

# 数据专题分析

华院分析技术（上海）有限公司 咨询顾问 单晖

随着华院项目经验的不断积累，以及通信行业客户对数据挖掘认识的不断深入。无论是我们的项目工作方法还是客户对我们的期望都在不断发生变化。

和以往的工作和项目相比，当前客户对我们的要求逐渐彰显出以下几个特点：

1. 有不少即时的突发的小问题、小任务希望项目组能够解决，并且问题解决周期短
2. 需求无法归类到我们项目中，也无法归类到几大类模型中
3. 期望我们提交的工作成果能够对应解决实际问题，或者可以直接归纳出解决方案。

这一类的客户需求，可以称作为数据专题分析。数据分析要解决实际问题，就必须脱离纯粹的技术手段回到实际问题上来。

下面我列举一个在某公司遇到的实际例子对这类工作作一个说明：

- 数据分析的角色  
当一个决策或者结论需要用事实来说话时，这个课题多半会落到数据分析的头上。

移动公司有很强大的数据库系统和规范的数据源，这使移动公司的管理层和市场部已经习惯于借助数据来进行决策支持和研究问题，所以移动公司的相关人员经常会接到基于数据的统计报表或者任务。

这一次，某公司领导收到一个报表（如图），2003年某省的高价值用户（月平均消费120元以上）在2004年的10月份发生严重流失，有61%的人变成了低价值用户（月平均消费120元以下），流失用户中的一半用户的ARPU甚至低于50元。这种情况的出现，极大影响了某公司的收入，必须调查清楚，产生这种情况的原因是什么。显然，这是需要用事实来说话的，问题的答案需要从数据中来，很自然的，这个问题落到华院头上。

	西安	铜川	宝鸡	咸阳	渭南	汉中	安康	商洛	延安	榆林	合计
50元以下	97084	1956	6417	12390	11352	2998	2712	1723	7445	7037	151114
50-80元	48853	895	3007	4339	3334	1800	1067	663	2773	2859	69590
80-100元	22320	427	2051	2130	1585	1242	709	484	1819	1986	34753
100-120元	17745	355	1980	1812	1439	1285	742	495	1724	2013	29590
120-150元	19467	483	2657	2227	1759	1506	847	701	2223	2739	34609
150-200元	21064	574	3579	2835	2107	2092	1298	945	3020	3889	41403
200-300元	21942	752	4014	3404	2505	2662	1390	1084	3768	5841	47362
300-400元	10161	402	1839	1763	1500	1371	630	472	1955	3087	23180
400-500元	5082	245	923	987	674	641	254	240	1044	1539	11629
500-600元	3562	119	481	521	336	383	137	139	587	940	7205
600-800元	2691	93	470	509	333	343	112	110	626	798	6085
800-1000元	1206	40	191	173	100	122	39	41	240	317	2469
1000-1500元	991	27	154	144	96	111	35	24	172	244	1998
1500-2000元	242	5	32	35	20	23	9	7	32	64	469
2000元以上	214	3	20	16	7	20	8	4	24	25	341
合计	272624	6376	27815	33285	27147	16599	9989	7132	27452	33378	461797

- 数据分析是我们的特长  
如果我们在数据分析方面的特长得到了客户的认可，客户遇到此类棘手问题的第一反映就是找华院的同志们来解决问题。

某公司领导把问题丢给了我们。这个问题和我们的项目工作有一定的关系，同时解决这个问题对我们的项目后期进展也非常有帮助，而且替客户解决棘手问题也有助于我们的客户关系。因此，我们接下了客户的问题需求。

- 分解问题、获取数据资源、做好工作准备  
接受了客户的请求，客户会授权我们提取需要的数据。

这个时候需要注意两点：一是这个问题突发性很强，需要我们短时间内整理完数据需求、明确所需数据来源，确定数据的沟通对象；二是问题解决周期很短，几个工作日内客户就希望得到有用的工作结果。

所以，分析人员对客户的数据资源（包括人员，数据源，数据结构等）应当非常熟悉，并且具备获取所需数据的能力。并且，应当提供适当的数据需求，这样可以有效控制客户数据工作人员的工作时间，和我们自己的数据理解、校验和处理时间。

这一步工作中，通过短时间的调研，我们迅速形成数据专题分析的工作思路，进而形成计划。

- 获取可以得到的数据和技术支持  
在这一步，我们将获取所有需要的数据；但是，由于各种原因（服务器满负荷、数据人员满负荷、数据保密等）我们无法获得部分需要的数据，这时候应当及时寻找其他可以替代的数据，并且适当修正我们的工作计划。同时要保证数据方面客户可以提供足够的技术支持，例如，数据的技术说明文档，或者和客户数据人员保持密切的合作联系。

由于这个问题提出的时候正好是 12 月初，计费中心的服务器正在月初出账阶段，无法支持我们的工作。于是使用另一个独立于 BOSS 系统的内部查询系统来提取数据，同时，由于系统的规模和数据滞后的原因，原定全省的数据只提取了西安市的用户，原定 11 月的数据改为 10 月的数据。

因为涉及一个全新的系统，我们和客户的一个技术人员保持了密切的联系，他在以后的分析工作中提供了重要的技术支持。

- 校验数据，验证问题  
取得了所有数据以后，我们必须校验数据，保证数据的正确性。除了我们通常的校验手段之外，针对这类问题，还有一个校验方法，那就是将客户提出的问题在我们的数据中间模拟一遍。校验的同时，我们把制定好的分析计划就很自然的映射到手头的数据中间了。

某公司的高价值客户价值流失严重的问题在西安市的数据中也非常明显的体现出来，从客户 ARPU 的交叉统计可以看到，同样存在着高价值用户向低价值

的大量迁徙。因此可以认为获取的数据符合这次数据专题分析的要求。

7月统计	ARPU		
	人均ARPU	人数	比例
小于50	0.0	0	0.0%
50-120	0.0	0	0.0%
120-200	153.0	166669	52.2%
200-300	239.9	79611	24.9%
300-500	372.8	49106	15.4%
500-800	603.5	16966	5.3%
800-1500	1024.8	5648	1.8%
大于1500	2343.2	1203	0.4%
总计	256.1	319203	100.0%

10月统计	ARPU		
	人均ARPU	人数	比例
小于50	24.9	38418	12.0%
50-120	87.3	80543	25.2%
120-200	156.5	84439	26.5%
200-300	240.2	56570	17.7%
300-500	371.8	39634	12.4%
500-800	601.6	14014	4.4%
800-1500	1021.0	4729	1.5%
大于1500	2317.1	856	0.3%
总计	202.9	319203	100.0%

- 上一步制定的工作计划只是一个初步计划  
需要注意的是，到此为止我们所制定的工作计划是一个初步的计划。因为，工作计划是根据已有的工作经验制定的，但是现在面对一个很具体、很个体的问题，应当量身定做分析计划。这样，原先的计划就必然是一个初步的，考虑不全面的计划。

于是，实际工作中，分析工作肯定会一定程度上偏离预先的设计，或者增加了大量的额外分析工作，甚至推翻最初的计划。

刚拿到某公司客户的问题时，马上就会想到应该重点分析价值流失用户的每一个账单科目，究竟是本地通话价值降低了，还是长途通话价值降低了，或者是什么别的原因。

在使用实际数据统计的时候，发现下降最多的本地通话费和漫游通话费。联想到某公司 2004 年的市场活动主要围绕降低资费展开，这个时候很自然就会想到从用户套餐变更情况入手，展开我们的分析。

- 初步分析，进一步理解我们的项目背景  
统计分析一些最基本的内容，可以明确问题的框架，反过来，通过统计分析的结果有助于我们理解项目工作目前的实际情况和实际背景。这时候往往会产生一种感觉：原来客户的实际情况是这样的呀，没想到我们还面临这样的问题。数据的统计结果会告诉我们很多未知的现象，或者说很多没有考虑周全的地方，同时会提示我们，其实实际情况中还是有很多异常情况出现的，不要忽视它们。这时候应当和项目组相关人员详细讨论，进一步理解项目背景，同时也应当修正分析思路，分析计划。

按照前面的思路，我们从用户套餐变更的角度来切入问题。通过观察相应的统计分析结果，我们发现，由于某公司在 04 年大力推广各种大幅度降低资费的新套餐和叠加 V 网套餐，某公司的老用户改选或者叠加适合自己行为特点的优惠套餐的行为非常普遍。于是，这样我们就发现了，今年某公司的市场活动在争取市场份额的同时，对网内用户转网产生的影响也是非常巨大的，初步统计每个月会有 7% 左右的用户选择新的套餐资费。

同时，分析结果中还找出一些某公司计费管理上的异常问题。例如，有的用户

新加入了月费 268 元的某品牌套餐，但是他原来套餐“月话费满 200 则当月月费只收 10 元”并没有取消，于是当月该用户的应收总额少计 258 元，给移动公司造成了巨大的损失。

- 逐一解决分析中显露出来的问题  
对分析所显露出来的问题必须逐一解决，因为这些异常情况都可能成为项目后续工作的隐患；每一个问题都应当至少有项目上的解决办法，同时和整个项目组充分沟通，一起达成共识：我们的团队的处境原来是这样的。

拿上文说的计费中出现的问题来说，发现这个问题后，我们在项目工作中，原先涉及这类用户的统计分析都会预先将这些用户排除在外，同时将这个情况向移动公司通报，并提出整改的参考意见。

- 及时修正和补充分析计划  
任何时候，数据告诉我们的结果和预先设计的分析思路、分析专题有矛盾或者补充的时候，我们都应当修正、或者补充分析计划。如果数据分析的结果和前面的项目工作对立起来的话，在排除数据问题和方法问题后（例如定义），整个项目组需要重视这个问题，找出解决或者妥协的方法。

在这次价值流失专题分析的工作中，至少两次修改了原来的分析计划。

一次是定义价值流失用户。老的计划是根据经验把贡献减少大于 10% 的用户定义为价值流失用户；但是在实际工作中发现，这样定义价值流失的话，ARPU 下降用户的比例高达 68.5%，和移动公司给我们的问题中的流失比例不是很匹配，于是调整价值流失的定义，发现设定减少比例为 25% 的时候，流失用户比例为 49%，和移动公司提供的问题中的比例匹配，于是在最终的分析中，将这个定义定为 25%。

还有一次就是确定以套餐为切入点，分析用户资费下降的原因。老计划原定的分析重点是用户的账单科目的变化，统计发现本地通话费和漫游通话费两者下降的幅度最厉害，于是联系到今年某公司的市场营销活动，觉得新资费套餐导致的网内转网可能是最主要的原因，于是调整分析重点在用户的套餐变更上，结果发现，新的分析计划确实抓住了问题的要点，找出了问题的主要原因。

- 从客户最感兴趣的角度来和客户讨论问题  
一般来说，一个观点通常可以有好几种角度来说明它，遇到这种情况，我们应当选择最直观和客户最感兴趣的角度来说明这个问题。

在分析价值流失用户的账单科目时，发现用户在各个科目上都有不同程度的价值流失，一开始使用人均损失来告诉客户，这个科目你一个人要损失好几十元，非常多；但是在和客户的沟通交流中，发现从他的角度来讲，他指关心总的损失，因为总的收入/损失是他上级考核他的指标，而不是人均损失或者损失比例。所以，最后分析结果当然全部改成了总量。

- 抓住问题主线，只分析主要问题

从技术的角度上看，使用数据来分析问题的角度和深度可以任意排列组合（比如使用 10 个变量两两交叉分析就可以生成 90 张表），因此可供分析的专题和潜在的发现可能是无穷多的。这个时候不应该拘泥于数据呈现出来的千变万化，只以我们的实际问题为基准，一切遵循“抓主弃次，紧扣问题，结论自洽”的分析原则。

在开始分析用户账单之前，根据用户的价值变化将用户分成了“ARPU 下降”、“ARPU 不变”和“ARPU 上升”三个群组；但是在接下去的分析中，我们所有的分析只针对“ARPU 下降”这群用户，因为只有这群用户才是这次分析的问题所在；如果把另外两群用户也纳入分析，那么实际工作量和最后的提交的报告都将增加到原来的三倍，并且也脱离了客户原先给我们提出的问题。

- 数据专题分析应当为我们的项目服务

我们和客户的合作基础是基于项目的合作，所以，任何专题分析，所有的分析工作都应当围绕我们项目中现阶段遭遇的问题进行。一次数据专题分析结束之后，必须让工作成果成为支撑现阶段项目工作的一块基石，否则，工作对项目没有意义。

例如这次价值流失专题分析的大项目背景是一个某品牌新套餐的全省推广项目。价值流失看起来和我们向用户推广新的套餐没有直接关系。但是，通过这次详细的专题分析，我们发现以往套餐推广中存在的很多问题，并且直观感觉到用户是怎样逐月从高端逐渐流失的，对某省原有的套餐系统也有了更清晰的认识。

这些经验和知识，对当前的推广项目起了独特的支撑作用。

- 分析结果整理备案

数据分析的成果应当总结成条理清晰，美观易懂的 PPT 文档。有这样一份文档可以有效和项目组其他成员沟通，同时可以作为提交客户的交付品，同时留做项目的存档备案。