

大数据分析在保险行业中的应用

朱俊嘉 杨 杉

四川大学锦城学院 四川 成都 611731

【摘要】本文通过对某保险公司退保数据进行 SPSS 数据分析。对于退保数据,选择了可视化分析,比较均值分析,方差分析三种方法。得出结论与建议:青年人退保占大部分,保险公司应该重点关注和服务青年人。险种不理想和公司信誉对投保人退保有着极大的影响,建议保险公司提高信誉,适当调整险种。此次分析可以通过数据反馈给保险公司一些建议,保险公司可以适当调整,有利于公司的发展。

【关键词】SPSS; 数据分析

1 引言

随着大数据的迅猛发展和普及应用,许多行业可以通过数据分析得出一定结论。国内保险业的快速发展,银行保险作为一种新型保险产品与销售途径出现在大众的视野里。如今银行保险在我国已发展十余年,但由于银保双方合作深度不足、业务内容单一等因素,近年持续退保的问题较多^[1]。本文通过对某保险公司退保数据进行 SPSS^[2] 数据分析。可以通过数据反馈给保险公司一些结论及建议,保险公司可以适当调整,有利于公司的持久发展。

2 研究思路

以某保险公司退保数据为研究对象,通过 SPSS 进行数据分析。运用可视化分析、比较均值分析、方差分析]针对不同因素的退保频率、不同险种与退保原因各种统计量、年龄与退保金额的均值是否有显著性差异三个方面进行分析。

3 数据说明

3.1 数据来源

数据来源于某保险公司,数据表中包含了机构、险种、缴费方式、缴费期限、投保份数、总保费、保额、退保金额、退保原因、客户号、性别、年龄、婚姻状况、过去三年平均年收入、教育程度、等 19 个字段共 16000 余条数据。

3.2 数据清洗

数据中,很多数据对于数据分析是没有用的,因此对数据进行了数据预处理[3],筛选出了我们需要的数据,把不需要的数据进行了剔除,对数据中明显无效的数据进行清洗:投保时间、退保时间、教育程度、家庭人口,选中整列,将整列删除。

接着查找是否有空白数据,单击查找,没有找到空白数据,剩下的数据均为有效数据。

为了便于后续分析的观察,将退保原因处查找“退保:”并替换为,便于分组。

添加险种分类列,将险种第一个字符相同的划为同一类,以第一个字符作为共同的类型。

将险种、退保原因、年龄、退保金额进行降序排列。

4. 数据分析

4.1 可视化主题数据分析

通过可视化的方法,研究不同年龄层的退保人数的分布、不同收入退保人数的分布,不同婚姻状况退保人数的分布。

首先将年龄和过去三年平均收入进行重新编码。

年龄分组: 25 岁以下、25-35 岁、35-45 岁、45 到 55 岁、55-64 岁、65 岁以上。收入: 5k 以下、5k 到 1w、1w-5w、5w-10w、10w 以上。

通过绘制,得到下列饼图。

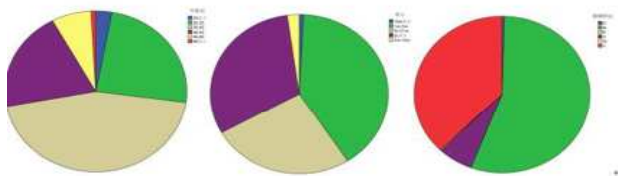


图 4.1 不同年龄、收入、婚姻状况退保人数饼图

由饼图的数据分布可得知:

年龄在 35-45 岁的退保人数占大部分, 26-35 岁阶段的位居第二, 其次是 46-55 岁。可以得知经济独立并且具有良好判断能力的青年人存在退保的可能性最大, 所以保险公司应该重点关注和服务青年人。

平均三年年收入在 1w 到 5w 的退保人数占大部分, 其次是收入在 5k 以下和 5k 到 1w。可以得到低收入人群退保占大多数, 退保原因可能为经济原因。保险公司可以适当的调整险种的金额。

婚姻状况为已婚的退保较多。已婚的人购买保险的人也占绝大多数, 基数越大, 退保的也越多。

4.2 比较均值主题数据分析

通过比较均值分析得出各险种退保情况和主要退保原因的统计量。

表 4.2 退保险种统计量 退保原因统计量

险种类型	平均值	个案数	标准差	峰度	偏度	中位数
4 类	4441.6073	2507	7675.51498	211.686	11.268	2600.0
6 类	263.4487	896	59.80532	3.896	1.369	250.0
B 类	1431.8192	2725	1396.49241	93.827	7.804	1039.0
L 类	715.5000	4	266.79518	-3.082	.515	666.0
S 类	3066.1329	152614	9896.63173	417.226	14.674	990.0
Y 类	5014.4282	8975	11639.01138	72.066	6.938	1740.0

退保原因	平均值	个案数	标准差	峰度	偏度	中位数, L
被保险人出国移居	2374.2874	87	6577.70103	70.682	8.138	1260.0
服务不理想	2570.1477	308	9820.37050	75.372	8.380	915.0
公司信誉	2516.6534	206	7405.40012	103.539	9.547	1115.0
经济原因	3253.0990	129909	9926.74477	315.004	12.780	1000.0
失效退保	1194.3291	3134	2476.38988	271.094	14.353	810.0
险种不理想	2141.8437	5513	6593.83216	194.054	11.888	950.0
因转保及移出困难而退保	932.7273	11	236.22408	-1.778	.398	810.0
正常退保	3100.5609	28553	10659.30788	609.031	18.393	980.0

通过表 4.2 退保险种统计量可以得出：

申请退保的客户中购买的险种为 S 类的频率最多，为 152614，高达 91.0%。L 类险种被退保的频率最少，频率为 4。4 类为 1.5%，6 类为 0.5%，Y 类保险占比 5.7%，B 类保险占比 1.4%。

Y 类险种的退保金额最高，中位数为 2124.37，平均值为 5282.4310，6 类退保金额最低，中位数为 65，平均值为 75.0243。但可以观察到 Y 类保险的退保金额的均值最高，而 S 类保险的均值比 Y 类保险要低，由此可得被退保次数最多的险种为 S 类险种，但是退保金额相对于较少。Y 类险种虽然退保频率较低，但退保金额较高。6 类保险是被退保次数和退保金额都最少的险种。

标准差最大为 Y 类说明退保金额波动较 [4] 大。除了 L 类保险其余几类险种的峰度都大于零说明这几类保险相较于正态分布更为陡峭，为尖峰分布。峰度高就意味着方差增大是由低频度的大于或小于平均值的极端差值引起的。说明样本之间差异较大，右边（高赔付金额区域）极端值较多，赔付金额更多的集中在众数周围。几种险种的偏度也都大于零，说明其退保金额分布形态相较于正态分布为右偏，也就是数据右端有较多的极端值，数据均值右侧的离散程度强，其中 S 类险种的峰度和偏度都最大，为 427.216 和 14.550，说明 S 类险种高赔付金额区域极端值较多，波动、范围更大。

通过表 4.2 退保原因统计量可以得出可以得出：

退保原因排名前三的是：经济原因（129909）、正常退保（28553）、险种不理想（5513）。但比较均值得，其中经济原因的平均退保金额水平最大为 3422.4156 元，其次是正常退保（3181.1958 元）> 公司信誉（2659.3023 元）> 被保险人出国移居（2640.9318 元）> 服务不理想（2594.0232 元）> 险种不理想（2412.9741 元）> 失效退保（909.2995 元），最小为因转保及移出困难而退保（844.9600 元）。说明因为险种不理想退保的频率和公司信誉退保的退保金额均值都在前三，因此保险公司应该及时调整险种，并提高公司信誉。

除因转保及移出困难而退保外，其他几种退保原因的峰度都大于零，峰度越大，说明该数据系列中的极端值越多，其中，正常退保（646.357）的波动更大，范围更广，其次是经济原因（332.482），其他几种费用类型的退保金额波动较小，极端值较少，分布较为均匀。

所有退保原因偏度均为正数，即右偏，尖峰分布，即高退保金额方分布相对稀疏，大部分集中于平均金额附近。其中，正常退保（19.135）的波动更大，范围更广，其次是经济原因（12.660），但几种退保原因的差异值也不大。其他几种费用类型的退保金额波动较小，极端值较少，分布较为均匀。

4.3 方差分析主题数据分析

通过方差分析探究不同年龄他们的退保金额的均值是否显著的差异。

通过方差齐次性检验得到显著性值为 0.000 小于显著性水平 0.05，拒绝原假设，也就是方差不具有齐次性。所以在多重比较表中，只看 Tamhane 部分。

通过方差分析表研究不同年龄他们的退保金额的均值是否显著的差异。显著性值为 0.000 小于显著性水平 0.05，拒绝原假设，也就是至少有两种年龄层的退保金额均值是有显著差异的。

表 4.3-2 多重因素方差分析

(I) 年		(J) 年 龄	显著性
龄 层		层	
T a m - h a n e	2 5 以 下	26-35	.010
		36-45	.024
	26-35	46-55	.998
		36-45	1.000

通过两两比较可以得出以下结论：

25 岁以下与 26-35 岁和 36-45 岁作比较，显著性值大于 0.05，可以得出 25 岁以下与 26-35 岁和 36-45 岁的退保金额均值是没有显著差异的。

其余两两比较都是年龄段小的退保金额均值是显著高于年龄高的退保金额均值。

表 4. 3-3 对比检验方差分析

	显著性(双尾)
假定等方差	.000
不假定等方差	.000

将 45 岁以下的人群分为一组，45 岁以上的人群分为一组，比较他们退保金额均值是否有显著差异。由于方差齐次性检验得到方差不具有齐次性。所以在对比检验中，只看假设等方差部分。显著性值为 0.000 小于 0.05，45 岁以下与 45 岁以上的退保金额均值是有显著差异的。45 岁以下的人退保金额均值是显著更高的。

5 结论及建议

5.1 结论

- 1. 退保人数前三的年龄层为：35-45 岁、26-35 岁、46-55 岁。青年人退保占大多大户。
- 2. 退保人数前三的平均三年年收入为：1w 到 5w、5k 以下、5k 到 1w。可以得到低收入人群退保占大多数，退保原因可能为经济原因。
- 3. 婚姻状况为已婚的退保较多。已婚的人购买保险的人也占绝大多数，基数越大，退保的也越多。
- 4. Y 类险种的退保金额最高，S 类的频率最多。
- 5. 退保原因频率排名前三的是：经济原因、正常退保、险种不理想。
- 6. 退保原因均值排名前三的是：经济原因、正常退保、公司信誉。

5.2 建议

- 1. 经济独立并且具有良好判断能力的青年人存在退保的可能性最大，所以保险公司应该重点关注和服务

青年人。要着重于了解潜在客户的真实需求与现实条件等，在给客户推荐险种时应具有针对性和相关性。

- 2. 低收入人群退保占大多数，退保原因可能为经济原因。保险公司可以适当的调整险种的金额，适应客户的多样化需求。
- 3. Y 类险种的退保金额最高，S 类的频率最多。建议保险公司根据客户反馈调整 Y 类和 S 类险种，根据数据分析实时优化公司的险种。
- 4. 因为险种不理想退保的频率和公司信誉退保的退保金额均值都在前三，因此保险公司应该及时调整险种，并提高公司信誉。

【参考文献】

[1] 杨森. 基于时间序列的 N 公司退保率分析及相关对策研究 [D]. 昆明理工大学, 2017.

[2] 丰靖淼. 寿险业退保行为及其影响因素研究 [J]. 企业导报, 2015, (13):194+101. DOI:10.19354/j.cnki.42-1616/f.2015.13.141

[3] 陈祺琦. 大数据分析在保险行业中的应用 [J]. 电子世界, 2020, (10):36-38. DOI:10.19353/j.cnki.dzs.2020.10.017

[4] 刘勇军. 多期模糊投资组合优化模型及算法研究 [D]. 华南理工大学, 2013.