第 5 章 可解释模型：没有实际场景的模型是劣质模型

在数据分析过程中，无论是 0 到 1 的业务还是慢慢成形的业务，到了把主要的业务问题解决之后都会把业务流程做成模型、把归因体系做成模型、把盈利方式做成模型，本身做模型这件事是数据分析工作中重要的内容。很多模型（像正态分布模型、漏斗模型）都有其本身的含义和理论基础，并且可以实际地反映问题。

5.1 串联业务的可解释模型

5.1.1 案例 32：面包质量问题

举个例子，在战争时期，德国为了应对战时粮食危机决定全国面粉统一由政府管理，每天发放固定的面粉，由指定工厂制作面包发给市民，并规定面包质量为 400g，这对于一个工厂来讲可是天大的好事，但由于政府的强压，使他们在定价和原料获取上都没有充分的利润，因此有些工厂就在面包上做了手脚，一个统计学家对每天发放的面包质量进行了记录，见表 5-1。

表 5-1 每天发放面包质量表

天　数	1	2	3	4	5	6	7	8	9	10
面包质量/g	393	407	393	395	389	409	389	390	394	415
天数	11	12	13	14	15	16	17	18	19	20
面包质量/g	410	409	384	388	414	414	392	393	408	387
天数	21	22	23	24	25	26	27	28	29	30
面包质量/g	399	393	414	404	398	400	399	407	406	411

表中为面包质量的数据，工厂制作面包使用面包模具进行生产，虽然模具的大小不一，但需要满足面包质量大小达到 400g 这个标准，在数据足够多的情况下，面包质量应该符合正态分布，如图 5-1 所示，也就是 68.2%的数据应该在 400g 左右，高于 410g 或者低于 390g 的数据应该不多于 31.8%。

而实际使用数据表中的数据进行作图后发现，图像出现了右偏分布，如图 5-2 所示，也

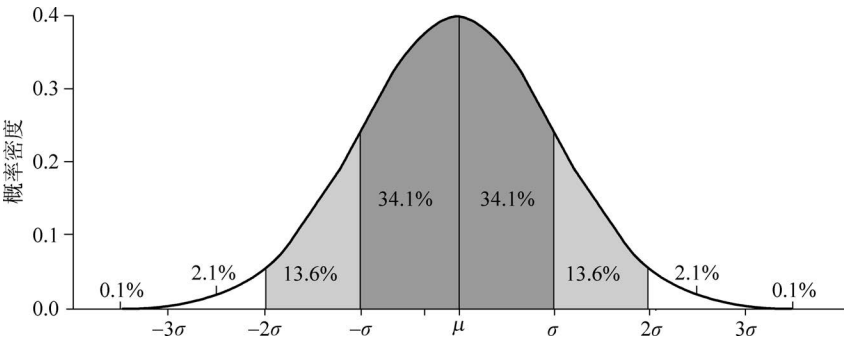

图 5-1 面包质量正态分布图

就是更多的面包其实处于 400g 的左边，并且低于 390g 的面包数量相对较多，因此可以判断工厂在模具上刻意选择了质量较小的模具，以此来赚取利润。

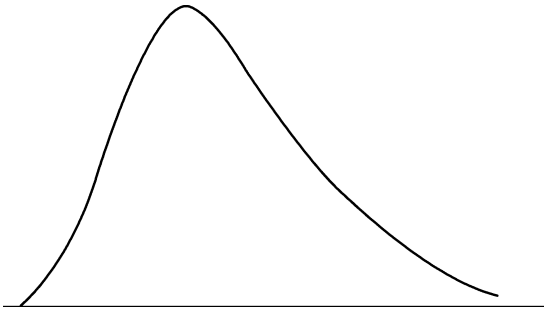

图 5-2 右偏的面包质量正态分布图

在这个例子中，由于工厂生产面包几乎与模具选择有关，所以最终的“结果”面包也只跟模具相关，可以应用可解释模型的原理去评估业务的效果，以及时发现问题，然而在实际业务中往往会出现更多的干扰因素，例如假设面包生产工作并没有那么简单，不容易可标准化，不同的员工、不同的时间点、不同的天气都会影响最终的面包质量，这些因素会使面包质量的分布并不按照正态分布，导致模型不可解释，失去评估业务的作用，因此模型的可解释性决定了模型是否能够与业务匹配，从而帮助提升业务水平。

5.1.2 案例 33：TikTok 商家成长模型问题

通过一个实际的业务场景来了解搭建一个业务模型都涉及什么问题。在国内，有很大部分网购用户已经逐渐熟悉和适应了直播电商，不管是在淘宝还是抖音，现在各大互联网公司又把矛头对准了海外市场，其中抖音的海外版 TikTok 本身在海外拥有海量的用户群体，因此 TikTok 电商也在快速地进行全球扩张，短时间内开辟了大量的新市场，海外扩张依赖于国内成熟的业务体系，如何套用国内的业务模型，帮助处于不同阶段的商家根据自身情况快速找到在 TikTok 上经营的稳定模式和最佳实践是扩张能否成功的关键。依靠海内外商家的优秀案例，分析商家的成长路径，找出成功商家不同成长阶段中的关键动作及阈值，形

成商家成长模型，以制定相应扶持政策和产品帮助新入驻商家快速高效地达到产生动销，稳定动销，直至温饱的里程碑，是数据分析乃至业务的重要命题。

就以如何建立商家成长模型为例，商家是怎么成长的是一个很大的概念，当具体到某个个体商家时，有的商家是有明确的经营目的的，有的商家仅仅做做看，还有的是由别的平台迁移过来且有着成熟经验的，所以首先需要明确不同类型的商家，从资质上看可以分为个人型商家和公司型商家，从行业上可以分为美妆商家、3C 商家、食品商家等，从品牌效力上可以分为知名品牌商家、普通品牌商家和无品牌商家，从商家角色上可以分为是否为机构商家，从经验上可以分为有无电商经验等。之所以把这些分类切出来，是因为商家成长情况会受到这些属性的影响，例如头部品牌的商家运营起来起码比没有品牌的商家更为高效，有电商经验的商家比没有电商经验的商家也是天差地别。

对不同类型的商家还可以划分出来针对每种类型商家的成长阶段，统一地可以分为破冰期，指从入驻到稳定有销量；成长期，指从稳定有销量到初级的头部商家；成熟期，指从初级的头部商家到顶级的头部商家，这种成长阶段的分层对于不同类型的商家是不一样的，具体体现在（例如同样处于破冰期），对于有电商经验的商家来讲，本身的分布数量比起没有电商经验的商家要小很多，并且有电商经验商家的破冰期定义可能是入驻到稳定每天卖 50 单，而对于没有电商经验的商家来讲破冰只需每天卖 5 单就可以了。

继续区分，不同类型的商家在同成长阶段也可能有不同的成长路径。商家有很多种经营模式，有的是自己准备货源自己发货，有的是找厂家一件代发，甚至有的是直接从别的店铺下单后发货到消费者手中，并且有的商家是依靠货架模式把商品都上架到店铺中，依靠橱窗展示来卖货，有的则是通过创作短视频，把商品绑定在短视频上带货，而有的是通过直播带货，不同商家选择的卖货载体也不同，因此可以把商家的主流成长路径划分成自卖型和非自卖型，把带货形式分成直播型、短视频型和橱窗型。把商家切分清楚后，可以开始对成长路径进行建模，在商家的成长过程中有很多关键行为对商家有极大影响和明确区分的，例如首次上传商品、首次有销量，并且在不同节点里又会分别处于不同阶段和分布区间，如图 5-3 所示，即为在直播电商平台上的商家成长情况，对应到不同环节都会有该阶段的关键动作和商家分布情况。

针对每部分的问题都需要可解释的实现方法，例如如何进行商家分类，并不是说把商家的所有属性字段作为特征输入，然后用类似聚类等一些算法把商家分层就完成任务了，那样在实际业务使用时可能不具有参考意义，并且也不知道怎么去使用，类似这种商家分类问题，更多的是根据业务经验判断，选择出一个确实影响到业务并可对一个商家进行判断的指标，然后以这样的指标做组合分层，从而对商家进行分类。成长阶段的分类也没有办法通过机器的方法解决，成长阶段其实是一个人为设定的概念，通常是通过统计分布情况进行划分的，例如想知道一个电商业务中什么样的用户才是在业务中长期停留的忠诚用户，常用的方法是观察用户下单数量的函数曲线，并在斜率发生较大变化时进行切分，人为地定义每个区间对应的阶段。主流成长路径划分由很多种方法共同切分完成，像通过业务经验和定性定量的方法可以去划分出一些关键性的区分动作，例如首次上传商品时间、绑定账号时间、开

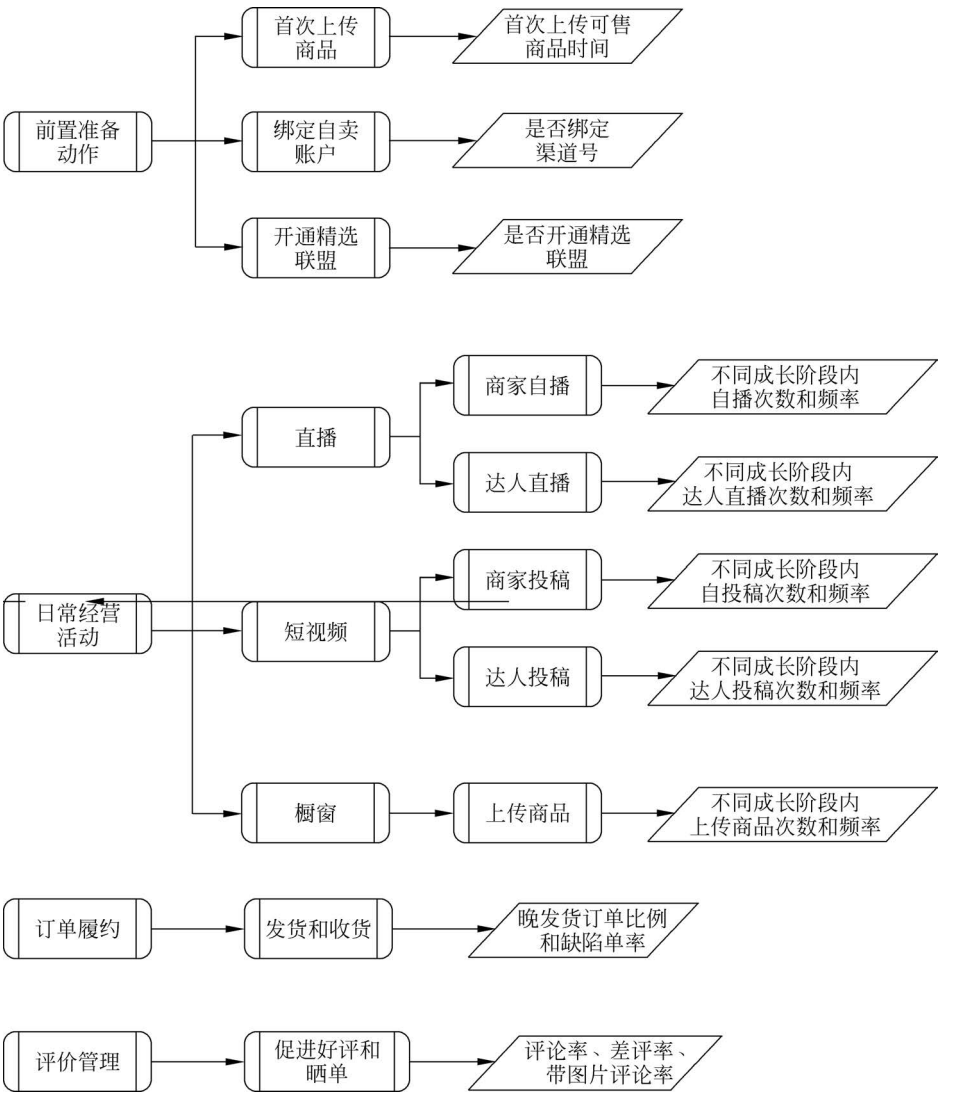

图 5-3　商家成长路径图

通联盟时间，也有通过统计方法去界定的，例如发货的及时率、退货订单的占比、订单评论率、好评率、差评率等，通过对这些数据的统计可以按照数值大小划分出类似高、中、低的分级，对于不同成长阶段内的商家，通过对自播次数、频率进行聚类划分，也可以对不同阶段内商家目前处于什么层级做出划分。这样做的核心优势在于可以清楚地知道划分的依据，以及对不同商家进行分层的不同理由，也就是具有可解释性。

当然有时业务上提出的一些关键因素确实从业务上明确地知道跟业务结果有关系，但是影响的方向和程度不是很明确，也没有得到量化，因此在构建可解释模型时会有几种需要经常用到的分析方法。

5.2　常用分析方法

1. 相关性分析

“万物皆有联”是数据分析最重要的核心思维。所谓联指的就是事物之间的相互影响、相互制约、相互印证的关系，而事物这种相互影响、相互关联的关系，在统计学上就叫作相关关系，简称相关性。世界上的所有事物都会受到其他事物的影响。HR 经常会问：影响员工离职的关键原因是什么？是工资还是发展空间？销售人员会问：哪些要素会促使客户购买某产品？是价格还是质量？营销人员会问：影响客户流失的关键因素有哪些？是竞争还是服务等？产品设计人员会问：影响汽车产品受欢迎的关键功能有哪些？价格还是动力等？将所有这些商业问题转换为数据问题，不外乎就是评估一个因素与另一个因素之间的相互影响或相互关联的关系，而分析这种事物之间关联性的方法就是相关性分析方法。

当然，有相关关系，并不一定意味着是因果关系，但因果关系一定是相关关系。在过去，传统的统计模型主要用来寻找影响事物的因果关系，所以过去也叫影响因素分析，但是，从统计学方法来讲，因果关系一定会有统计显著，但统计显著并不一定就是因果关系，所以准确地说，影响因素分析应该改为相关性分析，所以在不引起混淆的情况下，也会用影响因素分析。

客观事物之间的相关性，大致可归纳为两大类：一类是函数关系，另一类是统计关系。统计关系指的是两事物之间的非一一对应关系，即当变量 x 取一定值时，另一个变量 y 虽然不唯一确定，但按某种规律在一定的可预测范围内发生变化。例如，子女身高与父母身高、广告费用与销售额的关系等是无法用一个函数关系唯一确定其取值的，但这些变量之间确实存在一定的关系。在大多数情况下，父母身高越高，子女的身高也就越高；广告费用花得越多，其销售额也相对越多。这种关系就叫作统计关系。

进一步，统计分析如果按照相关的形态来讲，则可分为线性相关和非线性相关（曲线相关）；如果按照相关的方向来分，则可分为正相关和负相关等。描述两个变量是否有相关性，常见的方式有可视化相关图（典型的如散点图和列联表等）、相关系数、统计显著性。如果用可视化的方式来呈现各种相关性，则常见的相关性如图 5-4 所示。

对于不同的因素类型，采用的相关性分析方法也不相同，如图 5-5 所示，简单总结一下所选用的相关性分析方法。

简单地说，相关分析用于衡量两个数值型变量的相关性，以及计算相关程度的大小。相关分析常用的方法有简单相关分析、偏相关分析、距离相关分析等，其中前两种方法比较常见。简单相关分析是直接计算两个变量的相关程度。偏相关分析是在排除某个因素后计算两个变量的相关程度。距离相关分析是通过两个变量之间的距离来评估其相似性（这个应尽量少用），如图 5-6 所示，在没有特别说明的情况下，下文所讲述的相关分析指的是简单相关分析。

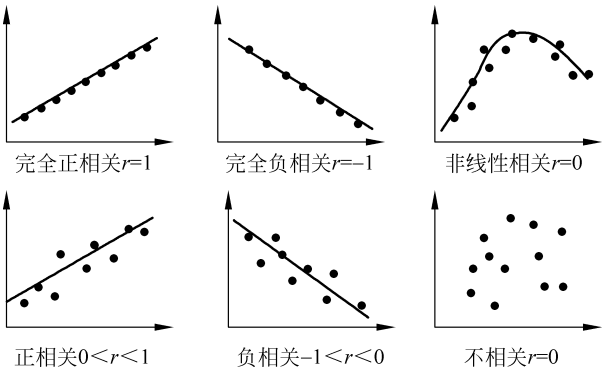

图 5-4　变量相关性图

解释变量类型	被解释变量类型	方法	作用
数值型变量	数值型变量	相关分析	衡量两个变量的相关程度
类别型变量	数值型变量	方差分析	评估因素对目标变量是否有显著影响
类别型变量	类别型变量	列联分析	评估两个因素是否相互独立

图 5-5　变量对应的相关性分析方法

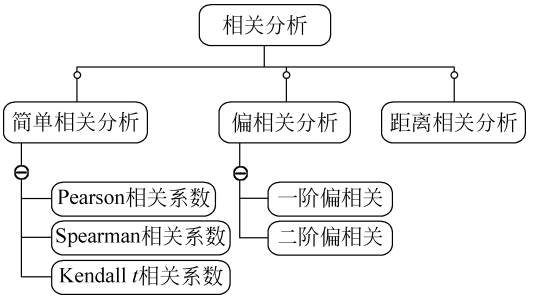

图 5-6　相关分析方法及系数图

第 1 种相关分析方法是对数据进行可视化处理，简单地说就是绘制图表。单纯从数据的角度很难发现其中的趋势和联系，而将数据点绘制成图表后趋势和联系就会变得清晰起来。对于有明显时间维度的数据，选择使用折线图，如图 5-7 所示。

为了更清晰地对比这两组数据的变化和趋势，使用双坐标轴折线图，其中主坐标轴用来绘制广告曝光量数据，次坐标轴用来绘制费用成本数据。通过折线图可以发现，费用成本和广告曝光量两组数据的变化和趋势大致相同，从整体的大趋势来看，费用成本和广告曝光量两组数据都呈现增长趋势。从规律性来看，费用成本和广告曝光量数据每次的最低点都出现在同一天。从细节来看，两组数据的短期趋势的变化也基本一致。

经过以上这些对比，可以说广告曝光量和费用成本之间有一些相关关系，但这种方法在整个分析过程和解释上过于复杂，如果换成复杂一点的数据或者相关度较低的数据就会出现很多问题。

比折线图更直观的是散点图，如图 5-8 所示。散点图去除了时间维度的影响，只关注广告曝光量和费用成本这两组数据间的关系。在绘制散点图之前，将费用成本标识为 x，也就是自变量，将广告曝光量标识为 y，也就是因变量。下面是一张根据每天广告曝光量和费用成本数据绘制的散点图，x 轴是自变量费用成本数据，y 轴是因变量广告曝光量数据。从数

据点的分布情况可以发现，自变量 x 和因变量 y 有着相同的变化趋势，当费用成本增加后，广告曝光量也随之增加。

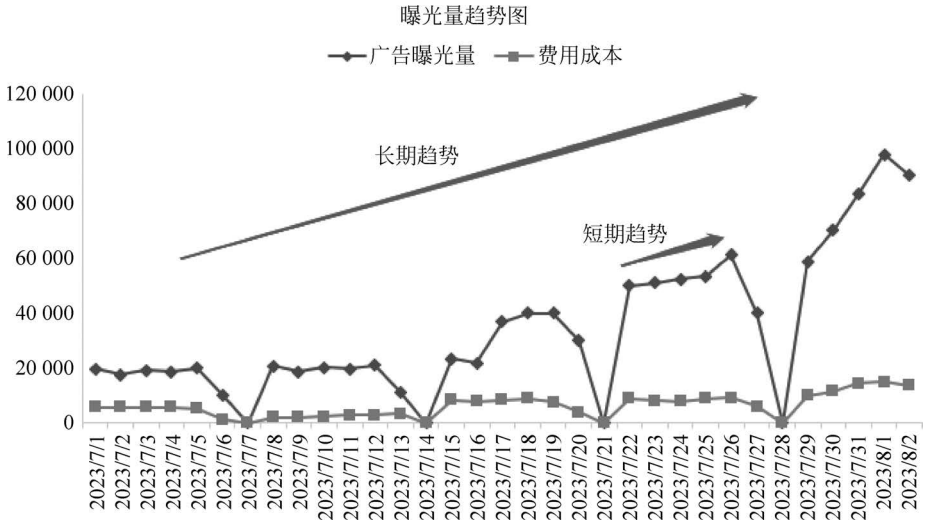

图 5-7　数据趋势图

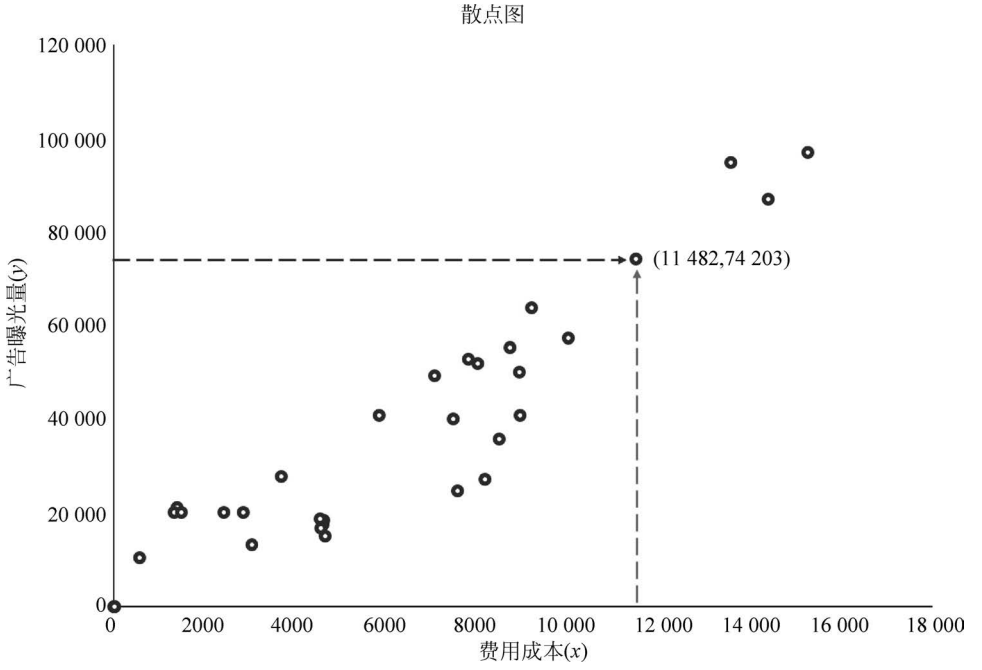

图 5-8　数据散点图

折线图和散点图都清晰地表示了广告曝光量和费用成本两组数据间的相关关系，优点是对相关关系的展现清晰，缺点是无法对相关关系准确地进行度量，缺乏说服力，并且当数

据超过两组时也无法完成各组数据间的相关分析。若要通过具体数字来度量两组或两组以上数据间的相关关系，则需要使用第 2 种方法：协方差。

第 2 种相关分析方法是计算协方差。协方差用来衡量两个变量的总体误差，如果两个变量的变化趋势一致，则协方差就是正值，说明两个变量正相关。如果两个变量的变化趋势相反，则协方差就是负值，说明两个变量负相关。如果两个变量相互独立，则协方差就是 0，说明两个变量不相关。以下是协方差的计算公式：

$$\mathrm{Cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1} \tag{5-1}$$

这是广告曝光量和费用成本间协方差的计算过程和结果，经过计算，我们得到了一个很大的正值，因此可以说明两组数据间是正相关的。广告曝光量随着费用成本的增长而增长。在实际工作中不需要按下面的方法来计算，可以通过 Excel 中的 COVAR()函数直接获得两组数据的协方差值。

协方差只能对两组数据进行相关性分析，当有两组以上数据时就需要使用协方差矩阵。下面是三组数据 x、y、z 的协方差矩阵计算公式：

$$\boldsymbol{C}=\begin{bmatrix}\mathrm{Cov}(x,x) & \mathrm{Cov}(x,y) & \mathrm{Cov}(x,z)\\ \mathrm{Cov}(y,x) & \mathrm{Cov}(y,y) & \mathrm{Cov}(y,z)\\ \mathrm{Cov}(z,x) & \mathrm{Cov}(z,y) & \mathrm{Cov}(z,z)\end{bmatrix} \tag{5-2}$$

协方差通过数字衡量变量间的相关性，正值表示正相关，负值表示负相关，但无法对相关的密切程度进行度量。当我们面对多个变量时，无法通过协方差来说明哪两组数据的相关性最高。要衡量和对比相关性的密切程度，就需要使用另一种方法：相关系数。

第 3 种相关分析方法是计算相关系数。相关系数(Correlation Coefficient)是反映变量之间关系密切程度的统计指标，相关系数的取值在−1 到 1 之间。1 表示两个变量完全线性相关，−1 表示两个变量完全负相关，0 表示两个变量不相关。数据越趋近 0 表示相关关系越弱。以下是相关系数的计算公式。

$$r_{xy}=\frac{S_{xy}}{S_xS_y} \tag{5-3}$$

其中，r_{xy} 表示样本相关系数，S_{xy} 表示样本协方差，S_x 表示 X 的样本标准差，S_y 表示 Y 的样本标准差。由于是样本协方差和样本标准差，因此分母使用的是 $n-1$。样本协方差 S_{xy} 的计算公式：

$$S_{xy}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1} \tag{5-4}$$

样本标准差 S_x 的计算公式：

$$S_x=\sqrt{\frac{\sum(X_i-\overline{X})^2}{n-1}} \tag{5-5}$$

样本标准差 S_y 的计算公式：

$$S_y=\sqrt{\frac{\sum(Y_i-\overline{Y})^2}{n-1}} \tag{5-6}$$

下面是计算相关系数的过程，分别计算变量 X 和 Y 的协方差及各自的标准差，并求得相关系数值为 0.93。0.93 大于 0，说明两个变量间正相关，同时 0.93 非常接近于 1，说明两个变量间高度相关。

在实际工作中，不需要上面这么复杂的计算过程，在 Excel 的数据分析模块中选择相关系数功能，设置好变量 x 和 y 后可以自动求得相关系数的值。从表 5-2 的结果中可以看到，广告曝光量和费用成本的相关系数与我们手动求的结果一致。

表 5-2　广告曝光量和费用成本的相关系数

	广告曝光量(Y)	费用成本(X)
广告曝光量(Y)	1	0.936 447 666
费用成本(X)	0.936 447 666	1

相关系数的优点是可以通过数字对变量的关系进行度量，并且带有方向性，1 表示正相关，−1 表示负相关，可以对变量关系的强弱进行度量，越靠近 0 相关性越弱。缺点是无法利用这种关系对数据进行预测，简单地说就是没有对变量间的关系进行提炼和固化而形成模型。要利用变量间的关系进行预测，需要用到另一种相关分析方法，即回归分析。

第 4 种相关分析方法是回归分析。回归分析(Regression Analysis)是确定两组或两组以上变量间关系的统计方法。回归分析按照变量的数量分为一元回归和多元回归。两个变量使用一元回归，两个以上变量使用多元回归。进行回归分析之前有两个准备工作，第 1 个工作是确定变量的数量，第 2 个工作是确定自变量和因变量。我们的数据中只包含广告曝光量和费用成本两个变量，因此使用一元回归。根据经验，广告曝光量是随着费用成本的变化而改变的，因此将费用成本设置为自变量 x，将广告曝光量设置为因变量 y。

以下是一元回归方程，其中 y 表示广告曝光量，x 表示费用成本。b_0 为方程的截距，b_1 为斜率，同时也表示两个变量间的关系。我们的目标就是 b_0 和 b_1 的值，知道了这两个值也就知道了变量间的关系，并且可以通过这个关系在已知成本费用的情况下预测广告曝光量：

$$y=b_0+b_1x \tag{5-7}$$

这是 b_1 的计算公式，我们通过已知的费用成本 x 和广告曝光量 y 来计算 b_1 的值。

$$b_1=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} \tag{5-8}$$

通过最小二乘法计算 b_1 值的具体计算过程和结果见表 5-3。同时也可以获得自变量和因变量的均值。通过这 3 个值可以计算出 b_0 的值。

表 5-3 变量差值计算表

投放时间	广告曝光量(y)	费用成本(x)	$y_i-\bar{y}$	$x_i-\bar{x}$	$(x_i-\bar{x})(y_i-\bar{y})$	$(x_i-\bar{x})^2$
2016/7/1	18 481	4616	−16 344	−1283	20 966 307	1 645 712
2016/7/2	15 094	4649	−19 731	−1250	24 663 380	1 562 532
2016/7/3	17 619	4600	−17 206	−1299	22 350 167	1 687 435
2016/7/4	16 825	4557	−18 000	−1342	24 154 482	1 800 838
2016/7/5	18 811	4541	−16 014	−1358	21 741 416	1 843 330
2016/7/6	10 430	568	−24 395	−5331	130 058 373	28 424 497
2016/7/7	18	—	−34 807	−5899	205 327 475	34 799 533
2016/7/8	…	…	…	…	…	…
2016/7/9	…	…	…	…	…	…
	$\bar{y}=$ 34 825	$\bar{x}=$ 5899			$\sum(x_i-\bar{x})(y_i-\bar{y})=$ 3 508 979 770	$\sum(x_i-\bar{x})^2=$ 600 651 674
						$b_1=5.841\ 954\ 536$

以下是 b_0 的计算公式，在已知 b_1 和自变量与因变量均值的情况下，b_0 的值很容易计算。

$$b_0=\bar{y}-b_1\bar{x} \tag{5-9}$$

将自变量和因变量的均值及斜率 b_1 代入公式中，求出一元回归方程截距 b_0 的值为 374。这里 b_1 保留两位小数，取值 5.84。

$$b_0=\bar{y}-b_1\bar{x}=34\ 825-5.84\times 5899=374.84 \tag{5-10}$$

在实际工作中不需要进行如此烦琐的计算，Excel 可以帮我们自动完成并给出结果。在 Excel 中使用数据分析中的回归功能，输入自变量和因变量的范围后可以自动获得 b_0 (Intercept)的值 362.15 和 b_1 的值 5.84。这里的 b_0 和之前手动计算获得的值有一些差异，因为前面用于计算的 b_1 值只保留了两位小数。

这里还要单独说明下 R Square 的值 0.87。这个值叫作判定系数，用来度量回归方程的拟合优度。这个值越大，说明回归方程越有意义，自变量对因变量的解释度越高。将截距 b_0 和斜率 b_1 代入到一元回归方程，得到结果见表 5-4。

表 5-4 回归统计表

Multiple R	0.936 447 666
R Square	0.876 934 23
Adjusted R Square	0.873 088 425
标准误差	9481.556 867
观测值	34

方差分析如图 5-9 所示。

方差分析

	df	SS	MS	F	Significance F
回归分析	1	20499300283	20499300283	228.0235638	4.12012E-16
残差	32	2876797460	89899920.62		
总计	33	23376097743			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	362.1503948	2802.246201	0.129235752	0.897980008	-5345.838329	6070.139119	-5345.838329	6070.139119
费用成本(x)	5.841954536	0.386872899	15.10044913	4.12012E-16	5.053920227	6.629988845	5.053920227	6.629988845

图 5-9　方差分析图

这样在回归方程中就获得了自变量与因变量的关系。费用成本每增加 1 元，广告曝光量会增加 374.84 次。通过这个关系可以根据成本预测广告曝光量数据。也可以根据转化所需的广告曝光量来反推投入的费用成本。获得这个方程还有一个更简单的方法，就是在 Excel 中对自变量和因变量生成散点图，然后选择添加趋势线，在添加趋势线的菜单中选中显示公式和显示 R 平方值即可。

$$y = 374.84 + 5.84x \tag{5-11}$$

以上介绍的是两个变量的一元回归方法，如果有两个以上的变量使用 Excel 中的回归分析，则可选中相应的自变量和因变量范围。下面是多元回归方程。

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n \tag{5-12}$$

最后一种相关分析方法是信息熵与互信息。前面我们一直在围绕消费成本和广告曝光量两组数据展开分析。在实际工作中影响最终效果的因素可能有很多，并且不一定都是数值形式。例如站在更高的维度来看之前的数据。广告曝光量只是一个过程指标，最终要分析和关注的是用户是否购买，而影响这个结果的因素也不仅是消费成本或其他数值化指标。可能是一些特征值。例如用户所在的城市、用户的性别、年龄区间分布，以及是否是第 1 次到访网站等。这些都不能通过数字进行度量。

度量这些文本特征值之间相关关系的方法就是互信息。通过这种方法可以发现哪一类特征与最终的结果关系密切。表 5-5 中是我们模拟的一些用户特征和数据。在这些数据中忽略了之前的消费成本和广告曝光量数据，只关注特征与状态的关系。

表 5-5　用户特征和数据表

城市	消费成本/元	广告曝光量	性　别	新用户	年龄分布/岁	状　态
杭州	13 588	78 844	男	是	25～34	未购买
杭州	20 738	120 473	男	否	25～34	未购买
北京	18 949	111 982	女	否	25～34	购买
上海	30 908	167 093	男	是	35～45	未购买
北京	27 822	167 897	男	否	＜25	购买
北京	30 100	185 418	男	否	35～45	未购买
南京	23 317	129 550	女	是	25～34	未购买

续表

城市	消费成本/元	广告曝光量	性　　别	新用户	年龄分布/岁	状　　态
广州	19 057	120 861	女	否	<25	未购买
北京	16 091	101 915	女	否	25～34	购买
⋮	⋮	⋮	⋮	⋮	⋮	⋮

这里直接给出每个特征的互信息值及排名结果。经过计算可知，城市与购买状态的相关性最高，所在城市为北京的用户购买率较高。

由于上述相关系数是根据样本数据计算出来的，所以上述相关系数又称为样本相关系数（用 r 来表示）。若相关系数是根据全部数据计算的，则称为总体相关系数，记为 ρ，但由于存在抽样的随机性和样本较少等原因，通常样本相关系数不能直接用来说明两总体（两变量）是否具有显著的线性相关关系，因此还必须进行显著性检验。相关分析的显著性检验，经常使用假设检验的方式对总体的显著性进行推断。显著性检验的步骤如下。

假设：两个变量无显著的线性关系，即两个变量存在零相关。

构建新的统计量 t，如下：

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{5-13}$$

在变量 X 和 Y 服从正态分布时，该 t 统计量服从自由度为 $n-2$ 的 t 分布。计算统计量 t，并查询 t 分布对应的概率 P 值。最后判断（α 表示显著性水平，一般取 0.05）：①如果 $P<\alpha$，则表示两变量存在显著的线性相关关系。②否则不存在显著的线性相关关系。

简单相关分析的基本步骤如下：对整体流程绘制散点图，然后选择系数类别，再计算相关系数，进行显著性检验，最后进行业务判断，如图 5-10 所示。

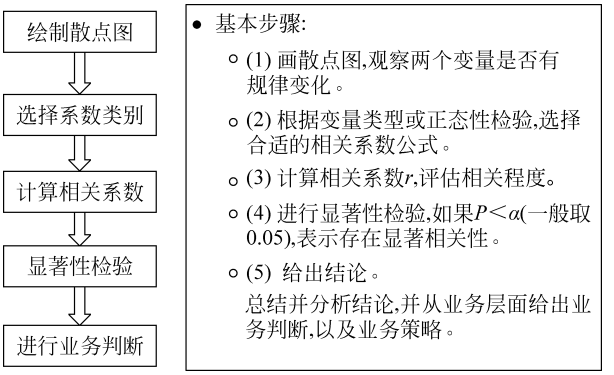

图 5-10　简单相关分析流程图

2. 决策树

在构造可解释模型的过程中，决策树也是一种常用方法，决策树的生成就像在回答是否问题，通过对具体特征值的判断来区分不同的个体，还是拿商家来讲，通过对商家的一些具体属性的区分可以形成一棵树状的结构，如图 5-11 所示，有些分支回答的是“是否”问题，有

些分支对数值是否达到阈值进行判断。当然当用决策树来构造一个可解释的模型时，对每个分支也有概率和分布的问题，这里需要引用信息熵的概念。

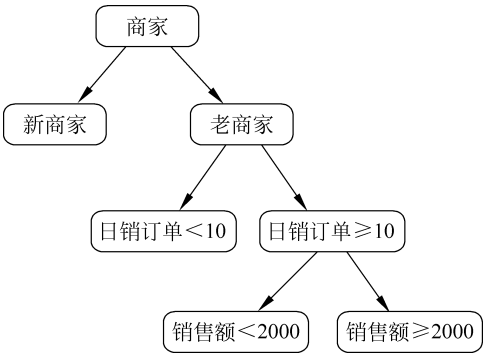

图 5-11　商家属性树状图

熵这个概念来源于信息论，用来度量事物的不确定性，越不确定的事物，它的熵就越大，随机变量 X 的熵的表达式如式(5-14)：

$$H(X)=-\sum_{i=1}^{n}p_i\log p_i \tag{5-14}$$

如抛一枚硬币为事件 T，P(正)$=1/2$，P(负)$=1/2$，则

$$H(T)=-\left(\frac{1}{2}\log\frac{1}{2}+\frac{1}{2}\log\frac{1}{2}\right)=\log 2 \tag{5-15}$$

抛一枚骰子为事件 G，$P(1)=P(2)=\cdots=P(6)=\frac{1}{6}$，则

$$H(G)=-\sum_{1}^{6}\frac{1}{6}\log\frac{1}{6}=\log 6 \tag{5-16}$$

$H(G)>H(T)$，显然掷骰子的不确定性比投硬币的不确定性要高，有了单一变量的熵，很容易推广到多个变量的联合熵，这里给出两个变量 X 和 Y 的联合熵表达式：

$$H(X,Y)=-\sum_{i=1}^{n}p(x_i,y_i)\log p(x_i,y_i) \tag{5-17}$$

有了联合熵，可以得到条件熵的表达式 $H(X|Y)$，条件熵类似于条件概率，它度量了在知道 Y 以后 X 剩下的不确定性，表达式为

$$H(X\mid Y)=-\sum_{i=1}^{n}p(x_i,y_i)\log p(x_i\mid y_i)=\sum_{j=1}^{n}p(y_j)H(X\mid y_i) \tag{5-18}$$

按照定义，$H(X)$度量了 X 的不确定性，条件熵 $H(X|Y)$度量了在知道 Y 以后 X 剩下的不确定性，那么 $H(X)-H(X|Y)$就度量了 X 在知道 Y 以后不确定性的减少程度，这个度量在信息论里称为互信息，记为 $I(X,Y)$。信息熵 $H(X)$、联合熵 $H(X,Y)$、条件熵 $H(X|Y)$和互信息 $I(X,Y)$之间的关系如图 5-12 所示。

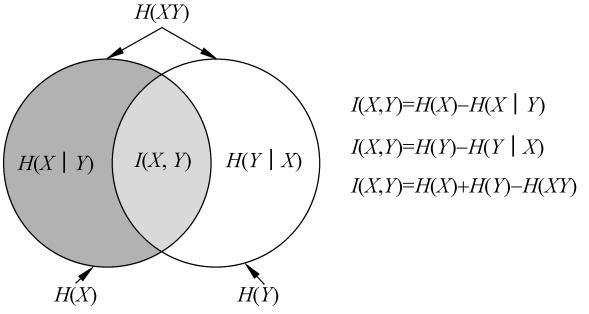

图 5-12　互信息关系图

下面就决策树的典型算法来展示一下决策树在构造可解释模型中的做法，实际上决策树算法有很多，但基本理念都是相同的，即都通过一个个分支去做结构，不同的只不过是数值处理上的方法不同，主要为了应对不同的数值类型，此处就不展开讲解了。

5.2.1 案例 34：决策树的 ID3 算法

在决策树的 ID3 算法中，互信息 $I(X,Y)$ 被称为信息增益。ID3 算法就是用信息增益来判断当前节点应该用什么特征来构建决策树，信息增益越大就越适合用来进行分类，社交网络当中经常会有机器人账号或者水军用来刷评论、恶意攻击，一般平台方会使用一些识别算法来屏蔽掉这些账号，当然在实际业务中可以有非常多的维度来鉴别是不是真实账号，如登录地、登录时长、发送消息间隔等，举个例子，假如主要参考发送内容的日志密度和好友密度及是否真实头像来判断，收集到的信息见表 5-6。

表 5-6 社交网络账号信息表

日志密度	好友密度	是否使用真实头像	账号是否真实
S	S	No	No
S	L	Yes	Yes
L	M	Yes	Yes
M	M	Yes	Yes
L	M	Yes	Yes
M	L	No	Yes
M	S	No	No
L	M	No	Yes
M	S	No	Yes
S	S	Yes	No

设 L、F、H、D 分别表示日志密度、好友密度、是否使用真实头像和账号是否真实，LA、M、S 代表程度上的大、中、小，通过计算有

$$H(D)=-(0.7\log_2 0.7+0.3\log_2 0.3)=0.879$$

$$\begin{aligned}H(D\mid L)=-\Big[&0.3\Big(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}\Big)+0.4\Big(\frac{1}{4}\log_2\frac{1}{4}+\frac{3}{4}\log_2\frac{3}{4}\Big)+\\&0.3\Big(\frac{0}{3}\log_2\frac{0}{3}+\frac{3}{3}\log_2\frac{3}{3}\Big)\Big]=0.603\end{aligned}\tag{5-19}$$

$$I(D,L)=H(D)-H(D\mid L)=0.879-0.603=0.276$$

因此日志密度的信息增益是 0.276，用同样的方法得到是否使用真实头像和好友密度的信息增益分别为 0.033 和 0.553，因此好友密度具有最大的信息增益，也就是说如果从好友密度去区分账号类型，由于大部分非真实账号的好友密度很低，所以决策树的第 1 次分裂就选择好友密度，假设对 LA、M 的好友密度的账号都被认为是真实账号，予以通过。对好友密度是 S 的账号可以对剩下的 3 个属性再进行一次树结构的延伸，见表 5-7，用该表再次递归计算子节点的分裂属性，最终就可以得到整棵决策树。

表 5-7　真实账号信息表

日 志 密 度	是否使用真实头像	账号是否真实
S	No	No
M	No	No
M	No	Yes
S	Yes	No

当然这种算法也有不足之处，像一些连续的特征，例如长度、密度这样的连续值就无法在 ID3 算法中运营，并且由于 ID3 采用信息增益大的特征优先建立决策树的节点，那么在相同条件下取值比较多的特征比取值少的特征信息增益大，例如一个变量有两个值，各为 1/2，另一个变量为 6 个值，均为 1/6，其实都是完全不确定的变量，但是取 6 个值的比取两个值的信息增益大，例子就是此前提到的抛硬币和骰子的区别。从概念上来讲信息增益反映的是给定一个条件以后不确定性减少的程度，必然是分得越细的数据集确定性越高，也就是条件熵越小，信息增益越大，但是在算法上也需要矫正。应对连续和信息增益问题就有 C4.5 的决策树算法。

对于 ID3 中的不能处理连续特征的问题，C4.5 算法的思路是将连续的特征离散化。例如有 n 个样本的连续特征 A，从小到大的排列为 $a_1,a_2,\cdots,a_n$。则 C4.5 取相邻两样本值的平均数，一共可以取到 $n-1$ 个划分点，其中第 i 个划分点 T_i 表示为 $T_i=\dfrac{a_i+a_{i+1}}{2}$。对于这 $n-1$ 个点，分别计算以该点作为二元分类点时的信息增益。选择信息增益最大的点作为该连续特征的二元离散分类点。例如取到的增益最大的点为 a_t，取大于 a_t 为类别 1，小于 a_t 为类别 2。这样就做到了连续特征的离散化。对于第 2 个问题，信息增益作为标准容易偏向于取值较多的特征，C4.5 中使用了信息增益比 $I_R(D,A)$来消除相应的影响。

$$I_R(D,A)=\frac{I(A,D)}{H(A)} \tag{5-20}$$

特征 A 对数据集 D 的信息增益与特征 A 信息熵的比，信息增益越大的特征和划分点，分类效果越好，某特征中值的种类越多，特征对应的特征熵越大，它作为分母，可以校正信息增益导致的问题。回到上面的例子中：

$$\begin{aligned}&I(D,L)=0.276,I(D,F)=0.533,I(D,H)=0.033\\&H(L)=-(0.3\log_2 0.3+0.4\log_2 0.4+0.3\log_2 0.3)=1.571\\&I_R(D,L)=\frac{I(D,L)}{H(L)}=\frac{0.276}{1.571}=0.176\end{aligned} \tag{5-21}$$

同样可得 $I_R(D,F)=0.35$，$I_R(D,H)=1$。因为 F 具有最大的信息增益比，所以第 1 次分裂选择 F 作为分裂属性，再递归使用这种方法计算子节点的分裂属性，最终就可以得到整棵决策树。

在上述的两种方法中，ID3 算法使用信息增益来选择特征，信息增益大的优先选择，在 C4.5 算法中使用信息增益比来选择特征，以减少信息增益容易选择特征值种类多的特征的

问题，但是无论是 ID3 还是 C4.5 都是基于信息论的熵模型的，并且都用到复杂的对数计算，因此也有不使用对数的简化模型，称为分类回归树模型，分类回归树中使用基尼系数来代替信息增益比，基尼系数代表模型的不纯度，基尼系数越小则不纯度越低，特征越好，和信息增益正好是相反的。在这类处理方法中，假设有 K 个类别，第 k 个类别的概率为 p_k，则基尼系数为

$$\text{Gini}(p)=\sum_{k=1}^{K}p_k(1-p_k)=\sum_{k=1}^{K}p_k^2 \tag{5-22}$$

对于给定的样本 D，假设有 K 个类别，第 k 个类别的数量为 C_k，则样本的基尼系数为

$$\text{Gini}(D)=1-\sum_{k=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2 \tag{5-23}$$

特别地，对于样本 D，如果根据特征 A 的某个值 a 把 D 分成 D_1 和 D_2 两部分，则在特征 A 的条件下，D 的基尼系数为

$$\text{Gini}(D,A)=\frac{|D_1|}{|D|}\text{Gini}(D_1)+\frac{|D_2|}{|D|}\text{Gini}(D_2) \tag{5-24}$$

回到上面的例子：

$$\begin{aligned}\text{Gini}(D)&=2\times0.3\times(1-0.3)=0.42\\ \text{Gini}(D,L)&=0.3\left(1-\frac{1^2}{3}-\frac{2^2}{3}\right)+0.4\left(1-\frac{1^2}{4}-\frac{3^2}{4}\right)+0.3\left(1-\frac{3^2}{3}-\frac{0^2}{3}\right)\\ &=0.283\end{aligned} \tag{5-25}$$

同理得 $\text{Gini}(D,F)=0.55$，$\text{Gini}(D,H)=0.4$，因为 L 具有最小的基尼系数，所以第 1 次分裂选择 L 作为分裂属性。再递归使用这种方法计算子节点的分裂属性，最终就可以得到整棵决策树。

还有一些主流的可解释性的模型，例如 LIME，LIME 是 Local Interpretable Model-agnostic Explanations 的缩写，它的关键目标就是让模型以可解释的方式呈现出来，如图 5-13 所示。

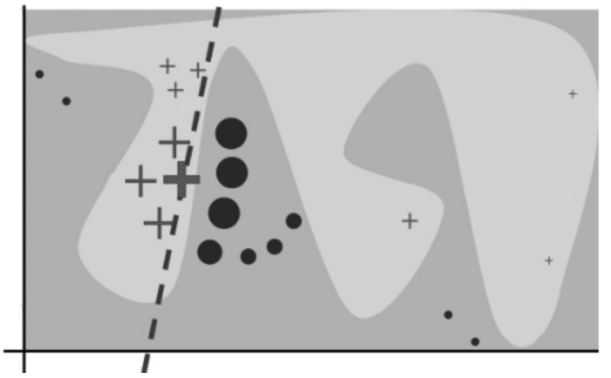

图 5-13　LIME 模型图

不同的颜色区域代表的是模型的 Decision Function(f)针对不同预测对象的输出结果，这一结果相对而言是黑盒的，很明显这是一个比较复杂的模型，并不能用一个 Linear Model 来轻易拟合它，但是可以做到局部可解释，假设图中大的十字叉表示的是被预测对象，那么通过一些扰动方法可以产生若干其他相邻的对象，并利用 f 获取模型对它们的预测结果，然后就可以基于这些数据集得到一个 Locally Faithful 的线性模型了。还有例如 PDP，部分依赖图简称 PDP 图，能够展现出一个或两个特征变量对模型预测结果影响的关系函数：近似线性关系、单调关系或者更复杂的关系。PDP 的分析步骤是训练一个机器学习模型，用代替列，利用训练的模型对这些数据进行预测，求所有样本的预测的平均值，遍历特征的所有不同值，PDP 的 x 轴为特征的各个值，而 y 轴是对应不同值的预测平均值。例如数据集包含 3 个样本，每个样本各有 3 个特征 A、B、C，想要知道特征 A 是如何影响预测结果的，假设特征 A 一共有 3 种类型，训练数据的其他特征保持不变，将特征 A 依次修改为各个特征值，然后对预测求平均值，最后 PDP 需要的是针对不同特征值的平均预测值。

5.2.2　案例 35：Shapley 值法

还有一种方法是 Shapley 值法，这种方法是指所得与自己的贡献相等，是一种分配方式，普遍用于经济活动中的利益合理分配等问题，最早由美国洛杉矶加州大学教授罗伊德·夏普利(Lloyd Shapley)提出，Shapley 值法的提出给合作博弈在理论上的重要突破及其以后的发展带来了重大影响。简单地来讲，Shapley 值法就是使分配问题更加合理，用于为分配问题提供一种合理的方式。例如 n 个人合作创造了 $v(N)$的价值，如何对所创造的价值进行分配，假设全集 $N=\{x_1,x_2,\cdots,x_n\}$有 n 个元素 x_i，任意多个人形成的子集 $S\subseteq N$，有 $v(S)$表示 S 子集中所包括的元素共同合作所产生的价值。最终分配的价值(Shapley Value)$\phi_i(N,v)$其实是求累加贡献的均值。

例如 A 单独工作产生价值 $v(\{A\})$，后加入 B 之后共同产生价值 $v(\{A,B\})$，那么 B 的累加贡献为 $v(\{A,B\})-v(\{A\})$。对于所有能够形成全集 N 的序列，求其中关于元素 x_i 的累加贡献，然后取均值即可得到 x_i 的 Shapley 值，但枚举所有序列可能性的方式的效率不高，注意到累加贡献的计算实际为集合相减，对于同样的集合当计算次数过多时效率低下。公式中 S 表示序列中位于 x_i 前面的元素集合，进而 $N\backslash S\backslash\{x_i\}$表示的是位于 x_i 后面的元素集合，而满足只有 S 集合中的元素位于 x_i 之前的序列总共有$|S|!(|N|-|S|-1)!$ 个，其内序列中产生的 x_i 累加贡献都是 $v(S\cup\{x_i\})-v(S)$；最后对所有序列求和之后再取均值。假设特征全集为 F，则有

$$\phi_i(v)=\sum_{S\subseteq N\{i\}}\frac{|S|!(|N|-|S|-1)!}{|N|!}(v(S\cup\{x_i\})-v(S)) \tag{5-26}$$

求得每维特征的 Shapley 值，值越大对目标函数的影响越正向，值越小对目标函数的影响越

负向。举个例子，若 $N=\{1,2,3\}$，$v(\{1\})=0$，$v(\{2\})=0$，$v(\{3\})=0$，$v(\{1,2\})=90$，则 $v(\{1,3\})=80$，$v(\{2,3\})=70$，$v(\{1,2,3\})=120$，如图 5-14 所示。

	1	2	3
1←2←3	0	90	30
1←3←2	0	40	80
2←1←3	90	0	30
2←3←1	50	0	70
3←1←2	80	40	0
3←2←1	50	70	0
总数	270	240	210
Shapley值	45	40	35

图 5-14　Shapley 值图