HZ BOOKS
华章科技

SAS
SAS 大学
技术丛书

SAS
金融数据挖掘与建模
系统方法与案例解析

陈春宝 徐筱刚 田建中◎著

SAS大学授权编写，SAS大学指定参考书！

以金融客户生命周期管理为主线，纯实战导向，通过5个经典案例详细讲解金融数据挖掘与建模的方法与技巧

SAS Financial Data Mining and Modeling
Practical Applications and Analytics

机械工业出版社
China Machine Press

SAS大学
技术丛书

SAS
金融数据挖掘与建模
系统方法与案例解析

陈春宝 徐筱刚 田建中◎著

SAS Financial Data Mining and Modeling
Practical Applications and Analytics

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

SAS 金融数据挖掘与建模：系统方法与案例解析 / 陈春宝，徐筱刚，田建中著 . —北京：机械工业出版社，2017.9
（SAS 大学技术丛书）

ISBN 978-7-111-58047-8

I. S… II. ①陈… ②徐… ③田… III. 金融统计 - 统计分析 - 应用软件 IV. F830.2

中国版本图书馆 CIP 数据核字（2017）第 232741 号

SAS 金融数据挖掘与建模：系统方法与案例解析

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：何欣阳 责任校对：李秋荣
印　　刷：北京市荣盛彩色印刷有限公司 版　　次：2017 年 10 月第 1 版第 1 次印刷
开　　本：186mm × 240mm 1/16 印　　张：14
书　　号：ISBN 978-7-111-58047-8 定　　价：59.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

Praise 赞　誉

数字化转型是银行业未来十年的发展趋势，大数据的分析与应用能力至关重要。本书系统归纳了数字化客户经营的主要场景，并选取最有代表性的案例介绍分析建模过程，写法通俗易懂，对技术和业务人员提升数字化客户经营能力都很有帮助。

——吴纯杰 上海财经大学统计与管理学院副院长

SAS 是统计分析软件领域的标杆，以功能强大著称，已经有超过 40 年的历史。由 SAS 技术大学精英学院组织撰写的这套书，有 SAS 编程的主题，有 SAS EG 和 SAS EM 这样的重要工具，还有金融数据挖掘与建模这样的行业应用，内容系统、立体、丰富，强烈推荐！

——俞章盛 上海交通大学教授 / 博士生导师

SAS 是全球分析领域的引领者，数十年来一直致力于应用分析解决最困难的业务问题，在统计分析、商业智能、客户智能、数据管理、风险管理、欺诈与安全智能等多个领域独树一帜。相信由 SAS 技术大学官方编写的系列参考书，一定能将 SAS 的技术精华和优秀实践总结、提炼出来，奉献给广大的 SAS 技术、产品的支持者和使用者们。

——宇传华 武汉大学教授 / 博士生导师

在开源软件大行其道的今天，作为商业软件，SAS 不仅没有没落，反而正爆发出更强大的生命力，这与 SAS 公司与时俱进的创新能力是分不开的。SAS 的技术和产品在不断改进，SAS 的教育工作也一直做得很好，每年一度的“高校 SAS 数据分析大赛”在教育界的影响力越来越大。他们出版的“SAS 大学技术丛书”一定能再为 SAS 教育工作添砖加瓦。

——杨启贵 华南理工大学教授 / 数学学院副院长 / 博士生导师

Foreword 序

大数据的浪潮正渐渐平静，整个行业已逐渐趋于成熟和理性。喧嚣与嘈杂渐远之时，才能更清晰地透过表象，看清事情的本质。大数据需要精挖掘，好客户需要勤耕耘，再好的故事，再炫的包装，再酷的产品，分析基础还是以 CRISP 方法为主流，做好数据采集、清洗、整合、建模、分析、部署与调优；经营基础还是“客户为中心”的市场营销，通过信息对称与否的博弈来驾驭经营风险，再给予风险成本加权计算基础上的损益评价。大数据的世界有时确实没表面上那么“性感”，而是非常“感性”。对于从事大数据相关工作的大多数专业人士，尤其是年轻朋友而言，诗与远方虽可筑梦，但要真正走得长远，还是需要真本领，需要耐着寂寞，翻开书本，撸起袖子，在实践中学习，在学习中实践。

书如其人。陈春宝博士的这本书和他本人一样，不太容易评价。多样、丰富、立体，因此复杂，需要多花些时间去深入细品，才能发现如同一篇好的散文，贵在“形散而神不散”。全书围绕两条主线，一条是金融客户生命周期管理，另外一条是数据挖掘项目和模型的生命周期管理，这两条主线串接起了一个一个独立、完整的实战场景。这样的组织方法使读者学习各篇完整成章时，能按图索骥，实践参考；两条主线也使得知识点跳跃较大，给人点到即止的感觉。因此建议结合其他书籍同步研读，效果会更好。比如，客户关系管理方面的，V. 库马尔的《赢得盈利客户》；数据挖掘领域的，本书作者的另一部著作《大数据与机器学习：实践方法与行业案例》。

金融数据的价值密度之高，堪称数据中的黄金。以银行为代表的金融机构对数据的深入挖掘分析与应用，起步不可谓不早，投入不可谓不大，成果不可谓不多，但是在这波风口中，由于内因、外因和低调（网红经济时代，低调可真是“致命”的优点），却常常被怼

到了市场边缘，连市面上关于大数据与数据挖掘方面的实战类专业书籍，也大多出自互联网同仁之手。其实银行及各家金融机构藏龙卧虎，不乏像三位作者这样能够洞察业务，兼具丰富实操经验和扎实理论功底的高手。衷心期待陈春宝博士能继续坚持下去，并带动更多金融机构的同仁们积极行动进来，出版更多优秀的，特别是实战类的书籍，共同为大数据时代增添一抹属于金融数据科学家们的别样风采。

陆小勇

浦发银行信息科技部副总经理，信息服务中心主任

Preface 前　言

古之欲明德于天下者，先治其国；欲治其国者，先齐其家；欲齐其家者，先修其身；欲修其身者，先正其心；欲正其心者，先诚其意；欲诚其意者，先致其知；致知在格物。

——《礼记》

知之真切笃实处即是行，行之明觉精察处即是知。

——王阳明

大数据势不可挡。然而，对于多数公司来说，数据分析和建模能力尚未完全发展起来，虽主观意识上认同了大数据的潜在价值，也开始采集、储备数据，却不知如何才能让数据充分融入业务、帮助业务部门达成业务指标。

大数据是一种全新的业务和产品创新思维，是海量数据存储和计算的基础架构，但小数据的分析运用才是多数公司和业务领域必须关注和掌握的核心能力。本书将聚焦于实践应用，介绍数据分析、建模的方法和在业务领域的实际应用，原理和基础理论知识不是重点，因此数学公式极少，除非它比文字更能表达内容。总体上，本书不会详细罗列最热门的机器学习算法、数据挖掘方法以及人工智能，而是基于金融企业当前的实际需要，精选最具代表性的业务领域以及被广泛验证实用高效的分析建模技术，这些技术是数据分析人员必须掌握的技能。本书同时也是为掌握统计学知识和基本数据分析方法的业务专家所写，帮助他们实践、应用数据建模手段，提升对业务的引导和驾驭能力。

本书的目标读者是高级数据分析师、咨询顾问、企业内部的业务专家、高校学者和研究生，以及立志于夯实数据建模基本功，并希望不断提升的数据挖掘与数据建模人员。

内容提要

知者过之，愚者不及也；贤者过之，不肖者不及也。

——《中庸》

在学校和生活中，工作的最重要的动力是工作中的乐趣，是工作获得结果时的乐趣以及对这个结果的社会价值的认识。

——阿尔伯特·爱因斯坦

本书是一本介绍金融企业数据建模的专著。在内容上，书中以信贷（信用卡）客户的生命周期管理为主线，选取了 5 个在客户获取、提升、成熟和衰退环节的最经典的金融企业案例，来详细介绍最具价值与实用性的数据建模过程，每个案例既自成体系又前后呼应。

第 1 章介绍了数据挖掘和建模在信贷（信用卡）客户生命周期管理中的应用场景。

第 2 章结合信用卡客户反欺诈案例，介绍了常用的三类反欺诈手段以及欺诈评分模型的构建过程，模型采用机器学习集成算法的典范——随机森林，并给出 SAS 代码（各类书中绝无仅有），对回归类、决策树类、神经网络类三大类机器学习算法做了比对。

第 3 章结合信用卡客户精准营销案例，介绍了营销响应模型的构建、评估与应用，完整阐述从数据准备、清洗、变量粗筛选、变量压缩与转换、建模、模型评估、部署、监测与更新等模型构建过程中所涉及的操作方法。

第 4 章通过信用卡客户细分案例，介绍了完整的聚类过程，除快速、系统、两步聚类算法外，还详细介绍了实际分析过程中必不可少的数据预处理过程，并对聚类模型做了最完整的阐释。

第 5 章通过贷款违约预测案例，为零建模基础的读者提供了一个最简化的行为评分模型的构建过程，帮助零基础读者快速上手，同时简单介绍了金融企业的三大风险模型（评分卡）。

第 6 章结合信用卡客户流失预警与挽留案例，介绍客户价值（数值）预测与流失倾向（事件）预测两类问题的建模过程及组合应用，不拘泥于方法本身，彰显了以企业实际运用为导向的写作思路，让案例更具实用参考价值。

了解完五个案例之后，你会发现这些方法和模型在大部分业务场景中似曾相识，金融企业的数据挖掘与建模将变得易如反掌。

全书由陈春宝统稿，其中，第 1、2、5、6 章由陈春宝撰写，第 3 章由徐筱刚撰写，第 4 章由田建中撰写。

源代码下载

若你对书中源代码感兴趣，可与作者联系，邮箱：64346837@qq.com。

Contents 目　录

赞誉
序
前言

第1章　金融数据挖掘与建模应用场景 …… 1

1.1　客户数据挖掘的价值 …… 1

1.2　金融客户生命周期及数据应用场景 …… 3

1.3　最具代表性的数据应用场景 …… 7

第2章　客户获取：信用卡客户欺诈评分案例 …… 8

2.1　案例背景 …… 9

2.2　数据准备与预处理 …… 10

2.2.1　数据源 …… 10

2.2.2　变量设计 …… 11

2.3　构建评分模型 …… 13

2.3.1　算法选择 …… 13

2.3.2　模型训练 …… 14

2.3.3　模型评估 …… 16

2.4　评分模型的应用 …… 19

2.5　小结 …… 20

第3章　客户提升：信用卡客户精准营销案例 …… 21

3.1　案例背景 …… 21

3.2　建模准备 …… 21

3.2.1　准备数据 …… 22

3.2.2　数据预处理 …… 26

3.2.3　过度抽样 …… 27

3.2.4　构造训练集及测试集 …… 30

3.3　数据清洗及变量粗筛 …… 32

3.3.1　连续变量与连续变量之间 …… 33

3.3.2　分类变量和分类变量之间 …… 39

3.3.3　分类变量和连续变量之间 …… 43

3.3.4　数据的错误及缺失值 …… 47

3.3.5　数据离群值 …… 53

3.3.6　重编码 …… 59

3.4　变量压缩与转换变量 …… 61

3.4.1　分类变量的水平数压缩 …… 61

3.4.2　连续变量聚类 …… 65

3.4.3　连续变量的分箱 …… 77

3.4.4　变量的转换 …… 79

3.5　模型训练 …… 80

3.5.1 关于 Logistic 回归……80
3.5.2 变量筛选方法……81
3.6 模型评估……88
3.6.1 模型估计……88
3.6.2 模型评估……89
3.6.3 调整过度抽样……98
3.6.4 收益矩阵……98
3.6.5 模型转换为打分卡……100
3.7 模型的部署及更新……100
3.7.1 模型的部署……100
3.7.2 模型的监测及更新……101
3.8 本章小结……103

第4章 客户成熟：银行零售客户渠道偏好细分案例……104

4.1 案例背景……104
4.2 聚类分析流程……105
4.3 数据标准化……107
4.3.1 标准化介绍……107
4.3.2 标准化实现……110
4.4 变量聚类……111
4.4.1 变量聚类介绍……111
4.4.2 变量聚类基本步骤……112
4.4.3 SAS 实现变量聚类……113
4.5 变量降维与可视化……118
4.5.1 图形化探索……118
4.5.2 主成分分析法降维……120
4.6 ACECLUS 预处理过程……123
4.6.1 ACECLUS 介绍……123
4.6.2 ACECLUS 过程……123
4.6.3 ACECLUS 示例……123
4.7 系统聚类分析……128
4.7.1 系统聚类法……128
4.7.2 样本与样本之间的度量……129
4.7.3 距离定义与测量……129
4.7.4 相关系数……131
4.7.5 类与类之间的度量……131
4.7.6 系统聚类法……139
4.7.7 不同系统聚类法之间的比较……140
4.7.8 类个数的确定……158
4.8 快速聚类……159
4.8.1 快速聚类法……159
4.8.2 快速聚类法实现……160
4.8.3 快速聚类法优缺点……161
4.9 两步聚类法……161
4.9.1 两步聚类法……161
4.9.2 两步聚类法实现……161
4.10 本章小结……167

第5章 客户衰退：银行贷款违约预测案例……168

5.1 案例背景……169
5.2 维度分析……170
5.3 建模分析……177
5.4 业务应用……179
5.5 小结……179

第6章 客户挽留：信用卡客户流失管理案例……180

6.1 案例背景……181
6.2 数据准备……182

6.2.1 设定目标变量 …… 182
6.2.2 设定时间窗 …… 183
6.2.3 设计预测变量 …… 184
6.2.4 准备数据宽表 …… 185
6.3 流失倾向预警：用 Logistic 回归构建响应率模型 …… 186
6.3.1 粗分类 …… 187
6.3.2 计算分组变量的 WOE 值和 IV 值 …… 191
6.3.3 共线性检验 …… 194
6.3.4 模型训练：显著性检验 …… 195
6.3.5 模型评估 …… 196
6.4 潜在客户价值预测：两阶段建模法 …… 201
6.4.1 阶段 1 概率预测 …… 201
6.4.2 阶段 2 数值预测 …… 201
6.4.3 模型评估 …… 203
6.5 细分：差异化营销服务的基础 …… 204
6.6 小结 …… 208

第 1 章 Chapter 1

金融数据挖掘与建模应用场景

己欲立而立人，己欲达而达人。

——《论语·雍也》

一颗善良的心就是一席永恒的筵席。

——夸美纽斯

1.1 客户数据挖掘的价值

金融业属于数据密集型行业，数据在大量的业务场景中广泛应用并创造着价值。如图 1-1 所示，以美国经济为例，其金融和保险业的大数据获取能力和价值潜力均领先于其他各个行业。

在产品同质化的市场环境下，企业之间的竞争已经由产品品质的竞争转为顾客满意度的竞争，企业能够长远发展和领先市场的核心是针对不同人群的不同需求提供真正差异化的产品、服务以及营销策略，因此，对客户数据深层次的挖掘至关重要。客户数据就像一座金矿，不断发掘和提炼，方能超越表层价值。在数据爆炸的年代，人们经常置身于海量信息和产品之中，却不知道自己想要什么。通过对客户数据的深层次挖掘，可

以洞察客户的真实需求，获得超乎表面数据所能提供的价值，主要包括：

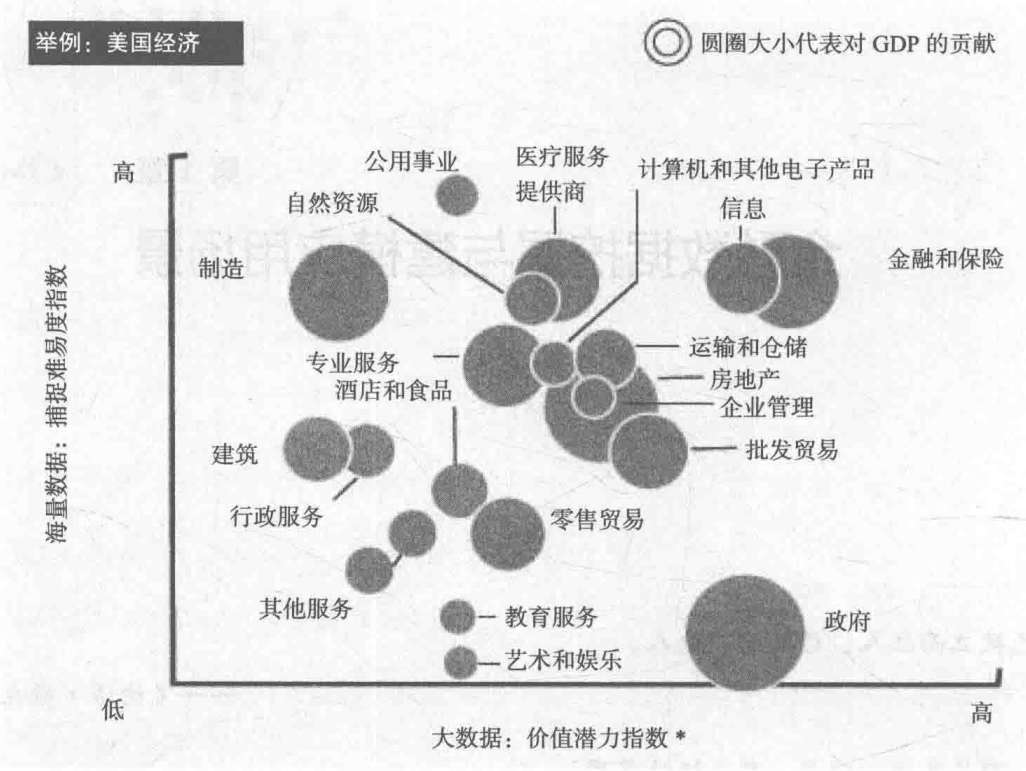

图 1-1　美国各行业大数据发展对比

资料来源：麦肯锡全球研究院。

（1）更完整的客户描述

与“以客户为中心”的管理模式相对应，企业正从传统面向群体的营销方式转向个性化营销方式，主动迎合客户需求，而前提就是要区分出不同的人群。在对客户更深刻了解的基础上，进行深层次的分析，可以绘制出更完整、更明确的客户画像，营销与服务人员也能够更形象地了解每一类人群。

（2）更深层次的客户需求洞察

挖掘客户的行为习惯和喜好，在凌乱纷繁的数据背后找出更符合客户需求的产品和服务，并对产品和服务进行针对性的调整和优化。同时，能够围绕客户需求对客户进行细分，真正做到个性化，而非简单地划分群体。

（3）更精细化的经营方式

这里的经营包括营销以及为客户提供产品与服务的过程。深层次挖掘客户数据能够帮助企业优化经营方案，在科学的客群细分基础上提供有针对性的服务与营销，从中获得更大的价值。比如：在一定周期内向客户发送他们最可能感兴趣的产品和优惠活动；基于历史交易记录，针对性地给他们推荐商户和餐馆优惠活动，并根据客户的回应不断优化推荐质量。

无论是360客户视图、客户标签体系，还是客户画像，都体现着数据对更好地理解客户需求、提升客户体验所发挥的有效价值。

1.2　金融客户生命周期及数据应用场景

客户是金融企业最重要的资产，与高价值的客户保持长期稳定的关系是企业获得持续竞争优势的关键，老客户更容易购买公司更多的产品、对价格也更不敏感；忠诚客户更愿意主动为公司传递好的口碑、推荐新的客户。为此，客户生命周期管理是任何一家金融企业的业务发展主线，客户关系管理、精细化经营、精准营销、风险管理等皆围绕该主线部署。那么，什么是客户生命周期？一个人会经历从出生、上学、毕业、工作、结婚、生子到退休等一系列生命阶段，而作为一个客户，与企业之间的关系会经历从潜在客户、响应者、新客户、稳定客户到流失客户等一系列生命周期，这也是企业与客户建立业务关系到关系终止的全过程。通常将客户生命周期划分为起始期、发展期、成熟期和衰退期四个阶段，称四阶段模型。起始期是客户关系的孕育期，发展期是客户关系的快速发展期，成熟期是客户关系相对稳定的时期，衰退期是客户关系发生退化、逆转的时期。

根据金融业的特点，这里将客户生命周期划分为五个阶段：客户获取、客户提升、客户成熟、客户衰退和客户挽留，客户经营中最重要的业务问题和大数据应用场景与各阶段一一对应，贯穿整个生命周期，如图1-2所示。

1. 客户获取阶段

客户获取阶段是指吸引潜在客户并将他们发展成客户的过程，发现潜在客户并过滤

具有欺诈、高风险特质的“坏客户”，是本阶段的核心业务问题，数据挖掘与建模的典型应用场景可以归纳为三个方面。

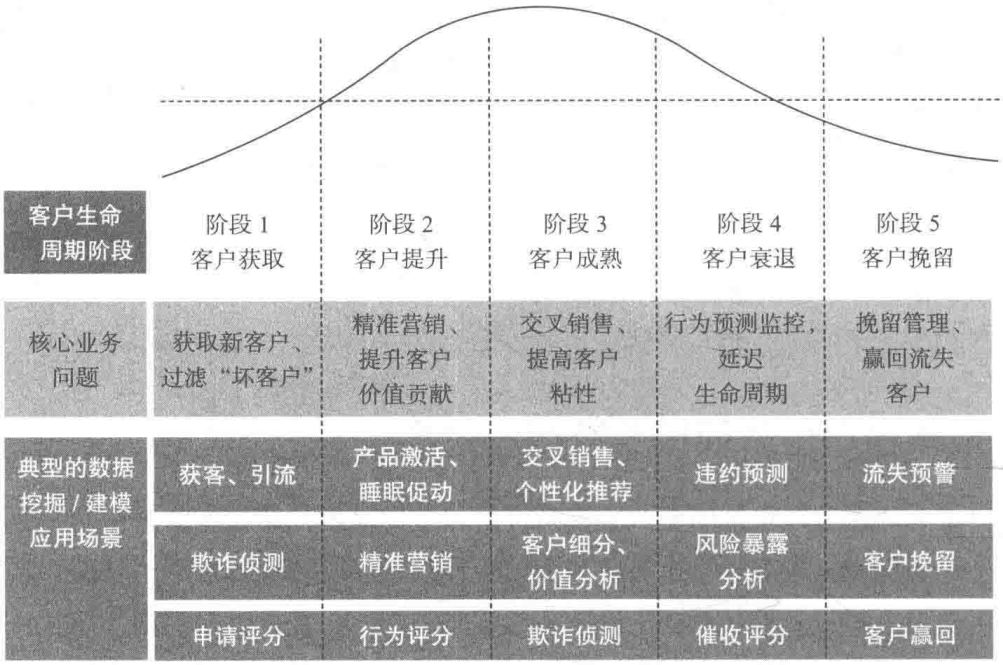

图 1-2　客户生命周期的数据挖掘与建模应用场景

1）获客与引流模型。从接触信息或外部数据中发现具有潜在需求的客户，定向发送信息，第三方引流是目前常见的大数据商业模式，广泛应用于信用卡发卡、商户合作等业务领域。

2）欺诈侦测。具体指申请阶段的反欺诈，通过构建欺诈黑名单库、欺诈规则库、欺诈评分模型、高风险人群关系网络等技术手段，为获客筑起第一道防火墙，将“坏客户”挡在门外。

3）申请评分 / 信用评分。这是银行的三大风险模型之一，通过客户申请时填写以及通过其他渠道查询到的信息，预测将来发生违约、逾期、坏账等的统计概率，例如信用卡公司决定是否向客户发卡、银行决定是否允许信贷审批；在另一个领域，大数据征信成为互联网金融的主要商业模式。

该阶段的客户都是新客户，企业尚不掌握客户的第一手信息，而仅靠客户申请时填

写的信息相对有限，因此，引进外部数据、开展第三方机构合作、采购大数据征信与反欺诈产品等成为传统金融企业的刚需，转而促成了大数据征信与相关数据产品的蓬勃发展。

2. 客户提升阶段

客户获取之后，如何促成客户交易、提升活跃度进而让他们贡献更多价值，是客户提升阶段要解决的核心业务问题，精准营销成为客户提升的代名词，本阶段数据挖掘与建模的典型应用场景同样归纳为三个方面。

1）产品激活与睡眠促动。信用卡发出去之后首先要解决的是让客户激活，银行给予客户授信之后要促进客户提款、用款，对于长期不动账的客户要让客户消费、转账、活跃起来，针对各类产品或客群，构建促动模型开展大促让客户动起来，是第一要务。

2）精准营销（营销响应模型）。如何最大化利用客户资源、提高单一客户的价值和利润？通过一组营销响应模型准确匹配关键营销要素，具体需要对的目标客户、对的接触渠道、对的产品与服务、对的价格、对的时间和对的接触方式，在深入了解客户需求及特征的基础上实施针对性营销，提高营销成功率，从而实现单一客户价值和利润贡献的提升。

3）行为评分。以贷款为例，通过客户的历史行为表现，预测其将来发生违约、逾期等的统计概率，防患于未然。狭义上的行为评分指风险评分模型，而实际上，营销响应模型是行为评分在营销领域的应用，评分结果为建立客户标签体系和精准营销提供依据。

精准营销不是新概念，但大数据对其作了新的阐释，催生出接触点营销、事件式营销、社交圈营销、大数据预测营销等新型方式，因此数据挖掘和建模技术将不断赋予传统的客户经营理念以新的抓手。

3. 客户成熟阶段

在客户成熟阶段，客户已经活跃起来并贡献着价值，此时企业要做的是让这种状态尽可能保持下去，即提高客户黏性，而通过交叉销售的营销策略售卖更多产品是最常用的手段。数据挖掘与建模的典型应用场景包括：

1）交叉销售与个性化推荐。为不同业务部门提供一致的客户信息视图，快速进行跨