



第 1 章

直 销

SPSS 中的直销模块，将多种统计方法组合使用，用来识别目标客户、对客户进行分类、生成潜在客户概要、比较活动效果、选择最有可能购买的联系人、标识最佳响应邮政编码、应用模型进行预测等。

1.1 识别最佳联系人

在 SPSS 中，识别最佳联系人通过 RFM 模型来实现。RFM 是三个英文单词首字母的缩写，其中 R 代表 recency（最近一次消费），F 代表 frequency（消费频率），M 代表 monetary（消费金额）。RFM 分析的基本逻辑是，最近购买产品较多、购买频率较高且消费金额较大的客户是需要维护的客户。

在菜单中选择“直销 / 选择方法”，进入“直销”模块选择方法界面，如图 1.1 所示。



图 1.1 选择方法界面

图 1.1 选择方法界面包括三个部分。第一部分为“了解我的联系人”，包括“帮助标识我的最佳联系人（RFM 分析）”“将我的联系人分段到聚类”“生成对产品做出响应的我的联系人的概要文件”；第二部分为“改进我的市场营销活动”，包括“标识最佳响应邮政编码”“选择最有可能购买的联系人”“比较活动效果”；第三部分为“对我的数据评分”，指根据“选择最有可能购买的联系人”和其他程序进行预测。

本节首先介绍第一部分中的“帮助标识我的最佳联系人（RFM 分析）”，该功能包括两种数据格式，具体要求各不相同。

1.1.1 交易数据格式

案例使用 SPSS 自带数据集 rfm_transtraction.sav，该数据集包含五个变量，如图 1.2 所示。

	ID	Product Line	ProductNumber	Date	Amount
1	955	C-300	384	16-Jul-2006	51
2	607	A-100	194	12-May-2005	27
3	791	A-100	131	18-Nov-2005	29
4	18	D-400	421	05-Aug-2004	123
5	65	A-100	130	01-Jun-2004	31
6	380	C-300	305	13-Aug-2004	70
7	209	E-500	504	29-Oct-2005	178

图 1.2 rfm_transtraction.sav 文件

其中 ID 为客户编号，ProductLine 为生产线，ProductNumber 为产品编号，Date 为购买日期，Amount 为购买金额。

在图 1.1 中选择“了解我的联系人帮助标识我的最佳联系人(RFM 分析)”选项，单击“继续”按钮，打开“RFM 分析：数据格式”对话框，如图 1.3 所示。



图 1.3 RFM 分析：数据格式

从图 1.3 可知：交易数据是指每行包含单个交易的数据，交易将按客户标识进行汇总以供分析；客户数据是指每行包含单个客户的数据，数据按已交易中的客户进行组合。

选择“交易数据”选项，单击“继续”按钮，打开“交易数据的 RFM 分析：变量”对话框，如图 1.4 所示。



图 1.4 交易数据的 RFM 分析：变量

图 1.4 中包括四个选项卡，第一个为“变量”选项卡。左侧为文件中所包含的变量，右侧分别包括“交易日期”“交易金额”“客户标识”和“摘要方法”。分别将 Purchase Date、Purchase Amount、Customer ID 选入相应的列表框。“摘要方法”用于选择汇总每个客户交易金额的方法，包含总计、最大值、中位数和均值四个统计量。在本案例中，交易金额是单次购买产品的单价，所以此处选择默认值“总计”。

在图 1.4 中，单击“分箱化”选项卡，可以设置分箱选项，如图 1.5 所示。

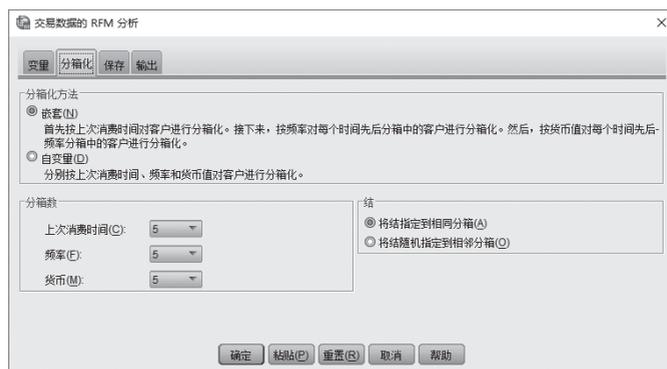


图 1.5 交易数据的 RFM 分析：分箱化

图 1.5 中，“分箱化”有两种方法，一种是“嵌套”，另一种是“自变量”。从统计学角度来说，“嵌套”就是复合分组，“自变量”就是平行分组。“分箱数”可以设置每个指标划分的组数，每个指标最多可划分为 9 组，最少可划分为 2 组，默认划分为 5 组。“结”指相同的数值。“将结指定到相同分箱”指将相同的数值划分到同一组；“将结随机指定到相邻分箱”指将相同的数值随机分配到相邻的两组。

在图 1.5 中，单击“保存”选项卡，可以设置保存选项，如图 1.6 所示。



图 1.6 交易数据的 RFM 分析：保存

在图 1.6 的“保存”选项卡，可以设置保存的变量，包括指标的原始变量、对原始变量分箱后的变量（以“_得分”为后缀名命名的变量）以及 RFM 得分，其中 RFM 得分为必须项，不能更改，其他变量则可以选择是否保存。另外还可以指定变量的保存位置，可以“创建新数据集”，也可以“写入新数据文件”。

在图 1.7 “输出” 选项卡，对于已分箱化数据，可以选择输出“按上次消费时间和频率绘制的平均币值热图”“分箱计数图表”和“分箱计数表格”。对于未分箱化数据，可以选择输出“直方图”和“变量对散点图”。默认只输出已分箱化数据的“按上次消费时间和频率绘制的平均币值热图”“分箱计数图表”，此处选择输出所有图表。



图 1.7 交易数据的 RFM 分析：输出

单击“确定”按钮后，即可得到系列输出结果，如图 1.8 所示。

ID	最近日期	交易计数	金额	上次消费时间得分	频率得分	消费金额得分	RFM得分
1	04-Sep-2006	5	485.00	4	3	4	434
2	10-Nov-2005	4	350.00	2	2	2	222
3	04-Jun-2005	2	233.00	1	2	4	124
4	18-Aug-2006	7	936.00	4	4	5	445
5	07-Jul-2006	3	359.00	4	1	5	415
6	16-Jul-2006	3	249.00	4	1	4	414
7	15-Feb-2006	7	1089.00	2	5	5	255
8	21-Aug-2006	4	423.00	4	2	4	424
9	31-Aug-2006	7	689.00	4	4	4	444

图 1.8 产生的新数据集

由于在图 1.6 选择了“创建新数据集”，因此会产生一个以默认名称“数据集 2”命名的新数据集。该数据集按 ID 进行排序，以 ID 为 1 的个案为例，其最近消费日期为 04-Sep-2006，共有 5 个交易记录，交易总金额为 485。上次消费时间得分为 4（即被分在第 4 组，消费日期越近等级越高），频率得分为 3（即被分在第 3 组，购买频率越高等级越高），消费金额得分为 4（即被分在第 4 组，消费金额越高等级越高），这 3 个数组合在一起即为最终的 RFM 得分 434。RFM 分值越高，客户价值越大。

图 1.9 为 RFM 分箱计数图。

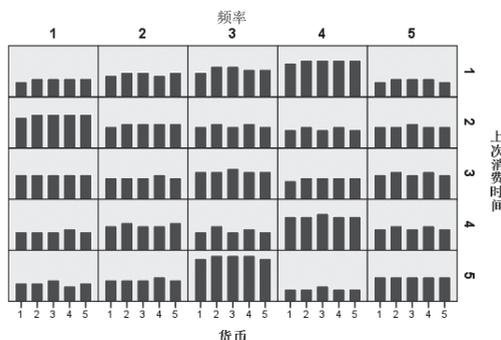


图 1.9 RFM 分箱计数图

RFM 分箱计数图显示根据分箱化方法设定的分箱数的分布。纵轴“上次消费时间”显示按最近一次消费时间将客户分为 1~5 级，横轴“频率”显示按购买频率将客户分为 1~5 级。根据最近一次消费时间和频率将客户分为 25 个组别。例如左上角的图表示最近一次消费级别为 1、购买频率级别也为 1 的客户群，右下角则表示最近一次消费级别为 5、购买频率级别也为 5 的客户群。在每个组别中，又按照消费金额把客户分为 1~5 级，图中显示为 5 个直条，从左到右表示消费金额级别依次增高，高度则代表相应直条所代表的 RFM 得分的客户数量的多少。如左上角的图反映了 RFM 得分为 111~115 的客户数量的分布，右下角的图反映了 RFM 得分为 551~555 的客户数量的分布。

可以结合交叉列表得到每个直条的具体数值，如图 1.10 所示。

上次消费时间得分		货币得分					总计
		1	2	3	4	5	
1	频率得分 1	5	6	6	6	6	29
	2	7	8	8	7	8	38
	3	8	10	10	9	9	46
	4	11	12	12	12	12	59
	5	5	6	6	6	5	28
	总计	36	42	42	40	40	200
2	频率得分 1	10	11	11	11	11	54
	2	7	8	8	8	8	39
	3	7	8	8	8	7	37
	4	6	7	6	7	6	32
	5	7	7	8	7	7	36
	总计	37	41	40	41	39	198
3	频率得分 1	8	8	8	8	8	40
	2	7	7	7	8	7	36
	3	9	9	10	9	9	46
	4	6	7	7	7	7	34
	5	8	9	8	9	8	42
	总计	38	40	40	41	39	198
4	频率得分 1	6	6	6	7	6	31
	2	8	9	8	8	9	42
	3	6	8	6	7	6	33
	4	11	11	12	11	11	56
	5	7	8	7	8	7	37
	总计	38	42	39	41	39	199
5	频率得分 1	6	6	7	5	6	30
	2	7	7	7	8	7	36
	3	14	15	15	15	14	73
	4	4	4	5	4	4	21
	5	8	8	8	8	8	40
	总计	39	40	42	40	39	200
总计	频率得分 1	35	37	38	37	37	184
	2	36	39	38	39	39	191
	3	44	50	48	48	45	235
	4	38	41	42	41	40	202
	5	35	38	37	38	35	183
	总计	188	205	203	203	196	995

图 1.10 交叉列表

交叉列表中的数值对应了图 1.9 中各个直条的高度。如交叉表第一行的数值对应了图 1.9 中左上角直条的高度，即 RFM 得分为 111 的有 5 个，得分为 112、113、114、115 的均有 6 个；第二行数值对应了图 1.9 中第一行第二个图，即 RFM 得分为 121、122、123、124、125 的分别有 7、8、8、7、8 个。依次类推，可知图 1.9 中右下角直条的高度均为 8。根据已分箱化数据还可以输出热图，如图 1.11 所示。

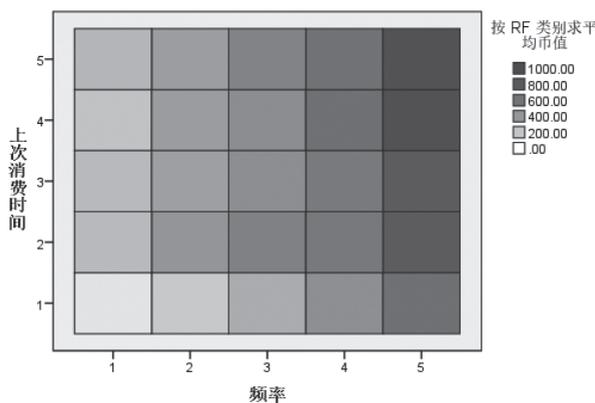


图 1.11 RFM 热图

RFM 热图的纵轴为“上次消费时间”，横轴为“频率”，均按分箱设定为 5 级。按照两个指标划分了 25 个客户群。对每个客户群的消费金额计算均值，并用颜色深浅表示消费金额均值的大小。默认颜色越深表示该客户群的消费金额的均值越高。

由于在图 1.7 中也选择了“未分箱化数据”，此处也会输出直方图和散点图。直方图如图 1.12 所示。

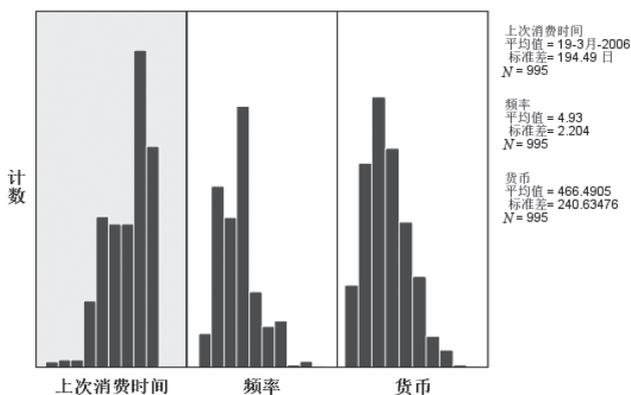


图 1.12 直方图

直方图显示了每个指标的均值、标准差和观测数。更为重要的是 SPSS 还给出了对各直方图横轴的注释：每个直方图的水平轴始终采用左侧为较小值、右侧为较大值的排序方法。但对于上次消费时间，图表的解释依赖于上次消费时间测量的类型：日期或时间间隔。对于日期，左侧条代表更“早”的值（即较远日期比较近日期的值更小）；对于时间间隔，左侧条代表更“近”的值（即时间间隔越小，交易离现在越近）。

RFM 散点图如图 1.13 所示。

图 1.13 显示了上次消费时间、频率、货币之间的两两散点图。对“上次消费时间”的解释依赖于最近一次消费时间的类型：日期或时间间隔。对于日期，越接近原点的点代表离现在越远的日期；对于时间间隔，越接近原点的点代表购买日期离现在越近的值。

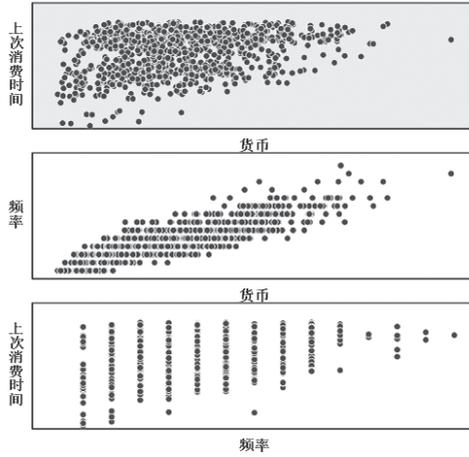


图 1.13 RFM 散点图

1.1.2 客户数据格式

打开 SPSS 自带数据集 rfm_customers.sav，该数据集为客户数据格式，如图 1.14 所示。

ID	TotalAmount	MostRecent	NumberOfPurchases	PurchaseInterval
1	1313.00	05/17/2006	10	229.00
2	1230.00	09/21/2005	11	467.00
3	1194.00	08/11/2006	13	143.00

图 1.14 数据集 rfm_customers.sav

该数据集包含 5 个变量：ID（客户编号）、TotalAmount（客户消费总金额）、MostRecent（最近一次购买日期）、NumberOfPurchases（购买次数）和 PurchaseInterval（最近一次购买时间间隔）。

在图 1.3 中选择“客户数据”选项，打开“客户数据的 RFM 分析”对话框，如图 1.15 所示。



图 1.15 客户数据的 RFM 分析

在图 1.15 “变量”选项卡中将 PurchaseInterval 选入“交易日期或间隔”，NumberOfPurchases 选入“交易数”，TotalAmount 选入“金额”，ID 选入“客户标识”。“分箱化”“保存”“输出”这 3 个选项卡和交易数据的 RFM 分析基本相同，在此不再介绍。

客户数据的 RFM 分析结果和交易数据的 RFM 分析结果类似，二者主要是数据格式的不同，在此不再赘述。

1.2 客户分类

直销模块中的“将我的联系人分段到聚类”可实现对客户的聚类，即根据客户购买行为及客户人口社会学特征将客户分类。本节使用 SPSS 自带数据集 dmdata.sav，如图 1.16 所示。

名称	类型	宽度	小数	标签	值	缺失	列	对齐	测量	角色
ID	字符串	5	0	Customer ID	无	无	10	右	名义(N)	输入
Responded	数值	8	0	Responded to t...	{0, No}...	无	11	右	名义(N)	输入
Previous	数值	8	0	Responded to p...	{0, No}...	无	10	右	名义(N)	输入
ControlPack...	数值	8	0	Control Package	{0, Control}...	无	16	右	名义(N)	输入
PostalCode	字符串	15	0	Postal Code	无	无	12	左	名义(N)	输入
Age	数值	8	0		无	无	10	右	度量	输入
Income	数值	8	0	Income categor...	{1, <25}...	无	10	右	有序(O)	输入
Education	数值	8	0		{1, Some hi...	无	11	右	有序(O)	输入
Reside	数值	8	0	Years at curren...	无	无	10	右	度量	输入
Gender	数值	8	0		{0, Male}...	无	10	右	名义(N)	输入
Married	数值	8	0		{0, No}...	无	10	右	名义(N)	输入
Children	数值	8	0		无	无	10	右	度量	输入
Region	数值	8	0		{1, North}...	无	10	右	名义(N)	输入
Sequence	数值	8	2		无	无	10	右	度量	输入

图 1.16 数据集 dmdata.sav

8 该数据集共包括 14 个变量，分别为 ID、Responded（做出响应）、Previous（上一个）、ControlPack（控制包装）、PostalCode（邮政编码）、Age（年龄）、Income（收入）、Education（教育程度）、Reside（目前居住年限）、Gender（性别）、Married（已婚）、Children（子女）、Region（地区）、Sequence（序列）。对于不容易理解的变量名，可以看变量标签。如“Responded”指是否对测试产品做出响应，“Previous”指是否对上一个产品做出响应。

在图 1.1 中选择“了解我的联系人/将我的联系人分段到聚类”选项，打开“聚类分析”对话框，如图 1.17 所示。



图 1.17 聚类分析：字段

图 1.17 字段选项卡左侧为候选变量集，右侧“创建段”指选择用于聚类分析的变量，ID、Responded、Previous、ControlPack、PostalCode、Sequence 对聚类没有帮助，将其他变量作为聚类变量。

单击“设置”选项卡可设置其他参数，如图 1.18 所示。



图 1.18 聚类分析：设置

图 1.18 主要包括三部分内容：一是是否显示图表和表格，默认显示；二是“段成员”，应为汉化有误，应翻译为“类成员”，默认保存类成员，默认变量名为“聚类组 1”，读者可自行设置需要的变量名；三是“段的数量”，应翻译为“类的数量”，可以自动确定，也可以指定固定值。如选择自动确定，默认最大值为 15；若选择指定固定值，默认为 5。选择默认设置，单击“运行”按钮可得到输出结果，如图 1.19 所示。

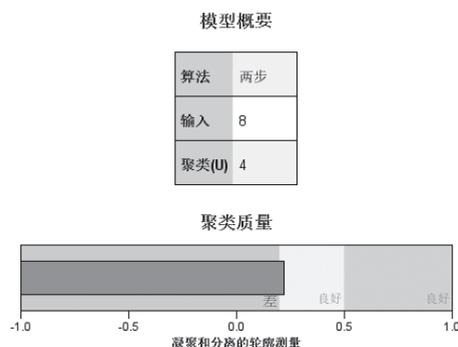


图 1.19 输出结果

图 1.19 显示使用了 8 个变量（输入），将客户分成了 4 类。双击图 1.19 “输出结果”界面，打开“模型查看器”界面，如图 1.20 所示。

图 1.20 模型查看器左侧与图 1.19 相同，右侧显示了聚类大小。从右侧饼状图可以看出，客户群被分成了四类，每类所占比重分别为 29.4%、22.4%、21.8%、26.3%，最小类包含 2182 个客户，最大类包含 2944 个客户。

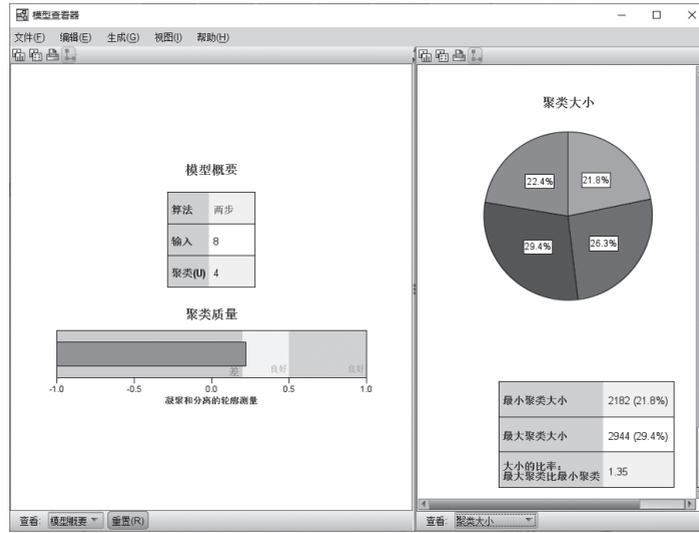


图 1.20 模型查看器

在图 1.20 右侧下方，单击“查看”下拉列表框，除“聚类大小”外，还可以选择“预测变量重要性”“单元格分布”和“聚类比较”。首先选择“预测变量重要性”，可得到各个变量的重要性，如图 1.21 所示。

从图 1.21 可知，预测变量重要性从大到小依次为 Children、Age、Married、Gender、Income category (thousa...)、Region、Years at current residence、Education，其中 Years at current residence、Education 基本没有预测力，去掉对聚类结果影响不大。

当选择“单元格分布”，左侧必须同时选择“聚类”，并需选中一个单元格，如图 1.22 所示。

由于在左侧选中了“Age”，右侧显示了年龄的分布，包括总体年龄分布和第 3 类客户群年龄的分布（左侧最上面列出了类序号为 3、2、4、1）。

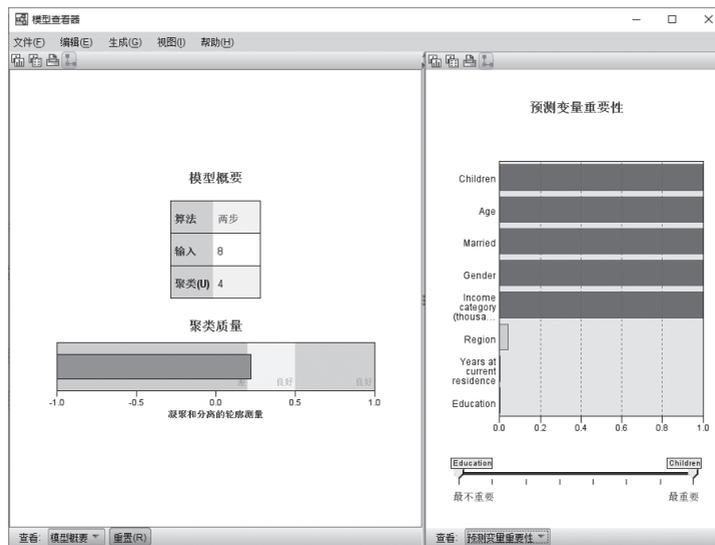
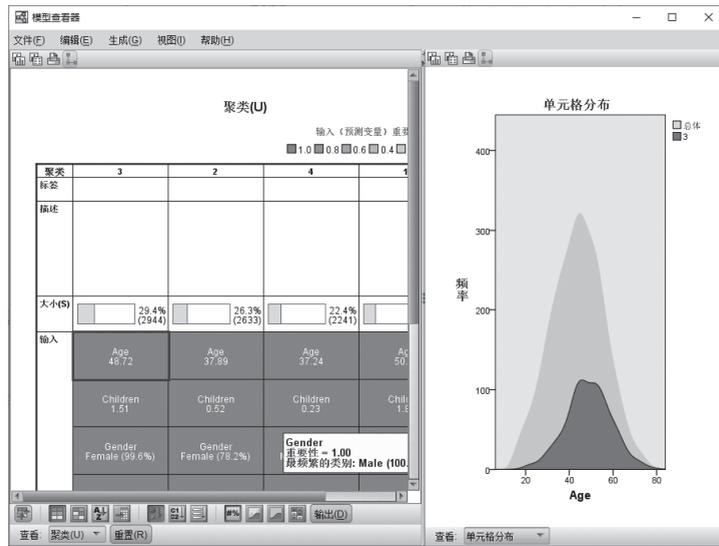


图 1.21 预测变量重要性



将图 1.22 左侧表格放大显示，如图 1.23 所示。

输入（预测变量）重要性
■ 1.0 ■ 0.8 ■ 0.6 ■ 0.4 ■ 0.2 ■ 0.0

聚类	3	2	4	1
标签				
描述				
大小(S)	29.4% (2944)	26.3% (2633)	22.4% (2241)	21.8% (2182)
输入	Age 48.72	Age 37.89	Age 37.24	Age 50.91
	Children 1.51	Children 0.52	Children 0.23	Children 1.86
	Gender Female (99.6%)	Gender Female (78.2%)	Gender Male (100.0%)	Gender Male (100.0%)
	Income category (thousands)	Income category (thousands)	Income category (thousands)	Income category (thousands)
	Married Yes (56.1%)	Married No (60.3%)	Married No (100.0%)	Married Yes (100.0%)
	Region East (26.3%)	Region West (28.4%)	Region South (28.5%)	Region West (27.5%)
	Years at current residence	Years at current residence	Years at current residence	Years at current residence
	Education	Education	Education	Education

图 1.23 聚类分布

图 1.23 从左往右依次列出第 3 类、第 2 类、第 4 类、第 1 类客户群在各个预测变量上的特征。根据预测变量在各群的分布可以归纳出各个客户群的特点。如第 1 类客户：男性为主，年龄偏大，已婚且有多个子女，收入较高。第 2 类客户：女性为主，年龄较小，子女较少，收入较低。第 3 类客户：女性为主，年龄偏大，已婚且有多个子女，收入较高。第 4 类客户：男性为主，年龄较小，子女较少，收入较低。

双击图 1.23 中的“描述”右侧的单元格，可以将各个客户群的特点输入。双击“标签”右侧的单元格，可以为各个客户群添加标签。

也可以采用分析菜单中分类菜单项的各种聚类方法进行分类，但即使也分成 4 类，分类结果也不一定和本案例相同。“将我的联系人分段到聚类”像一个“黑箱”，并没有说明聚类的方法。聚类方法不一样，聚类结果就有可能不同。

1.3 生成潜在客户概要文件

潜在客户概要文件指使用先前测试活动的结果创建描述概要文件。该文件常被用于定位未来市场营销活动中的潜在客户特性，以提高公司市场活动的客户响应率。其基本流程是市场部门在客户数据库中随机抽取一定百分比的客户做测试，测试结束后，整理出包含客户响应率及客户信息的数据，以此生成潜在客户的概要文件，指导后续的市场活动。

本节仍然使用 SPSS 自带数据集 dmdata.sav。在图 1.1 中选择“生成对产品做出响应的我的联系人的概要文件”，打开“潜在客户概要文件”对话框，如图 1.24 所示。

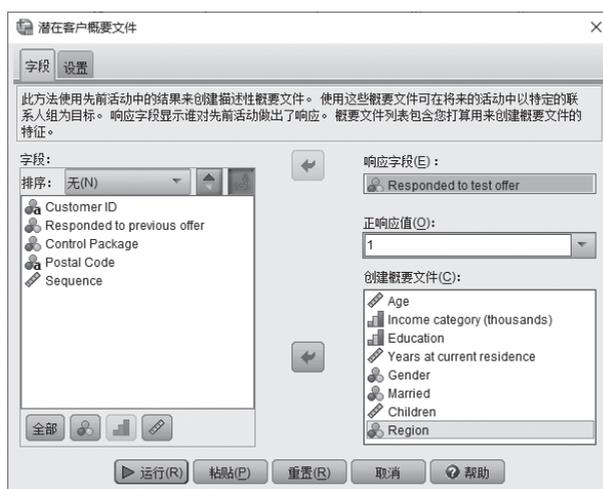


图 1.24 潜在客户概要文件：“字段”选项卡

在图 1.24 “字段”选项卡中，将 Responded to test offer（对测试产品做出响应）选入“响应字段”，正响应值选择“Yes”（是），此时显示为 1。将尽可能多的描述顾客特征的变量选入“创建概要文件”。单击“设置”选项卡，进一步设置参数，如图 1.25 所示。



图 1.25 潜在客户概要文件：设置

图 1.25 中“最小概要文件组大小”实际上是指聚类中每一个客户群所包含的最小人数，默认为 100 人。数值越大，划分的组越少；数值越小，划分的组越多。客户群包含的人数越多，结果越可靠；包含的人数越少，越不具有代表性，结果越不可靠。

勾选“在结果中包括最小响应率阈值信息”复选框，可以指定一个目标响应率，本例设为 7%。选择此项后，输出结果中会包含一个响应率表及累计响应率图，在响应率表中通过颜色编码显示哪些概要文件（实际就是目标客户群）超过设定的目标响应率。

单击“运行”按钮后得到输出结果，其中图 1.26 为响应率表。

目标类别：Yes

响应率				
数字	概要文件			
	描述	组大小	响应率	累积响应率
1	Region = "West","South","East" Gender = "Female" Married = "No"	379	9.23%	9.23%
2	Region = "West","South","East" Gender = "Female" Married = "Yes"	299	5.02%	7.37%
3	Region = "West","South","East" Gender = "Male"	722	4.71%	6.00%
4	Region = "North"	517	2.51%	5.06%

绿色（浅色区域）：满足目标响应率
红色（深色区域）：不满足目标响应率

图 1.26 响应率表

图 1.26 响应率表中显示了每个概要文件的信息。概要文件描述中只显示了为模型提供显著贡献的那些变量的特征，不包括那些对模型没有显著贡献的变量。因此在图 1.24 中的“创建概要文件”框中选中无关变量并没有什么影响，多选中变量不比少选中变量要好。概要文件按响应率降序显示，响应率是做出正面响应（购买产品）的客户的百分比。指定的目标响应率为 7.00%。绿色行的累积响应率大于或等于 7.00%，并且红色行的累积响应率小于 7.00%。尽管绿色区域中某些概要文件组可能有个别响应率小于 7.00%，但此处的累积响应率仍然大于或等于 7.00%。

如果针对西部，南部和东部的女性开展后续市场活动，也就是图 1.26 中绿色区域响应率大于 7% 的第一组和第二组类型客户，会得到比预期目标响应率 7.00% 更高的响应率。如果市场部门对目标响应率要求更高，则可以针对西部，南部和东部的未婚女性（第一组类型客户）开展后续市场活动，将会得到 9.23% 的响应率。

图 1.27 为累积响应率图。累积响应率图比响应率表反映的信息较为粗略。概要文件是按响应率降序排列，所以累积响应率随着概要文件增大而下降。累积响应率和目标响应率相交于 2 和 3 之间，因此针对第一组和第二组类型客户开展市场活动可以达到 7% 的目标响应率。

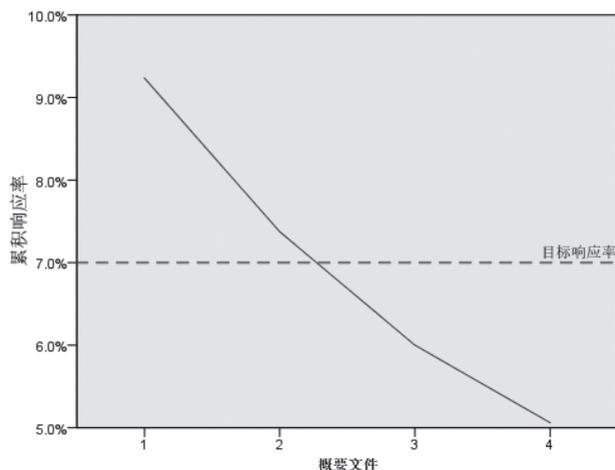


图 1.27 累积响应率图

1.4 基于邮政编码响应率分析的客户选择

当企业计划推出新产品时，向客户发送广告邮件，应用邮政编码响应率分析客户的响应数据，找出对新产品响应最高的客户所在区域，从而定位新产品的重点营销区域。

仍然使用 SPSS 自带数据集 `dmdata.sav`，选择“直销 / 选择方法 / 改进我的市场营销活动 / 标识最佳响应邮政编码”，打开“邮政编码响应率”对话框，如图 1.28 所示。

图 1.28 相对简单，只需设置“响应字段”和“邮政编码字段”。将 `Responded to test offer` 选入“响应字段”，正响应值高为“`Yes`”；将 `Postal Code` 选入“邮政编码字段”。单击“设置”选项卡，进行相应参数设置，如图 1.29 所示。

图 1.29 分为四个部分。第一部分为邮政编码分组方式，默认根据邮政编码的完整值进行分组，将具有相同邮政编码的数据作为一组计算响应率。也可以选择“前 3 个数字或字符”“前 5 个数字或字符”或“前 N 个数字或字符”进行分组，本例选择“前 3 个数字或字符”。



图 1.28 邮政编码响应率：字段



图 1.29 邮政编码响应率：设置

如果邮政编码字段为数值型，那么数值基于邮政编码的前 N 个数字进行分组，且需要在第二部分的数值邮政编码格式中指定原始邮政编码是如何记录的，即指定邮政编码原始值的位数。如果各地区邮政编码包含的位数不同，则应该指定为最大可能位数。比如某些地区的邮政编码包含 5 位数字，某些地区的邮政编码包含 9 位数字，则需指定 9 为邮政编码的位数。本例邮政编码为字符型，所以本部分未被激活，呈灰色显示。读者可以试着将邮政编码改为数值型，再进行比较。

第三部分用于设定响应率和容量分析。如果不勾选“响应率和容量分析”，此部分将不会被激活。在最低可接受响应率部分，默认无最小值，也可以选择目标响应率，在其右

侧文本框输入相应数据，比如 3.5。“通过公式计算收支平衡率”是基于下面公式计算最低累积响应率：

$$\text{最低累积响应率} = (\text{邮寄包装成本} / \text{每次响应的净收入}) \times 100$$

最大联系人数量默认为所有联系人，也可以设定联系人百分比或联系人数量。如果指定了最大联系人数量，生成的响应率表将通过颜色编码区分客户数量是否超过累积最大联系人数量，同时响应率图将在指定的最大联系人数量位置画出标线。

当同时指定最低可接受响应率和最大联系人数量时，表格颜色编码将基于二者中先被满足的条件。

第四部分用于设定是否将邮政编码响应率保存至 Excel，默认不保存。

为了显示更多的输出结果，本例将目标响应率设为 3.5%，联系人百分比设定为 50%。单击“运行”按钮后可得到输出结果。

图 1.30 为响应率表。响应率表按十分位数的降序排序（前 10%、前 20%，等等）摘要列出结果。累积响应率是当前和前面所有行中正响应的组合百分比。由于结果以响应率的降序显示，因此这是当前十分位数和前面所有较高响应率的十分位数的组合响应率。由于十分位数排序包含在新的数据集中，因此可以有效地确定满足特定累积响应率要求的邮政编码。在新的数据集中确定十分位数排序的字段称为“排序”，其中 1= 前 10%、2= 前 20%，等等。

百分位	响应率	联系人	累积响应率	累计联系人
顶端 10%	22.00%	100	22.00%	100
顶端 20%	13.48%	178	16.55%	278
顶端 30%	10.20%	196	13.92%	474
顶端 40%	7.25%	207	11.89%	681
顶端 50%	5.24%	229	10.22%	910
顶端 60%	3.51%	114	9.47%	1024
顶端 80%	0.00%	893	5.06%	1917

每个邮政编码的响应率列在由过程自动创建的新数据集中。
绿色（浅色区域）：满足目标响应率，且符合容量。
红色（深色区域）：不满足最低响应率和 / 或超出容量。

图 1.30 响应率表

图中的颜色编码基于首先被满足的设定条件达到阈值。绿色区域（浅色区域）显示满足设定的目标响应率和最大联系人数量条件的客户，红色区域（深色区域）显示不满足条件的客户，应重点关注绿色区域客户。表中可以看出对营销活动有响应的客户数量为 1917，指定的最大联系人数量为 958（ 1917×0.5 ）。最大联系人数量在最低目标响应率之前出现（累积联系人 $910 < 958 < 1024$ ，响应率 $0 < 3.5\% < 3.51\%$ ）。所以绿色和红色分界线画在顶端 50% 和顶端 60% 之间。绿色区域表示该部分客户组满足最低目标响应率且累积客户数量小于最大联系人数量。在本例中，绿色行的累积联系人数量小于或等于 958，累积响应率均大于或等于 3.50%。

在判断目标响应率和最大联系人数量哪个先出现时，一定要注意小于最大联系人数量为满足要求，大于最大联系人数量不符合要求。而目标响应率则是大于最低目标响应率为

满足要求，小于最低目标响应率为不符合要求。虽然顶端 60% 部分的客户组已满足目标响应率大于 3.5%，但累积客户数量已大于最大联系人数量，所以被标为红色。

邮政编码响应率图如图 1.31 所示。图 1.30 响应率表按十分位数的降序排列出累积响应率，因此对于后续十分位数而言累积响应率列始终处于下降趋势。由于累计联系人列代表累积联系人数量，因此它始终处于上升趋势。如果存在指定的最低累积响应率和 / 或最大累积联系人数量，图 1.31 则会显示水平参考线以指示这些值（图 1.31 中的红线）。由图 1.31 (b) 中可以看出最大联系人数量红色标线和联系人蓝色标线的交点落在顶端 50% 和顶端 60% 之间，这和图 1.30 反映的信息是一致的。实际上根据图 1.31 (a) 和 (b) 红色标线和蓝色标线交点的位置，即可判断哪个条件先出现。

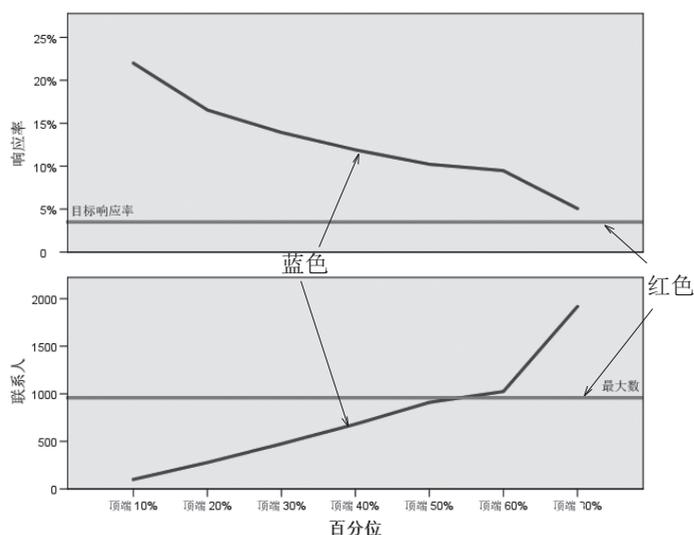


图 1.31 响应率图

图 1.32 为产生的新数据集。该数据集共包含 6 个变量。“PostalCode”显示了邮政编码分组，由于选择了分组方式为前 3 个数字或字符，因此该变量只显示了 3 个数字。“响应率”为每个邮政编码组中正响应客户所占百分比。“响应”为每个邮政编码组中正响应客户数量。“联系人”是每个邮政编码组中做出响应的所有客户数量。“索引”为基于公式 $N \times P \times (1-P)$ 的加权响应。“等级”为按降序排列的邮政编码响应率的十分位排序，该值和响应率表中（图 1.30）的“百分位”列对应，即 1 对应顶端 10%、2 对应顶端 20% 等。

PostalCode	响应率	响应	联系人	索引	等级
957	42.86%	3.00	7	1.71	1
74	33.33%	1.00	3	.67	1
623	28.57%	2.00	7	1.43	1
932	25.00%	1.00	4	.75	1
953	23.08%	3.00	13	2.31	1
762	22.22%	2.00	9	1.56	1

图 1.32 产生的新数据集

结合图 1.30 和图 1.32, 根据响应率高低, 针对不同邮政编码地区的客户进行营销活动。例如可以选取图 1.30 中“顶端 10%, 响应率 22.00%”的客户, 从生成的新数据集中找到秩为 1 的客户所在邮政编码地区, 进行新产品的促销活动。

1.5 基于控制包装检验的营销效果优化

营销部门试图通过发送邮件的方式测试产品新包装是否比现有包装获得更高的客户响应。测试邮件分别发给控制组和测试组, 控制组客户将得到现有包装的产品, 测试组客户得到新包装产品。然后对两组响应结果进行比较, 检验是否存在显著差异, 根据结果选择客户最喜欢的包装。

采用 SPSS 自带数据集 dmdata2.sav, 选择“直销 / 选择方法 / 改进我的市场营销活动 / 比较活动效果 (控制包装检验)”, 打开“控制包装检验”对话框, 如图 1.33 所示。

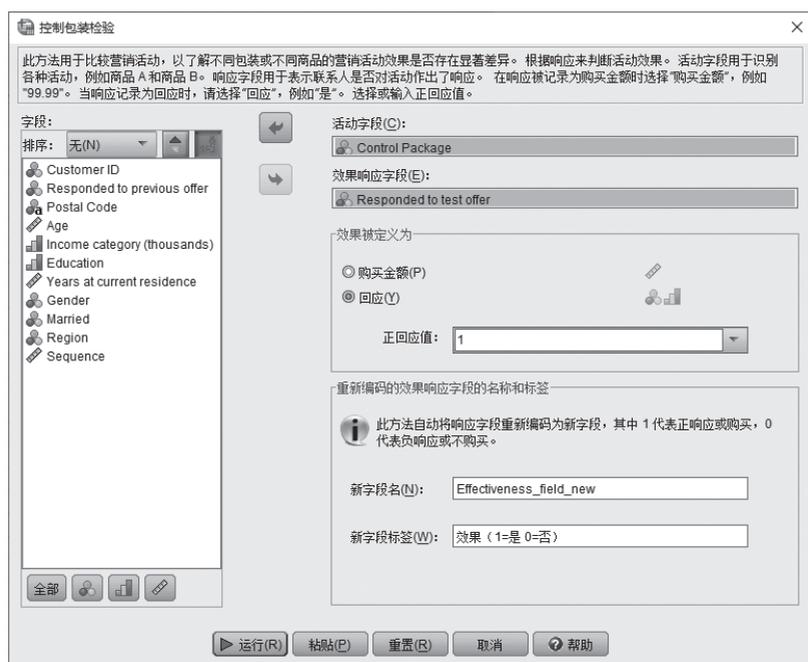


图 1.33 控制包装检验

“控制包装检验”对话框比较简单, 将 Control Package 选入“活动字段”, 将 Responded to test offer 选入“效果响应字段”。效果可被定义为购买金额, 也可被定义为回应, 默认定义为回应, 然后选择正回应值为“Yes”。重新编码的效果响应字段的名称和标签可以设置新字段名和新字段标签, 也可以使用默认设置。单击“运行”按钮, 可得到输出结果, 如图 1.34 所示。

输出结果很简单, 只简单列示了控制组和测试组的正向响应和响应率。如控制组

910人，其中35人正向响应，响应率3.8%；测试组1007人，其中62人正向响应，响应率6.2%。图1.34只显示了控制组和测试组在统计方面有显著差异，但并没有列示检验统计量和p值。从输出结果可知，产品新包装比现有包装有更高的响应率。

		Control Package			
		Control		Test	
效果 (1=是 0=否)	0	计数	列 N%	计数	列 N%
	0	875	96.2%	945	93.8%
	1	35	3.8%	62	6.2%

Control 与 Test 在统计方面有显著差异。

图 1.34 输出结果

该方法实际等价于独立样本的 T 检验。选择“分析 / 比较平均值 / 独立样本 T 检验”，打开“独立样本 T 检验”对话框，如图 1.35 所示。

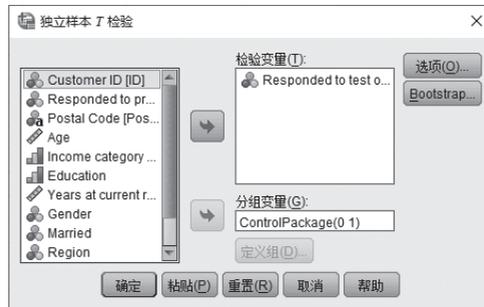


图 1.35 独立样本 T 检验

将 ControlPackage 选入“分组变量”，同时定义组，即输出 0 和 1 两个分类值。将 Responded to test offer 选入“检验变量”，单击“确定”后得到输出结果。

图 1.36 为组统计量，给出控制组 910 人，平均值 0.04，双击“0.04”所在的单元格，进入编辑模式，可见其值为 0.038 46，这就是图 1.34 中的 3.8%。对于二分类变量来说，均值就是取值为 1 的比例。类似的，测试组有 1007 人，平均值 0.06，双击其单元格进入编辑模式后为 0.061 569，等同于图 1.34 中的 6.2%。

Control Package	数字	平均值(E)	标准偏差	标准误差平均值
Responded to test offer				
Control	910	.04	.192	.006
Test	1007	.06	.240	.008

图 1.36 组统计

图 1.37 为检验结果。由于方差不等，应采用方差不等情况下的 T 检验， T 统计量值为 -2.333， p 值为 0.02，拒绝二者相等的原假设，结合均值，测试组有更高的响应率。

	列文方差相等性检验		平均值相等性的 T 检验						
	F	显著性	T	自由度	显著性 (双尾)	平均差	标准误差差值	差值的 95% 置信区间	
								下限	上限
Responded to test offer	21.571	.000	-2.307	1915	.021	-.023	.010	-.043	-.003
			-2.333	1887.724	.020	-.023	.010	-.043	-.004

图 1.37 独立样本检验

1.6 基于购买倾向分析的客户选择

基于购买倾向分析的客户选择是指根据已有的测试邮件结果或前期已有的少量客户购买信息，构建一个预测模型，然后对剩余客户的购买行为进行预测打分，以区分哪些客户最有可能购买产品。该过程包括两个步骤：第一步选择最有可能购买的联系人，需要使用已知数据集构建购买倾向模型并保存模型文件。第二步应用来自模型文件的得分，即应用模型到其他数据集（该数据集购买倾向未知）以获取预测结果，并对结果分析过滤得出最优的客户群。

本例使用 SPSS 自带数据集 dmdata2.sav，该数据集是从 dmdata.sav 的 1 万名客户中选出的对测试产品做出响应的数据。

选择“直销/选择方法/改进我的市场营销活动/选择最有可能购买的联系人”，打开“购买倾向”对话框，如图 1.38 所示。



图 1.38 购买倾向：字段

在图 1.38 字段选项卡中，将 Responded to test offer 选入“响应字段”，将个人属性特征变量选入“预测倾向”。在保存模型中一定要勾选“将模型信息输出到 XML 文件”，此处将文件名设定为“text”。

单击“设置”选项卡，设置相应参数，如图 1.39 所示。

图 1.39 包括四部分，第一部分用于设置验证样本集，如果不设置验证样本集，容易产生过拟合。在“训练样本分区大小”右侧文本框输入 0 ~ 100 的数值，用于划分训练样本集和验证样本集的比例，图中的“50”表示训练样本集和验证样本集各占一半。因为 50% 的划分是随机的，重新运行一次通常不会得到相同的结果，为了使结果重现，可以在“设置种子以复制结果”下侧文本框随意输入一个数，只要种子相同，就可以使结果重现。

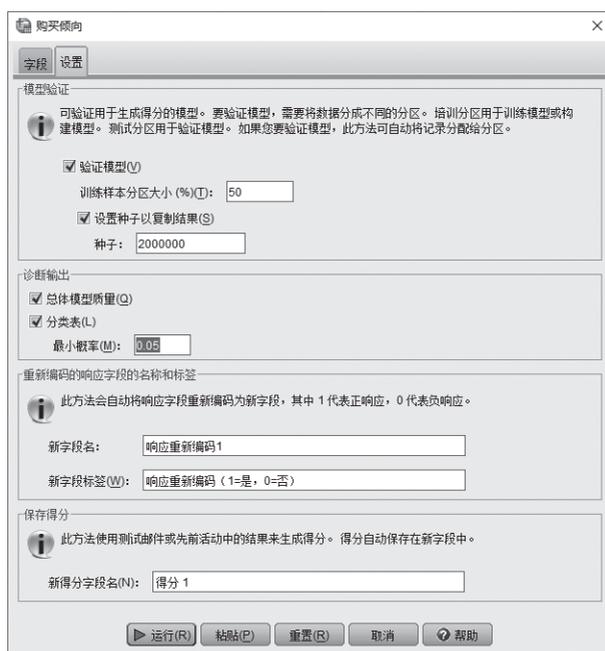


图 1.39 购买倾向：设置

第二部分为诊断输出，总体模型质量会给出模型的一个整体情况介绍，分类表会以表格形式分别给出训练样本集和验证样本集正负响应的正确预测率。SPSS 默认不输出分类表，建议勾选。最小概率是指输入一个概率值，当客户做出正面响应的可能性大于该概率值时，该客户将被指定到分类表中的预测正响应类别。

第三部分用于设定重新编码的响应字段的名称和标签，默认为“响应重新编码 1”和“响应重新编码 (1= 是, 0= 否)”。第四部分用于设定新得分字段名，默认为“得分 1”。“响应重新编码 1”的值和原始数据中的 Responded to test offer 相同，设置这个变量的目的只是为了方便和“得分 1”比较（运行后两个变量在数据文件的最右边）。

单击“运行”按钮，可得到输出结果。图 1.40 为分类表。在训练样本中，对于那些预测为正响应的，实际正响应的正确分类率为 7.43%；在样本中，对于那些预测为正响应的，实际正响应的正确分类率为 7.61%。均大于设定的 5% 的最小概率，因此该模型可以用于确定超过 5% 的响应率的一组客户。

观测值	预测值					
	训练样本			检验样本		
	响应重新编码 (1=是, 0=否)		百分比正确	响应重新编码 (1=是, 0=否)		百分比正确
否	是	否		是		
响应重新编码 (1=是, 0=否)	651	249	72.33	653	267	70.98
	19	20	51.28	36	22	37.93
总体百分比	2.84	7.43	71.46	5.22	7.61	69.02

图 1.40 分类表

图 1.41 为总体模型质量图。总体模型质量图提供有关模型质量的快速直观指示。作为一般规则，总体模型质量应大于 0.5。因为总体模型质量是根据得分 1 画出来的，而得

分 1 是对每个客户购买倾向的打分。如果不采用模型预测，采用随机猜测的方式，客户购买倾向的均值为 0.5。大于 0.5 即意味着模型的预测效果至少要比随机猜测好。该总体模型质量为 0.56，稍大于 0.5。

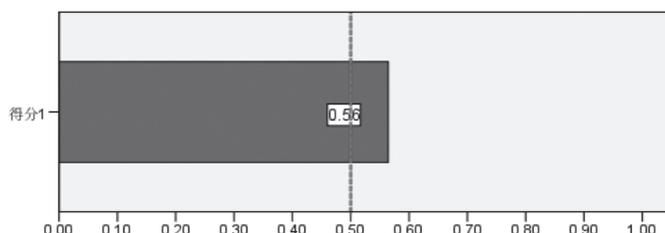


图 1.41 总体模型质量

SPSS 输出结果自带的注释说明，在解释图 1.41 时应该谨慎，因为它只反映了对总体模型质量的大致度量，即使正响应的正确预测率不满足于指定的最小可能性，总体模型质量也可视为良好。要使用分类表检查正确的预测率。

通过邮寄测试客户购买倾向分析产生的模型，可以用来对待确定购买倾向的客户购买行为进行打分，从而选择出最有可能购买产品的客户群。

选择“直销 / 选择方法 / 对我的数据评分 / 应用来自模型文件的得分”选项，打开“评分向导”对话框，如图 1.42 所示。

选择图 1.38 中生成的文件 test.xml 作为评分模型，在模型详细信息中可以看到模型方法、应用程序、目标、预测变量等。



图 1.42 评分向导：选择评分模型

在图 1.42 中单击“下一步”按钮，进行字段（变量）匹配，如图 1.43 所示。

图 1.43 默认自动匹配活动数据集中任何与模型具有相同名称和类别的字段。在模型中所有预测变量和活动数据集中的字段匹配之前，无法对活动数据集进行评分。对于缺失值可以使用值替换，也可以使用系统缺失值。值得注意的是，在对话框顶部给出了模

型类型：二元 Logistic 回归，其实图 1.42 中的 BLR 就是 binary logistic regression 首字母的缩写。



图 1.43 评分向导：字段匹配

在图 1.43 中单击“下一步”按钮，选择创建新字段，如图 1.44 所示。默认可以产生四个新字段：预测值、预测类别的概率、所选类别的概率、置信度，由于我们只对“所选类别的概率”感兴趣，此处只选择“所选类别的概率”。系统运行后包含正响应概率的新字段会被添加到待确定购买倾向数据集的末尾，其取值范围在 0 ~ 1 之间，表示每个客户购买该产品的概率大小。另外需要注意在“所选类别的概率”右侧的“值”处要选择一个类别，单击单元格选择“0”或“1”，本例选择“1”。如果不选择类别，“完成”按钮将无法激活。

23



图 1.44 评分向导：创建新字段

在图 1.44 中单击“完成”按钮，即可完成预测，预测结果主要体现在数据集的最后一列，如图 1.45 所示。

Previous	ControlPackage	PostalCode	Age	Income	Education	Reside	Gender	Married	Region	Sequence	SelectedProbability
0	1 96 600		67	3	5	10	1	0	4	26.00	.15
0	0 95 510		53	3	4	9	0	0	4	296.00	.08
0	1 92 590		56	3	2	11	1	0	4	512.00	.08
0	1 92 670		55	2	3	8	1	0	4	531.00	.07
0	1 92 690		56	2	1	6	1	0	4	534.00	.05
0	0 93 410		50	1	1	8	1	0	4	699.00	.03
0	1 93 480		67	4	2	9	1	0	4	719.00	.04
0	0 93 490		47	1	2	10	1	0	4	723.00	.03

图 1.45 预测完成后的数据集

在图 1.45 中，最后一列“SelectedProbability”即为新产生的预测变量，该变量显示了每个客户购买产品的可能性大小。可以使用该变量选择大于或等于正响应率的指定客户子集，用来开展营销活动。需要注意的是，本例使用了相同数据集建立模型并进行预测，在实际工作中用于预测的数据集应该和建模的数据集不同，尤其预测数据集在反应变量上是没有取值的（否则也就不需要进行预测了）。比如我们选择不低于 5% 的正响应率。

选择“数据 / 选择个案”，打开“选择个案”对话框，如图 1.46 所示。

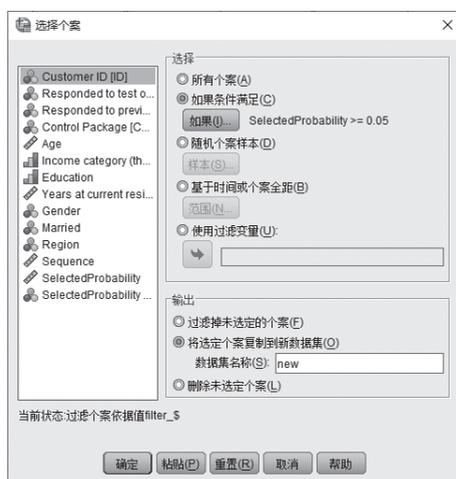


图 1.46 选择个案

在图 1.46 中选择“如果条件满足”，设定选择条件“SelectedProbability>=0.05”，在“输出”部分选择“将选定个案复制到新数据集”，并赋予新数据一个名称，如 new。单击“确定”按钮后，即可得到客户子集。

自测题

