w_{1q_1},w_{21},\cdots,w_{2q_2},\cdots,w_{pq_p}\right)^{\mathrm{T}}$；$T=\sum_{i=1}^{p}q_i$：总共分箱组数。

⑦ 好客户样本 m 个，分别为：$\boldsymbol{g}^{(1)},\cdots,\boldsymbol{g}^{(m)}$；坏客户样本 n 个，分别为：$\boldsymbol{b}^{(1)},\cdots,\boldsymbol{b}^{(n)}$。

$$\boldsymbol{g}^{(r)}=\left(g_{11}^{(r)},\cdots,g_{1q_1}^{(r)},\cdots,g_{p1}^{(r)},\cdots,g_{pq_p}^{(r)}\right)^{\mathrm{T}}\quad(1\le r\le m)$$

$$\boldsymbol{b}^{(r)}=\left(b_{11}^{(r)},\cdots,b_{1q_1}^{(r)},\cdots,b_{p1}^{(r)},\cdots,b_{pq_p}^{(r)}\right)^{\mathrm{T}}\quad((1\le r\le n)$$

注：对于每个样本第 $i(1\le i\le p)$ 个输入变量的分箱取值 $x_{i,1},x_{i,2},\cdots,x_{i,q_i}$ 中，有且仅有一个分箱值为 1，其他值为 0，表示该样本第 i 输入变量值落在取值为 1 的分箱区间内。

⑧ 样本中好客户占比 $p_g=m/(m+n)$，坏客户占比 $p_b=n/(m+n)$。

⑨ $S_{g^{(r)}}$：第 r 个好客户样本的信用评分 $(1\le r\le m)$；$S_{b^{(r)}}$：第 r 个坏客户样本的信用评分 $(1\le r\le n)$。

基于①、②、③的要求和假设，我们可以画出评分分布示意图如图 1 所示：

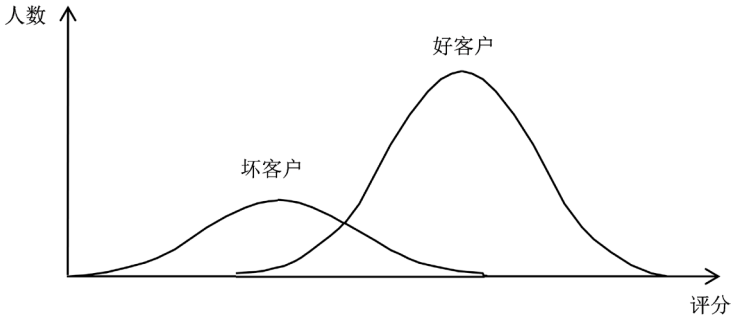

Figure 1. The diagram of score distribution
图 1. 评分分布示意图

为了能尽量区分好、坏客户两个群体，显然好客户的信用评分应该尽量大，坏客户的信用评分应该尽量小，我们以此建立我们评分模型的目标函数：

$$\max\sum_{r=1}^{m}S_{g^{(r)}}-\sum_{r=1}^{n}S_{b^{(r)}}\tag{1}$$

接下来，我们确定评分模型的约束条件：

首先，由贝叶斯判别定理[4] [5]，可得

$$p(g\mid S)=p_g\cdot f_g(S),p(b\mid S)=p_b\cdot f_b(S)$$

$$\ln\left(odds(S)\right)=\ln\left(p_g/p_b\right)+\ln\left(f_g(S)/f_b(S)\right)\tag{2}$$

为了在拟合阶段中，评分结果与客户好坏的概率值能建立函数关系，我们在模型建立时就要考虑评分 S 与 $p(g\mid S),p(b\mid S)$ 的关系。

由于 $\ln\left(p_g/p_b\right)$ 是个定值，不妨就假设：$\ln\left(odds(S)\right)=\ln\left(p_g/p_b\right)+S$

$$\text{即 } S=\ln\left(f_g(S)/f_b(S)\right)\tag{3}$$

这样就建立了 S 与 $p(g\mid S),p(b\mid S)$ 的函数关系，且 S 越大，$p(g\mid S)$ 越大，$p(b\mid S)$ 越小。

又因为 $S_g\sim N\left(\bar{S}_g,\sigma_g^2\right)$，$S_b\sim N\left(\bar{S}_b,\sigma_b^2\right)$，所以

$$\ln\left(f_g(S)/f_b(S)\right)=\ln\left(\frac{1}{\sqrt{2\pi}\sigma_g}\exp\left(-\frac{\left(S-\bar{S}_g\right)^2}{2\sigma_g^2}\right)\right)-\ln\left(\frac{1}{\sqrt{2\pi}\sigma_b}\exp-\frac{\left(S-\bar{S}_b\right)^2}{2\sigma_b^2}\right)$$

$$=\left(\frac{1}{2\sigma_b^2}-\frac{1}{2\sigma_g^2}\right)\cdot S^2+\left(\frac{\bar{S}_g}{\sigma_g^2}-\frac{\bar{S}_b}{\sigma_b^2}\right)\cdot S+\left(\frac{\bar{S}_b^2}{2\sigma_b^2}-\frac{\bar{S}_g^2}{2\sigma_g^2}\right)$$

由(3)得 $\sigma_g^2=\sigma_b^2$， $\bar{S}_g-\bar{S}_b=\sigma_g^2$， $\bar{S}_g+\bar{S}_b=0$。

这样我们就可以得到一个初步的评分模型：

$$\begin{aligned}&\max\quad \sum_{r=1}^{m}S_{g^{(r)}}-\sum_{r=1}^{n}S_{b^{(r)}}\\ &\text{s.t.}\quad \begin{cases}\sigma_g^2-\sigma_b^2=0\\ \bar{S}_g-\bar{S}_b=\sigma_g^2\\ \bar{S}_g+\bar{S}_b=0\end{cases}\end{aligned}\tag{4}$$

实际上，好、坏客户两类群体信用评分的方差一般不会完全相等，而且在数据测试中我们发现 σ_g^2、σ_b^2 不需要严格相等，效果会更好一些，这样我们可以把约束条件 $\sigma_g^2-\sigma_b^2=0$ 去掉，同时 $\bar{S}_g-\bar{S}_b=\sigma_g^2$ 改为 $\bar{S}_g-\bar{S}_b=0.5\cdot\left(\sigma_g^2+\sigma_b^2\right)$。

另外，如果对 S 不加约束，由(4)求得的目标函数会异常大，甚至求不出最优解，因此需要对信用评分 S 的取值范围加以约束，可以想到的方法有：

① 直接将 S 约束在某一区间范围内；

② 将每个分箱权重约束在某一范围内；

③ 设定各分箱权重的平方和小于某个阈值。

这 3 种方法都是有效的，但我们发现：第 3 种约束效果要好一些。另外，考虑到不同的评分模型其分箱组数是会不一样的，为了模型的普适性，我们采用“各分箱权重平方和的平均值小于某个阈值”来对 S 进行约束。这样，评分模型进一步优化为：

$$\begin{aligned}&\max\quad \sum_{r=1}^{m}S_{g^{(r)}}-\sum_{r=1}^{n}S_{b^{(r)}}\\ &\text{s.t.}\quad \begin{cases}\bar{S}_g-\bar{S}_b=0.5\cdot\left(\sigma_g^2+\sigma_b^2\right)\\ \bar{S}_g+\bar{S}_b=0\\ \dfrac{1}{T}\cdot\sum_{i=1}^{p}\sum_{j=1}^{q_i}w_{ij}^2\le K\end{cases}\end{aligned}\tag{5}$$

其中，T 为分箱组数，K 为阈值(在本文实例计算中，$K=2$ 效果比较理想)。

3. 评分模型的参数推导

下面我们进行具体的参数推导，由前面的假设我们可以得出：

第 r 个好客户样本的信用评分：

$$S_{g^{(r)}}=\left(\boldsymbol{g}^{(r)}\right)^{\mathrm{T}}\cdot\boldsymbol{w}=\sum_{i=1}^{p}\sum_{j=1}^{q_i}g_{ij}^{(r)}\cdot w_{ij}\quad\left(1\le r\le m\right)$$

第 r 个坏客户样本的信用评分：

$$S_{b^{(r)}}=\left(\boldsymbol{b}^{(r)}\right)^{\mathrm{T}}\cdot\boldsymbol{w}=\sum_{i=1}^{p}\sum_{j=1}^{q_i}b_{ij}^{(r)}\cdot w_{ij}\quad\left(1\le r\le n\right)$$

好客户样本信用评分之和：

$$\sum_{r=1}^{m} S_{g^{(r)}} = \sum_{r=1}^{m} \left(\boldsymbol{g}^{(r)}\right)^{\mathrm{T}} \cdot \boldsymbol{w} = \left(\sum_{r=1}^{m} \left(\boldsymbol{g}^{(r)}\right)^{\mathrm{T}}\right) \cdot \boldsymbol{w}$$

坏客户样本信用评分之和：

$$\sum_{r=1}^{n} S_{b^{(r)}} = \sum_{r=1}^{n} \left(\boldsymbol{b}^{(r)}\right)^{\mathrm{T}} \cdot \boldsymbol{w} = \left(\sum_{r=1}^{n} \left(\boldsymbol{b}^{(r)}\right)^{\mathrm{T}}\right) \cdot \boldsymbol{w}$$

好客户样本信用评分平均值：

$$\overline{S}_g = \frac{1}{m} \sum_{r=1}^{m} S_{g^{(r)}} = \left(\frac{1}{m} \sum_{r=1}^{m} \left(\boldsymbol{g}^{(r)}\right)^{\mathrm{T}}\right) \cdot \boldsymbol{w}$$

坏客户样本信用评分平均值：

$$\overline{S}_b = \frac{1}{n} \sum_{r=1}^{n} S_{b^{(r)}} = \left(\frac{1}{n} \sum_{r=1}^{n} \left(\boldsymbol{b}^{(r)}\right)^{\mathrm{T}}\right) \cdot \boldsymbol{w}$$

好客户样本信用评分方差：

$$\begin{aligned}
\sigma_g^2 &= E\left(S_g - \overline{S}_g\right)^2 = \frac{1}{m} \sum_{r=1}^{m} \left(S_g^{(r)} - \overline{S}_g\right)^2 = \frac{1}{m} \sum_{r=1}^{m} \left(\left(\boldsymbol{g}^{(r)} - \frac{1}{m} \sum_{r=1}^{m} \boldsymbol{g}^{(r)}\right)^{\mathrm{T}} \cdot \boldsymbol{w}\right)^2 \\
&= \frac{1}{m} \sum_{r=1}^{m} \left(\boldsymbol{w}^{\mathrm{T}} \cdot \left(\boldsymbol{g}^{(r)} - \frac{1}{m} \sum_{r=1}^{m} \boldsymbol{g}^{(r)}\right) \cdot \left(\boldsymbol{g}^{(r)} - \frac{1}{m} \sum_{r=1}^{m} \boldsymbol{g}^{(r)}\right)^{\mathrm{T}} \cdot \boldsymbol{w}\right) \\
&= \frac{1}{m} \boldsymbol{w}^{\mathrm{T}} \cdot \sum_{r=1}^{m} \left(\left(\boldsymbol{g}^{(r)} - \frac{1}{m} \sum_{r=1}^{m} \boldsymbol{g}^{(r)}\right) \cdot \left(\boldsymbol{g}^{(r)} - \frac{1}{m} \sum_{r=1}^{m} \boldsymbol{g}^{(r)}\right)^{\mathrm{T}}\right) \cdot \boldsymbol{w}
\end{aligned}$$

同理，坏客户样本信用评分方差：

$$\sigma_b^2 = E\left(S_b - \overline{S}_b\right)^2 = \frac{1}{n} \sum_{r=1}^{n} \left(S_b^{(r)} - \overline{S}_b\right)^2 = \frac{1}{n} \boldsymbol{w}^{\mathrm{T}} \cdot \sum_{r=1}^{n} \left(\left(\boldsymbol{b}^{(r)} - \frac{1}{n} \sum_{r=1}^{n} \boldsymbol{b}^{(r)}\right)\left(\boldsymbol{b}^{(r)} - \frac{1}{n} \sum_{r=1}^{n} \boldsymbol{b}^{(r)}\right)^{\mathrm{T}}\right) \cdot \boldsymbol{w}$$

令 $\boldsymbol{g} = \sum_{r=1}^{m} \boldsymbol{g}^{(r)},\ \overline{\boldsymbol{g}} = \frac{1}{m} \sum_{r=1}^{m} \boldsymbol{g}^{(r)},\ \boldsymbol{H}_g = \frac{1}{m} \sum_{r=1}^{m} \left(\boldsymbol{g}^{(r)} - \overline{\boldsymbol{g}}\right)\left(\boldsymbol{g}^{(r)} - \overline{\boldsymbol{g}}\right)^{\mathrm{T}}$，

$\boldsymbol{b} = \sum_{r=1}^{m} \boldsymbol{b}^{(r)},\ \overline{\boldsymbol{b}} = \frac{1}{n} \sum_{r=1}^{n} \boldsymbol{b}^{(r)},\ \boldsymbol{H}_b = \frac{1}{n} \sum_{r=1}^{n} \left(\boldsymbol{b}^{(r)} - \overline{\boldsymbol{b}}\right)\left(\boldsymbol{b}^{(r)} - \overline{\boldsymbol{b}}\right)^{\mathrm{T}}$

则 $\sum_{r=1}^{m} S_{g^{(r)}} = \boldsymbol{g}^{\mathrm{T}} \cdot \boldsymbol{w}, \overline{S}_g = \overline{\boldsymbol{g}}^{\mathrm{T}} \cdot \boldsymbol{w}, \sigma_g^2 = \boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{H}_g \cdot \boldsymbol{w}, \sum_{r=1}^{n} S_{b^{(r)}} = \boldsymbol{b}^{\mathrm{T}} \cdot \boldsymbol{w}, \overline{S}_b = \overline{\boldsymbol{b}}^{\mathrm{T}} \cdot \boldsymbol{w}, \sigma_b^2 = \boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{H}_b \cdot \boldsymbol{w}$

最后，我们的评分模型就可以表示为：

$$\begin{aligned}
&\max \quad \left(\boldsymbol{g}^{\mathrm{T}} - \boldsymbol{b}^{\mathrm{T}}\right) \cdot \boldsymbol{w} \\
&\text{s.t.} \quad \begin{cases}
\left(\overline{\boldsymbol{g}}^{\mathrm{T}} - \overline{\boldsymbol{b}}^{\mathrm{T}}\right) \cdot \boldsymbol{w} = 0.5 \cdot \boldsymbol{w}^{\mathrm{T}} \cdot \left(\boldsymbol{H}_g + \boldsymbol{H}_b\right) \cdot \boldsymbol{w} \\
\left(\overline{\boldsymbol{g}}^{\mathrm{T}} + \overline{\boldsymbol{b}}^{\mathrm{T}}\right) \cdot \boldsymbol{w} = 0 \\
\frac{1}{T} \cdot \boldsymbol{w}^{\mathrm{T}} \cdot \boldsymbol{w} \le K
\end{cases}
\end{aligned} \tag{6}$$

T 为分箱组数，K 为阈值，$\boldsymbol{w}$ 是我们要求解的分箱权重向量，模型是一个二次型最优化问题[6]。

4. 评分结果的拟合与尺度化

在模型建立中，我们假设：

$$\ln\left(odds\left(S\right)\right)=\ln\left(p_g/p_b\right)+S$$

因此，我们采取 $S,\ln\left(odds\left(S\right)\right)$ 进行线性拟合。

在信用评分的实际应用中，我们往往对某一具体的好坏概率比 $odds\left(S_0\right)$ 特别重视，期望该 $odds\left(S_0\right)$ 对应某个评分 S_0'，不仅如此，还要求信用评分 S' 每增加一个固定值 $\Delta S_0'$，好坏概率比 $odds\left(S'\right)$ 就增加一个 $odds\left(S_0\right)$。例如：我们期望好坏概率比为 100 时对应的信用评分为 500 分，且信用评分每增加 20 分，好坏概率比就增加 100，当信用评分为 700 分时，可以推算出好坏概率比为 1100。

在此做如下假设：

S'：S 对应的尺度化后评分。

$odd\left(S_0\right)$：预设的好坏概率比。

S_0：$odds\left(S_0\right)$ 对应的尺度化前评分。

S_0'：$odds\left(S_0\right)$ 对应的尺度化后评分。

$\Delta S_0'$：表示尺度化后的评分值每增加 $\Delta S_0'$，好坏概率比就增加 1 个 $odds\left(S_0\right)$。

$\boldsymbol{w}'=\left(w_{11}',\cdots,w_{1q_1}',\cdots,w_{p1}',\cdots,w_{pq_p}'\right)^{\mathrm{T}}$：尺度化后的各变量分箱权重。

进一步假设 $S,\ln\left(odds\left(S\right)\right)$ 拟合得到的线性方程为：

$$\ln\left(odds\left(S\right)\right)=b_0+b_1\cdot S\quad(b_0,b_1\text{是系数})\tag{7}$$

则我们可以得出如下关系：

$$\ln\left(odds\left(S'\right)\right)=\ln\left(odds\left(S\right)\right)=b_0+b_1\cdot S\tag{8}$$

$$\frac{\ln\left(odds\left(S'\right)\right)-\ln\left(odds\left(S_0'\right)\right)}{S'-S_0'}=\frac{\ln\left(2\cdot odds\left(S_0'\right)\right)-\ln\left(odds\left(S_0'\right)\right)}{\Delta S_0'}$$

$$\text{即 }\ln\left(odds\left(S'\right)\right)=\frac{\ln 2}{\Delta S_0'}\cdot S'+\ln\left(odds\left(S_0'\right)\right)-\frac{\ln 2}{\Delta S_0'}\cdot S_0'\tag{9}$$

由(8)进一步可得：

$$\ln\left(odds\left(S_0'\right)\right)=\ln\left(odds\left(S_0\right)\right)=b_0+b_1\cdot S_0\tag{10}$$

将(8)、(10)代入(9)，得：

$$b_0+b_1\cdot S=\frac{\ln 2}{\Delta S_0'}\cdot S'+b_0+b_1\cdot S_0-\frac{\ln 2}{\Delta S_0'}\cdot S_0'$$

$$S'=\Delta S_0'\cdot\frac{b_1}{\ln 2}\cdot S+S_0'-\Delta S_0'\cdot\frac{b_1}{\ln 2}\cdot S_0$$

令 $c_1=\Delta S_0'\cdot\frac{b_1}{\ln 2}$，$c_0=S_0'-c_1\cdot S_0$，则

$$S'=c_0+c_1\cdot S\tag{11}$$

$$\ln\left(odds\left(S'\right)\right)=b_0-\frac{b_1\cdot c_0}{c_1}+\frac{b_1}{c_1}\cdot S'\tag{12}$$

式(11)就是尺度化评分 S' 与原始评分 S 的尺度化关系，式(12)就是尺度化后评分 S' 与好坏概率比的关系。

需要补充说明是：利用式(7)进行拟合时，实际上并不知道每个原始评分 S 对应的好坏客户概率比，但是我们可以对原始评分进行排序分组，然后取每个分组原始评分的中间值作为 S，每个组的好坏客户数比作为 $odds(S)$，这样就可以进行拟合了。另外，考虑到按原始评分排序分组以后，S 值最大的几个组里可能没有坏客户，S 值最小的几个组里可能没有好客户，所以要剔除这些“特殊”组，然后再进行拟合。

最后，我们将尺度化评分 S' 拆分到每个变量分箱中。拆分时要遵循如下 2 条原则：

① 每个变量的各分箱权重非负。

② 各样本的尺度化分箱权重之和仍为 S'。

记第 i 个变量的尺度化前最小分箱权重 $\min(w_i)=\min\limits_{1\le j\le q_i}(w_{ij})\ (1\le i\le p)$，则

$$w'_{ij}=c_1\cdot\left(w_{ij}+\left|\min(w_i)\right|\right)+\frac{c_0-c_1\cdot\sum_{i=1}^{p}\left|\min(w_i)\right|}{p}\quad(1\le i\le p,1\le j\le q_i)\tag{13}$$

其中 $w_{ij}+\left|\min(w_i)\right|$ 是为了将变量的最小分箱权重由负值变为零值，乘以 c_1 表示的是每个分箱权重的尺度化也服从 S 到 S' 的线性关系，加上 $\left(c_0-c_1\cdot\sum_{i=1}^{p}\left|\min(w_i)\right|\right)\Big/p$ 是为了保证尺度化后的评分值仍然等于尺度化后的分箱权重之和。

例如：假设 $c_0=115.8, c_1=23, \Delta S'_0=20$，有 3 个分箱变量，尺度化过程可用如表 1 所示：

Table 1. Weight scaling for binning variables
表1. 分箱权重尺度化步骤说明表

变量	w_{ij}	$\left\|\min(w_i)\right\|$	c_1	$c_1\cdot\left(w_{ij}+\left\|\min(w_i)\right\|\right)$	$\left(c_0-c_1\cdot\sum_{i=1}^{p}\left\|\min(w_i)\right\|\right)\Big/p$	w'_{ij}
变量 1	2.1	1.2	23	76	11	87
	0.1			30	11	41
	−1.2			0	11	11
变量 2	−0.5	0.5	23	0	11	11
	0.5			23	11	34
	1			67	11	78
变量 3	0.1	1.9	23	46	11	57
	−1.9			0	11	11

5. 数据测试与对比

数据来源：SPSS自带的bankloan.sav数据，包含：517位拖欠贷款客户(坏客户)，183位不拖欠贷款客户(好客户)。

输出变量：default (1：坏客户；0：好客户)。

输入变量及分箱结果：见表2。

尺度化要求：500 分对应的好坏概率比是 100:1，且尺度化后的评分每增加 20 分，好坏概率比增加 100。

阈值设置：$K=2$。

利用样本数据计算结果如表 3 所示：

Table 2. The binning of input variables
表2. 输入变量及分箱表

变量	*address* 地 址	*age* 年龄	*creddebt* 信用卡欠款	*debtinc* 贷款收入比	*ed* 学历	*employ* 工龄	*income* 收入	*othdebt* 其他债务
分箱1	(−, 7)	(−, 34)	(−, 1.0615)	(−, 5]	ed = 1	(−, 3)	(−, 27)	(−, 1.926)
分箱2	[7, +)	[34, +)	[1.0615, 1.9607)	[5, 8.5)	ed = 2	[3, 7)	[27, 37)	[1.926, +)
分箱3			[1.9607, +)	[8.5, 13.8)	ed = 3	[7, 12)	[37, 56)	
分箱4				(13.8, +)	ed = 4	[12, +)	[56, +)	
分箱5					ed = 5			

Table 3. Calculating results of the scoring model
表3. 评分模型计算结果

变 量	变量分箱	g	b	$\bar{g}$	$\bar{b}$	原始权重	尺度化权重
address	(−, 7)	223	117	0.43	0.64	−1.23	36
	[7, +)	294	66	0.57	0.36	1.16	58
age	(−, 34)	218	110	0.42	0.60	−0.26	36
	[34, +)	299	73	0.58	0.40	0.19	40
creddebt	(−, 1.0615)	322	75	0.62	0.41	2.15	76
	[1.0615, 1.9607)	92	40	0.18	0.22	−0.09	55
	[1.9607, +)	103	68	0.20	0.37	−2.14	36
debtinc	(−, 5]	152	17	0.29	0.09	2.01	86
	[5, 8.5)	145	28	0.28	0.15	1.17	78
	[8.5, 13.8)	130	45	0.25	0.25	0.09	68
	(13.8, +)	90	93	0.17	0.51	−3.35	36
ed	ed = 1	293	79	0.57	0.43	1.34	57
	ed = 2	139	59	0.27	0.32	−0.15	43
	ed = 3	57	30	0.11	0.16	−0.94	36
	ed = 4	24	14	0.05	0.08	−0.32	41
	ed = 5	4	1	0.01	0.01	−0.01	44
employ	(−, 3)	79	76	0.15	0.42	−3.37	36
	[3, 7)	119	52	0.23	0.28	−0.49	63
	[7, 12)	140	30	0.27	0.16	1.51	82
	[12, +)	179	25	0.35	0.14	2.26	89
income	(−, 27)	153	77	0.30	0.42	−0.56	36
	[27, 37)	108	39	0.21	0.21	−0.20	39
	[37, 56)	116	35	0.22	0.19	−0.22	39
	[56, +)	140	32	0.27	0.17	0.91	49
othedebt	(−, 1.926)	265	77	0.51	0.42	−0.31	36
	[1.926, +)	252	106	0.49	0.58	0.24	41

尺度化评分 S' 与 $\ln\left(odds\left(S'\right)\right)$ 如图 2 所示：

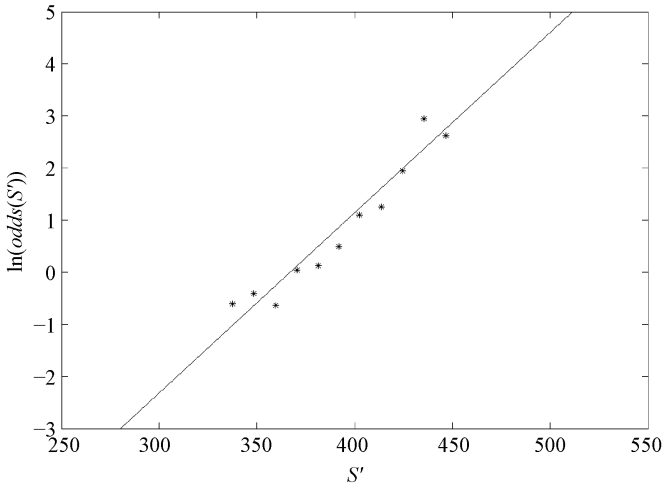

Figure 2. Linear fitting chart by S' and $\ln\left(odds\left(S'\right)\right)$

图2. S' 与 $\ln\left(odds\left(S'\right)\right)$ 拟合直线图

本文评分模型与Logistic回归ROC曲线比较，如图3所示：

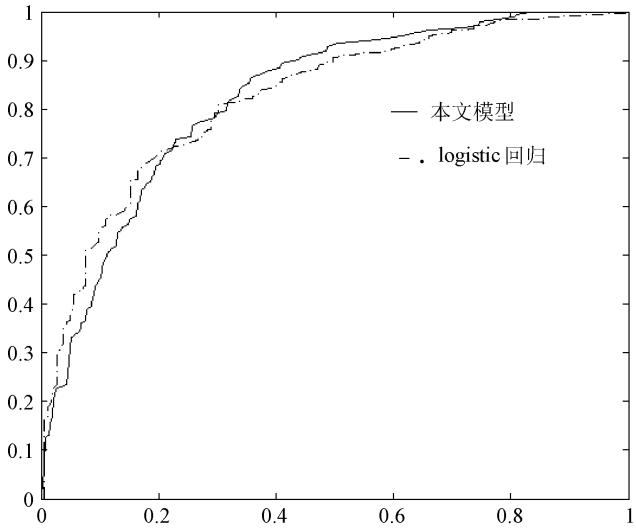

Figure 3. ROC curve comparison chart

图3. ROC曲线对比图

可以看出：文中的评分模型跟 Logistic 回归模型相比也是一种非常有效的评分方法。另外，基于以往的项目经验：本文评分模型的稳健性是非常好的，利用本文方法建立的评分卡应用 3 年后仍然有很好的预测性，而 Logistic 回归、决策树要逊色很多。

参考文献

[1] 陈建. 信用评分模型技术与应用[M]. 北京: 中国财政经济出版社, 2005: 1-286.

[2] 杨静. 信用评分卡的建立与应用[D]: [硕士学位论文]. 天津: 天津商业大学, 2018.

[3] 石勇, 孟凡. 信用评分基本理论及其应用[J]. 大数据, 2017(1): 24-31.
[4] 马达. 基于贝叶斯的判别理论及其算法实现[D]: [硕士学位论文]. 北京: 中国地质大学, 2011.
[5] 宋云鹏, 武钰. 数据挖掘技术在信用评分中的应用研究[J]. 征信, 2013(10): 24-28.
[6] 袁亚湘. 非线性优化计算方法[M]. 北京: 科学出版社, 2018.