信用风险评分卡研究

基于SAS的开发与实施

马姆杜·雷法特 (Mamdouh Refaat) 著
王松奇 林治乾 译

CREDIT RISK SCORECARDS: DEVELOPMENT AND IMPLEMENTATION USING SAS

科学文献出版社
SCIENCES ACADEMIC PRESS (CHINA)

信用风险评分卡研究

基于SAS的开发与实施

马姆杜·雷法特 (Mamdouh Refaat) 著
王松奇 林治乾 译

CREDIT RISK SCORECARDS: DEVELOPMENT AND IMPLEMENTATION USING SAS

图书在版编目(CIP)数据

信用风险评分卡研究：基于SAS的开发与实施/（美）雷法特（Refaat，M.）著；王松奇，林治乾译. —北京：社会科学文献出版社，2013.7
ISBN 978-7-5097-4771-1

Ⅰ.①信… Ⅱ.①雷… ②王… ③林… Ⅲ.①贷款风险管理-统计分析-应用软件 Ⅳ.①F830.5-39

中国版本图书馆CIP数据核字（2013）第142614号

信用风险评分卡研究
——基于SAS的开发与实施

著　　者／马姆杜·雷法特
译　　者／王松奇　林治乾

出 版 人／谢寿光
出 版 者／社会科学文献出版社
地　　址／北京市西城区北三环中路甲29号院3号楼华龙大厦
邮政编码／100029

责任部门／经济与管理出版中心（010）59367226
责任编辑／王婧怡　许秀江
责任校对／张　羡
电子信箱／caijingbu@ssap.cn
责任印制／岳　阳
项目统筹／恽　薇
经　　销／社会科学文献出版社市场营销中心（010）59367081　59367089
读者服务／读者服务中心（010）59367028

印　　装／三河市尚艺印装有限公司
开　　本／787mm×1092mm　1/20
印　　张／17.8
版　　次／2013年7月第1版
字　　数／179千字
印　　次／2013年7月第1次印刷
书　　号／ISBN 978-7-5097-4771-1
著作权合同登记号／图字01-2013-3849号
定　　价／58.00元

Credit Risk Scorecards: Development and Implementation Using SAS

ISBN 978-1-4475-1119-9

Printed by LULU.COM-USA

目　录

第 1 章
评分卡的开发过程

1.1 标准评分卡

1.1.1 评分卡的类型和目的

信用评分卡主要分为两类：

1. 申请评分卡，对新贷款申请进行筛选并判断其违约风险。

2. 行为评分卡，对审批通过的贷款账户进行覆盖整个贷款周期的管理。

通常，申请评分卡被用来对新贷款申请进行一次性信用评分。其评分结果将决定以下几个方面：

- 估计的信用状况，即正常还是违约，并据此决定批准还是拒绝该笔贷款申请。
- 为了获得审批通过需要提供的抵押物。
- 贷款金额（信用额度）。
- 贷款定价（利率水平）。

行为评分卡被用来对已经通过审批并进入执行阶段的账户，即已经进行了一定交易的账户，进行信用评分。评分过程将反复进行，以监测和管理业务账户。其评分结果将用于：

- 审查信用重建。

- 审查信用额度。
- 制定清收策略（如果违约或逾期付款）。
- 审查贷款定价和贷款条件。

两种评分卡的开发过程都遵循同样的基本方案。但是，两种评分卡的开发过程中也存在两个主要的差别：

1. 通常，行为评分卡要比申请评分卡更为精确。因为行为评分卡在对账户状态进行预测时基于更多的数据要素（交易产生的）。

2. 拒绝演绎技术只在申请评分卡的开发过程中使用。关于拒绝演绎技术的详细介绍见本书第 13 章。

1.1.2 “正常”、“违约”和“不确定”

本书的目的是帮助分析人员以所谓的标准格式开发出强有效的评分卡。评分卡的目的是按照一定的标准分配分数，从而确保违约可能性高的账户的得分一定低于表现为正常的账户。

由于某些原因，正常和违约的概念是主观的并取决于企业。例如，可以将信用卡申请中的违约定义为一个账户的敞口金额未偿还期限达到 90 天或超过 90 天的状态。这种特定的选择通常被称为逾期 90 天。但是，也可以将违约定义为在特定的时间窗口内，比如 6 个月，一个贷款账户超过 2 次未能及时偿还到期金额的情况。关键在于，对于违约和正常并不存在统一的定义。但是，大多数评分卡的开发都是基于 60 天、90 天或 180 天逾期。

明确定义正常和违约的含义后，用于评分卡开发的数据中必须包含一个表示账户观测值的变量，即所谓的状态变量或指标。通常，状态变量分别用数值 1 表示违约，0 表示正常。

不确定账户状态可以被定义为介于违约和正常之间的另外一种状态变量。这种情况下，可以允许评分卡开发所用模型中的因变量分为两种以上的类别。

本书中，假设历史数据仅被分为两种类别：违约和正常。该状

态指标用一个二元标识表示，即 1 = 违约和 0 = 正常。

需要注意的是，上述选择并没有牺牲普遍性。账户状态的不确定值可以在后处理阶段考虑。例如，可以用以下的简单策略解释不确定，即介于违约和正常之间的状态：

- 如果得分小于 S_l，该账户被标记为违约；
- 如果得分大于 S_h，该账户被标记为正常；
- 如果得分在 S_l 和 S_h 之间，该账户需要人工干预。例如，可能要求提供额外的文件，如就业或收入的证明。

因此，尽管评分卡开发过程中使用的是一个二元状态指标，但可以将其用于多层次决策策略的设计。

1.1.3　标准评分卡格式

标准评分卡采用的格式如表 1.1 所示。

表 1.1 是使用三个决定变量（dicision variables）的标准评分卡：

- Age：账户持有人的年龄；
- TmAtAddress：在当前住址的居住年限；
- EmpStatus：就业状况。

使用该评分卡对以下记录进行评分：

- Age = 37；

表 1.1　标准评分卡示例

变　量	条　　件	分值
基础分值		485
Age	如果 Age < 25	19
	如果 Age ≥ 25 且 Age < 33	28
	如果 Age ≥ 33 且 Age < 48	39
	如果 Age ≥ 48 且 Age < 56	24
	如果 Age ≥ 56	20
	如果 Age = “缺失值”	19

续表

变　量	条　　件	分值
TAR	如果 TmAtAddress < 1	12
	如果 TmAtAddress ≥ 1 且 TAR < 3	24
	如果 TmAtAddress ≥ 3 且 TAR < 5	36
	如果 TmAtAddress ≥ 5	41
	如果 TmAtAddress = “缺失值”	17
ES	如果 EmpStatus = “全职”	38
	如果 EmpStatus = “兼职”	19
	如果 EmpStatus = “自由职业”	25
	如果 EmpStatus = “失业”	7
	如果 EmpStatus = “缺失值”	3

- TmAtAddress = 3.5;
- EmpStatus = “全职”。

可以得到总的信用评分是 485 + 39 + 36 + 38 = 598。

除了为抽象的统计模型提供简明的表现形式外，标准评分卡还有其他重要优势。这些优势包括：

1. 容易理解，因为采用了为人熟知的表格形式。

2. 将评分卡中每个变量的贡献加总得到一个账户的总信用评分。这使得最终分值对普通大众来说更加透明。当信贷产品的监管法规要求客户有了解其信用评分及评分理由的权利时，这将变得十分关键。

3. 比率（odds），依据可计算得到的比率，信用状况是正常或违约的概率与总的评分直接相关。全部细节将在本书第 11 章介绍。

4. 决定变量（预测变量）采用表格形式，总信用评分的计算采用简单加法，这些使得评分卡容易在众多实施平台和编程语言环境下实施。

5. 对于每个预测变量，根据不同类别或范围分别赋予一定分值，使得最终消费者清楚如何提高其信用评分。例如，表 1.1 中的

评分卡表明，如果一个申请人在申请表中未提供就业状态，则就业状态一项在其信用评分中的贡献只有 3 分。如果在随后阶段，该申请人能够提供全职工作的证明材料，其信用评分将提高 35 分（38 - 3）。

1.2　评分卡开发流程

图 1.1 概括了典型的评分卡开发流程。该流程中各个步骤的顺序可以根据具体情况的不同进行调整，也可以根据需要重复某些步骤。

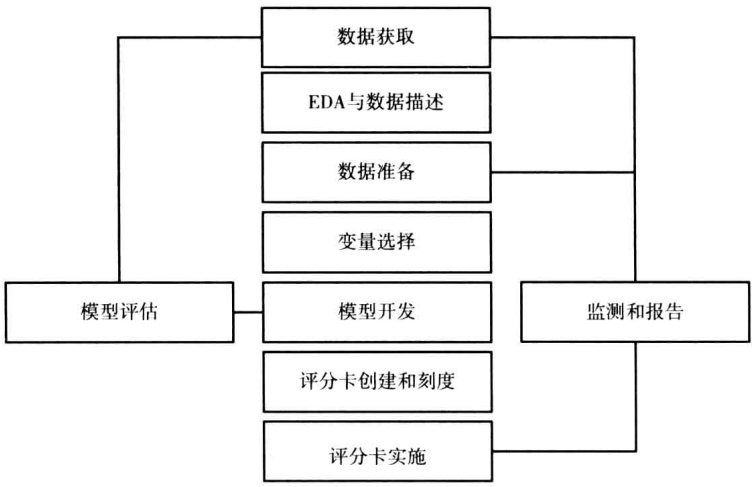

图 1.1　典型的评分卡开发流程

该流程的主要步骤包括：

1. 问题准备
2. 数据获取与整合
3. EDA（探索性数据分析）与数据描述
4. 数据准备

5. 变量选择

6. 模型开发

7. 模型检验和评价

8. 评分卡创建和刻度

9. 评分卡实施

10. 监测和报告

1.3 问题准备

在此阶段，需要做出下列决策和解决下列问题：

1. 在特定业务重点、财务结果和具体信贷产品历史表现的基础上，确定违约和正常的定义。

2. 确定计划的评分卡的范围、开发和实施窗口。

3. 识别数据的范围和来源，内部还是外部，并确保能够取得这些数据。

4. 设计主要项目管理计划，对时间、资源、人员等进行管理。

1.4 数据获取与整合

本阶段从识别与评分卡开发相关并能够获取到的数据开始。数据获取的过程包括取得这些数据项并将其整合为适合进一步数据准备的形式。这些数据可能并不在一个表格或 SAS 数据集中。实际上，如果数据规模很大，甚至建议将这些数据分成几个表格以加快数据准备阶段的处理速度。通常，数据表中的每一行代表一个账户。但是，每一行并不一定对应一个客户记录。在很多贷款申请中，多个客户同时对应一个账户（联名账户，joint accounts）。因此，基于客户开发的评分卡通常适用于客户级的评分。

除了企业内部数据，还可以获取外部资源，如征信局数据和

外部评分。随着外部数据提供商和供应商数量的增加，产品数量和评分数据急剧增加，选择合适的、最有用的外部资源变得更加重要。

第 2 章将对数据获取问题进行概述，并介绍一些常用的数据联结和整合方法。

1.5　EDA 与数据描述

探索性数据分析（EDA）和数据描述是检查数据并理解其特征的一系列过程的名称。在评分卡开发过程中，需要进行下列分析：

- 候选预测变量单变量统计特征的评价，及其取值在变量范围内的分布。
- 计算每个候选预测变量分类或分段条件下的违约率分布，也被称为要素分析。
- 通过列联表、关联性和相关性指标确定不同变量之间的检验关系。

EDA 和数据描述的过程将在第 3 章进行详细介绍。

1.6　数据准备

数据准备是整个评分卡开发过程中最重要，也是最耗时的工作。实际上，数据准备阶段消耗的时间占整个项目时间的 80% 以上。数据准备的目的就是创建所谓的数据挖掘或建模视图，即包含开发评分卡模型所需的所有要素的唯一数据集。为了准备该视图，通常需要进行大量的数据清洗和转换工作，以创建具有较强预测力的预测指标或自变量。

关于数据准备问题的详细介绍不包含在本书的范围之内。第 5

章仅限于介绍评分卡开发的数据准备中最常见的任务。对数据准备的详细内容感兴趣的读者可以参阅参考文献［10］。

证据权重（WOE）转换是评分卡开发过程中一个特有的数据准备过程。评分卡中使用的所有变量都需要进行 WOE 转换。然而，对变量进行转换前，需要减少分类变量的基数，需要将连续变量分段。分段和降低基数与 WOE 转换一道，被称为粗分类。

数据准备过程，包括降低基数和连续变量的最优分段，将在第 5 章进行介绍。粗分类将在第 8 章进行介绍。

读者或许会对本书中将数据准备放在 EDA 和数据描述之后的章节安排感到疑惑。实际上，这并不意味着必须要按照这种顺序进行。数据准备和 EDA 是两个密切相关的步骤。通常，EDA 和数据描述揭示出需要的具体数据变量的转换，而数据准备生成需要进行分析和描述的新变量。因此，对这两个步骤的上述顺序安排是主观的，对这两个主题进行划分的目的仅仅是为了避免一个太大的章节。

1.7 变量选择

数据准备和转换过程的成果就是包含众多候选自变量的建模视图。并不是所有这些候选自变量都会在模型中得到实际应用。大多数信贷供应商都拥有丰富的数据，有数百甚至上千的建模变量。处理如此大量数据的最好方法就是只选择那些表现出较强预测力的变量，以减少变量的数量。

评分卡开发中使用的衡量候选自变量预测力的指标以及选择评分卡中使用的最优变量的程序将在第 4 章介绍。

尽管已经用衡量预测力的指标减少了候选预测变量的数量，但并不是所有被选中的变量都会出现在最终模型中。Logistic 回归，