数据分析与决策技术丛书

Game Data Analysis and Mining

R语言游戏数据分析与挖掘

谢佳标◎著

乐逗游戏高级数据分析师撰写，资深R语言技术工程师近10年数据分析与挖掘的经验总结

以解决游戏行业的具体问题为目标，技术和业务双重导向，系统阐述游戏数据分析与挖掘的技术、方法论和工具，以及对游戏业务的理解与思考

机械工业出版社
China Machine Press

数据分析与决策
技术丛书

R语言游戏数据分析与挖掘

谢佳标◎著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

R 语言游戏数据分析与挖掘 / 谢佳标著 . —北京：机械工业出版社，2017.6
（数据分析与决策技术丛书）

ISBN 978-7-111-57308-1

I. R… II. 谢… III. 程序语言－程序设计 IV. TP312

中国版本图书馆 CIP 数据核字（2017）第 141146 号

R 语言游戏数据分析与挖掘

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：殷 虹
印 刷：北京诚信伟业印刷有限公司
版 次：2017 年 7 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16
印 张：25.75
书 号：ISBN 978-7-111-57308-1
定 价：79.00 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
购书热线：（010）68326294 88379649 68995259
投稿热线：（010）88379604
读者信箱：hzit@hzbook.com

华章IT
HZBOOKS | Information Technology

Preface 前　言

为什么要写这本书

随着大数据的概念越来越流行，越来越多的企业开始重视数据，期待从数据中寻找有价值的结论，以指导公司管理层决策，最终创造更大的价值。但是在游戏行业，数据分析的发展相对缓慢，很多游戏公司是在发现人口红利消失后才逐渐重视数据，希望利用数据驱动产品。而在各种数据分析技术中，R 语言作为一个可进行交互式数据分析和探索的强大平台，拥有举足轻重的作用。R 语言的免费开源使得很多公司用它来处理数据、展示数据、分析数据、完成模型。

使用 R 语言可以进行游戏数据分析系统的搭建，可以对累积的海量游戏数据进行挖掘，找出其中的特征和规律。对于有志成为互联网数据挖掘 / 分析师的读者来说，R 语言将成为他们未来必备的技能之一。

笔者在历届中国 R 语言会议演讲时，都会遇到一些同学问类似这样的问题：“是否学好数据挖掘工具就能胜任数据分析工作？”虽然这些学生都具备很好的理论和工具使用能力，但是缺乏对实际生产数据的处理能力，即学生们很少接触到企业的真实数据，不知道如何将脏数据处理为可以建模的数据集。这也是笔者写这本书的初衷。在本书中，笔者希望结合自己多年的数据挖掘实战经验，将 R 语言与游戏数据分析有机结合，真正做到“授之以渔”。

本书特色

本书从实际应用出发，结合实例及应用场景，通过对大量案例进行详细阐述和深入分析，进而指导读者在实际工作中通过 R 语言对游戏数据进行分析和挖掘。

本书的核心是游戏数据分析实战，所以在案例讲解过程中均会对分析结果进行业务解读，进而帮助数据分析师提高“利用结果数据指导实际商务决策”的能力。

基于对业务的思考，本书从解决问题入手，以游戏为最佳切入点，辐射整个数据分析

领域，并完成数据分析和挖掘建模工作，对其他行业的数据分析师如何做数据分析 / 挖掘也具有很大的启发性。同时，本书内容涵盖了 R 语言基础、数据挖掘理论与实战、交互式绘图和 Web 网页开发等，故也可以作为数据挖掘的入门书籍。

本书适用对象

- 游戏产品运营人员
- 游戏数据分析人员
- 各行各业的数据分析师
- 数据分析爱好者
- 具有数据分析背景的数据科学家
- 进行数据挖掘应用研究的科研人员
- 相关专业的在校生

如何阅读本书

全书一共 13 章，分为三篇：基础篇、实战篇和提高篇。基础篇介绍了游戏数据分析的基本理论知识、R 语言的安装与使用、R 语言中的数据结构、常用操作和绘图功能。实战篇主要介绍了游戏数据的预处理、常用分析方法、玩家路径分析和用户分析。提高篇介绍了 R 语言图形界面工具 Rattle 和 Web 开发框架 shiny 包。

第一篇是基础篇（第 1～4 章）：第 1 章主要介绍了游戏数据分析的必要性和流程；第 2 章讲解了 R 语言和 RStudio 的安装及使用方法，并对数据对象和数据导入进行了介绍；第 3 章介绍了 R 语言绘图基础，包括常用图形参数设置、低级绘图函数和高级绘图函数；第 4 章介绍了 lattice 和 ggplot2 绘图包，并详细介绍了一些基于 R 语言可用于生成交互式图形的软件包，包括 rCharts、recharts、rbokeh、plotly 等。

第二篇是实战篇（第 5～11 章）：第 5 章介绍了游戏数据预处理常用的手段，包括数据抽样、数据清洗、数据转换和数据哑变量处理；第 6 章介绍了游戏数据分析的常用方法，包括指标数据可视化、游戏数据趋势分析、游戏数据相关性分析和游戏数据中的降维技术；第 7 章介绍了事件点击行为常用的漏斗分析和路径分析；第 8 章介绍了留存指标的计算、留存率计算与预测、常用分类算法原理和模型评估；第 9 章介绍了常用用户指标计算、LTV 计算与预测、用户物品购买关联分析、基于用户物品购买智能推荐和社会网络分析；第 10 章介绍了渠道数据分析的必要性和对渠道用户进行质量评级；第 11 章介绍了常用收入指标计算、利用用户活跃度衡量游戏经济状况、RFM 模型研究。

第三篇是提高篇（第 12～13 章）：第 12 章介绍了 R 语言的图形界面工具 Rattle，该工具能够在图形化的界面上完成数据导入、数据探索、数据可视化、数据建模和模型评估

整个数据挖掘流程；第 13 章介绍了 Web 开发框架 shiny 包，使得 R 的使用者不必太了解 CSS、JS，只需要了解一些 HTML 的知识就可以快速完成 Web 开发。

勘误和支持

由于笔者的水平有限，书中难免会出现一些错误或者不准确的地方，恳请读者批评指正。你可以把意见或建议直接发至我的邮箱（jiabiao1602@163.com）。如果你有什么问题，也可以发邮件来提问，我将尽力为读者提供满意的解答，期待你们的反馈。书中全部数据及源代码都可以从 GitHub 网站（登录网站 https://github.com/jiabiao 1602/Game_DataMining_With_R 或扫描下方二维码）进行下载。

致谢

首先，感谢乐逗游戏 CEO 陈湘宇的支持，让笔者能把这几年在游戏行业中的一些数据挖掘实战写进本书，使读者能完整地看到如何对原始的数据源进行清洗转换以达到建模需求。书中介绍了对游戏行业付费用户行为研究的几种模型算法，相信对其他行业进行付费用户挖掘分析也可以起到很好地借鉴作用。

其次，感谢机械工业出版社华章公司副总编杨福川的信任，同时，也要感谢编辑李艺审阅本书的全部章节，有了他们的支持、鼓励和帮助，本书才能得以顺利出版。

最后，感谢家人，感谢你们一直以来的理解、陪伴和支持。

谨以此书献给我最亲爱的家人以及众多 R 语言的爱好者和数据分析师们！

目　录 *Contents*

第二篇 实战篇

第一篇 Part 1

基 础 篇

Chapter 1 第 1 章

什么是游戏数据分析

随着游戏市场竞争的日趋激烈，在如何获得更大收益延长游戏周期的问题上，越来越多的手机游戏开发公司开始选择借助大数据，以便挖掘更多更细的用户群来进行精细化、个性化的运营。数据分析重要的不是提供历史和现状，而是通过分析发现手机游戏现状，以及对未来进行预测。一切以数据出发，用数据说话，让数据更好地指导运营服务好玩家，对玩家的行为和体验不断进行分析和调整，使玩家可以在虚拟世界中得到各方面的满足。要实现这个目的，需要搭建专业的数据化运营团队。此外，游戏数据分析与其他行业的数据分析不同的是，游戏综合了经济、广告、社交、心理等方面的内容，这就对数据分析师提出了更高的要求。

1.1 为什么要对游戏进行分析

伴随着游戏互联网的快速发展和智能终端的普及，移动游戏进入了全民时代。越来越多的玩家利用碎片化时间进行游戏，使得游戏数据呈现井喷式增长，同时也对数据存储技术、计算能力、数据分析手段提出了更高的要求。海量数据的存储是必须面对的第一个挑战，随着分布式技术的逐渐成熟，越来越多的互联网企业采用分布式的服务器集群＋分布式存储的海量存储器进行数据的存储和计算，从而解决数据存储和计算能力不足的问题。如何在海量的、复杂高维的游戏数据中发掘出有价值的知识，将是很多公司下一步亟待解决的难题。

虽然积累了海量的玩家数据，很多公司也开发了自己的 BI 报表系统，但是多数停留在“看数据”阶段，还是用传统的数据分析方法对数据进行简单的加工、统计及展示，并没有

进行深度挖掘发现数据背后的规律和把握未来趋势。正是在这样的大背景下，游戏数据分析逐渐在游戏行业中变得重要。公司需要从传统的粗放型运营进化到精细化运营，从而了解如何有效地获取用户、评估效果；如何激活用户、评估产品质量；如何提升收益，并挖掘潜在的高价值用户。要满足精细化运营的需求，数据化运营就应运而生了。数据化运营就是在以海量数据的存储、分析、挖掘和应用的核心技术支持的基础上，通过可量化、可细分、可预测等一系列精细化的方式来进行的。

数据化运营是飞速发展的数据存储技术、数据挖掘技术等诸多先进数据技术直接推动的结果。数据技术的飞速发展，使数据存储成本大大减低，同时提供了成熟的数据挖掘算法和工具让公司可以去尝试海量数据的分析、挖掘、提炼和应用。有了数据分析、数据挖掘的强有力支持，运营不再靠“拍脑袋”，可以真正做到运营过程自始至终都心中有数。比如，在玩家的细分推送中，数据分析师利用数据挖掘手段对玩家进行分群，运营根据不同的用户群制定差异化策略，数据分析师再根据推送效果进行评估。

1.2　游戏数据分析的流程

游戏数据分析、数据挖掘的价值一定要落实到具体的业务应用中才可以得到检验和实现，所以需要流程和制度来有效保障最终的业务实践效果。这些流程一方面可以促使各相关方在数据分析业务实践的不同阶段落实各自的角色、分工和价值，维护整个业务流的畅通和效率；另一方面可以有效达成数据分析项目中各环节的阶段性目标。

游戏数据分析整体流程可以参考跨行业的数据挖掘标准流程 CRISP-DM 方法论，它是一种业界认可的用于指导数据挖掘工作的方法。按照 CRISP-DM 方法论，一个游戏数据分析的完整流程包括 6 个阶段，分别是业务理解、数据理解、数据准备、建立模型、模型评估和模型发布。这 6 个阶段的顺序并不是固定不变的，在不同的业务场景中，可以有不同的流转方向。但是总体来说，业务理解是第一位的，是游戏数据分析流程中的第 1 环节，制定了业务目标后，就可以针对业务目标进行数据收集、数据清洗、数据转换、数据建模及模型评估等流程。

图 1-1 是 CRISP-DM 方法论的示意图。它的外圈象征游戏数据分析自身的循环本质，数据分析过程可以不断循环、优化，后续的过程可以从前面的过程中得到借鉴和启发。

- 业务理解：该阶段的核心内容包括正确理解业务背景和业务需求，同时能把业务需求有效转化成合理的分析需求，并设计指标体系和拟定实施计划。例如，业务有关于核心用户画像的需求，但是由于休闲游戏的大多数玩家是游客登录，并没有注册，所以无法收集到玩家的基础属性信息，从而不能帮助业务对核心用户的性别、年龄、职业等属性进行画像，此时可以从游戏活跃、付费、行为等角度进行核心用户画像，将业务的需求转换成目前数据能支撑到的分析需求。

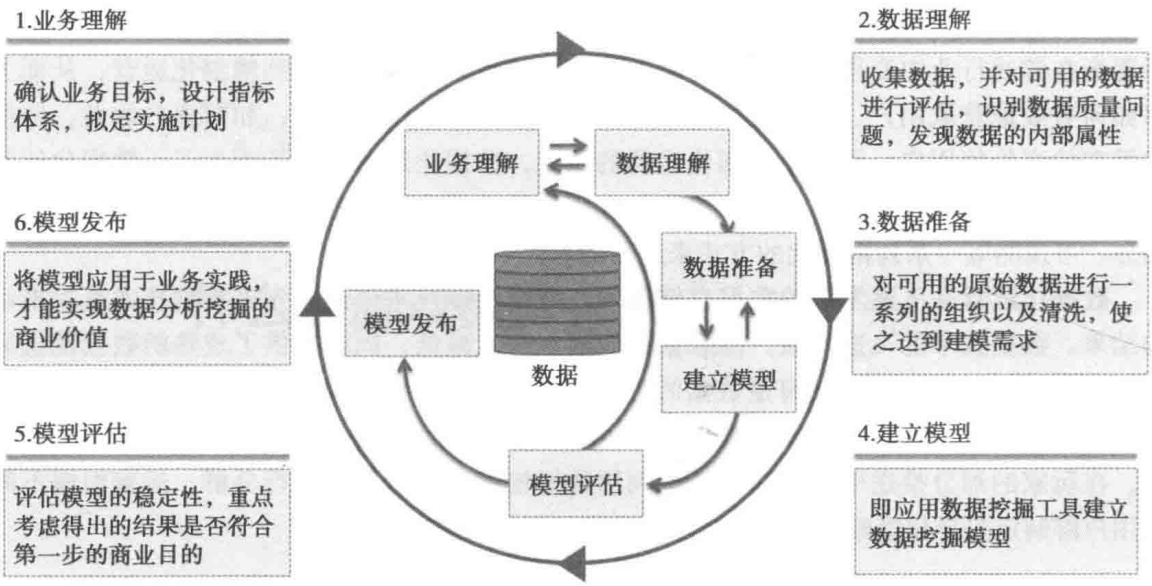

图 1-1 CRISP-DM 方法论示意图

- 数据理解：该阶段从数据收集开始，并对可用的数据进行数据探索和评估，识别数据质量问题，发现数据不同属性间的关系。
- 数据准备：这个阶段主要是做数据清洗和转换工作，包含数据缺失值和异常值的处理，保证建模前的数据质量；数据的重组、转换以及衍生等处理，比如对数据进行标准化处理、对某些指标进行分箱操作以便达到建模需求。
- 建立模型：这是游戏数据分析流程中技术含量最高的阶段，数据分析师应该根据项目需求和数据特点选择合适的算法，并使用专业的数据挖掘工具建立模型。
- 模型评估：评估建立模型的稳定性和有效性，常用的模型评估方法有混淆矩阵、ROC 曲线、K-S 曲线、交叉验证等。根据评估结果判断是否满足当初的业务需求，如果模型未满足需求，需要重复上一阶段的建模工作，有时甚至需要从数据收集阶段重新开始。如果现有数据不能满足分析需求，就需要业务和开发人员参与，在游戏中重新埋点收集数据。
- 模型发布：将模型应用于业务实践，才能实现数据分析挖掘的商业价值。根据业务反馈的结果，进而调整分析方法。

1.3 数据分析师的能力要求

因为在数据化运营中，数据分析师要深入业务背景，倾听和发现业务需求，走到业务第一线，与业务团队并肩作战，所以要求数据分析师具备很强的组织协调能力，具有项目大局观，懂得在不同阶段调用不同的资源。从这点来看，业务理解力和沟通能力的重要性

甚至要超过技术层面的能力（数据处理能力、数据统计分析能力、数据挖掘能力、数据应用能力）。图 1-2 是游戏数据分析师需要具备的关键能力。

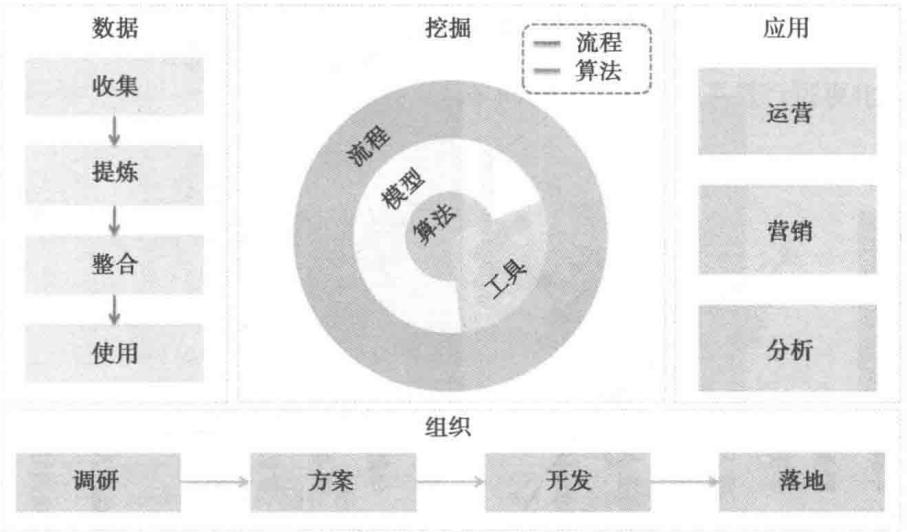

图 1-2　数据分析师关键能力示意图

首先数据分析师要具备组织能力。这体现在项目前期调研、方案制定、项目开发和项目落地的职责和能力要求。

- 调研：深入业务背景，发现、倾听业务需求。
- 方案：通过前期调研，有效判别分析需求价值，根据需求能有效提供分析解决方案。
- 开发：针对制定的解决方案，能通过技术手段进行项目开发。
- 落地：将开发成果结合业务场景进行落地，并持续跟踪落地应用效果，修正或优化方案和模型。

数据处理能力、数据挖掘能力和数据应用能力这三大块能力需要数据分析师通过时间、项目经验去磨砺，不断成长，懂得何时运用哪种数据挖掘技术解决相应的问题。

1.3.1　数据处理能力

刚刚收集上来的 raw data（原始数据）一般存在脏数据，不能达到直接建模的要求。我们不能直接利用 raw data 进行数据分析建模，所谓“垃圾进垃圾出”，这样得到的分析结果也不一定是可靠的。对于 raw data，我们需要评估数据质量，清洗脏数据，通常包括缺失值和异常值的处理，使之达到数据分析的需求。假如现在有一份 30 万的用户调研数据，由于某些玩家不愿意填写自己的性别、收入等，导致这些变量存在数据缺失的情况。现在利用数据分析技术对缺失值模式进行可视化探索，如图 1-3 所示。

由图 1-3 可知，有 2 万位玩家没有填写性别信息，其中有 609 位玩家同时缺失性别、年龄信息，31 位玩家同时缺失性别、年龄和收入的信息。掌握了数据缺失模式后，就知道

应该运用何种技术处理这些缺失值。

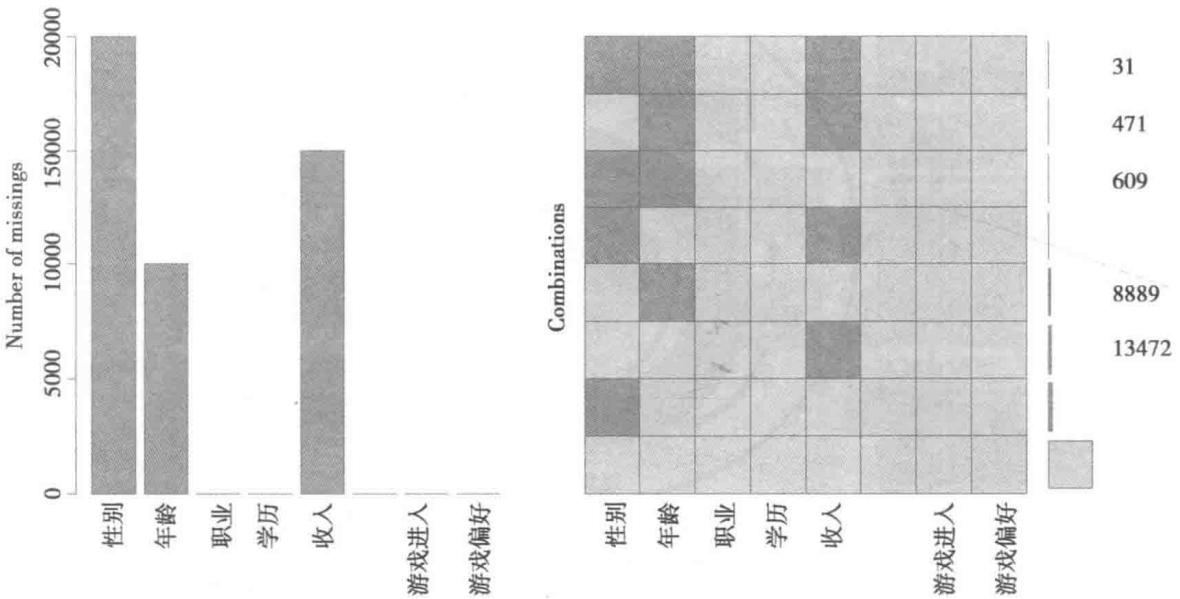

图 1-3 对数据调研数据进行缺失值可视化

针对异常值数据，我们同样希望能通过科学的方式甄别异常值并处理。例如，可以利用箱线图发现异常值，并在图上打印出异常值的样本号和数值，直观地对异常值进行可视化展示。比如现在有某个月日新增用户在第 30 日留存率的数据，通过普通曲线图很难发现是否有某些天的新增在第 30 日留存存在异常情况。此时可以借助箱线图的方式甄别异常值，如图 1-4 所示。

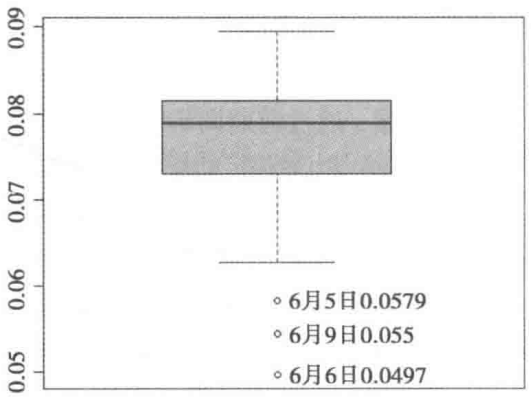

图 1-4 利用箱线图甄别异常值

由图 1-4 可知，这个月有三天的新增用户在第 30 日留存率低于正常水平，分别是 5 日、6 日和 9 日。

进行数据清洗后，有时候还需要对数据进行数据整合转换，使之符合建模前的数据需求，常用的一种方式是添加衍生变量。所谓衍生变量，其实就是指数据分析师在分析（建模）过程中人为增添的一些新变量，这些新变量产生之后，可以明显提升模型的效果，或者可以有效提炼出有价值的分析结论。

1.3.2 数据挖掘能力

数据分析师在建模的过程中，需要根据业务需求和数据特点选择合适的算法，利用专业的数据挖掘工具进行建模，并评估模型效果。比如在面对用户分析的需求时，可分别