基于 spss 的保险公司数据画像分析

周彦汐 杨 杉

四川大学锦城学院 计算机与软件学院 四川 成都 611731

【摘 要】随着人们的收入增长，健康、养老等安全保障的需求成为人们日常生活需求必不可少的部分。而在竞争激烈的保险市场环境下，保险公司为了寻求更长远的发展，必须分析市场需求，提供更优质的险种服务。所以除简单的 excel 分析工具外，我们利用 SPSS 软件对三份数据做更具体的数据分析。这三份数据均来自于四川人寿保险公司，分别是 90 万条的新投保数据、21 万多条的续保数据、16 万条的退保数据。本次分析主要运用了 SPSS 中线性回归分析、散点图、簇状条形图、单样本 T 检验、等方法。

【关键词】大数据分析；SPSS 软件；保险市场

1 引言

随着人民群众生活水平的提高和保险意识的增强，我国保险行业一直保持不断发展的上升势头，但是近年来发展速度有所放缓，未来的增长压力较大。但是我国保险业仍然有 6 ～ 11 倍的发展空间，发展潜力非常大。[1] 中国保险市场成为世界第二大保险市场并保持稳定增长趋势。人们的保险意识也在不断加强，开始主动的了解相关保险产品来保障自身的健康状况和财产安全。而保险公司面对竞争激烈的保险市场为了寻求更长远的发展，吸引到客户，必须要分析市场需求，赋能内生动力。保险业是利用精确的营销策略来发现目标客户的保险需求，并进行有针对性的营销；根据客户的实际需求推销相关的险种，而绝不能无差别的对所有潜在客户发送雷同信息。[2] 在此次数据分析中我们利用海量数据分析和市场风向标分析保险人群的相关情况。保险公司可以借鉴本次分析结果有针对性的为投保人和被保人提供更合适的险种服务，而客户们也能够根据自己的合适情况，选择最合适自己的保险产品。

2 数据准备及清洗

2.1 数据来源说明

数据来源于四川仁寿保险公司的新投保数据表（90 万条），续保数据表（21 万条），退保数据表（21 万条）[3]。

2.2 数据清洗

在分析三张表的数据之前，为保证分析结果的可靠性和有效性我们先对数据从以下几个方面进行清洗：首先剔除三张表中的重复值和缺失值以及删除与探讨问题无关的数据，在完成上述操作后为了后续的操作分析我们将续保数据表中“年龄”这一字段按未成年（0-18）、青年（18-35）、中年（35-60）、老年（60 以上）年龄段进行分类以及将险种按首字符分类：1、2、3、4、5、6 共六类险种类（4 为第 1 类，6 为第 2 类，B 为第 3 类，L 为第 4 类，S 为第 5 类，Y 为第 6 类）；。最后还需要注释表中部分行字段的含义，如：表中共用 6 个字母表示不同的婚姻状态，其中婚姻状态中的 D 是离婚、S 是单身、M 是结婚、R 是再婚、W 是丧偶、X 是未知。并且将关于客户号一列的科学计数法变为了常规计数法。再用 SPSS 软件将其打开并对数据进行区间估计以及双变量数据分析 [4]。

3 投保数据分析

3.1 新投保数据分析

研究新投保数据中客户的“总保费”和“年龄”，“过去三年平均年收入”的相关性。步骤如下：通过 spss 工具里面采用“双变量相关”分析工具，选择相关系数为 pearson 相关系数，并且统计量中选择均值和标准差，在显著性检验勾选双侧检验和标记显著性相关。

通过以上分析步骤，我们可以得出以下分析结果如图一。总保费的均值为 8376.815 元，客户过去三年平均年收入均值为 21225.6488 元，年龄的均值是 41.33 岁。在建立线性回归方程前，我们需要先确定年龄和过去平均三年年收入是否有相关关系，使用到了双变量相关分析，根据双变量相关分析得到，显著性小于 0.01 的同时两者的皮尔逊相关性有两个 * 号，所以判断两者具有显著相关性。通过 pearson 方法对相关性分析可得出，总保费、过去三年平均年收入和年龄之间呈现正向显著相关。这说明人们在选择险种进行投保时，会考虑自身的收入和年龄情况，选择适合自己并且在经济能力范围之内的险种进行投保。同样，我们通过 pearson 相关系数发现，年龄和总保费以及过去三年平均年收入之间也呈现正向显著相关关系。由此我们可以得出结论，随着收入的增加，人们也开始更关心自己的健康，注重提高生活保障，所以在有一定经济支持的基础上，所投入的的投保金额比例也会增加。

相关性

		过去三年平均年收入	年龄	总保费
过去三年平均年收入	Pearson 相关性	1	-.059**	.081**
	显著性（双侧）		.000	.000
	N	900647	900647	900647
年龄	Pearson 相关性	-.059**	1	.091**
	显著性（双侧）	.000		.000
	N	900647	900647	900647
总保费	Pearson 相关性	.081**	.091**	1
	显著性（双侧）	.000	.000	
	N	900647	900647	900647

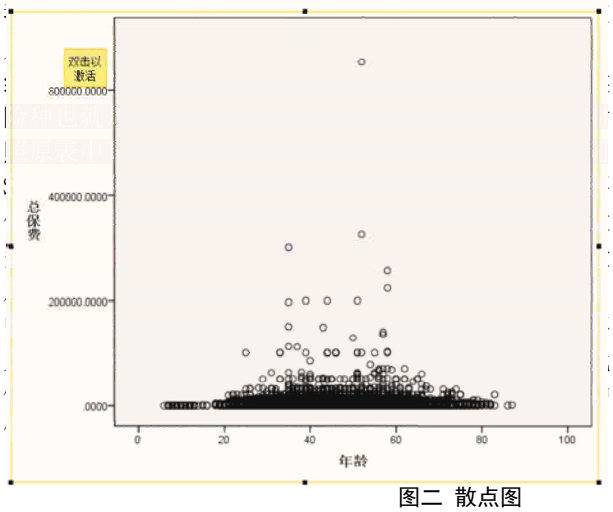
**.在 .01 水平（双侧）上显著相关。

图一 皮尔逊相关系数

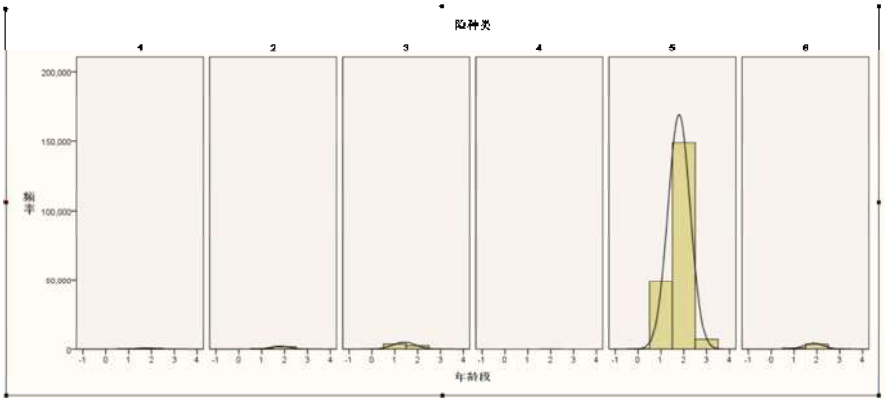
3.2 续保数据分析

在续保数据中，我们同样运用新投保数据中的分析方法。利用线性回归分析年龄和过去平均三年年收入对保额的影响，得到保额 =17672.727-98.032* 年龄 +0.044* 过去三年平均年收入，而在新投保数据中，得到保额 =21480.321-165.532* 年龄 =0.059* 过去三年平均年收入。对比两张表的线性回归方程，续保数据的常量系数和过去三年平均年收入系数小于新投保数据。年龄系数大于新投保数据，但系数值差别不大，所以两份数据保额变化趋势不明显，以及和新投保的人群相比该公司的续保人群的没有太大的变化。

接着我们使用 spss 工具，通过对总保费和年龄数据进行绘制简单散点图，直观分析二者关系。步骤为：选择工具栏中的图形工具，再选择使用散点图进行结果可视化，最后将总保费设为 Y 轴数据，年龄设为 X 轴数据。通过对绘制出的图表，如图二所示，根据散点的分布密集程度我们可以直观看出，总保费较多的分布在 0-200000 这一区间，年龄也主要处在 20-80 岁这些阶段，其中 40-60 年龄段分布最明显。根据以上结果我们可以得出不同年龄段的人群都有需要保险保障的需求，只是选择的投保金额不同，而处于 40~60 这个年龄段的人群有较多的购买保险需求，因此对于这类客户保险公司应该定向、定期的进行保险产品的维系与销售，实时跟进他们的商品体验和需求。在得出上述结论后为了进一步的探讨不同年龄段的人群更倾向于选择哪个险种，我们先利用 Excel 将年龄进行年龄段分层，分为未成年、青年、中年、老年（此步骤已在数据清洗工作中完成），接着我们利用 spss 软件中的簇状条形图来可视化数据结果，步骤为：设置类别轴为险种（在数据清洗工作中



图二 散点图



图三 簇状条形图

3.3 退保数据分析

首先我们对总保费进行独立样本 T 检验，结果如表一所示，计算出总保费的均值为 3149，将其设置为退保金额的检验值。我们再次通过对退保金额经行独立样本 T 检验得出检验的总体均值 3298，标准差 10309，以及 t 统计量 5.926，双侧 Sig 值为 0.000<0.05。根据以上结果我们不难得出在 0.05 的显著性水平下，退保金额与总保费有显著性差异，也就是以 95% 的概率接受退保金额平均高于总保费的结论。

接下来我们使用数据的汇总，对不同退保原因进行分析。可得到 8 种退保原因，分别是：被保人出国移居、服务不理想、公司信誉、经济原因、失效退保、险种不理想、因转保及移出困难而退保和正常退保。在 8 种退保原因中因“经济原因”退保的平均退保金额是最大的，而因为“失效退保”和“因转保及移出困难而退保”的两类退保金额相比其他都较少。

根据以上结果，我们可以得出结论有 95% 的概率接受退保金额平均高于总保费的结论，人们普遍会因为服务不理想、经济愿意、险种不理想等原因选择退保，而这其中因“经济原因”退保的平均退保金额是最大的。那么保险公司为了留住客户，持续发展应该对于退保频率较高的险种进行产品的升级或改革，减小退保率。此外，对于退保客户，在其退保前，采取相关有效办法或措施，尽量挽留客户，减少企业损失，具体深入的分析客户画像。

表一：T 检验

	单个样本检验（检验值 =0）					
	均值差值	t	df	Sig（双侧）	差分的 95% 置信区间	
				下限	上限	
总保费	3149.36725	130.560	167720	.000	3102.08879	3196.645712
	单个样本统计量					
	N	均值	均值的标准差误		标准差	
	167721	3149.267755	9878.839569		24.1219388	

结论

通过分析新投保数据、续保数据、退保数据三张表，我们可以得出以下三点结论，一是：总保费、过去三年平均年收入和年龄之间始终呈现正向显著相关，那么保险公司在针对不同收入和不同年龄的人群时，可以推荐不同类别人群所适合的险种，适当为客户建议其投保份数和投保额等，进而提高客户满意度。而在购买保险的总人群中，中年人居多，未成年人较少，故该公司可调整销售策略，主要针对中年人销售保险，减少对未成年的保险销售。二则：对于不同理赔需求的客户，推荐相应理赔价位的保险产品，能达到较好的保险销售与理赔结果，使得客户满意。对于销量较高的险种，如 S81 产品，可适当进行产品的升级或改革，加大宣传力度，提升购买量。三是：对于退保客户，在其退保前，保险公司应该采取相关有效办法，如更换险种，附赠礼品等等，尽量挽留住客户，减少企业损失。综上，各大保险公司应当抓住机遇浪潮，不断加强保险的保障能力，开发新险种，不断满足我国未来保险市场对养老保险、寿险、长期护理保险、重大疾病保险等的需求，抓住机遇，获取客户信任，便能促进保险行业的未来快速持续发展。[5]

【参考文献】

贺超 . 我国保险业发展影响因素的实证分析 [J]. 河北企业 ,2020(11):77-81.

李弘熙 . 浅析大数据背景下保险业发展 [J]. 中国集体经济 ,2020(21):98-99

[3] 吴华斌 . 大数据的统计分析浅议 [J]. 经济师 ,2014(10):277-280.

[4] 文颖 . 大数据时代数据分析中统计软件的运用研究 [J]. 信息与电脑 (理论版),2017(06):97-98.

[5] 贺超 . 我国保险业发展影响因素的实证分析 [J]. 河北企业 ,2020(11):77-81.