arXiv:2303.01227v1 [stat.AP] 2 Mar 2023

信用风险评分中现有及新型总体稳定性检验程序的批判性评述

预印本

J. du Pisanie
西北大学数学与统计科学学院，南非。
johan.dupisanie@monoclesolutions.nl

J. S. Allison 西北大学数学与统计科学学院，南非。james.allison@nwu.ac.za

C. J. Budde 自由州大学数学与应用数学系，南非。buddecj@ufs.ac.za

I. J. H. Visagie 数学与统计科学学院，西北大学，南非。
jaco.visagie@nwu.ac.za

2023年3月3日

摘要

信用评分卡是用于对客户违约概率进行建模的模型。是否向申请人提供信贷，以及信贷定价，通常都以这些模型为依据。为确保评分卡长期保持准确性，需要定期检验人口稳定性假设；也就是说，要检验评分卡开发时客户属性的分布在审查时是否仍具有代表性。实践中采用了多种总体稳定性度量方法，近期文献中也提出了若干新方法。本文对用于检验上述假设的多种程序进行了批判性审查。文中讨论了广泛使用的总体稳定性指数，以及近期提出的两种方法。此外，本文还探讨了经典拟合优度方法的应用，并考察了大样本带来的问题。除了现有检验程序外，我们还提出了两种可用于检验总体稳定性的新方法。第一种方法基于效应量的计算，在面对大样本时不会遭遇经典拟合优度方法所面临的同类问题。第二种方法称为重叠统计量。我们认为，由于这种简单度量易于直观解释，因此具有实用价值。为展示各种度量方法的用法，并突出其优缺点，文中还列举了若干数值示例。

关键词和短语：信用风险评分卡；效应量；假设检验；重叠；总体稳定性。

1 引言与研究动机

在银行业中，模型被广泛用于量化与客户或申请人相关的各种风险。此类模型中最知名的包括违约概率（PD）、违约损失率（LGD）和违约风险暴露（EAD）模型。需要注意的是，违约的定义由银行的建模团队设定，但违约通常是指客户在一个12个月的周期内有3次未还款。这些定义上的差异超出了本文的范围，本文其余部分均采用上述定义进行讨论。

PD、LGD和EAD模型既可以单独使用，也可以结合使用，以评估风险并确定与特定客户打交道时所需的谨慎程度。在共同使用时，PD模型通常被认为是三种模型中影响最大的一种。这是因为，与LGD或EAD相比，与客户相关的违约概率通常变化更大，因此PD成为单个客户的主要风险驱动因素。根据所考虑的金融产品不同，1%至10%的违约概率十分常见，而平均违约概率的微小变化可能会对风险暴露产生重大影响。用于违约概率的主要模型之一是逻辑回归模型，通常称为评分卡。该模型根据客户的各种属性提供预测违约概率（本文将模型中纳入的客户特征统称为属性）。评分卡广泛应用于银行及其他金融机构。如 [3], 所强调的，评分卡主要用于以下三个领域:

1. 信用风险评分卡——此类评分卡用于估算新申请人希望申请某种金融产品时的违约概率（PD），以及现有客户申请额外信贷时的违约概率。此外，这些模型在制定信贷策略和计算贷款损失准备时也发挥着重要作用。2. 催收评分卡——此类评分卡用于确定拥有一个或多个逾期账户的客户的还款概率。评分卡的结果用于确定在哪些方面开展催收工作最为有效。3. 营销评分卡——此类评分卡用于估算特定潜在客户接受某项邀约的概率。应用此类评分卡，业务便能集中营销工作，通常通过外呼或互动，面向最有可能接受特定邀约的客户开展营销。

通过比较预测结果与观察结果，可以衡量评分卡的平均准确性。不过，只有在结果期结束后才能进行这种比较。例如，假设某项业务在未来12个月内接纳（批准并纳入）100名客户，其平均违约概率为5%。在这种情况下，如果评分卡准确，我们预计这100名客户中有5名会在12个月的结果期内违约。12个月后，可以审查这100名客户，并将违约人数与预测值5进行比较。如果发现违约人数与5相差显著，则表明模型无法准确预测违约。这种现象的一种可能解释是，具有特定属性的客户所对应的风险水平发生了内在变化。另一种可能的解释是，客户总体中观察到的特征结构发生了重大变化；也就是说，属性分布可能发生了如此显著的变化，以至于模型不再能代表所考察的总体。我们关注的是上述第二种可能原因，即识别客户属性分布是否发生了显著变化。如果这种分布随时间发生变化，就有理由怀疑模型在之后阶段所作的预测。为确保模型仍可使用，从业者会定期检验这一假设：当前的属性分布与模型拟合时观察到的分布相同。

在详细介绍用于识别属性分布变化的方法之前，应先区分属性分布的变化与属性测量的不稳定性。属性分布发生变化，表明通过模型评分的客户类型发生了变化。例如，用于开发评分卡（即拟合逻辑回归模型）的数据中，男性申请人占比较高，而当前客户构成中女性申请人占比较高。这正是我们希望通过某种统计上可靠的程序识别的变化。不过，某些外部影响也可能导致特征测量变得不稳定，而这种影响实际上并不会对模型的准确性造成任何影响。举一个具体的例子，假设某个变量将过去两个月内信用查询次数超过两次的客户评为高风险。如果有一家新的信贷提供方进入市场，并为营销目的对总体进行评分，那么每位潜在客户名下都会多出一条查询记录，因而会被视为风险更高，尽管该潜在客户本身的风险并未改变。本文余下部分不讨论后一类变化。

如上所述，评分卡的准确性可以在结果期结束后轻松衡量。然而，银行通常希望持续衡量评分卡的准确性，以确保估计违约概率尽可能准确。为此，银行会使用人口稳定性度量（PSM）。这些工具至关重要，能让业务部门在实际结果出炉前，在一定程度上验证评分卡。这些度量旨在判断当前客户属性的分布是否与用于开发评分卡的客户属性分布足够相似，从而确保模型仍具有代表性。本文概述了PSM在评分卡中的应用。

首先，对PSM的使用作几点总体说明；

- PSM 无法识别评分卡之外的因素。例如，营销评分卡的总体分布和结果分布可能保持稳定，但负责致电销售的人员可能会影响结果。
- 经济因素对催收评分卡的影响可能不会立即显现，因为影响可能有所延迟。因此，即使催收评分卡所代表的总体正面临经济困难，例如通货膨胀或利率上升（这可能会降低他们的偿付能力），催收评分卡仍可能保持稳定。
- PSM 无法区分总体转移与不稳定的群体特征。PSM 一旦显示基础数据的分布发生变化，业务就需要通过进一步调查了解变化的原因。
- 本文不考虑检测到的变化对结果的影响。本文讨论的 PSM 旨在检测属性分布是否发生变化，而非考察这种变化对估计违约概率的影响。

尽管存在上述注意事项，PSM 可以向业务提供早期信号，表明总体的底层分布正在发生变化，而且计算得出的违约概率可能与账面客户的实际情况不一致。本文审查了总体稳定性的替代衡量方法，并就如何选择 PSM 提供见解。此外，我们还探讨了两种可能有助于衡量总体稳定性的新方法。第一种基于计算得出的效应量；第二种是一种称为重叠统计量的衡量指标。该统计量已应用于心理学领域，我们将展示它也可以作为衡量总体稳定性的直观指标。

本文其余部分的结构如下。第2节简要概述评分卡构建过程，第3节则详细介绍与 PSM 相关的统计学考量。本文特别关注大样本量带来的问题，并概述各项指标的优缺点。第4节提供数值示例，说明如何在实践中使用 PSM。第5节给出结论和建议。

2 评分卡构建过程审查

本节概述了用于构建评分卡的过程。所采用的程序将在 [10] 中进行更详细的介绍。下面我们将简要介绍理解评分卡构建和监控所需的主要主题。

2.1 关键概念

评分卡是用于对客户的违约概率进行建模的工具。对于给定客户，会为每项属性分别分配一个分数。这些单项分数汇总为所谓的信用评分，评分卡也因此得名。模型将信用评分转换为估计的违约概率。信贷提供方通常会以该违约概率为依据作出信贷决策。

2.2 评分卡构建

评分卡构建过程通常包括以下步骤：

1. 获取数据——确定用于构建评分卡的数据集。该数据集应包含原始决策时的属性值；例如，客户申请贷款时的属性值。同时还应包含客户的结果属性，以表明客户是否违约。2. 考虑分群——如果有足够的数据，且客户可以划分为不同的群组，则应考虑进行分群。需要通过分析来了解最佳群组。所选群组应使各群组中属性的底层特性存在显著差异。确定分群方案后，可以为每个群组分别构建评分卡。3. 对属性值进行分组——应检查每个属性的取值范围，以进行分组。创建组的过程具有双重作用。首先，它可以提高评分卡的稳定性。例如，考虑客户的收入。根据收入将客户划分为若干组通常是可行的，因为合理的预期是，模型的违约概率会随收入逐渐变化。因此，将属性划分到关于违约概率具有相同行为的分箱中是有意义的。为了

确定与某一属性相关的合理分组，需要将信息值（如下文所述）与实际考量结合使用。

4. 计算证据权重（WOE）——计算证据权重的目的是对每个属性的取值范围进行标准化。在创建统计模型时，可以比较模型中估计系数的大小。考虑一个具有 k 个水平的离散属性。设 D 为数据集中的违约数量， D_j 为与该属性的 j^{th} 水平相关联的违约数量，设 n_j 为与该属性的 j^{th} 水平相关联的观测值总数。证据权重的计算如下：

$$WOE_j = \log\left(\frac{D(n_j - D_j)}{D_j(n - D)}\right).$$

5. 选择评分卡中使用的属性——根据可用属性的数量，可以采用逐步淘汰的方式移除部分属性。需要考虑的因素包括未来是否能够获得该属性；未来可能无法获得的属性应予以移除。此外，还应考虑与使用各属性相关的法律限制。种族或性别等属性通常不允许用于信用评分卡，因为这可能会导致决策过程中产生偏差。最后，还应考虑属性与违约概率（PD）之间关联的强度；只有与违约概率具有实际显著关系的属性才应纳入模型。评分卡中某个属性的预测能力可通过信息值（IV）进行衡量，见 [9]:

$$\text{IV} = \sum_{j=1}^{k}\left(\frac{n_j - D_j}{n - D} - \frac{D_j}{D}\right)\log\left(\frac{D(n_j - D_j)}{D_j(n - D)}\right).$$

Siddiqi [9] 提出了以下用于解释信息价值的经验法则：

- 如果信息价值小于 0.02，则认为该属性不具备预测能力。
- 信息价值在 0.02 至 0.1 之间，表示其预测能力较弱。
- 如果信息价值在 0.1 至 0.3 之间，则该属性的预测能力被归类为中等。

信息价值达到或超过 0.3，表示其预测能力较强。

如果两个变量之间存在较强的相关性，建议最终评分卡中只纳入其中一个，以避免多重共线性带来的问题。

6. 执行逻辑回归——通过对所有选定属性的证据权重拟合逻辑回归，构建评分卡模型。

2.3 基础数据和分组数据的评分卡监控

评分卡监控如今可以同时考察基础数据（可能为连续数据）和分组数据。建议在检验总体稳定性时，同时考察两者的分布。对分组数据进行检验，有助于了解总体的基础稳定性对评分卡和结果的影响。这应作为主要指标，用于判断属性的分布是保持稳定还是已变得不稳定。若分布已不稳定，可根据数据稳定性决定是否重建或停用评分卡。

此外，也建议对未分组的基础数据进行稳定性检验。此类检验可作为早期指标，提示即使模型分组保持稳定，基础数据仍可能正在偏离原始分布。在这种情况下，需要进一步调查，以了解这一趋势是否可能持续；如果会持续，那么使用评分卡进行的预测将受到何种影响。

2.4 总体不稳定性对总体违约概率结果的影响

我们始终假设预测违约概率是在账户级别计算的。这在实践中通常如此，尤其是在数据可用性不存在问题的发达信贷市场中。计算这些违约概率时，通常会计算整个总体的平均违约概率，以量化总体整体所固有的风险水平。

当总体中的某些属性变得不稳定时，这些属性对总体违约概率的影响可能在一定程度上相互抵消。假设有两个属性变得不稳定，其中一个会使申请人的违约概率评分降低，另一个则会使申请人的违约概率评分升高。根据这两种变化的幅度，其影响可能相互抵消，这意味着总体风险水平可能仍然

保持不变。分析师如果仅考虑总体的平均违约概率来判断总体是否仍然稳定，就无法正确识别这种变化，并会错误地得出结论，认为总体及其所有属性都处于稳定状态。

3 关于总体稳定性指标的文献审查

本节审查了可用于检验总体稳定性假设的现有PSM。首先，在详细介绍这些PSM之前，我们将对总体稳定性假设进行一般性讨论。讨论内容包括待检验假设的性质，以及为解释这些度量的计算值而需要采用的方法。我们讨论了两种不同的解释方法：一种旨在确定统计显著性，另一种基于经验法则。对于前一种方法，我们需要计算临界值。下文列出了一种算法，可用于近似计算与所讨论的各种PSM相关的临界值。

在对待检验假设进行一般性讨论之后，我们将注意力转向文献中现有的各种PSM。其中包括已逐渐成为行业标准的总体稳定性指数（PSI）。我们还考察了最近提出的两种PSM；这些度量由Taplin and Hunt [11]以及Du Pisanie and Visagie [3]提出。随后，我们转向经典统计技术，考察Kolmogorov-Smirnov拟合优度检验以及一种基于效应量的度量。本节最后讨论所谓的重叠统计量，该统计量用于衡量分布的相似性。迄今为止，这一度量尚未被用于检验信用评分中的总体稳定性。然而，我们认为，这一实用且直观的度量能够为分析师深入理解总体稳定性相关问题提供有价值的参考。

3.1 待检验的假设

如上所述，我们会持续监测属性和结果的分布性质。有些属性是连续的（例如收入），而有些则是离散的（例如客户使用的信用卡数量）。上文已说明，连续属性通常会划分为若干组，然后在评分卡中作为离散属性处理。因此，我们主要关注检验一个离散随机变量是否服从指定分布这一假设。

为继续讨论，我们引入与 [3] 中类似的符号。考虑一个具有 k 个水平的属性；用 $\mathbf{q} = (q_1, \ldots, q_k)$ 表示评分卡开发时与该属性可能取值对应的比例。分别用 $\mathbf{p} = (p_1, \ldots, p_k)$ 和 $\mathbf{P} = (P_1, \ldots, P_k)$ 表示待检验总体稳定性时该属性各 k 个水平对应的总体比例和样本比例。我们关注的假设是

$$H_0 : p_j = q_j, \forall j \in \{1, \ldots, k\}, \tag{1}$$

该假设将针对一般备择假设进行检验。

在构建评分卡时，通常会将预测结果划分为所谓的 PD 组。这些组的规模通常会设置得相近；也就是说，将 PD 最高的 x% 归为一组，接着将下一档 x% 归为一组，依此类推。对违约概率进行分组，可用于制定相关产品的策略，例如为较高 PD 组的客户设定特定额度或提供优惠。本文讨论的 PSM 可用于检验 PD 组的稳定性，其方式与检验评分卡中任何属性的方式相同。

由于无法直接观测到 $\mathbf{p}$，因此要基于 $\mathbf{P}$（以及 $\mathbf{q}$）检验（1）中的假设。因此，所考虑的大多数 PSM 都是衡量 $\mathbf{q}$ 与 $\mathbf{P}$ 之间差异的度量。也就是说，对于给定的特定属性，PSM 通常旨在量化评分卡开发时与检验总体稳定性时的分布差异。形式上，PSM 通常是差异度量，或 $\mathbf{q}$ 与 $\mathbf{P}$ 之间的不相似性度量。令 $\Delta(\mathbf{q}, \mathbf{P})$ 表示基于 $\mathbf{q}$ 的指定值和 $\mathbf{P}$ 的观测值计算的某种 PSM。

一般而言，在解读 $\Delta(\mathbf{q}, \mathbf{P})$ 的观测值时，有两种常用方法。第一种是根据指定的经验法则来解读该值，第二种则是在人口稳定性假设下近似求得临界值。这两种方法都受到相关文献的广泛批评，但各有其可取之处。虽然从统计学角度看，经验法则难以论证，但实践中经常遇到的大样本量，也令人质疑计算所得 p 值是否适用。然而，计算临界值和 p 值有助于研究假设 (1) 成立时 $\Delta(\mathbf{q}, \mathbf{P})$ 的行为。下文将介绍求得临界值和 p 值所需的计算。

令 m 表示待检验总体稳定性的当前样本中的客户数量。在检验某项属性的总体稳定性时，q 是一个固定常向量。假设 (1) 成立，则有 p $=$ q。在这种情况下， mP 服从参数为 m 和 q 的多项分布。因此，我们可以轻松模拟零假设下 P 的实现值。请注意，进行这项模拟无需观测到 P 的值，只需已知 q 和样本量 m 。现在可以使用以下算法近似求得 Δ 的临界值。

1. 给定 q 和 m，从参数为 m 和 q 的多项分布中模拟一个实现值，例如 M*。计算 P* $=$ M*$/m$。 2. 计算 $\Delta^* = \Delta$(q,P*)。 3. 将上述前两个步骤重复 b 次，以获得零假设下 Δ 的 b 个实现值。将结果记为 Δ_1^*、 ..... 、 Δ_b^*。 4. 将观测值排序，得到 $\Delta_{(1)}^* \leq \cdots \leq \Delta_{(b)}^*$。 5. 显著性水平为 100α% 时的近似临界值为 $\Delta_{(\beta)}$，其中 $\beta = \lfloor b(1-\alpha) \rfloor$ ，而 $\lfloor \cdot \rfloor$ 是下取整函数。

给定 P 的一个观测值，计算与 Δ(q, P) 相关的 p值很简单。这个 p-值为

$$\frac{1}{b}\sum_{j=1}^{b} \mathrm{I}\left(\Delta(\mathbf{q},\mathbf{P}) \leq \Delta_j^*\right),$$

其中 I 表示指示函数。

请注意，上文用于计算经验临界值和 p 值的算法，可用于下文讨论的各类 PSM，但重叠度量除外。不过，只需对上述公式稍作调整，即可将其用于该度量。令 ξ(q, P) 表示重叠统计量。设定 Δ(q, P) $= 1 - \xi$(q, P) 后，即可使用上述计算方法求得 Δ 的临界值和 p 值，并由此推得 ξ 的相应值。

3.2 总体稳定性指数（PSI）

PSI 是使用最广泛的 PSM，可见于多部文献，包括 [7] 和 [9]。采用上文定义的符号，PSI 定义如下：

$$\Psi(\mathbf{q},\mathbf{P}) = \sum_{j=1}^{k} (P_j - q_j)\log\left(\frac{P_j}{q_j}\right). \tag{2}$$

Lewis [7] 提出了对 PSI 实现值的以下解释：

1. $0 \leq \Psi$(q, P) < 0.1 表明总体没有发生显著变化。2. $0.1 \leq \Psi$(q, P) < 0.25 表明总体发生了小幅变化。3. Ψ(q, P) ≥ 0.25 表明总体发生了显著变化。

近期有几篇论文探讨并讨论了 PSI；例如参见 [13] 以及 [3]。需要注意的是，前述论文对 PSI 持批评态度，并指出，尽管这一指标有助于检验总体稳定性，但使用时应谨慎。

Yurdakul 和 Naranjo [13] 探讨了 PSI 的统计特性。除了证明经过适当重新缩放的 PSI 渐近服从 χ-平方分布外，作者还主张在解释 PSI 时使用渐近临界值，而非使用上文所述的经验法则。研究表明，渐近临界值取决于样本量以及与相关属性对应的组数。

Du Pisanie and Visagie [3] 对 PSI 持批评态度，因为对 Ψ 计算所得数值的解读没有考虑样本量或相关属性的水平数量。该论文提供了数值示例，说明 PSI 的临界值取决于样本量和水平数量；随着样本量增加，PSI 的临界值降低；随着与该属性相关的组数增加，PSI 的临界值升高（至少在各组对应的比例相等时如此）。在极端情况下，开发总体或审查总体中的某个水平未出现在另一个总体中，即在一个总体中观察到某个属性水平，而在另一个总体中没有观察到时， Ψ 的计算会失效，且该指标会被计算为无穷大。

3.3 Taplin and Hunt [11]的度量

Taplin and Hunt [11]提出了一种评分卡稳定性检验的替代方法。作者提出了一项称为总体准确性指数（PAI）的指标，用于衡量自开发以来估计平均响应方差的变化。PAI 定义为两个量的比率：分子是审查评分卡时估计平均响应方差的平均值，分母是开发评分卡时估计平均响应方差的平均值。在当前语境中，响应是计算所得的违约概率（PD）。

令 $\mathbf{X} = (X_1, \ldots, X_n)$ 和 $\mathbf{Y} = (Y_1, \ldots, Y_m)$ 分别表示开发阶段和审查阶段所考察属性的观测值。在使用简单线性回归对 PD 建模的情况下，可以证明 PAI 为

$$\pi(\mathbf{X}, \mathbf{Y}) = \frac{1}{2}\left(1 + \frac{\frac{1}{m}\sum_{j=1}^{m}(Y_j - \bar{X})^2}{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2}\right), \tag{3}$$

其中，$\bar{X}$ 和 $\bar{Y}$ 分别表示 X_1、…、X_n 以及 Y_1、…、Y_m 的样本均值。如果采用多元线性回归模型，PAI 同样可以得到闭式表达式；详情请参阅 [11] 。较小的 PAI 值被认为更理想，并表明分布未发生显著变化。具体而言，对 PAI 的建议解释如下：

1. $0 \leq \pi(\mathbf{X}, \mathbf{Y}) < 1.1$ 表示该属性的分布没有显著变化。2. $1.1 \leq \pi(\mathbf{X}, \mathbf{Y}) < 1.5$ 表示该属性的分布略有变化。3. $\pi(\mathbf{X}, \mathbf{Y}) \geq 1.5$ 表示该属性的分布发生了显著变化。

虽然考虑违约概率的预测或计算准确性确有价值，但人口稳定性指数（PAI）在实际应用中存在一些难以实施的方面。首先，PAI 是为评分卡提出的，尽管该指标也可用于逻辑回归模型，但在这种情况下，PAI 没有闭式表达式。事实上，要计算这种情况下各个违约概率的方差，需要采用参数自助法程序。这种复杂性可能会带来问题，尤其是考虑到实践中对 PSM 的解释赋予了很高的重要性。其次，PAI 隐含地假设平均方差越小越好。这一假设是任意的，因为较小或较大的方差都可能表明底层总体不稳定。

为了说明该度量可能存在的一种陷阱，请看下面这个简单示例。设开发阶段和审查阶段观测到的属性值分别为 $\mathbf{X} = (-2, 1, 1)$ 和 $\mathbf{Y} = (1.1, 1.2, 1.3)$。在此示例中，X 和 Y 的取值完全没有重叠，这表明总体以可想象到的最极端方式处于不稳定状态。然而，考虑使用（3）计算得到的 PAI：

$$\pi(\mathbf{X}, \mathbf{Y}) = \frac{1}{2}\left(1 + \frac{\frac{1}{m}\sum_{j=1}^{m}(X_j - \bar{X})^2}{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X})^2}\right) = \frac{1}{2}\left(1 + \frac{1.1^2 + 1.2^2 + 1.3^2}{2^2 + 1^2 + 1^2}\right) = 0.862.$$

根据 [11], 中提出的建议，计算得到的 PAI 值应被视为表明所考虑属性的分布没有发生实质性变化。

虽然 PAI 作为总体稳定性的度量具有一定价值，但我们建议不要孤立地看待这一度量。

3.4 Du Pisanie & Visagie 的度量 [3]

Du Pisanie and Visagie [3] 认为，（1）中所述的人口稳定性假设过于严格。也就是说，传统的人口稳定性要求是属性的分布随时间保持完全不变。鉴于实践中经常遇到大样本量，使用标准统计方法可以检测出分布随时间发生的微小变化。上述论文指出，在实践中，更值得关注的问题是，观测到的分布变化是否足够显著，以至于会对评分卡所作的预测产生负面影响。

作者提出了一种替代性的 PSM，旨在检验观察到的分布变化是否超过用户定义的某个阈值。采用与上文相同的符号表示法，所提出的 PSM 为

$$T(\mathbf{q}, \mathbf{P}) = \max_{1 \leq j \leq k} \frac{|P_j - q_j|}{q_j}. \tag{4}$$

将 T的实际值与用户定义的参数（例如 δ）进行比较。如果 T(q, P) 超过 δ，则认为分布发生了显著变化。δ的取值根据评分卡的相对重要性来确定。若评分卡不准确所造成的损失风险越大，则应选择越小的 δ 值。这种 PSM 的一个优势是，用户定义的 δ 值让业务能够针对不同评分卡，设定测试对不同水平的敏感度。

通常采用总体稳定性技术来判断评分卡的预测是否仍然准确。对于信用风险评分卡，往往有一些属性构成获批信贷的最低要求。因此，这些属性的一个或多个水平上的预测违约概率准确性只具有学术意义，因为它不会影响银行的盈利能力。因此， [3] 认为，应从分析中剔除相关属性的这些类别。具体而言，设 k^* 表示对于某个 $k^* < k$，与相关属性中可获批信贷的水平相对应的组数。对统计量（4）进行简单修改，即可取该属性前 k^* 个水平上的最大值。

当前PSM的一个潜在问题是其对稳定性的评估非此即彼。也就是说，如果 T(q, P) $\leq \delta$，则认为总体稳定。然而，观测值P稍有变化，就可能导致 T(q, P)的值略微增大，从而使 T(q, P) $> \delta$；在这种情况下，总体就会被判定为不稳定。可以通过将 T(q, P)与 δ的不同水平进行比较来解决这一问题：较小的值可能需要进一步调查，而较大的值则表明需要重新开发评分卡。
使用 T 可能还存在一个问题：当存在大量小组时，该度量更容易判定某个属性不稳定。也就是说，与小组相关的比例发生微小变化时，更容易被判定为重要变化，因为该变化会按比例 q_j进行缩放。实际使用该度量时，应将这一点考虑在内。

3.5 柯尔莫哥洛夫-斯米尔诺夫检验

柯尔莫哥洛夫-斯米尔诺夫检验统计量用于衡量两个分布函数之间的最大距离，见 [6]。这项广为人知的检验已在文献中得到广泛研究；对于假设分布为离散分布的情形，感兴趣的读者可参阅 [1] 。请注意，尽管这里我们仅关注柯尔莫哥洛夫-斯米尔诺夫检验，但在考察其他经典拟合优度检验（例如克拉默-冯·米塞斯检验）时，也会得出类似观察，见 [2] 和 [12]。
在当前设定下，柯尔莫哥洛夫-斯米尔诺夫检验的检验统计量衡量两个经验分布函数之间的上确界距离；这两个分布函数分别对应于 q 和 P。令 $x(j)$ 表示所考虑属性的 j^{th} 水平。与 q 相关的经验分布函数定义为

$$F_{\mathbf{q}}(x(j)) = \sum_{l=1}^{j} q_l,$$

其中 $I(\cdot)$ 是指示函数。类似地，与 P 相关的经验分布函数为 s

$$F_{\mathbf{P}}(x(j)) = \sum_{l=1}^{j} P_l.$$

柯尔莫哥洛夫-斯米尔诺夫检验的检验统计量定义为

$$D(\mathbf{q}, \mathbf{P}) = \max_{j \in \{1,\ldots,k\}} |F_{\mathbf{q}}(x(j)) - F_{\mathbf{P}}(x(j))|.$$

银行业的性质导致了这样一种情况：每个月都有大量客户通过银行模型进行评分。南非的一家典型银行每月可以轻松地为超过 10 000 名客户评估一款新产品。在一家大型金融机构的贷款损失准备金部门，每位客户每月都会通过 PD模型重新评分，通常每月会产生数百万个新评分。在使用标准统计技术时，这带来了挑战。由于下文讨论的现象，这些技术依赖于 p-值的做法被证明存在问题。

为了说明上述问题，考虑这样一种情况：我们希望检验性别这一属性的人口稳定性假设。举一个具体例子，假设观察到一个规模为 100 000 的开发样本，其中 50% 的客户为男性，50% 为女性。在审查时，又观察到一个规模为 100 000 的样本；但这次样本中 50.5% 为男性，49.5% 为女性。使用第 3.1 节概述的算法，基于 1 000 000 个参数自助法样本，我们估计与 Kolmogorov-Smirnov 检验相关的 p 值

用于检验总体稳定性的该值为 0.15%。因此，在所有合理的名义显著性水平下，都会拒绝总体稳定性假设。拒绝这一假设是基于相关属性分布发生了非常微小的变化；我们认为，大多数从业者都会认同，观察到的变化不应被视为该属性不稳定的迹象。

需要指出的是，在实践中，经典拟合优度检验并不常用于评估总体稳定性。作者认为，这些统计技术在当前情境下不受欢迎，原因在于实践中经常会遇到大样本量。对于式（1）中的零假设，这些检验对固定备择假设是一致的。这意味着，只要样本量足够大，这些检验就会因任何偏离零假设的情况而拒绝总体稳定性，即使这种偏离微乎其微。

关于使用经典拟合优度检验程序，还有最后一点需要说明。虽然信用评分卡中使用的大多数属性通常以序数尺度进行测量，但也有一些属性以名义尺度进行测量。如果确实如此，且相关属性有两个以上的水平，那么使用经典拟合优度检验技术就并非易事。不过，由于这些测量方法在当前情境下用途有限，所需的调整超出了本文的讨论范围。举例来说，可以考虑潜在客户申请信贷时所采用的申请方式。该属性可分为以下四个水平：

- 分行 - 在分行完成的申请。
- 在线 - 通过在线申请渠道提交的申请。
- 电话 - 通过非直接渠道提交的申请。
- 营销电话 - 在信贷提供方提示后提交的申请。

例如，使用 PSI 检验总体稳定性时，计算所需的只是评分卡开发和审查时各水平的观测比例。然而，计算 D 需要计算经验分布函数，而经验分布函数本质上假设各组是有序的，但这里并非如此。因此，在实践中很难使用柯尔莫哥洛夫-斯米尔诺夫检验来检验总体稳定性。

由于上述困难，我们未将经典拟合优度检验纳入第4节所列的数值示例中。

3.6 基于效应量的度量

一种可能的总体稳定性检验方法不受实践中遇到的大样本量影响，它基于效应量。假设总体稳定；也就是说，假设某个属性的分布没有变化，我们知道，对于 ，

$$mPs_j \sim Bin\ m\ p_j \qquad j \in (\ ,\ )$$
$$mP_j \quad mq_j \quad mq_j\ 1-q_j$$

。因此，$\{1,\ldots,k\}$ 如果总体稳定，则 的期望值和方差分别为 和 ( )。

我们先考虑单变量情境下的效应量。也就是说，属性的 j^{th} 水平在两个总体之间的变化效应量为

$$\gamma_j = \frac{|P_j - q_j|}{\sqrt{q_j(1-q_j)}}.$$

请注意，这个量不依赖于 n 或 m。也就是说，给定 q 和 P 的一组观测值，计算 γ_j 时并未考虑所用的样本量。在进行此项计算时，我们会得到一组结果：γ_1、.....、γ_k，可以通过加权平均将它们合并为单个 PSM；

$$\Gamma(\mathbf{q},\mathbf{P}) = \sum_{j=1}^{k} q_j\gamma_j = \sum_{j=1}^{k} \frac{\sqrt{q_j}|P_j - q_j|}{\sqrt{1-q_j}},$$

其中，分配给与属性的 j^{th} 水平相关的效应量的权重为 q_j；即该水平在零假设下的比例。选择这种加权方式，是为了确保水平 j 的计算效应量对 Γ 的贡献，与观察到该属性水平的似然成正比。由于 Γ 由效应量构成，因此我们可以按照解释效应量的方式来解读 Γ 的值。不过，由于 Γ 是加权平均，而且我们希望该指标能够识别出仅少数几个相关属性水平的比例发生变化的情况，因此建议使用低于效应量阈值的阈值。一般而言，效应量超过 0.3 时被视为具有实际显著性；详情请参阅，

参见 [7]。根据第4节给出的数值结果，我们建议将 $\Gamma > 0.1$ 视为 q 与 P 之间具有实际显著性差异的指标。

为了说明这一度量的用法，以及它不受大样本量问题影响的特点（经典拟合优度方法会受到此问题影响），我们重新考察第3.5节中的示例。如果观测到的比例从 q $=$ (50%,50%) 变为 P $=$ (50.5% 49.5%)，那么与两个水平相关的各个效应量为 $(\gamma_1, \gamma_2) = (0.01, 0.01)$，而 $\Gamma(\mathbf{q}, \mathbf{P}) = 0.01$。这一较小的 Γ 值表明，q 与 P 之间的差异不具有实际显著性，这意味着我们不拒绝人口稳定性假设。相比之下，柯尔莫哥洛夫-斯米尔诺夫检验拒绝了总体稳定性假设。

应当指出，已有多种将效应量扩展至多变量情形的方法，参见 [4]。然而，我们未能找到任何关于将效应量用作 PSM（即使是在简单情形下）的论述（但 [3] 提到了这种可能性，却未作进一步讨论）。

3.7 重叠

文献 [8], 探讨了一种比较分布的不同方法，即一种称为重叠统计量的相似性度量。在该文献中，对于连续分布，重叠被解释为两个相应密度函数交集的面积。也就是说，设 f 和 g 为两个密度函数。这些密度函数之间的重叠指数 $\eta(f, g)$: $\mathbb{R} \times \mathbb{R} \to [0, 1]$ 为

$$\eta(f, g) = \int_{\mathbb{R}} \min\{f(x), g(x)\} \mathrm{d}x.$$

图3.7展示了这一概念。图中显示了两个正态密度函数，并用阴影标出了这两个密度函数之间的“重叠”区域。

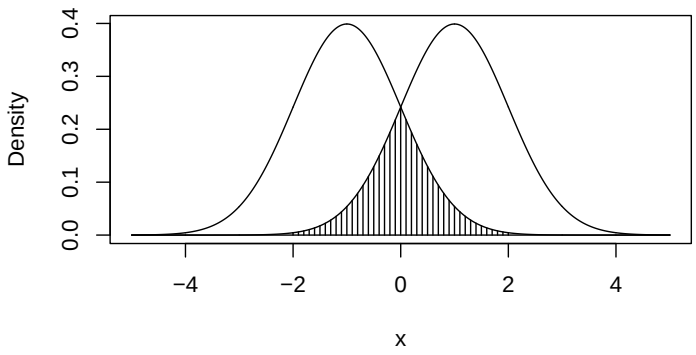


图1：标出重叠区域的正态密度图。

在总体稳定性的背景下，我们关注的是离散情形；具体而言，是比较与 q 和 P 对应的质量函数。在这种情况下，η 的定义只需重新定义为

$$\eta(\mathbf{q}, \mathbf{P}) = \sum_{j=1}^{k} \min\{q_j, P_j\}.$$

重叠统计量的一项主要优势在于其解释简单；η 表示两个分布之间相对应的概率质量比例。重叠概念简单明了，因此这一指数特别适用于金融领域，因为该领域高度重视结果的可解释性。无需解释与例如 p 值、临界值或效应量相关的概念，就能轻松地向非技术受众说明重叠的概念。重叠指数为 0.8，可以简单地理解为，在所考察的两个分布中，80% 的概率质量对应于所讨论属性的相同水平。鉴于信用评分模型强调可解释性，且要求向非技术受众解释所用模型，我们认为，这种简单直观的度量在总体稳定性检验中可能非常有用。

4 数值结果

本节包含三个数值示例，用于比较第 3 节讨论的 PSM。这些示例如下：

1. 稳定总体——所有特征的分布保持不变。2. 稳定结果——所考察的两个属性的分布发生了变化；然而，这些变化相互抵消，整体违约概率基本保持不变。3. 不稳定总体——评分卡中包含的所有属性的分布均发生了变化。

需要注意的是，这三个示例均基于模拟数据。在当前情境下，使用模拟数据优于使用观测数据，因为前者能够让我们明确指定是否发生分布变化。为生成所需的数据集，我们采用了 [5] 中提出的模拟技术。

对于每个示例，我们模拟一个开发总体，其中包含十个按不同测量尺度测量的属性。表1显示了评分卡中包含的每个属性的名称、测量尺度和信息价值。我们谨慎选择评分卡中经常出现的属性，并为所考察的这些属性的水平采用现实的配置。我们使用与 [5] 中相同的属性及其水平。由于无需了解这些属性的确切分布即可解释下文的结果（同时也为了避免打断本文的主要论述），我们将这些分布的具体设定推迟至附录A。总体规模始终设定为10,000名客户。

Tabl表1：模型中包含的属性名称、测量尺度和信息值 el

Name	尺度	信息值
性别	顺序尺度	0.499
现有客户	顺序尺度	0.441
查询次数	比率	0.394
其他机构的信用卡	比率	0.515
居住省份	顺序尺度	0.284
申请方式	顺序尺度	0.222
Age	比率	0.164
未偿总额	比率	0.083
收入	比率	0.182
近期违约余额	比率	0.192

4.1 示例1：稳定总体

在第一个示例中，我们模拟了一个审查总体，其分布与开发总体所设定的分布相同。开发数据集和审查数据集的平均违约率分别为 9.92% 和 10.08%。表2展示了第3节讨论的各种 PSM 的计算结果。

表2：稳定总体

变量	PSI	DPV	效应量	重叠
性别	0.0004	0.0241	0.0198	0.9903
Age	0.0007	0.0575	0.0102	0.9889
查询次数	0.0010	0.0639	0.0133	0.9874
现有客户	0.0001	0.0153	0.0077	0.9969
其他信用卡	0.0008	0.1206	0.0087	0.9924
未偿余额	0.0008	0.1066	0.0136	0.9878
Prov	0.0016	0.1538	0.0163	0.9838
AppMethod	0.0010	0.0458	0.0175	0.9857
收入	0.0012	0.0633	0.0128	0.9868
RecDef	0.0007	0.0734	0.0132	0.9909
PD 组	0.0017	0.0690	0.0109	0.9837

4.1.1 PSI

根据 PSI 通常采用的经验法则，所有变量都被视为稳定，其 PSI 均低于 0.1。根据 PSI，我们不拒绝总体稳定性假设。

4.1.2 DPV

使用 [3], 中提出的方法时，需要指定重要性常数 δ 的值。尽管建议结合业务考量来选择该值，但我们注意到， [3] 中给出的示例采用了 $\delta = 0.2$。为了说明这种 PSM 的解释，我们也采用该值。对于所考虑的所有属性，我们注意到，总体稳定性的假设均未被拒绝。不过，如果将 δ 设为 0.15，属性“省份”就会被判定为不稳定。请注意，该属性有九个水平，其中一些水平对应的比例相当小。如上所述，这种组合可能会导致错误地认为分布发生了变化，而实际上并没有。因此，在检验水平数量较多且对应比例较小的属性的稳定性时，我们不建议将 δ 设为小于 0.2 的值。

4.1.3 效应量

基于效应量的 PSM 对每个属性以及 PD 组得出的值都较小。同样，我们不拒绝总体稳定性假设。

4.1.4 重叠

对于所考察的每个属性和 PD 组，重叠度量得出的值都接近 1。事实上，观测到的重叠统计量的最小值超过 0.98。这表明重叠统计量未能提供反对总体稳定性假设的证据。

4.2 不稳定结果

第二组数据是在改变两个属性的分布后模拟得出的，改变方式使它们对违约概率的影响相互抵消。也就是说，开发总体和审查总体的平均违约概率大致相等。“查询次数”以及“其他提供商的信用卡数量变化”的各水平所对应的比例，分别按照表3和表4中指定的方式进行了调整。表5展示了各种PSM的结果。表中显示，“未偿余额”属性的PSI被计算为无穷大。正如第3.2节的讨论所指出的，这是因为在一个总体中观察到该属性的某个水平，而在另一个总体中没有观察到。

表3：查询次数

组	描述	开发比例	审查比例
0	无咨询	30%	40%
1	一次咨询	25%	25%
2	两次咨询	20%	10%
3	三次咨询	15%	15%
4	四次咨询	5%	5%
5	五次或更多咨询	5%	5%

表4：其他机构的信用卡

组	描述	开发比例	审查比例
0	其他提供商处没有信用卡	50%	30%
1	其他提供商处的信用卡	30%	50%
2	其他服务提供商处的信用卡	15%	15%
3	其他服务提供商处的三张或更多信用卡	5%	5%

开发总体和审查总体的平均违约概率分别为 9.92% 和 9.69%。这一变化可视为相对较小，是各属性的分布变化相互抵消的结果。“查询次数”的变化使总体中转入低风险类别的比例增加，而“其他机构的信用卡”这一属性则倾向于使总体中高风险类别的比例增加。

表5：稳定结果

变量	PSI	DPV	效应量	重叠
性别	0.0002	0.0156	0.0128	0.9937
Age	0.0014	0.0659	0.0135	0.9848
查询次数	0.0952	0.4948	0.1170	0.8961
现有客户	0.0001	0.0222	0.0112	0.9955
其他信用卡	0.2046	0.6729	0.3323	0.7963
未偿余额	∞	0.0410	0.0097	0.9918
Prov	0.0021	0.0775	0.0251	0.9787
AppMethod	0.0018	0.0529	0.0269	0.9796
收入	0.0021	0.2400	0.0094	0.9878
RecDef	0.0016	0.1210	0.0250	0.9828
PD 组	0.0191	0.2610	0.0372	0.9442

4.2.1 PSI

对于分布发生变化的两个属性，计算得出的 PSI 指标明显高于其他属性的 PSI 指标。“查询次数”属性对应的 PSI 值为 0.0952，而 “ 其他机构的信用卡 ” 对应的 PSI 值为 0.2041。根据本文前文所述的经验法则，前一个属性被认为没有发生实质性变化，而后一个属性则被认为发生了轻微的分布变化。

“未偿余额”属性有一个值得关注的现象。该属性包含一个用于表示极大余额的分箱。在基准集中，该分箱没有观测值；然而，在测试集中却出现了极少量观测值。PSI 公式的设置会导致这样一种情况：如果某个分箱的值趋近于零，PSI 的值就会趋近于无穷大。

4.2.2 DPV

Du Pisanie 和 Visage 指标成功识别出了分布发生变化的两个属性。“查询次数和其他机构的信用卡”属性的计算值分别为 0.4948 和 0.6729。如 [3], 所述，用户需要选择一个 δ 值，作为检验该指标数值的基准。在这种情况下，如果 δ 为 0.5，则查询次数仍被视为稳定（尽管仅以微弱优势），而其他机构的信用卡则发生了足够大的变化，因此被判定为不稳定。与前一个示例一样，将 $\delta = 0.2$ 设为相应值，可以正确识别出这两个分布发生变化的属性。然而，收入的检验统计量值为 0.24，这意味着该属性被判定为不稳定，尽管其分布并未发生变化。这可以解释为该属性中有一个水平对应的比例很低。

4.2.3 效应量

效应量度量以 0.1 为实际显著性阈值，能够有效识别发生变化的属性：“查询次数”的值为 0.1170，“其他机构的信用卡”的值为 0.3323，显著高于其他度量计算出的值。与 PSI 以及 Du Pisanie and Visagie 提出的度量一样，在此度量中，“其他机构的信用卡”的值远高于“查询次数”。这与前者属性的变化幅度较大、后者属性的变化幅度较小相符。

4.2.4 重叠

重叠度量清晰地显示了分布变化的确切幅度。值为 0.8961 表明“查询次数”属性的变化略高于 10%。“其他机构的信用卡”的重叠度量值为 0.7963，与数据中相关比例变化 20% 的情况一致。

4.3 一个不稳定的总体

最终数据集通过模拟分布生成，其中六个属性的各分箱比例发生了5%的变化，具体变化详见附录B。四个属性（“性别”、“查询次数”、“现有客户”和“其他机构的信用卡”）均向较高风险分箱移动了5%。两个属性

（“省份”和“申请方式”）向较低风险分箱移动了5%。各属性移动的具体情况详见附录B。其余四个属性（“年龄”、“未偿余额”、“收入”和“近期违约”）的分布没有变化。平均违约概率从基准数据集中的9.92%升至测试数据集中的10.67%。此示例计算得出的PSM可见表6。

表6：不稳定结果

变量	PSI	DPV	效应量	重叠
性别	0.0079	0.1092	0.0897	0.9560
Age	0.0014	0.0659	0.0135	0.9848
查询次数	0.0175	0.1765	0.0638	0.9461
现有客户	0.0160	0.262	0.1321	0.9469
其他信用卡	0.0105	0.1556	0.0738	0.9529
未偿余额	∞	0.0410	0.0097	0.9918
Prov	0.0102	0.1462	0.0608	0.9534
AppMethod	0.0244	0.1869	0.1001	0.9297
收入	0.0021	0.2400	0.0094	0.9878
RecDef	0.0016	0.1210	0.0250	0.9828
PD 组	0.0055	0.1430	0.0197	0.9705

4.3.1 PSI

PSI 指标能够识别分布发生变化的所有属性。值得注意的是，尽管各属性分布变化的幅度相近，但计算得出的 PSI 值在不同属性之间却存在显著差异。以“省份”和“申请方式”这两个属性为例。两者的变化幅度均为 5%，但省份的 PSI 为 0.0102，而“申请方式”的值为 0.0244，相差近 2.5 倍。此外，值得注意的是，尽管指标值存在如此大的差异，根据 PSI 经验法则，这两个属性都会被归类为未发生实质性变化。

4.3.2 DPV

Du Pisanie 和 Visagie 指标能够正确识别所有发生变化的属性。与 PSI 指标一样，有些属性的计算值远高于其他属性。“现有客户”属性的值是“性别”属性值的 2.5 倍，尽管两者的实际变化幅度相近。因此，我们注意到，就这一点而言，Du Pisanie 和 Visage 指标也容易出现与 PSI 类似的问题。

“收入”属性的分布保持不变，但其 Du Pisanie and Visagie 指标值却位居第二高。仔细检查后发现，在属性分组中，有一个组与其他组相比非常小，在基准集中仅占总体的 3%。在测试集中，该组增加了 0.7%，达到 3.7%。鉴于其基数较小，这一增长幅度相当显著，因此按照该指标，该变量被判定为非常不稳定。

4.3.3 效应量

基于效应量的度量能够有效识别分布发生变化的属性，前提是将实际显著性的阈值设为 0.05。在此示例中，若将截断值设为 0.1，则会有若干分布变化无法被识别。

4.3.4 重叠

重叠度量给出的结果最具可预测性，并且能够在大多数属性中识别出 5% 的变化。与预期值 0.95 偏差最大的属性是申请方式；在此情况下，该统计量的计算值为 0.9297。

5 总结

本文概述了几种用于检验人口稳定性假设的方法。本文讨论了广泛使用的总体稳定性指数（PSI），见 [9], , 并探讨了使用该指数时的一些相关问题；尤其关注大样本量带来的挑战。

除了 PSI，本文还考察了新近提出的方法，包括 Du Pisanie 和 Visage 提出的度量方法，见 [3], ，以及 Taplin and Hunt 提出的人口准确性指数，见 [11]。前一种度量方法具有理想的经验特性，但容易受到分组过细所引发问题的影响。Taplin and Hunt 提出的方法未能针对逻辑回归明确定义；本文还指出，在某些特定情况下，该方法可能得出错误结果。
本文还探讨了效应量，并指出效应量能够有效克服总体规模较大的挑战。最后，本文探讨了重叠统计量。该统计量直观易懂、使用简单，但能让使用者清楚了解数据偏移的程度。

本文考察的各种方法均无法帮助理解变化对模型整体结果的影响。所有方法都能确定单个属性发生了何种变化，却无法将这种变化与整体结果联系起来。要开发一种指标来衡量各属性分布变化对预测结果造成的变化幅度，仍需进一步研究；作者目前正在开发此类指标。

6 附录A

下文列出了各属性每个水平对应的比例。对于这些水平，我们还给出了相应的不良率。该比率用于通过 [5] 中开发的模拟技术，确定各客户对应的违约概率。

6.1 性别

我们假设申请人中女性占60%、男性占40%，并将男性与女性的不良率之比设为3。

6.2 年龄

表7显示了假设的“年龄”属性分布及坏账率。

表7：年龄

组	比例	不良率
18 - 21	5%	1.00
22 - 25	7%	0.85
26 - 30	9%	0.78
31 - 45	26%	0.66
46 - 57	21%	0.50
58 - 63	11%	0.43
64 - 75	21%	0.31

6.3 查询次数

表8列出了该属性各个水平的比例和违约率。

表8：查询次数

组	描述	比例	不良率
0	无咨询	30%	1.0
1	一次咨询	25%	1.3
2	两次咨询	20%	1.8
3	三次咨询	15%	1.9
4	四次咨询	5%	2.1
5	五次或更多咨询	5%	2.7

6.4 现有客户

指定80%的申请人为现有客户。我们假设新客户与现有客户的坏账率之比为2.7

6.5 其他机构的信用卡

表9显示了其他机构信用卡数量的假定分布，以及指定的坏账率。

表9：其他机构的信用卡

组	描述	比例	不良率
0	在其他服务提供商处无信用卡	50%	1.0
1	在其他服务提供商处有信用卡	30%	1.2
2	在其他服务提供商处有多张信用卡	15%	1.7
3	在其他服务提供商处有三张或更多信用卡	5%	2.5

6.6 未偿还总额

我们假设客户的未偿总额服从标准对数正态分布，并按 10 000 的因子进行缩放。评分卡所含各水平的比例和坏账率见表10。

表10：未偿总额

组	分组	比例	不良率
0	0 - 5 000	24.4%	1.0
1	5 000 - 10 000	25.6%	1.2
2	10 000 - 25 000	32.0%	2.0
3	25 000 - 100 000	16.9%	2.1
4	超过 100 000	1.1%	0.8

6.7 居住省份

表11显示了南非9个省份各自所占的比例，以及相应的不良率。

表11：居住省份

组	描述	比例	不良率
0	豪登省	40%	1.0
1	西开普省	30%	0.7
2	夸祖鲁-纳塔尔省	7%	1.8
3	普马兰加省	5%	1.5
4	西北省	5%	3.0
5	林波波省	4%	2.5
6	东开普省	4%	2.0
7	北开普省	3%	4.0
8	自由邦省	2%	1.2

6.8 申请方式

表12列出了此属性的分布和坏账率。

表12：申请方式

组	描述	比例	不良率
0	分行	30%	1.0
1	在线	40%	0.5
2	电话	15%	1.5
3	营销电话	15%	0.4

6.9 收入

收入采用混合分布进行建模，其中包含若干局部模型。相关比例和坏账率见表13。

表13：收入

组	分组	比例	不良率
0	0 - 5 000	3.2%	3.0
1	5 000 - 11 000	15.6%	2.5
2	11 000 - 20 000	20.4%	2.0
3	20 000 - 30 000	21.8%	1.4
4	30 000 - 70 000	24.0%	1.2
5	超过70 000	15.0%	1.0

6.9.1 近期违约余额

表14列出了近期违约值的假设比例及相关坏账率。

表14：近期违约余额

组	分组	比例	不良率
0	0 - 1 000	60.0%	1.0
1	1 000 - 3 000	1.1%	1.1
2	3 000 - 5 000	2.1%	2.0
3	5 000 - 30 000	18.9%	2.5
4	30 000 - 1 000 000	18.0%	3.0
5	超过 1 000 000	0.0%	3.3

7 附录B

本附录详细介绍了第4.3节中所考察的审查总体分布变化。

7.1 性别

开发总体中女性占60%，男性占40%。审查总体中女性占55%，男性占35%。

7.2 查询次数

表15列出了查询次数属性在开发总体和审查总体中的比例。

表15：查询次数

组	描述	开发比例	审查比例
0	无咨询	30%	25%
1	一次咨询	25%	30%
2	两次咨询	20%	20%
3	三次咨询	15%	15%
4	四次咨询	5%	5%
5	五次或更多咨询	5%	5%

7.3 现有客户

在开发总体中，80% 的申请人被指定为现有客户，20% 的申请人为新客户。在审查总体中，75% 的申请人为现有客户，25% 为新客户。

7.4 与其他服务提供商合作的信用卡

表16规定了开发总体和审查总体中“其他机构的信用卡”属性的比例。

表16：其他机构的信用卡

组	描述	开发比例	审查比例
0	在其他提供商处没有信用卡	50%	45%
1	在其他提供商处有一张信用卡	30%	35%
2	在其他提供商处有多张信用卡	15%	15%
3	在其他提供商处有三张或更多信用卡	5%	5%

7.5 居住省份

表17规定了南非9个省份的开发总体和审查总体中的比例。

表17：居住省份

组	描述	开发比例	审查比例
0	豪登省	40%	35%
1	西开普省	30%	35%
2	夸祖鲁-纳塔尔省	7%	7%
3	普马兰加省	5%	5%
4	西北省	5%	5%
5	林波波省	4%	4%
6	东开普省	4%	4%
7	北开普省	3%	3%
8	自由邦省	2%	2%

7.6 申请方式

表18列出了 “申请方式”属性在开发总体和审查总体中的比例。

表18：申请方式

组	描述	开发比例	审查比例
0	分行	30%	25%
1	在线	40%	45%
2	电话	15%	15%
3	营销电话	15%	15%

参考文献

[1] W. J. Conover. 不连续分布的柯尔莫哥洛夫拟合优度检验。美国统计学会期刊，67:591– 596，1972。[2] H. Cramér. 关于基本误差的构成。I. 数学推导。II. 统计应用。斯堪的纳维亚精算期刊，11:13–74，141–180，1928。[3] J. Du Pisanie 和 I.J.H Visagie. 关于检验信用风险评分卡总体稳定性假设。ORiON期刊，36(1):19–34，2020。[4] R. J. Grissom 和 J. J. Kim. 研究效应量。Lawrence Erlbaum Associates，伦敦，英格兰，2005。[5] I.J.H Visagie J. Du Pisanie 和 J.S. Allison. 一种用于信用风险评分卡总体稳定性检验的模拟技术。《数学》，11(2)，2023。[6] A. Kolmogoroff. 关于分布规律的经验确定。意大利精算师协会公报，4:83–91，1933。[7] E M Lewis.《信用评分导论》。Athena Press，伦敦，英格兰，1994。[8] M. Pastore 和 A. Calcagnì. 样本间分布相似性的测量：一种无分布假设重叠指数。《心理学前沿》，10，2019。[9] N. Siddiqi.《信用风险评分卡：开发与实施智能信用评分》。John Wiley and Sons, Inc.，新泽西州，美国，2006。[10] N. Siddiqi.《智能信用评分：构建并实施更优的信用风险评分卡》。John Wiley and Sons, Inc.，新泽西州，美国，2016。

[11] R. Taplin 和 C. Hunt. 总体准确性指数：一种用于模型监测的总体稳定性新度量。Risks，7(2)，2019。
[12] R. von Mises.《概率、统计与真理》。维也纳：J. Springer（《科学世界观丛书》第3卷）。VII，189页（1928）。, 1928。[13] B. Yurdakul 和 J. Naranjo. 总体稳定性指数的统计特性。风险模型验证期刊，14(4)，2020。

此图“Rplot01.jpeg”以“JPEG”格式提供，网址为：

http://arxiv.org/ps/2303.01227v1