JIYU R YUYAN SHUJU WAJUE DE TONGJI YU FENXI

基于R语言数据挖掘的统计与分析

韦鹏程　邹　杨　冉　维　著

电子科技大学出版社

基于R语言数据挖掘的统计与分析

韦鹏程　邹　杨　冉　维　著

电子科技大学出版社
University of Electronic Science and Technology of China Press

图书在版编目(CIP)数据

基于R语言数据挖掘的统计与分析/韦鹏程，邹杨，冉维著. -- 成都：电子科技大学出版社，2017.12
ISBN 978-7-5647-5409-9

Ⅰ.①基… Ⅱ.①韦…②邹…③冉… Ⅲ.①程序语言-程序设计 Ⅳ.①TP312

中国版本图书馆CIP数据核字(2017)第288905号

基于R语言数据挖掘的统计与分析

韦鹏程 邹 杨 冉 维 著

策划编辑 李述娜 卢 莉

责任编辑 卢 莉

出版发行 电子科技大学出版社

成都市一环路东一段159号电子信息产业大厦九楼 邮编 610051

主 页 www.uestcp.com.cn

服务电话 028-83203399

邮购电话 028-83201495

印 刷 北京一鑫印务有限责任公司

成品尺寸 170mm × 240mm

印 张 15.75

字 数 326千字

版 次 2017年12月第一版

印 次 2017年12月第一次印刷

书 号 ISBN 978-7-5647-5409-9

定 价 57.00元

版权所有，侵权必究

前言
PREFACE

现在的社会是一个高速发展的社会，科技发达，信息流通，人们之间的交流越来越密切，生活也越来越方便，大数据就是这个高科技时代的产物。有人把数据比喻为蕴藏能量的煤矿。煤炭按照性质有焦煤、无烟煤、肥煤、贫煤等分类，而露天煤矿、深山煤矿的挖掘成本又不一样。与此类似，大数据并不在“大”，而在于“有用”。价值含量、挖掘成本比数量更为重要。对于很多行业而言，如何利用这些大规模数据成为赢得竞争的关键。

在时下商界的流行语中，很难找出一个比“大数据”更吸引眼球的术语了。大数据的颠覆和创新作用几乎在每个行业都有体现。

20 世纪 90 年代末，美国航空航天局的研究人员创造了大数据一词，自诞生以来，它一直是一个模糊而诱人的概念，直到最近几年，才跃升为一个主流词汇。但是，人们对它的态度却仍占据了光谱的两端，一些人对它抱有近乎宗教崇拜的热情，认为大数据时代将释放出巨大的价值，是通往未来的必经之途。在一些观察者眼中，大数据已成为劳动力和资本之外的第三生产力。

大数据时代已经来临，它将在众多领域掀起变革的巨浪。但我们要冷静地看到，大数据的核心在于为客户挖掘数据中蕴藏的价值，而不是软硬件的堆砌。因此，针对不同领域的大数据应用模式、商业模式研究将是大数据产业健康发展的关键。我们相信，在国家的统筹规划与支持下，通过各地方政府因地制宜制定大数据产业发展策略，通过国内外 IT 龙头企业以及众多创新企业的积极参与，大数据产业未来发展前景十分广阔。

在这个信息爆炸的时代，人们随时随地受到各种数据的“轰炸”，有限的注意力被各种碎片化的内容“瓜分”。那么如何才能获得想要的数据信息呢？这需要数据足够“聪明”，海量的原始数据只有经过分类、加工、整理、分析才能满足不同的要求。而要处理海量的数据，就需要大数据技术的支持。

R 语言作为如今最热门的编程语言之一，它由统计学家开发，在解决数据分析问题时具有先天优势。它是一门新兴的语言，掌握它，就是掌握了一门高校的数据分析软件。随着大数据的轰炸，R 语言的功能越来越丰富，越来越多的人对 R 语言产生了兴趣。R 语言的特点主要是开源性、全面性、操作简便性、可扩展性等。

本书的编写是为了让对 R 语言有兴趣的读者能更加了解 R 语言，了解大数据时代的数据挖掘等，还可以解决一些人员遇到的困难。

本专著由重庆第二师范学院韦鹏程教授，邹杨，冉维等三位教师完成，并得到重庆市交互式教育电子工程技术研究中心、重庆市儿童大数据工程实验室和重庆第二师范学院计算机科学与技术重点学科支持，在此表示感谢

由于时间的仓促，编者水平有限，本书难免存在不足之处，在此出版之际，我们真诚地希望读者对本书提出宝贵的意见和建议。

试读结束：需要全本请在线购买： www.ertongbook.com

目录 CONTENTS

第一章

大数据时代数据挖掘

第一节　大数据概述

全球知名咨询公司麦肯锡在研究报告中指出，数据已经渗透到每一个行业和业务职能领域，逐渐成为重要的生产因素；而人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来。近年来，“大数据”概念的提出为中国数据分析行业的发展提供了无限的空间，越来越多的人认识到了数据的价值。

那么，什么是“大数据”呢？大数据是规模非常巨大和复杂的数据集，传统数据库管理工具处理起来面临很多问题，如获取、存储、检索、共享、分析和可视化，数据量达到 PB、EB 和 ZB 的级别。大数据有四个特点：

1. 数据量（Volume）是持续快速增加的；
2. 高速度（Velocity）的数据 I/O；
3. 多样化（Variety）的数据类型和来源；
4. 数据价值（Value）大。

一、大数据的应用

大约从 2009 年开始，“大数据”才成为互联网信息技术行业的流行词汇。美国互联网数据中心指出，互联网上的数据每年将增长 50%，每两年便翻一番，而目前世界上 90% 以上的数据是最近几年才产生的。此外，数据又并非单纯指人们在互联网上发布的信息，全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化，也产生了海量的数据信息。

数据充斥所带来的影响远远超出了企业界。贾斯汀·格里莫将数学与政治科学联系起来，他研究的内容涉及对博客文章、国会演讲和新闻稿进行计算机自动化分析等，希望借此洞察政治观点是如何传播的。在科学和体育、广告和公共卫生等其他许多领

域中，也有着类似的情况——也就是朝着数据驱动型的发现和决策的方向发生转变。

在公共卫生、经济发展和经济预测等领域中，“大数据”的预见能力正在被开发中，而且已经崭露头角。研究者发现，曾有一次他们发现“流感症状”和“流感治疗”等词汇在谷歌上的搜索查询量增加，而在几个星期以后，到某个地区医院急诊室就诊的流感病人数量就有所增加。

二、大数据的战略意义

大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。且中国物联网校企联盟认为，物联网的发展离不开大数据，依靠大数据可以提供足够有利的资源。

随着云时代的来临，大数据（Big data）也吸引了越来越多的关注。《著云台》的分析师团队认为，大数据（Big data）通常用来形容一个公司创造的大量非结构化和半结构化数据，这些数据在下载到关系数据库用于分析时会花费过多时间和金钱。大数据分析常和云计算联系到一起，因为实时的大型数据集分析需要像 Map Reduce 一样的框架来向数十、数百或甚至数千的计算机分配工作。

大数据分析相比于传统的数据仓库应用，具有数据量大、查询分析复杂等特点。

三、大数据的作用

大数据最主要的作用是服务，即面向人、机、物的服务。对机器来说，需要数据有一些关联，能够从中分析出有用的信息，非结构化、半结构化、结构化等。人、机、物对数据的贡献和参与度非常高，从数据规模上可看到，人到物理世界是从小到大，从数据质量来讲，人提供的数据质量是最高的。

四、大数据与传统数据库

传统数据库 / 数据仓库是 GB/TB 级高质量、较干净、强结构化、Top-down、重交易、确定解。大数据是 PB 级的，有噪声、有冗余、非结构化、Bottom-up、重交互、满意解。大数据出现后，NoSQL 模式变得非常流行。大数据引发了一些问题，如对数据库高并发读写要求、对海量数据的高效存储和访问需求、对数据库高可扩展性和高可用性的需求，传统 SQL 主要性能没有用武之地。互联网巨头对于 NoSQL 数据模式应用非常广泛，如谷歌的 Big Table、Facebook 的 Cassandra、Oracle 的 NoSQL 及亚马逊的 Dynamo 等。从大数据处理角度来看，Map Reduce 已成为事实的标准。大数据的存储和处理，已有了成熟解决方案，对于在系统软件中占较大比重的操作系统来

说没有太大变化，一些重要的命题还没有解决，例如，操作系统对新兴计算资源的直接抽象的调度（GPU、APU），分布式文件系统下的统一数据视图、全数据中心范围内的能耗管理、大数据下的安全性等，但还不成熟，需要研发。

五、大数据与 Web

大多数研究大数据的商业公司，都有明确的商业目的，即更好地支撑 Web 服务，如谷歌搜索引擎服务、Facebook SNS 网站、新浪微博网站等。在大数据驱动下的 Web 服务特征：更加流畅的网页交互体验，更加快速的社会资讯获取，更加便捷的日常工作和生活，更加深入的人、机、物融合。

回顾一下 Web 的发展，也是国际上比较通用的说法，Web 1.0 时代 Web 内容主要由网站服务商提供，Web 2.0 时代用户大量参与 Web 内容的贡献，像博客和微博。到了 Web 3.0 时代，特征就是人、机、物共同参与 Web 内容贡献，使 Web 形成对真实世界的全面映射。

大数据来源于人、机、物，同时服务于人、机、物，大数据时代系统软件，特别是操作系统有待进一步发展，人、机、物融合大数据将推动 Web 进入崭新 Web 3.0 时代。

第二节　数据挖掘形式与特点

一、数据挖掘概述

大数据对于数据挖掘既是挑战更是机遇。褪去了发展初期的浮躁与喧哗，数据挖掘在理论方法与软件工具上都有了长足的进步，并在诸多领域积累了成熟的应用案例，取得了扎实的应用成果。人们曾经将数据挖掘形象地比喻为从数据“矿石”中开采知识“黄金”的过程，如今面对数据的“矿山”，数据挖掘充分汲取机器学习、统计学、分布式和云计算等技术养分，在方法研究、算法效率、软件工具集成环境和创新应用等方面不断开拓，正将昔日的数据“矿锤”升级为现代化的数据“挖掘机”，成为大数据时代最有效的数据分析利器。所以，数据挖掘具有多学科综合性、方法性和工具性的特征。对此，初学者应具有较强的数据操作能力和学习领会能力，能够举一反三，触类旁通，边学边做，边做边学。

数据挖掘的发展过程是一个兼容并包的成长过程。一般来说，数据挖掘经历了两个主要发展阶段，从初期局限于数据库中的知识发现，发展到中期内涵不断丰富完善

以及多学科的融合发展，乃至今天成为大数据时代的关键分析技术，数据挖掘已经取得了实质性的跨越。

目前，对数据挖掘的理解已达成如下共识。

首先，数据挖掘是一个利用各种方法，从海量的有噪声的各类数据中，提取潜在的、可理解的、有价值的信息的过程。这里，信息可进一步划分为两大类：一类是用于数据预测的信息，另一类是用于揭示数据内在结构的信息。

其次，数据挖掘是一项涉及多任务、多学科的庞大的系统工程，涉及数据源的建立和管理、从数据源提取数据、数据预处理、数据可视化、建立模型和评价以及应用模型评估等诸多环节。

针对复杂问题且涉及海量数据的数据挖掘任务，往往是一项大规模的系统工程。为更加规范地开展数据挖掘工作，NCR、SPSS和Daimler-Benz三家公司联合制定了跨行业数据挖掘标准CRISP-DM，SAS公司也发布了相关数据挖掘标准SEMMA。这些标准希望对数据挖掘过程中各处理步骤的目标、内容、方法、应注意的问题等提出可操作性建议，从而帮助学习者从方法论的高度深入理解并掌握数据挖掘的一般规律。

数据挖掘的诸多环节本质上可归纳为两个具有内在联系的阶段：数据的存储管理阶段和数据的分析建模阶段，涉及计算机科学和统计学等众多交叉学科领域。

当前，数据挖掘的对象是大数据系统。大数据往往来自不同的采集渠道以及不同的数据源，数据量庞大且杂乱有噪声。如何高效合理地存储数据，如何有效地保障数据的一致性等，在数据挖掘中尤为重要，也始终是数据挖掘的难点，涉及计算机学科中的数据库和数据仓库计算、分布式计算、并行处理等多个研究领域。大数据的存储管理有两个层面：一个是基础设施层面，包括对存储设备、操作系统、数据库、数据仓库、分布式计算等方面的整体评估，需求的客观理解，系统架构、技术和产品的选择，稳定、高效的数据基础设施体系的建立等一系列问题；另一个是数据管理工具层面，包括数据的抽取检索、集成清洗，以及其他预处理的软件、技术和管理等诸多方面。数据的存储管理是数据分析的基础和保障，也在某种程度上为采用怎样的数据分析方法提供依据。

数据挖掘中的数据预处理、数据可视化、建立和评价模型等环节，其核心目标是发现数据中隐藏的规律性，这是统计学和从属计算机科学的机器学习以及具有跨学科（统计和计算机）特点的可视化研究的主要任务。事实上，从统计学视角看数据挖掘会发现，数据挖掘与统计学有着高度一致的目标——数据分析，正因如此，数据挖掘对统计学而言似乎并不陌生。然而，目标尽管一致，但仍提出数据挖掘概念的重要原因是：数据分析对象是大数据。大数据特征决定了数据处理需要多学科的共同参与，数据分析需要一种集中体现多学科方法和算法优势的理论和工具，这就是数据挖掘。

试读结束：需要全本请在线购买： www.ertongbook.com

二、数据挖掘基本特征

数据挖掘是一个从大数据中挖掘出有用信息的过程。数据挖掘结果具有不同的呈现方式，这些是数据挖掘结果外在的特征，而对于其内在内容，数据挖掘结果（有用信息）还具有以下三个重要特征：潜在性、可理解性和有价值性，如图 1-1 所示。

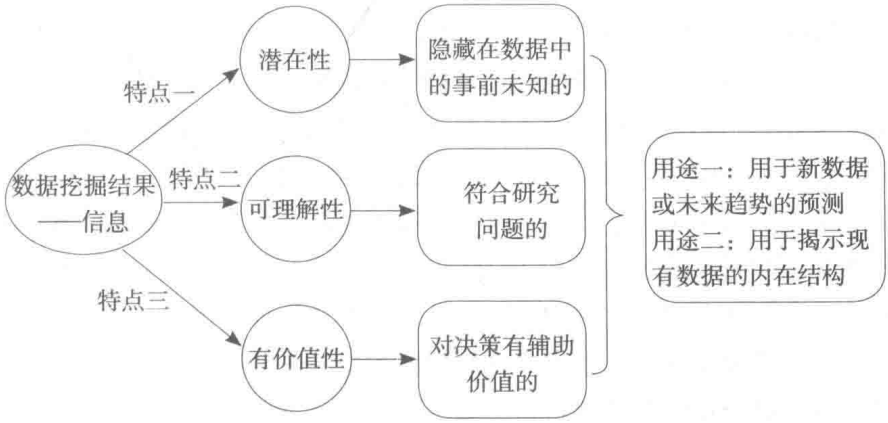

图 1-1　数据挖掘结果示意图

（一）潜在性

发现大量数据中隐含的变量相关性、数据内在结构特征等，是数据挖掘的重要任务，也是数据挖掘的核心成果。研究变量相关性以及数据内在结构特征是统计学的长项，其传统分析思路是：基于对研究问题的充分理解，依据经验或历史数据，首先预设数据中存在某种相关性假定，然后验证这种假定是否显著存在于当前数据中。这是一种典型的验证式分析思路。然而，大数据分析中的数据量庞大，变量个数多且类型复杂，以传统方式预设假定将非常困难，甚至不可能。所以数据挖掘通常会倾向采用一种归纳式的分析思路，即不事先对数据中是否存在某种关系做任何假定，而是通过“机械式”的反复搜索和优化计算，归纳出所有存在于数据中的规律。

这样的分析思路有优势也存在问题。优势在于它既可能找到隐藏于数据中的人们事先知道的规律，也可能发现那些人们事先未知的规律。存在的问题是由此得到的分析结果，一方面，可能是类似传说中“尿布和啤酒”的典型案例；另一方面，也可能是令人无法理解和没有价值的。

（二）可理解性

数据挖掘结果的可理解性是指分析结论具有符合研究问题的可解释性。例如，在消费者行为偏好的数据挖掘中，若分析结果是一段时间内顾客的消费金额与其身高有密切关系，那么这样的结论就不具有可解释性。事实上，数据挖掘揭示出的不可理解的相关

性，一部分可能是一种虚假相关，另一部分可能是因其他相关因素传递而导致的表象。

（三）有价值性

数据挖掘结果是否有价值体现在是否对决策有意义。对决策没有指导意义的结果是没有价值的。例如，在居民健康管理的数据挖掘中，若分析结论是 90% 的居民每日就餐次数是 3 次，且三餐的平均就餐时间是早上 7 点、中午 12 点、晚上 7 点。那么这种分析结论的价值很低，因为它是个常识。

谁是导致数据挖掘结果有可能无法理解和没有价值的“元凶”？答案是：海量大数据。事实上，发现海量大数据中隐藏的可理解的、有价值的信息，难度要远高于小数据集，会出现分析小数据集时不曾出现的诸多新问题。其中的一个主要问题就是“机械式的挖掘”给出的“信息”很可能只是数据的某种“表象”，而非“本质”。用统计术语讲就是，很可能并不是数据真实分布或关系的反映，而仅是海量数据自身的某种无意义的随机性的代表。

为此，人们试图借助统计学对“表象”和“本质”加以区分。作为数据挖掘成员的一分子，统计学确实在区分信息是系统性的本质还是随机性的表象上可见一斑。其通常的做法是：以分析数据是随机样本为前提，采用统计推断式的假设检验。统计推断以随机样本为研究对象，通过找到样本的某些特征并计算这些特征将以多大的概率出现在总体中，进而判断它们是否具有统计上的显著性，即这些特征是系统性的还是样本的随机性所致。事实上，数据挖掘发展初期也确实采纳了这种方式，所以某些数据挖掘方法貌似就是统计方法也很正常。但问题在于随着大数据的出现以及数据挖掘应用的不断拓展，这样的思路出现了如下主要问题。

第一，大数据的海量特性极大限制了上述分析思路的可行性。

若认为数据挖掘的数据对象是个样本，那么这个样本通常是大样本。对以小规模数据集为研究对象发展起来的统计推断而言，小样本表现出的某些特征，如果确实是由随机性导致的，那么在统计推断过程中，会因样本量小、在总体中出现的概率很小而被正确地确认为是随机性。这种分析思路在小数据集上是可行的，但在数据挖掘中的海量大样本集上就不再奏效。因为任何统计不显著的随机性都可能因样本量大而被倾向性地误断为显著，即误断为系统性的、有意义的，即使是“表象”也会被误判为“本质”。

第二，数据挖掘的研究对象往往是总体而非随机样本。

数据挖掘对象一般是现有数据集，它们通常就是人们关注的总体而不是样本。从这个角度来讲，统计推断不再必要。当然，数据挖掘并不否认统计推断的重要作用。若将现有数据放到一个更大的时空中去，那么目前数据这个总体也可以视为更大时空中的一个样本。但问题是能否确保样本是个随机样本，否则统计推断还会因丧失原本的理论基础而不再适用。

另外，有些数据挖掘应用问题只能基于总体而不能基于样本来研究。例如，在信用卡欺诈甄别研究中，若确实存在极少数的恶意透支行为，这些交易数据会因数量很小而不易或无法进入随机样本。若以样本为研究对象，样本中的欺诈特征会变得不再明显甚至消失，从而得到不存在欺诈行为的分析结论。

基于上述原因，数据挖掘不再以统计推断方式验证数据挖掘的结果是否有意义，而是采用一种“退而求其次”的做法，即强烈要求行业专家深度参与数据挖掘过程，并由行业专家负责判断数据挖掘结果的意义和价值。例如，“所有前列腺癌患者都是男性”，“加油站的信用卡刷卡金额通常在个位为零上出现峰值”，这些结论是否可理解和有价值，完全由行业专家去评估。

第三节　R 语言数据挖掘的应用

数据挖掘的应用极为广泛。易观智库以应用成熟度和市场吸引力作为两个维度，给出了当前数据挖掘的十大典型应用及其分布。

数据挖掘在电子商务领域的应用是最成熟和最具吸引力的，金融和电信行业紧随其后。政府公共服务领域的数据挖掘将有较大的发展潜力，其未来的应用成熟度将会有巨大的提升空间。

进一步，数据挖掘在电子商务中的应用价值主要体现在市场营销和个性化导购等方面。有效实现用户消费行为规律的分析，制订有针对性的商品推荐方案，根据用户特征研究广告投放策略并进行广告效果的跟踪和优化；金融行业中，数据挖掘主要应用于客户金融行为分析以及金融信用风险评估等方面；数据挖掘在电信企业的应用主要集中在客户消费感受等分析方面。目的是通过洞察客户需求，有针对性地提升网络服务的质量和安全；在政府公共服务中，数据挖掘的作用主要体现在智慧交通和智慧安防等方面，旨在实现以数据为驱动的政府公共服务；医疗行业的数据挖掘应用价值集中在药品研发、公共卫生管理、居民健康管理以及健康危险因素分析等方面。

尽管上述典型数据挖掘应用所解决的问题不同，但研究思路类似，且问题的切入也有很多共同点。若对上述各个应用问题分别展开论述，内容难免冗余、雷同。因此，这里仅对金融、电子商务、电信中的典型商业数据挖掘共性问题进行梳理并做详尽讨论。主要包括客户细分研究、客户流失预测、交叉销售、营销响应、欺诈甄别等方面。

一、数据挖掘在客户细分中的应用

客户细分的概念是美国著名营销学家温德尔·史密斯于 20 世纪 50 年代中期提出

的。客户细分是经营者在明确其发展战略、业务模式和市场条件下，依据客户价值、需求和偏好等诸多因素，将现有客户划分为不同的客户群，属于同一客户群的消费者具有较强的相似性，不同细分客户群间存在明显的差异性。

在经营者缺乏足够资源应对客户整体时，由于客户间价值和需求存在异质性，有效的客户细分能够辅助经营者准确认识不同客户群体的价值及需求，从而制定针对不同客户群的差异化的经营策略，以资源效益最大化、客户收益最大化为目标，合理分配资源，实现持续发展新客户、保持老客户、不断提升客户忠诚的总体目标。

客户细分的核心是选择恰当的细分变量、细分方法以及细分结果的评价和应用等方面。

（一）客户细分变量

客户细分的核心是选择恰当的细分变量。不同的细分变量可能得到完全不同的客户细分结果。传统的客户细分是基于诸如年龄、性别、婚姻状况、收入、职业、地理位置等的客户基本属性。此外，还有基于各种主题的，如基于客户价值贡献度、需求偏好、消费行为的客户细分等。

不同行业因其业务内容不同，客户价值、需求偏好以及消费行为的具体定义也不同。需选择迎合其分析目标的细分变量。例如，电信行业 4G 客户细分，主要细分变量可以包括使用的手机机龄、自动漫游业务、月平均使用天数、月平均消费额、月平均通话时间、月平均通话次数、月平均上网流量等。再例如，商业银行为研发对不同客户有针对性的金融产品和服务，对于金融客户个人主要关注年龄、家庭规模、受教育程度、居住条件、收入来源、融资记录等属性。对金融客户企业主要关注行业、企业组织形式、企业经营年限、雇员人数、总资产规模、月销售额、月利润等。同时，关注的贷款特征包括贷款期限、贷款用途、抵押物、保证人等；对于电子商务的客户细分，除关注其收入资产、职业特点、行业地位、关系背景等基本属性外，还需关注喜好风格、价格敏感、品牌倾向、消费方式等主观特征，以及交易记录、积分等级、退换投诉、好评传播等交易行为特征等。

能否选择恰当的细分变量，取决于对于业务需求的认知程度。不同领域的客户细分问题中，客户的“好坏”标准可能不同。随着业务的推进以及外部环境的动态变化，这个标准也可能随时发生变化。所以，确定客户细分变量应建立在明确当前的业务需求的基础之上。细分变量的个数应适中，以能否覆盖业务需求为准，同时各细分变量之间不应有较强的相关性。

（二）客户细分结果的评价和应用

客户细分的结果是多个客户群。在合理的客户群基础上制定有针对性的营销策略，才可能获得资源效益的最大化以及客户收益的最大化。客户群的划分是否合理，

一方面依赖于细分变量的选择，另一方面也依赖于所运用的细分方法。细分方法的核心是数据建模，而数据建模通常带有“纯粹和机械”的色彩。尽管它给出的客户群划分具有数理上的合理性，但并不一定都是迎合业务需求的。所以还需从业务角度评价细分结果的实际适用性。例如，各个客户群的主要特征是否具有业务上的可理解性；客户群所包含的人数是否足够大，能否足以收回对其营销的成本；客户群的营销方案是否具有实施上的便利性，等等。

二、数据挖掘在客户流失分析中的应用

客户流失是指客户终止与经营者的服务合同或转向其他经营者提供的服务。通常，客户流失有如下三种类型。

第一，企业内部的客户转移，即客户转移到本公司的其他业务上。例如，银行因增加新业务或调整费率等所引发客户的业务转移，如储蓄账户从活期存款转移至整存整取，理财账户从购买单一类信托产品转移到集合类信托产品等。企业内部的客户转移，就某个业务来看存在客户流失现象，可能对企业收入产生一定影响，但就企业整体而言，客户并没有流失。

第二，客户被动流失，即经营者主动与客户终止服务关系。例如，金融服务商由于客户欺诈等行为而主动终止与客户的关系。

第三，客户主动流失，包括两种情况：一种情况是客户因各种原因不再接受相关服务；另一种原因是客户终止当前服务而选择其他经营者的服务。例如：手机用户从中国联通转到中国移动。通常客户主动流失的主要原因是客户认为当前经营者无法提供所期望的价值服务，或希望尝试其他经营者所提供的新业务。

数据挖掘的客户流失分析主要针对上述第三种类型，是以客户基本属性和历史消费行为数据为基础，通过适当的数据挖掘方法而进行的各种量化建模。主要围绕以下两个目标。

（1）客户流失原因的分析，目的是为制订今后的客户保留方案提供依据。

即找到与客户流失高度相关的因素，如哪些特征是导致客户流失的主要特征，具有哪些属性值或消费行为的客户容易流失等。例如，抵押放款公司需了解具有哪些特征的客户，会因为竞争对手采用低息和较宽松条款而流失；保险公司需了解取消保单的客户通常有怎样的特征或行为。只有找到客户流失的原因，才可能依此评估流失客户对经营者的价值，分析诸如哪类流失客户会给企业收入造成严重影响，哪类会影响企业的业务拓展，哪类会给企业带来人际关系上的损失。客户流失原因分析的核心目的是为制订今后的客户保留方案提供依据。

（2）客户流失的预测。目的是为测算避免流失所付出的维护成本提供依据。

客户流失预测主要有以下两个方面。

第一，预测现有客户中哪些客户流失的可能性较高，给出一个流失概率由高到低的排序列表。由于对所有客户实施保留的成本很高，只对高流失概率客户开展维护，将大大降低维护成本。对流失概率较高的客户，此时还需进一步关注他们的财务特征，分析可能导致其流失的主要原因是财务的还是非财务的。通常非财务原因流失的客户是高价值客户，这类人群一般正常支付服务费用并对市场活动做出响应，是经营者真正需要保留的客户。给出流失概率列表的核心目的是为测算避免流失所付出的维护成本提供依据。

第二，预测客户可能在多长时间内流失。如果说上述第一方面是预测客户在怎样的情况下将流失，这里的分析是预测客户在什么时候将会流失。

统计学中的生存分析可有效解决上述问题。生存分析以客户流失时间为响应变量建模，以客户的人口统计学特征和行为特征为解释变量，计算每个客户的初始生存率。客户生存率会随时间和客户行为的变化而变化，当生存率达到一定的阈值后，客户就可能流失。生存分析一般不纳入数据挖掘的范畴。

三、数据挖掘在营销响应分析中的应用

为发展新客户和推广新产品，企业经营者通常需要针对潜在客户开展有效的营销活动。在有效控制营销成本的前提下，了解哪些客户会对某种产品或服务宣传做出响应等，是提高营销活动投资回报率的关键，也是营销响应分析的核心内容。

营销响应分析的首要目标是确定目标客户，即营销对象。对正确的目标客户进行营销，是获得较高客户响应概率的前提。因营销通常涉及发展新客户和推广新产品两方面，所以营销响应分析中的关注点也略有差异。

（一）发展新客户

在推广新客户的过程中，可以根据已有的现实客户数据，分析其属性特征。通常具有相同或类似属性特征的很可能是企业的潜在客户，应视为本次营销的目标客户。

（二）推广新产品

在推广新产品的过程中，若新产品是老产品的更新换代，或与老产品有较大相似度，则可通过分析购买老产品的客户数据，发现他们的属性特征。通常可视这类现实客户为本次营销的目标客户，同时具有相同或类似属性特征的潜在客户也可视为本次营销的目标客户，他们很可能对新产品感兴趣。

若新产品是全新的，尚无可供参考的市场和营销数据，可首先依据经验和主观判断确定目标客户的范围，并随机对其进行小规模的试验性的营销。然后，依据所获得的营销数据，找到对营销做出响应的客户属性特征。具有相同或类似属性特征的现实客户和潜在客户，通常可视为本次营销的目标客户。

试读结束：需要全本请在线购买： www.ertongbook.com