O'REILLY®

TURING 图灵程序设计丛书

数据科学实战

DOING DATA SCIENCE

[美] Rachel Schutt Cathy O'Neil 著
冯凌秉 王群锋 译

中国工信出版集团 人民邮电出版社 POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

数据科学实战

DOING DATA SCIENCE

[美] Rachel Schutt Cathy O'Neil 著
冯凌秉 王群锋 译

O'REILLY®
Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo
O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北 京

图书在版编目（CIP）数据

数据科学实战 /（美）舒特（Schutt,R.），（美）奥尼尔（O'Neil,C.）著；冯凌秉，王群锋译. -- 北京：人民邮电出版社，2015.3
（图灵程序设计丛书）
ISBN 978-7-115-38349-5

Ⅰ. ①数… Ⅱ. ①舒… ②奥… ③冯… ④王… Ⅲ. ①数据管理 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2015)第013752号

内 容 提 要

本书脱胎于哥伦比亚大学“数据科学导论”课程的教学讲义，它界定了数据科学的研究范畴，是一本注重人文精神，多角度、全方位、深入介绍数据科学的实用指南，堪称大数据时代的实战宝典。本书旨在让读者能够举一反三地解决重要问题，内容包括：数据科学及工作流程、统计模型与机器学习算法、信息提取与统计变量创建、数据可视化与社交网络、预测模型与因果分析、数据预处理与工程方法。另外，本书还将带领读者展望数据科学未来的发展。

本书适合所有希望通过数据分析解决问题的人阅读参考，包括数据科学家、金融工程师、统计学家、物理学家、学生及其他对数据科学感兴趣的人。

◆ 著 [美] Rachel Schutt Cathy O'Neil
译 冯凌秉 王群锋
责任编辑 李松峰 毛倩倩
执行编辑 周宇宁
责任印制 杨林杰

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号
邮编 100164 电子邮件 315@ptpress.com.cn
网址：http://www.ptpress.com.cn
三河市海波印务有限公司印刷

◆ 开本：800×1000 1/16
印张：19.75 彩插：8
字数：487千字 2015年3月第1版
印数：1-4 000册 2015年3月河北第1次印刷
著作权合同登记号 图字：01-2014-5612号

定价：79.00元

读者服务热线：(010)51095186转600 印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京崇工商广字第 0021 号

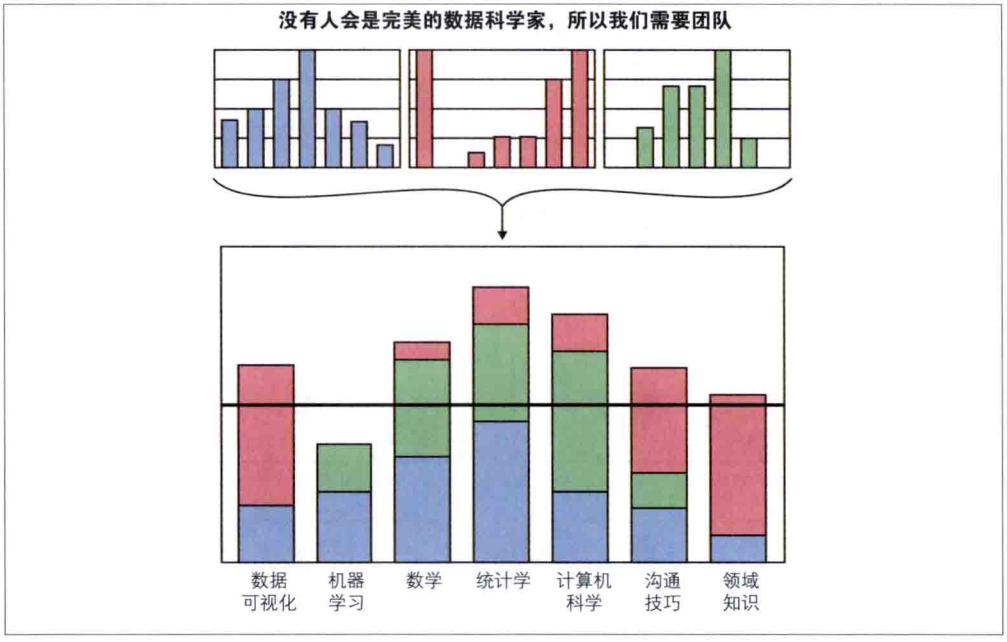

图 1-3：数据科学团队的知识结构由每个成员的知识结构叠加而来，在组建团队时，要让团队技能与所解决的问题大致匹配

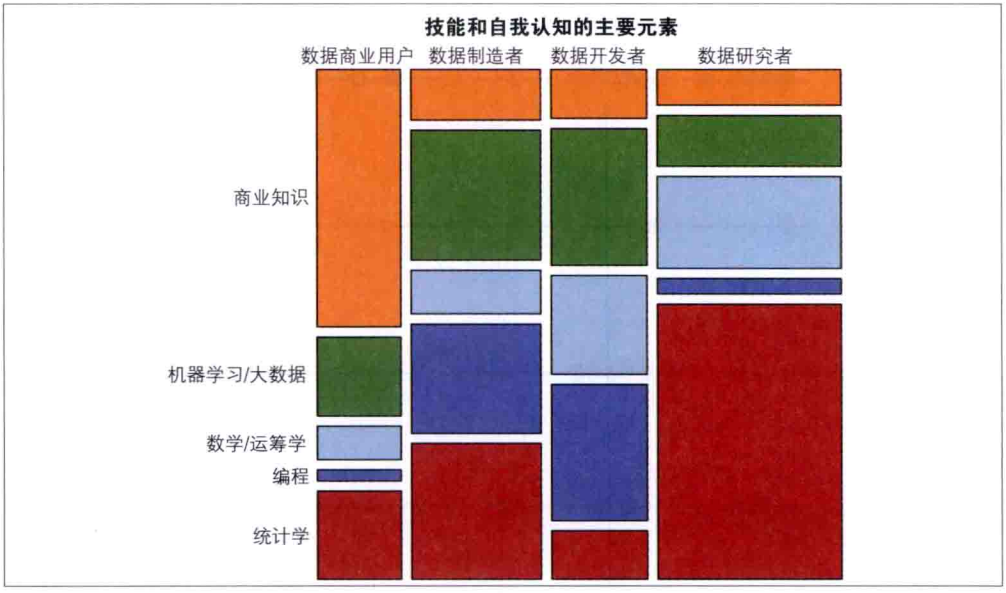

图 1-4：此图使用聚类算法描述了数据科学的子领域，源自 Harlan Harris、Sean Murphy 和 Marck Vaisman 基于 2012 年年中对数百名数据科学从业者的调查所著的 *Analyzing the Analyzers*（O'Reilly）

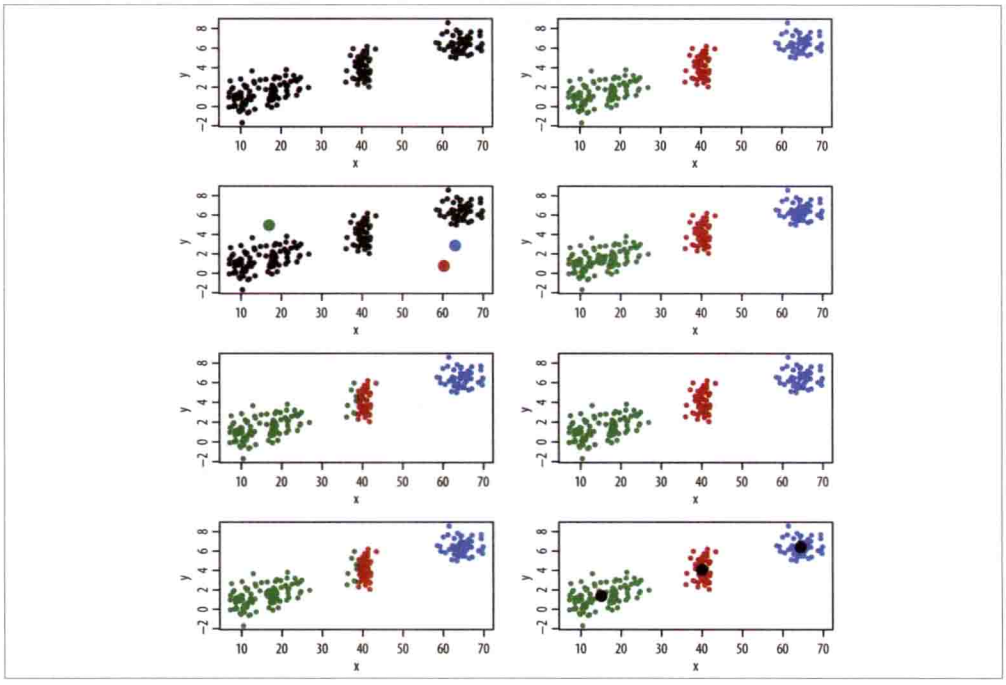

图 3-9：二维空间上的聚类过程，先看左半边从上往下，再看右半边从上往下

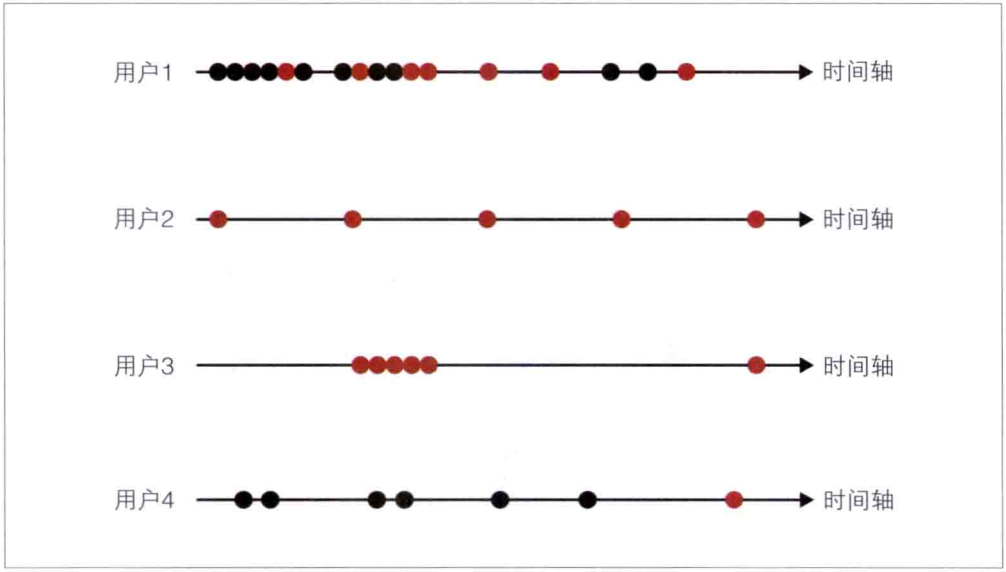

图 6-3：在用户的时间序列图中，用不用的颜色代表用户不同的动作类型。红色表示“点赞”，黑色表示“点衰”

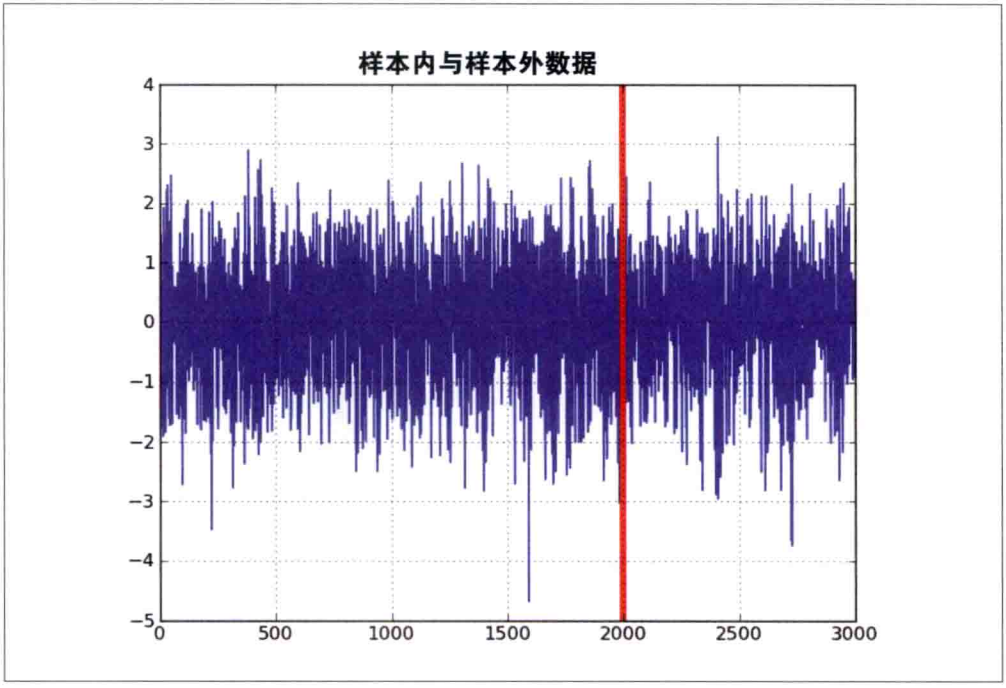

图 6-7：样本期内的数据永远发生在样本期外数据之前，红色线代表了模型建立的时点

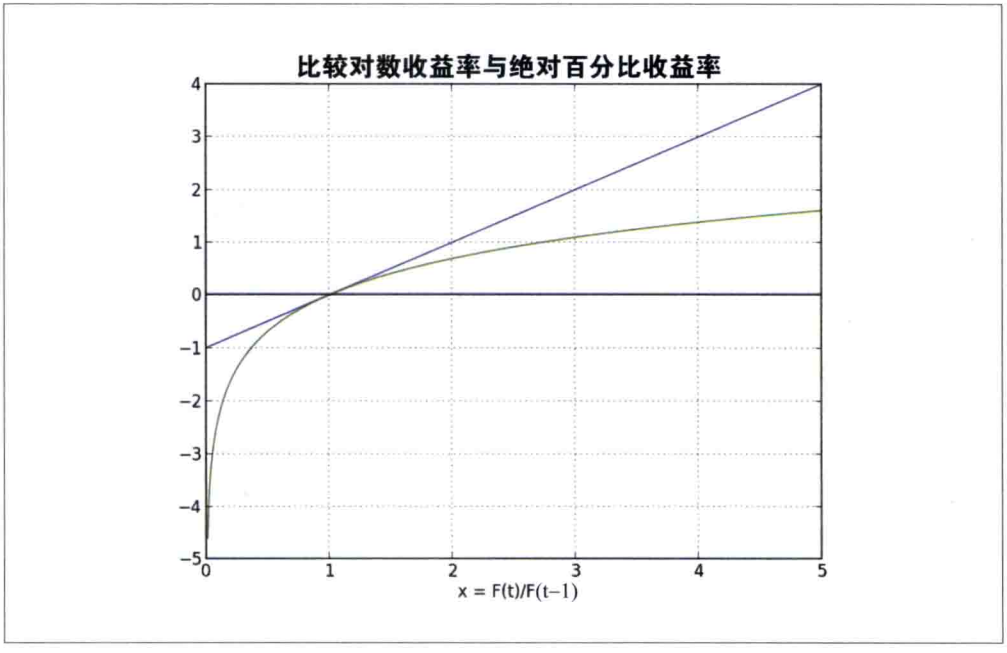

图 6-8：对数和绝对百分比收益率曲线对比图

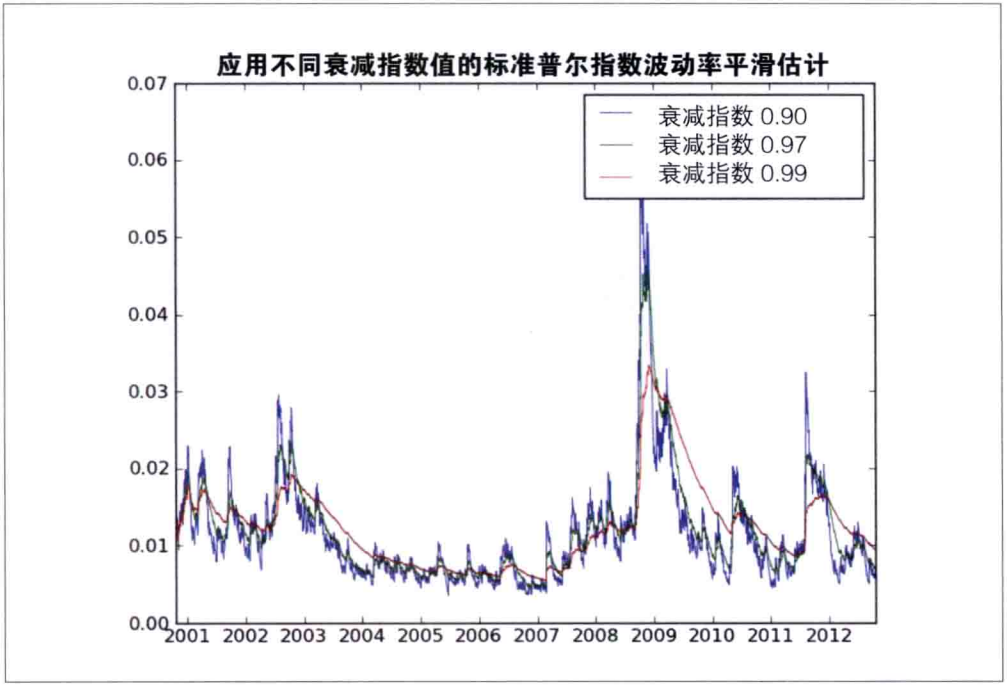

图 6-12：标准普尔指数的波动率的指数平滑估计：使用了三个不同大小的指数值

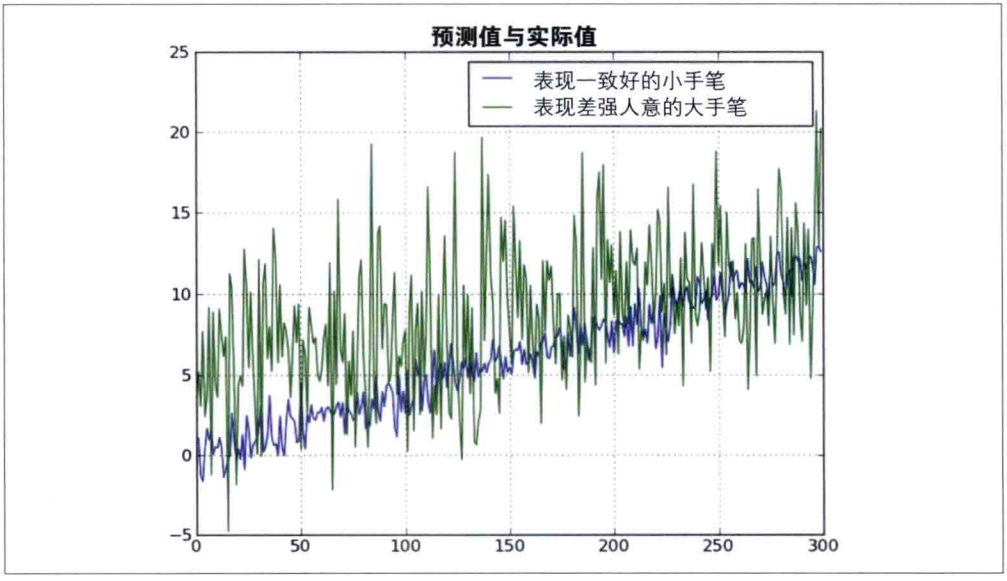

图 6-13：两个理论模型的累积 PnL 值对比图

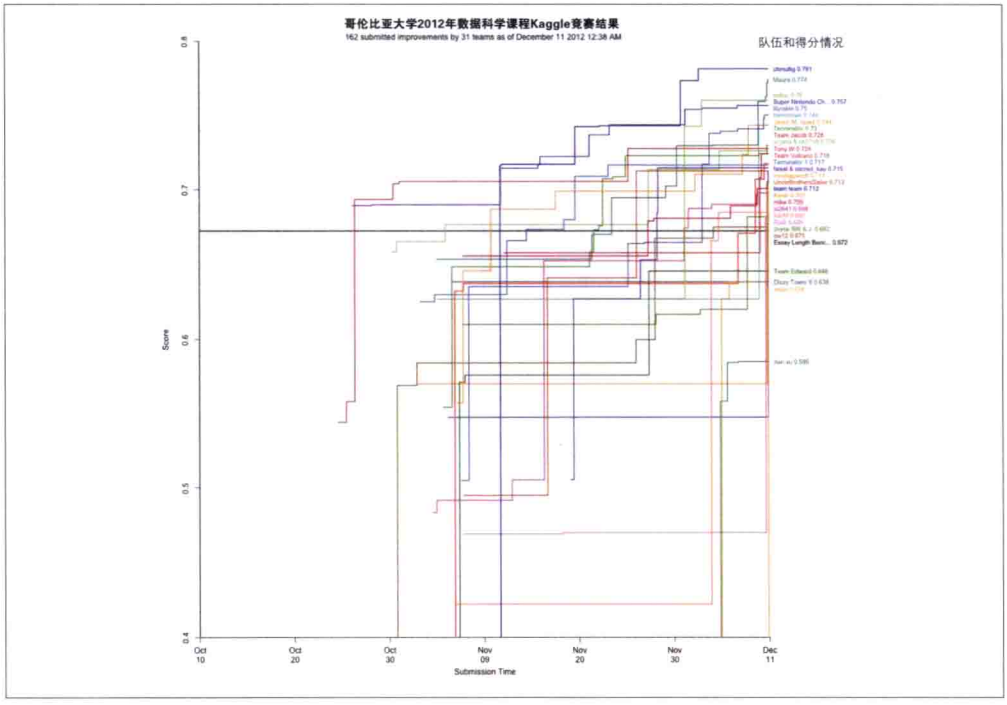

图 7-1：该图出自 Chris Mulligan，他是 Rachel 班里的学生。该图很好地描述了每个参赛个人 / 队伍在比赛期间，模型的进化情况

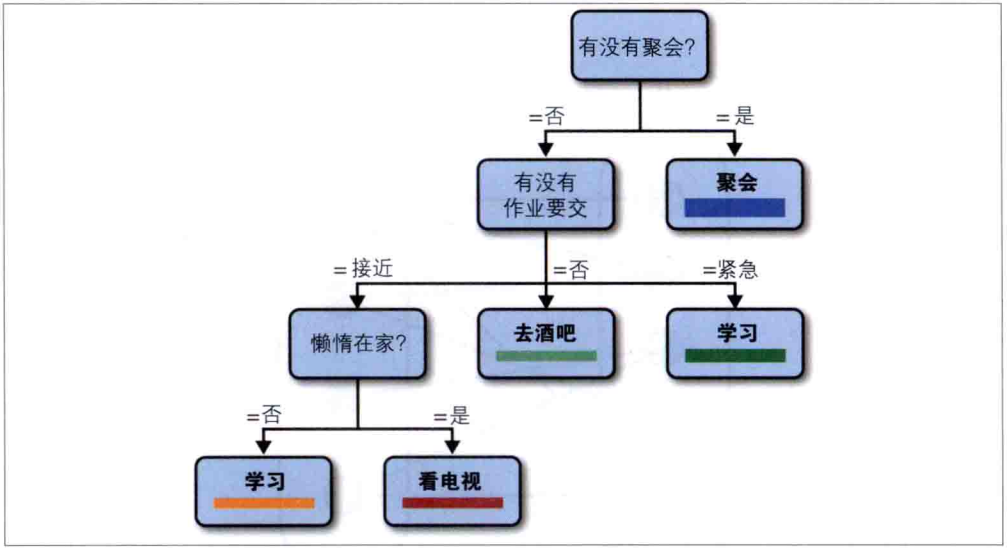

图 7-3：一个大学生在解决自己的时间分配问题时用到的决策树（原图摘自 Stephen Marsland 的著作 *Machine Learning: An Algorithmic Perspective*（《基于算法的机器学习》，Chapman and Hall/CRC），并获得了作者的许可

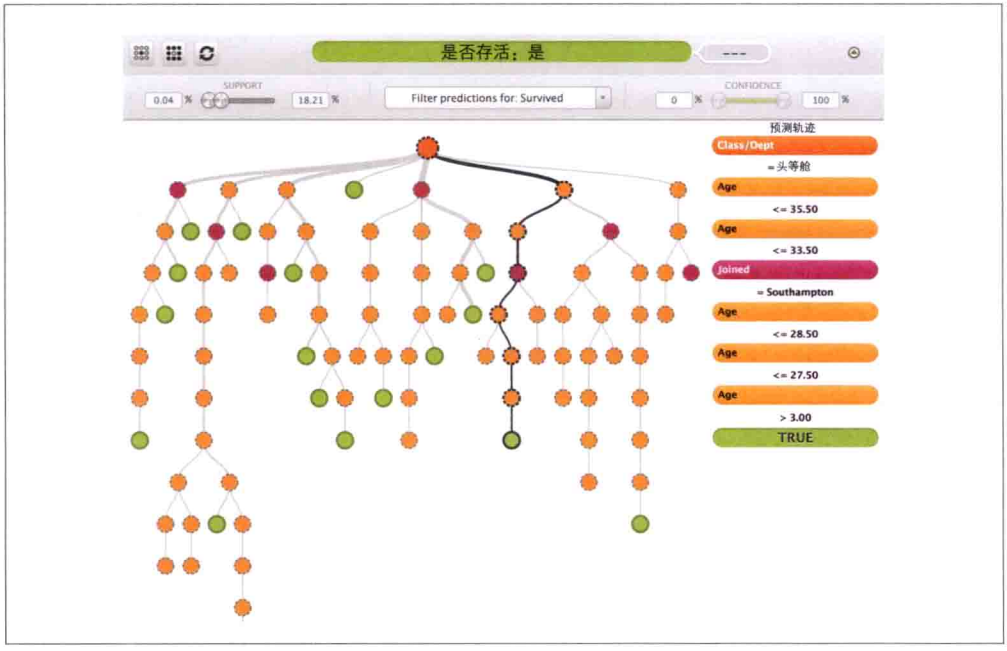

图 7-6：泰坦尼克号乘客生存模型的决策树图

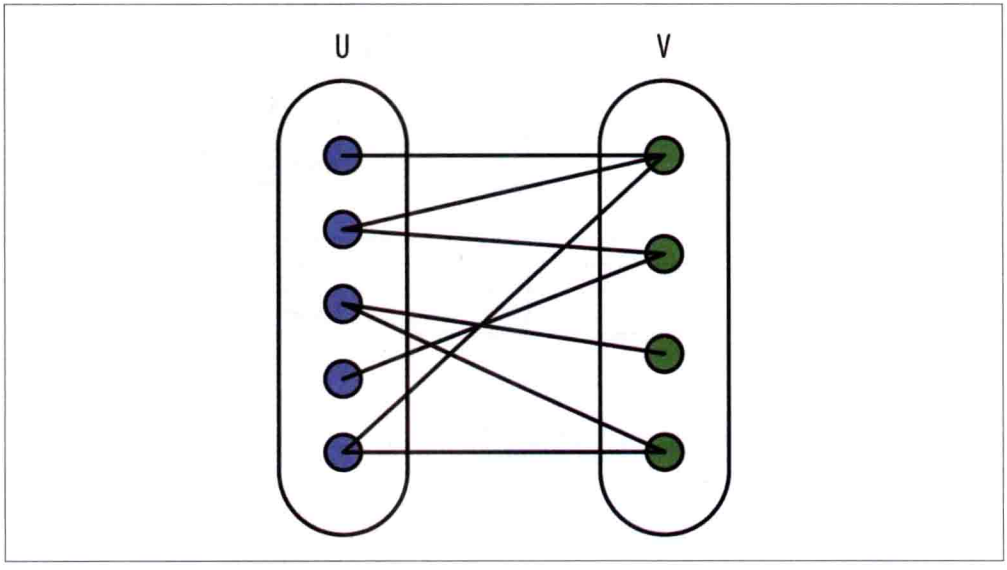

图 8-1：推荐系统的二分图：左侧是用户，右侧是推荐的项目，比如电视节目

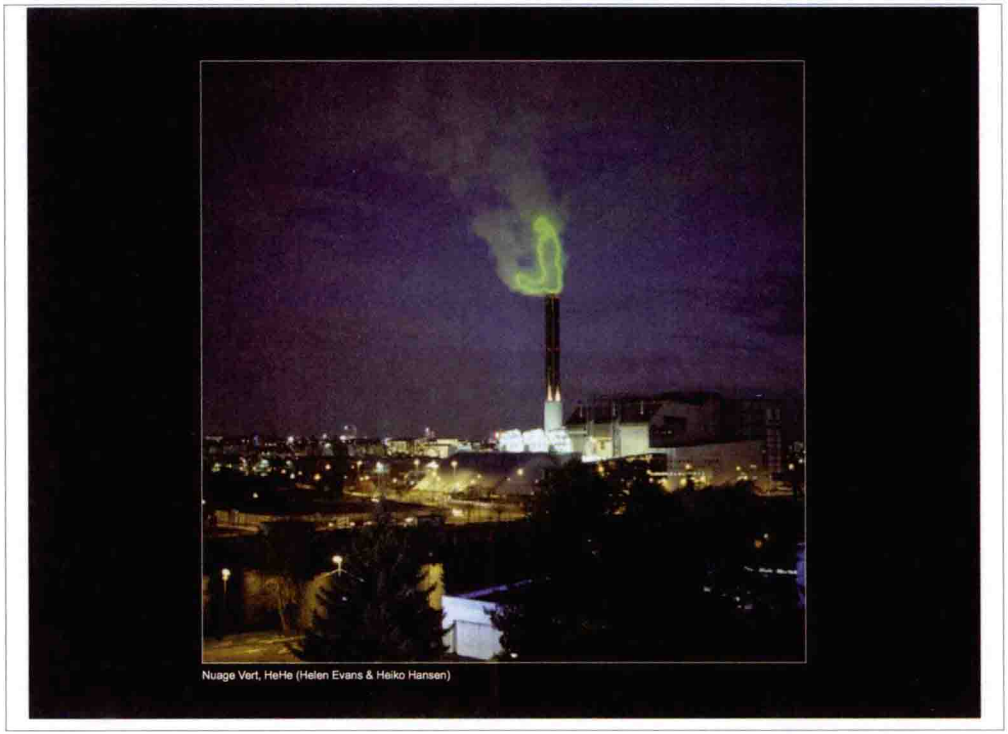

图 9-1：Helen Evans 和 Heiko Hanse 的 Nuage Vert 可视化案例（http://youtu.be/l_4rTQCWItw）

图 9-3：New Territories 的 Dusty Relief 项目（http://www.new-territories.com/roche2002bis.htm）

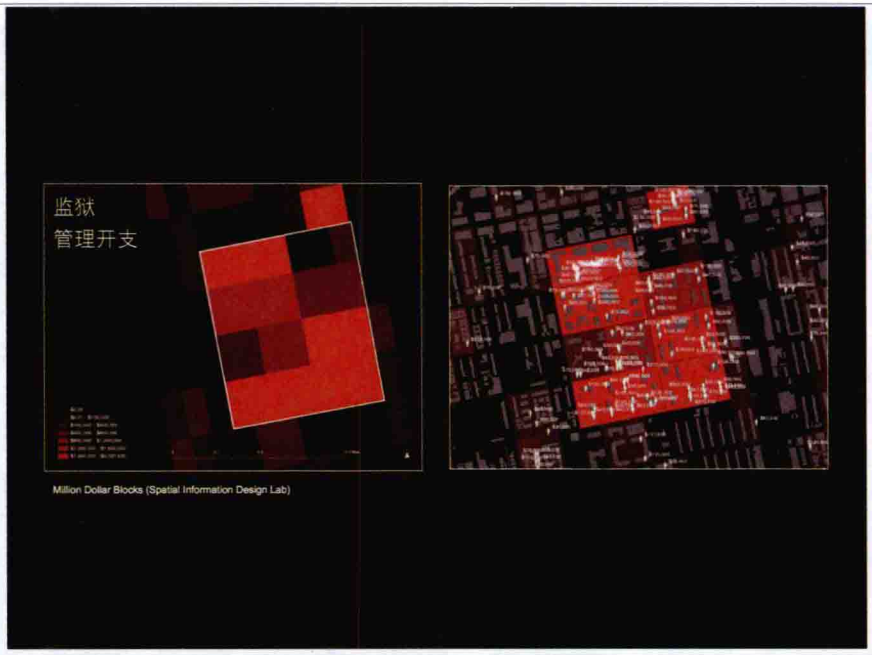

图 9-5：地理信息设计实验室（SIDL）设计的“百万美元街区”项目（http://www.spatialinformationdesignlab.org/）

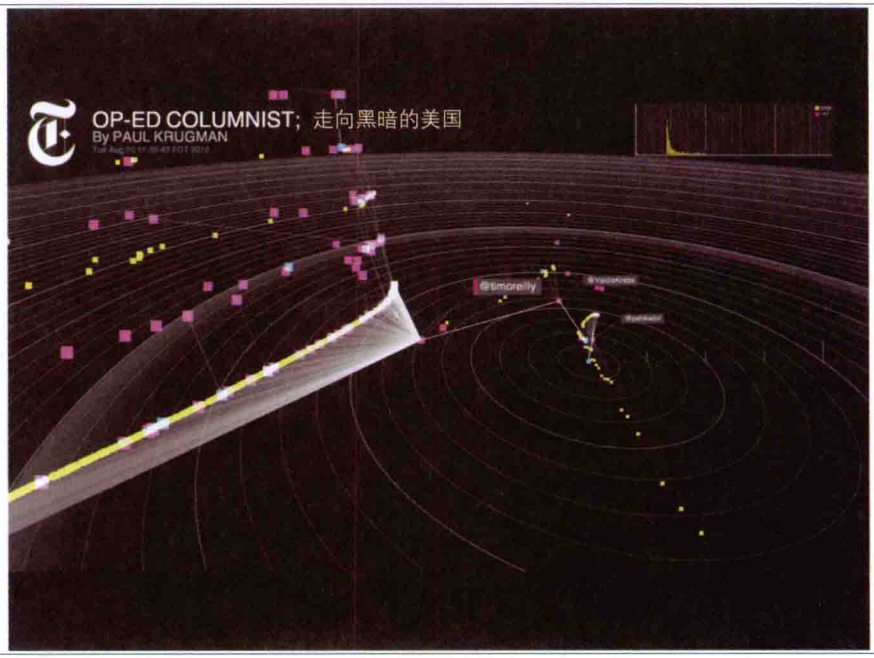

图 9-8：Jer Thorp 和 Mark Hanse 的作品：Cascade 项目

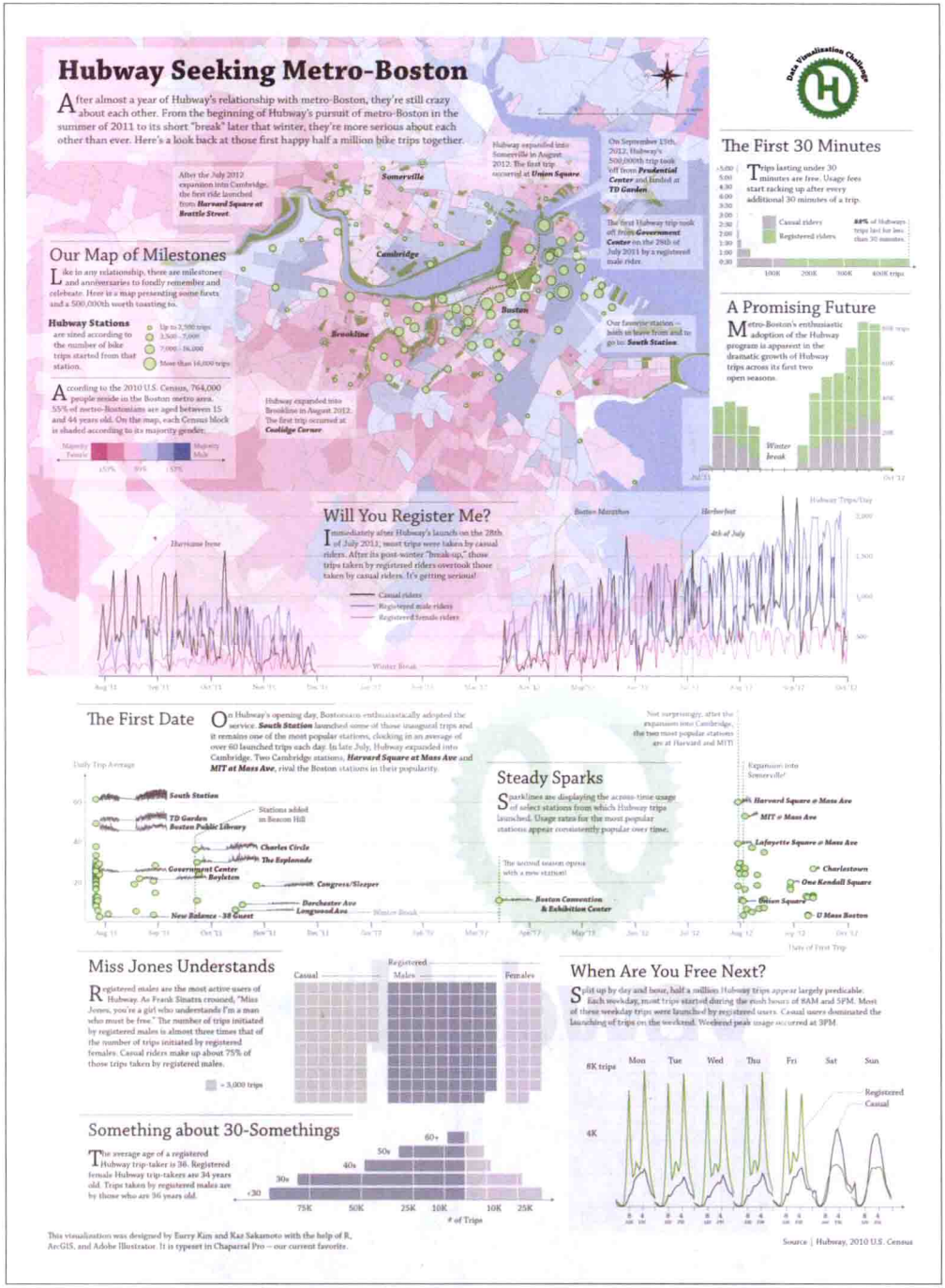

图 9-17：Eurry Kim 和 Kaz Sakamoto 参加 Hubway 公共自行车项目可视化竞赛的最终作品，以及这个项目在波士顿中心区实施的情况

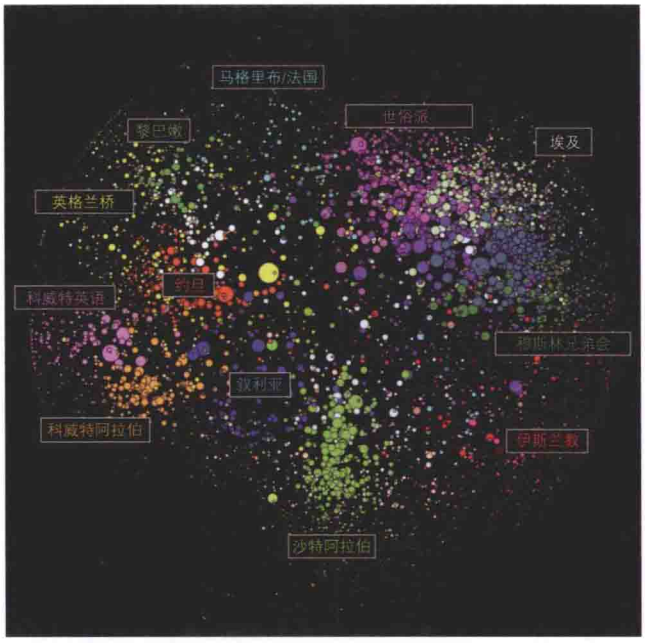

图 10-1：阿拉伯博客圈

图 10-2：英语圈博客分类图

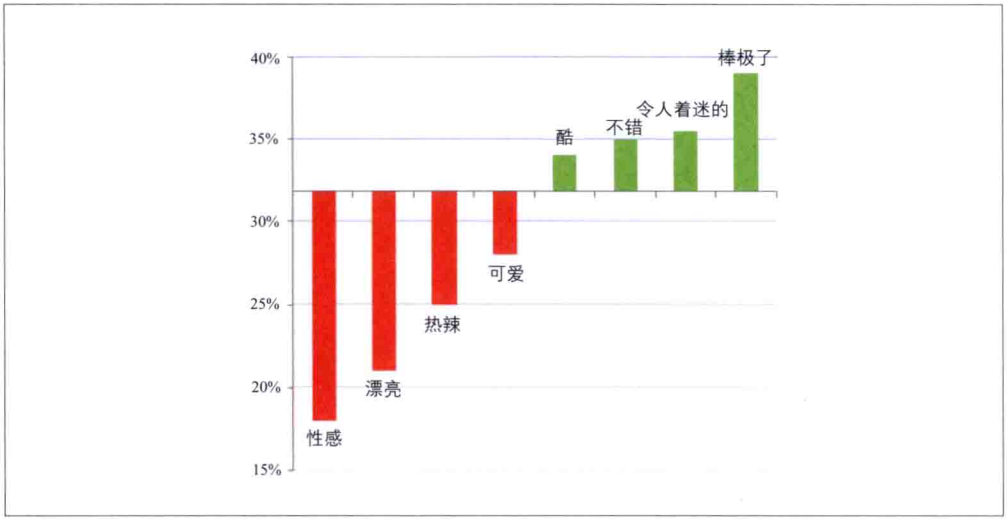

图 11-2：OK Cupid 的研究发现，在第一次接触性对话中使用“漂亮”一词不利于得到积极的答复

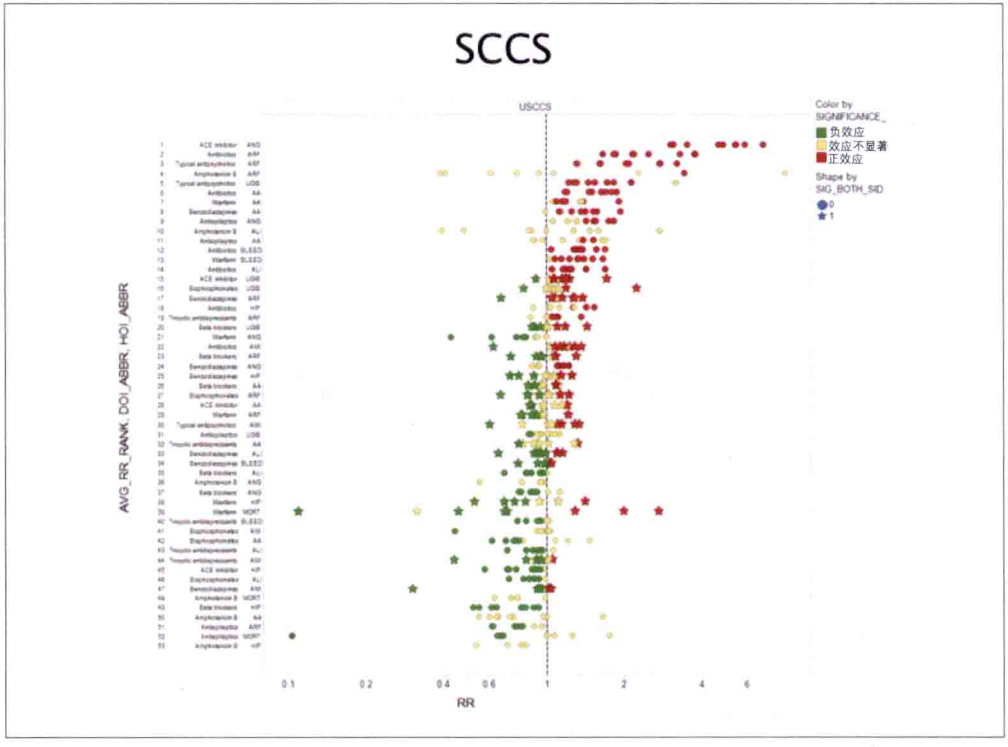

图 12-1：50 个课题的统计显著性分布图

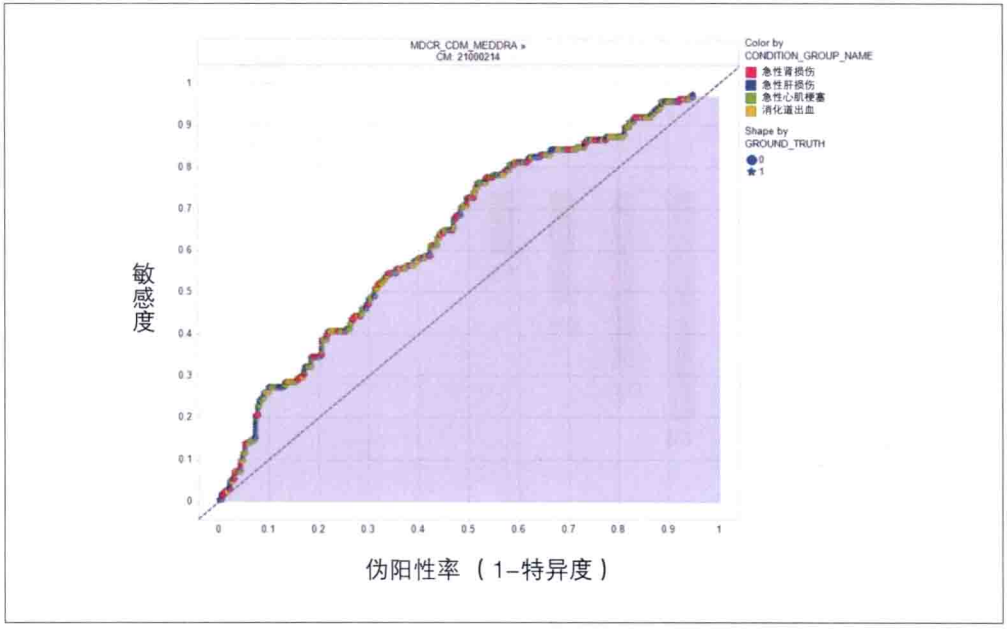

图 12-2：价值 2500 万美元的 ROC 曲线图

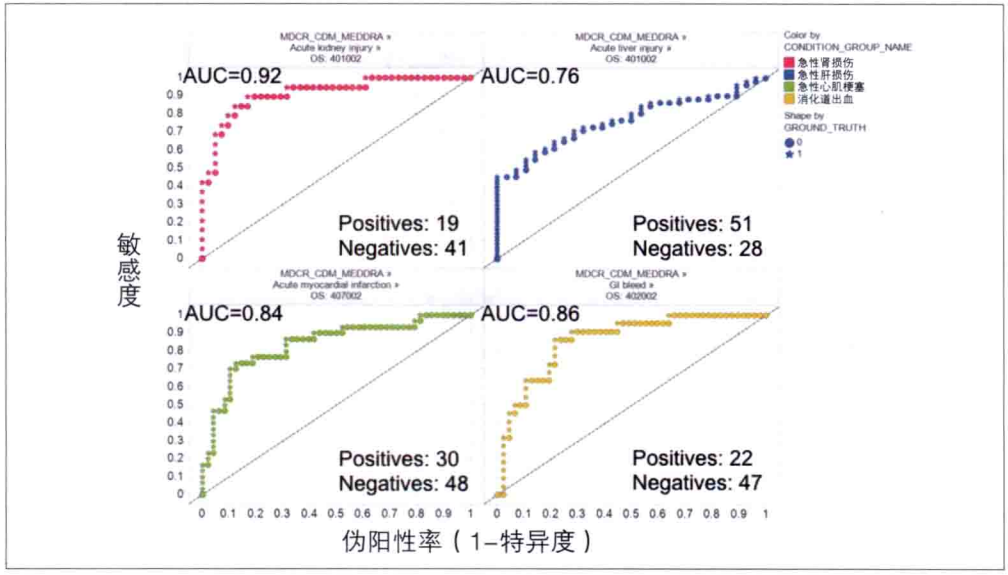

图 12-3：针对性的选取模型后，模型的预测效果可以得到较大改善

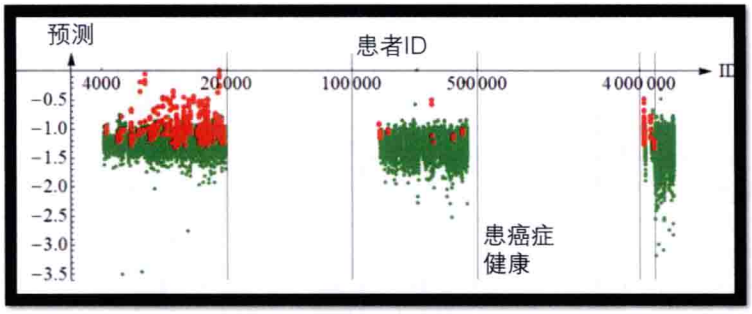

图 13-1：依患者的 ID 排序，红色代表得癌症的患者，绿色代表未得癌症的人

icd1x	icd2x	icd3x	icd4x
786	285	459	-1
401	486	-1	-1
401	486	780	-1
599	-1	-1	-1
V22	650	-1	-1
V56	492	586	-1
786	493	285	459

icd1x	icd2x	icd3x	icd4x
786	285	459	-1
401	-1	-1	-1
401	780	-1	-1
599	-1	-1	-1
V22	650	-1	-1
V56	492	586	-1
786	493	285	459

图 13-2：INFORMS 竞赛中数据是如何准备的

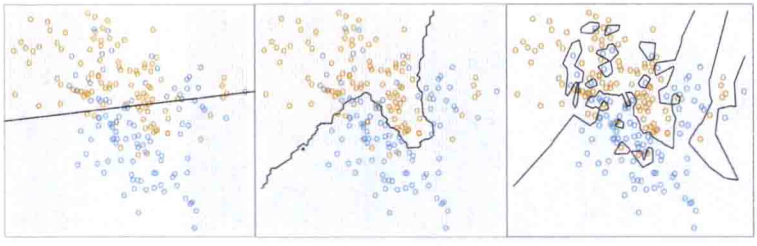

图 13-3：这幅经典的图片来摘自 Hastie 和 Tibshirani 合著的 *Elements of Statistical Learning*（《统计学习基础》，Springer-Verlag，参见 http://stanford.io/17szrYz），展示了同一份数据，对二值响应拟合线性回归模型时，采用 15 个最近邻和 1 个最近邻得到的不同结果

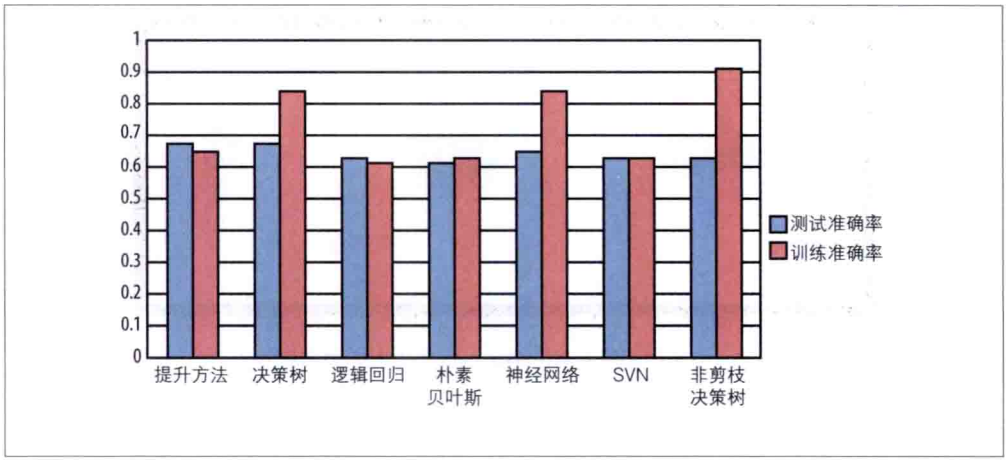

图 13-4：模型的差别

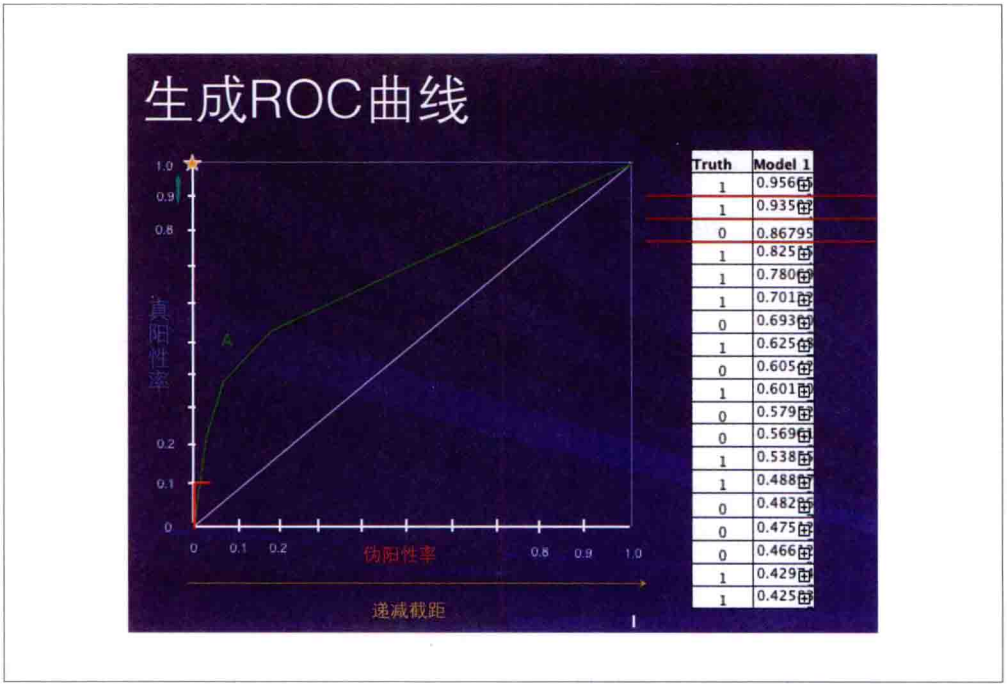

图 13-5：一个绘制 ROC 曲线的例子