

Predictive Analytics

The Power to Predict Who Will Click,
Buy, Lie, or Die

大数据预测

告诉你谁会点击、购买、死去或撒谎

[美] 埃里克·西格尔 (Eric Siegel) © 著

周 昕 © 译



中信出版社 · CHINA CITIC PRESS

Predictive Analytics

The Power to Predict Who Will Click,
Buy, Lie, or Die

大数据预测

告诉你谁会点击、购买、死去或撒谎

[美] 埃里克·西格尔 (Eric Siegel) 著
周 昕 译



图书在版编目 (CIP) 数据

大数据预测：告诉你谁会点击、购买、死去或撒谎 / (美) 西格尔著；周昕译. —北京：中信出版社，2014.4

书名原文：Predictive Analytics：The Power to Predict Who Will Click, Buy, Lie, or Die

ISBN 978-7-5086-4460-8

I. ①大… II ①西… ②周… III. ①经济预测学 IV. ①F201

中国版本图书馆CIP数据核字 (2014) 第 032675 号

Predictive Analytics：The Power to Predict Who Will Click, Buy, Lie, or Die, Eric Siegel

Copyright © 2013 by Eric Siegel

All Rights Reserved. This translation published under license.

Simplified Chinese translation copyright © 2014 by China CITIC Press

本书仅限中国大陆地区发行销售

大数据预测——告诉你谁会点击、购买、死去或撒谎

著 者：[美]埃里克·西格尔

译 者：周 昕

策划推广：中信出版社 (China CITIC Press)

出版发行：中信出版集团股份有限公司

(北京市朝阳区惠新东街甲4号富盛大厦2座 邮编 100029)

(CITIC Publishing Group)

承 印 者：北京画中画印刷有限公司

开 本：787mm×1092mm 1/16

版 次：2014年4月第1版

京权图字：01-2013-2470

书 号：ISBN 978-7-5086-4460-8 / F · 3144

定 价：58.00 元

印 张：18.5 字 数：211千字

印 次：2014年4月第1次印刷

广告经营许可证：京朝工商广字第8087号

版权所有·侵权必究

凡购本社图书，如有缺页、倒页、脱页，由发行公司负责退换。

服务热线：010-84849555 服务传真：010-84849000

投稿邮箱：author@citicpub.com

序言

Predictive Analytics

本书旨在通过量化方法来预测人类的行为。人类在此方面的最初实践是在第二次世界大战时期。1940年，“控制论之父”诺伯特·维纳（Norbert Wiener）便开始尝试预测德国空军飞行员的行为，目的是消灭这些纳粹空中力量。其预测方法是，观测德国飞机运动的轨迹，推测飞行员可能采取的机动规避动作，由此推断飞机接下来所处的位置并用高射炮将其击落。然而，维纳只能推断出飞机下一秒的运动轨迹，要想精确炮击飞机，必须预测飞机至少 20 秒的运行轨迹。

在埃里克·西格尔的书中，读者将看到许多预测案例，这些案例与维纳预测德国飞机的案例相比要精准许多。与“二战”时期相比，目前计算机的运算性能有了极大的提升，数据的丰富程度也非维纳之时可比。因此，银行、零售商、政治团体、医院以及其他众多机构，都在通过计算机数据处理来预测某些特定人群的行为，进而实现赢取客户、赢得选举或治愈疾病的最终目标。

在本人看来，这些预测行为是有益于人类发展的。在疾病治疗、打击犯罪以及反恐等领域，预测能挽救生命；在商业广告领域，预测能让广告定位更加精准，从而达到保护森林（减少无效纸质广告和宣传册的发放）、节省受众的时间和精力的目的；在政治领域，那些相信科学预测方法的政治候选人会拥有更大的胜算。

然而，正如西格尔在本书开篇中坦诚指出的那样，这些方法也可能产生问题。西格尔引述了电影《蜘蛛侠》中的台词“力量越大，责任越大”来说明这一点。其引申意义是，人类必须谨慎运用预测模型，否则其效用和益处就会大打折扣。与其他重要发明或革命性创新成果相似，预测分析本身并无是非对错之分，但作为工具，它却会带来或善或恶的后果。要想避免预测分析的不正当应用，我们首先必须知晓预测分析究竟能做什么，随着对本书阅读的深入，相信读者会对此问题形成自己的见解。

本书的重点是预测分析，这是诸多分析方法中的一种，是最有趣味和最重要的分析方法。在我看来，纯粹的描述性分析已经过时了，因为它记录的是过去发生的事情，无法真正说明这些事情为何会发生。此外，我也经常在自己的书里提到第三种分析方法，即规范性分析，也就是通过实验监测或定向优化来告诉人们应该怎么做。但这些数理分析方法的应用范围较预测分析要小许多。

本书内容及其背后的思想与纳西姆·尼古拉斯·塔勒布（Nassim Nicholas Taleb）的思想恰恰相反。塔勒布在其《黑天鹅》等书中提到，由于世界充满着偶然性且复杂事物的发展总是具有内在的不可预测性，因此预测行为注定会有失误。毫无疑问，塔勒布的话是有道理的，世界上总会有不可预测的“黑天鹅事件”，但我们认为，大部分人类行为都具有惯常性和可预测性。西格尔在本书中所给出的大量成功预测的案例表明，世界上大部分天鹅都是白色的。

同时，西格尔也在试图避免陷入“大数据”的陈词滥调。尽管书中的某些案例具有“大数据”分析的特征，即数据量庞杂无序且难以用传统关系数据库进行分析，但预测分析的关键点不在于数据的规模或繁复程度，而在于其如何对待或取舍数据。

我认为，通常，“大数据不过是小算术”，某些大数据实践者所做的不过是用宏大数据来装点门面。因此，其价值与真正的预测模型相比，自然有云泥之别。

西格尔在本书中所阐述的是复杂精巧的理念，但其行文却浅显易懂，无论读者是否熟悉数理分析，都可读懂本书。书中包含了大量的实际案例和分析图表，并用通俗诙谐的笔触剖析预测分析。即便是非数理分析专业人士，也应该好好阅读本书，因为在现实生活中，任何人的行为都免不了成为他人分析和预测的对象。此外，随着信息社会的发展，非数理分析专业人士也必然要在实践中学习预测模型、评估模型效果并根据预测模型的结果采取适当的行动。

总而言之，我们所处的是讲究预测的社会。要想在这样的社会中生存发展，最好的方法就是去理解预测的目标、方法以及限制，要想做到这一点，最好的方法莫过于阅读本书。

托马斯·H·达文波特

哈佛商学院访问教授、巴布森学院杰出教授和国际数据分析研究所创始人

《数据分析竞争法》作者

预测分析的职业风险

昨天已经过去，明天全然未知，我们能够把握的，只有今天。

——英国儿童文学作家米尔恩（A.A. Milne）、

美国漫画家比尔·基恩（Bill Keane）以及《功夫熊猫》中乌龟大师的名言

每当我告诉别人我的职业时，人们总是用异样的眼光看我。这也算是职业危害吧。

信息时代其实存在着巨大的不确定性。这样的论断可能会使许多人感到吃惊，因为当前，我们几乎可以把世界上发生的每一件事情记录下来。如果说历史书上仅仅记载那些重大事件，那么现在的信息系统如此发达，以至于人类的每次点击、每次支付、每个电话、每次交通事故、每次犯罪行为以及每次求医问诊都会被记录在案。在如此完备的海量数据面前，数据爱好者即便没有觉得自己是天之骄子，至少也应该感到心满意足吧。

但如此巨大的信息库中所缺乏的恰恰是最值得人类知晓的事：未来之事。

人人都渴望拥有预知未来的能力，他们对预测几乎无法抗拒。我们对先知神明顶礼膜拜；我们为算命先生慷慨解囊；我们热衷于占卜问卦，崇拜占星之术，对那些“讨口彩”的食品甘之如饴。

在狂热追求非理性预测行为的同时，人们却又常常鄙夷符合科学规律的预测。他们对科学预测的直接反应是“敬而远之”——科学预测显得深奥而乏味。对于许多人而言，或许预测是只有凭借超能力才可以做到的事情。美国喜剧《灵异妙探》（*Psych*）中的故事颇能说明这一点，这部剧的主角是一位具有敏锐的数据推断力的侦探，这位现代版的福尔摩斯具有超强的观察力，他对事实的精准描述常常令警察误认为他就在犯罪现场。为此，这位“神探”给出了一个最合乎情理的解释：自己有通灵能力。警察信以为真，而他也得以继续四处侦探，打击犯罪。绝对的喜剧！

我也有过类似的经历，比如，当别人友善地问起我的星座属性时，我不会假装自己相信这套东西，但我会用自相矛盾的方式回答他们：“我是天蝎座，天蝎座的人从不相信星座属性。”

在各类聚会中，经常有人问我到底是做什么的。每当此时，我都会调整姿态，看着对方略带困惑的表情，准确无误说出四个字：预测分析。大部分人的工作用一个单词就可以形容：医生、律师、服务员、会计或演员，但我的工作却不太容易描述。每次我都要费半天口舌向别人介绍我到底是做什么的。如果我含糊回答，对方更会打破砂锅问到底：

“我做技术领域的商业咨询。”如此回答后，对方会接着问：“什么技术？”

“我通过电脑来预测人的行为。”这种回答通常会引发更大的困惑，其中夹杂着怀疑和恐惧。

“研究数据来预测个体人类的行为。”对方还是不解，在聚会上，没人愿意谈论数据。

“分析数据来总结模式。”对方听后，表情更加困惑，在懵懂中陷入尴尬、沉默。

“帮助营销人员确定哪些客户会下单，哪些不会下单。”虽然对方能听得懂大概

的意思，但这种描述完全贬低了我的职业。毕竟，这只是我工作的一部分。

“预测客户行为，就像用孕试纸检测你是否怀孕了一样。”对方直接被吓跑。

为此，我写了这本书，想说明预测分析是直观的、有力的，是可以令人大开眼界的。

一点预测，无限可能。我称之为“预测效应”，这也是贯穿本书的主题。只要是预测而不是猜测，那么其力量就是显而易见的。预测效应表明，预测分析是可信的。我们只要顺势而为，就可以更好地看清未来。有一项令人激动又令人信服的发现：现在与未来之间隔着层层迷雾，但只要我们能将雾气稍微冲淡些，然后潜心研究较为清楚的那部分，我们就将创造出无限的价值。正因如此，预测分析可以帮助人们规避金融风险、改善医疗服务、清除垃圾邮件、加强打击犯罪或提升销售业绩。

你有科学家的好奇之心吗？你有不断进取的创业者的情怀吗？你是否对预测本身或预测能产生的价值感到着迷？

我对“知晓不可知之事”尤为热衷。预测似乎能打破这一自然规律：人不能知晓未来，因为未来尚未到来。但自从研发了能从历史经验中总结规律的计算机系统，人类现在可以更好地看清未来。通过严谨的方法来整理“已知”数据信息，人们就可越来越精准地预见未来之事。这是数学与科技的融合，两者之间不断地相互砥砺，最终开花结果，产生了科学的系统，由此连通现在与未来之间那个曾经不可逾越的鸿沟。

这是一项前无古人的事业！

有人做销售，有人搞政治，而我则做预测，这令我骄傲。

预测效应

我也是普通人，在生活中，有成功，也有失败；有时交好运，有时走霉运。人们总是想象，如果生活不是这样，那将会怎样。在此，我想简单说说我所遭遇的 6 次不幸。

1. 2009 年在犹他州滑雪时受伤，我的右膝盖差点儿残废。滑雪起跳时没有问题，但落地时却发生了偏差。膝盖要做手术，因为膝关节前交叉韧带断裂，所以要选择身体其他部位的韧带进行修复。这样的选择很痛苦，因为如果选择失误，我下半辈子就有可能变成瘸子。最后，我选择了用自己的腿后腱。那么，医院能否给我提供一个更好的治疗方案？

2. 我本人承受了极大的身体痛苦，但付钱付到肉疼的却是保险公司，因为膝盖手术相当昂贵。那么，面对我这类蹩脚的滑雪爱好者，保险公司能否更好地预见风险并把风险计算在保费里？

3. 早在 1995 年时，我就遭遇过事故，虽然那次事故并未对我造成

大的身体伤害。我的身份证号被盗，我不得不耗费大量时间在不同部门之间奔波，走那些烦琐的程序，填写各种表格，由此来消除错误的信用记录。那么，那些对我的账号放债的人，他们有没有办法在第一时间就判断出我的账户被盗了呢？

4. 在恢复了良好信用记录后，我以按揭方式购买了一套公寓。这算是明智的投资决策吗？或许我的理财顾问应该对我进行风险提示，因为这套房子在买入之后，很快就因跌价而变成负资产。

5. 其实我的职业生涯充满风险。虽然现在生意还可以，但作为企业，势必要面对经济环境变化或竞争加剧带来的风险。那么，我们能否预测：哪些营销行为会有效果，哪些投资活动会有良好回报，哪些行为只是烧钱呢？

6. 我们的日常生活是由小事构成的，这些小事的顺利与否决定了我们的命运。有效的垃圾邮件过滤系统可防止我们在工作时被打扰。有效的互联网搜索也很重要，不仅工作中要用到，还可用来搜索医疗信息（如膝盖手术的知识）、家居装潢以及其他信息。我们也信赖潘多拉网络电台以及Netflix（网飞公司）所推荐的个性化影片和音乐。但在许多年之后，我的邮箱却还是常常收到垃圾邮件。为何有些公司就不能多了解一些我的信息，来减少无效邮件呢（如果是纸质邮件，还可减少森林砍伐）？

这些问题并非无关紧要，它们决定着 we 每天、每年甚至这辈子生活的好坏。那么，这些问题有什么共同点呢？

与其相似的许多挑战和问题其实都可通过预测的方法加以解决。病人是否适合做这个手术？借款人是否会欠钱不还？这位购房者能及时还上按揭吗？这位消费者是否会对邮寄的营销材料感兴趣？如果能正确预测这些问题，那么，我们的生活将因此而得到极大的改观。

大企业的预测——资产的归宿

我们还可以从其他角度来看待这个问题。预测除了让你我这样的消费者获益之外，也可让企业脱胎换骨，形成全新的竞争力。因此，很多企业都在不遗余力地提升预测力。

20 世纪 90 年代中期，一位名叫丹·斯坦伯格（Dan Steinberg）的商业科学家走进了美国大通银行，他要帮助这家金融机构预测数百万份按揭的风险。大通银行采纳了斯坦伯格的预测技术，并借助斯坦伯格研发的系统来评估、处理大量的银行按揭。从此，斯坦伯格在金融界声名鹊起。

预测就是力量。如果某大型商业机构能预测个体资产的风险变化和价值，那么，它将形成不可撼动的市场竞争优势。在本案例中，大通银行精确预测了按揭申请人的未来还款行为，由此极大降低了放贷风险并增加了赢利——大通银行当年就获得了高达 9 位数的利润。

发明会学习的电脑

预测技术不断完善，渐成主流，现在几乎无处不在，时刻影响着我们的生活。预测技术正在不知不觉中影响着人类的体验，无论是开车、购物、学习、就医、沟通、看电视，还是赚钱、借钱甚至偷盗。

本书要讲述的是计算机预测技术中最具影响力和最有价值的成就，以及其背后的两大因素：技术背后的人和推动技术发展的神奇的科学。

做出精确的预测是件很难的事。每项预测都有若干先决条件，即首先要掌握每个病人、每个购房者以及每封邮件的不同特征信息。那么在每项预测中，我们该如何将这些分散的信息综合起来呢？

说起来容易做起来难。我们的应对之策就是，用系统化和科学化的方法来开发

并持续改善预测技术，即要让计算机系统自动“学习”如何预测。

这就是机器学习，也就是让电脑自动获取新知识和新能力，持续不断地输入现代社会最有价值和最重要的非自然资源：数据。

“喂我吧！”——机器思考的食物

数据是一种新型石油。

——欧洲消费者委员会委员梅格莱纳·库尼瓦 (Meglena Kuneva)

知识的唯一来源是经验。

——阿尔伯特·爱因斯坦

我只信数据。

——质量管理大师威廉·爱德华兹·戴明 (William Edwards Deming)

大部分人都对数据感到厌倦，它好像总是让人觉得乏味。数据仿佛是无数事实和数字的堆砌，每条数据都显得无聊，其乏味程度等同于“我买了双新鞋”之类的微博。这是一大堆尚未加工的索然无味的材料，只有企业才会去咀嚼。

但千万不要被数据的表象欺骗！其实，数据里凝结了极为珍贵的值得学习的经验。每一次医疗诊断、借款申请、Facebook（脸书网）发帖、影视推荐、欺诈行径、垃圾邮件，以及结果或好或坏的购买行为、失败或成功的电话推销、交通事故、事件或交易，都会被整理成数据并积累起来，由此缔造了海量的原始材料。它们的数量是如此庞杂，只有计算机才有可能从中总结出规律。如果应用得法，计算机就会像海绵吸水一样从原始材料的汪洋中汲取知识。

随着数据的不断累积，人们也开始掀起了从数据中获取财富的淘金热。但数据本身并不是黄金，作为原始材料的数据只是枯燥无味的代码组合。只有从数据中提炼出来的规律和知识才是黄金。



图 0-1 数据—机器学习—预测

计算机自动学习系统的研发使得数据资源的能量开始爆发。因为这一系统揭示出了人类的动机及行为，这是人类生存的印迹以及世界发展的奥秘所在。在获取这些新知识后，科学预测就成为可能。

计算机在自我学习过程中发现了如下有意思的规律：

- 提前退休会缩短寿命。
- 网上约会时，评分越高，配对成功率就越低。
- 女歌手蕾哈娜的粉丝大都支持民主党。
- 素食主义者不太容易误机。
- 公共体育活动会导致本地犯罪率上升。

机器学习会从这些发现中尝试建立预测能力，通过对数字的挖掘和试错，并用统计学方法和计算机科学方法实现这种预测。

我早就知道你会这么做

当拥有这些预测能力之后，我们想要预测什么呢？人的每个行为都值得预测，无论是消费、思考、工作、放弃、恋爱、生育、离婚，还是捣乱、撒谎、欺骗、盗窃、凶杀、死亡。让我们来看看其中某些行为。

人的消费行为

- 好莱坞的影视公司会预测，如果某剧本被制作成影片，它是否会受欢迎。
- Netflix 悬赏 100 万美元来改善其影片推荐系统，某研究团队因为成功地让系统推荐的影片更受消费者喜爱而赢得奖金。
- Energex（澳大利亚电力集团）预测电力需求，以此确定在何处建设电网；Con Edison（联合爱迪生电力公司）会预测在用电高峰时可能出现的系统故障。
- 华尔街的金融机构通过观察股价涨跌预测股价未来走势，AlphaGenius 和 Derwent Capital 两家投资机构会根据 Twitter（推特）上公众的表现来进行对冲基金交易。
- 企业会预测哪些消费者会购买其产品并选择其作为营销对象，无论是大型金融机构还是像港湾甜品（糖果）和佛蒙特乡村商场（提供“高品质且难以寻找的经典产品”），它们都在这样做。宝贵的营销预算要依照预测结果来分配。某些公司甚至会预测，如何才能提升对消费者的影响力，让消费者买得更多（详见第七章）。
- 超市收银台发放的优惠券，也是预测在起作用。英国零售业巨头乐购是全球第三大零售商，它会预测哪些优惠券被拿回来使用，由此提升每年在全球 13 个国家的门店里发放的上亿张优惠券的利用率。与此前的方法相比，科学预测将乐购优惠券的使用率提升了 3.6 个百分点。于是，凯马特超市（Kmart）、克罗格超市（Kroger）、拉尔夫超市（Ralph's）、喜互惠连锁超市（Safeway）、停车采购超市（Stop & Shop）、目标超市（Target）以及温迪克斯（Winn-Dixie）等零售商也纷纷效仿。
- 如果能预测网络用户的点击率，企业将获得丰厚的回报。通常，网站上的广告都是按点击率来收费的，因此网站会预测你最可能点击的广告，然后让广告在最显眼的位置弹出。实际上，网站替你选择了与你相关性更高的广告，并以此增加了数百万美元的收入。

人的恋爱、工作、生育和离婚

- 职场社交网站 LinkedIn (商务化人际关系网) 会预测你的未来职业选项。
- 婚恋网站 Match.com、OkCupid 和 eHarmony 会根据你的特征为你推荐相亲对象。
- 目标超市会通过预测客户的怀孕时间来推销母婴用品。
- 临床研究人员会预测不忠与离婚之间的关系, 现在甚至还有预测离婚的自助网站, 让你测测自己婚姻的长久程度 (www.divorce360.com), 据传信用卡公司也采纳了这套方法。

人的思考和决策

- 2012 年, 奥巴马在选民预测系统的帮助下再次当选为总统。奥巴马的竞选团队成功预测到, 通过与选民进行与竞选有关的接触 (例如电话、家访、横幅广告以及电视广告), 选民更容易支持己方, 反之, 选民便会倾向于支持对方。这套方法针对的是数百万的“摇摆选民”, 与传统的选民定位相比, 它让更多的选民选择了奥巴马。
- “你是什么意思?” 现在, 电脑系统已经学会了从书面词句中推测人的主观动机。花旗银行和贝宝支付 (PayPal) 能以此感知到用户对其产品的态度, 而某位研究人员研发的系统则可判断亚马逊网站上的书评哪些是在说反话。
- 学生论文评分系统现在已经问世, 可自动给论文评分, 系统精确程度与人相当。
- 现在, 计算机已经能参加美国最受欢迎的智力和知识挑战节目, 直接与人类选手对决。在《危险边缘》(Jeopardy!) 电视问答节目中, IBM (国际商业机器公司) 所研发的电脑“沃森”击败人类选手获得了胜利。这台电脑可准确处理英语信息并回答各类随机抽取的知识问题, 还击败了两位曾获得该节目冠军的人类选手。
- 现在的计算机已经可以读懂人的思想。科研人员研发出了这样的系统来解码人类大脑活动并确定大脑正在思考的对象。例如, 是关于某些工具、建筑, 还是

关于食物？现在，计算机预测的正确率已经超过 80%。2011 年，IBM 预测说，未来 5 年，读心术将成为主流。

人的放弃

- 惠普公司对其全球 33 万名员工都进行了“离职风险”评估，预测每位员工的离职可能性，这样管理层就可提前介入或及早采取应对措施。
- 你是否对手机运营商的服务感到厌倦？运营商也在努力预测用户的体验。世界上所有大型移动运营商都会判断用户终止服务合约并转投竞争对手的风险，因此或许在你决定更换运营商之前，他们就已经提前知晓，其预测的依据就是你的通话故障次数、电话使用频率、账单信息以及你的主要联系人是否已经选择了其他运营商。
- 联邦快递在快递市场上始终处于领先地位，因为这家公司能预测哪些客户会投向竞争对手，其预测成功率高达 65%-90%。
- 美国公立大学系统可预测辍学率并依据预测结果来积极管理学生；此外，亚拉巴马大学、亚利桑那州立大学、艾奥瓦州立大学、俄克拉何马州立大学以及荷兰爱因霍芬科技大学都在用计算机预测学生的辍学率。
- 维基百科的大部分编辑人员都在免费工作，他们维护着这份网络百科全书纯粹出于热爱，但维基百科也在预测，哪些编辑人员可能会中途退出而不再继续为其做出贡献。
- 哈佛医学院的研究人员预测，如果你的朋友戒烟，那么你戒烟的可能性就会提高。戒烟也具有“传染性”。

人的事故

- 保险公司会预测谁开车更容易出交通事故，或谁在滑雪时更容易受伤。好事达