数据分析在保险领域的运用——以退保数据为例

廖 丹 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】在获取的退保数据中包括险种、总保费、保额、退保金额、投保时间、退保时间、退保原因、过去三年收入等数据项。针对什么样的客户（年龄、年收入）最容易退保，客户的退保原因主要是什么的问题进行分析，得到分析结果为保险公司提供数据支撑，基于此，保险公司可根据结果调整险种、金额、投保规则等，达到保险公司和客户双赢的局面。

【关键词】大数据分析；保险；SPSS

1 引言

保险行业是当今社会重要的行业之一，人们常通过购买意外险、分红险、车险等弥补因意外造成的损失。但与之伴生的是保险公司逐年提高的退保率，影响了保险公司的利润，部分客户也造成了经济损失。[1] 我国人均保险消费额以及保险支出占 GDP 的比重仍有较大差距，我国保险行业的发展依然拥有广阔的市场前景 [2] 基于此可对退保数据进行研究，分析客户的忠诚度及退保原因，保险公司可根据结果调整投保金额、投保规则等，达到保险公司和客户双赢的局面。

2 数据预处理

2.1 数据预处理的意义

随着时代发展，大数据被应用到各行各业，也因此产生了许多数据质量问题，因此使得后续的数据分析产生偏差，得到的结论不够准确，可能会给社会或企业带来不可预估的后果，所以在进行数据分析前应对数据集进行预处理，保证数据质量 [3]

2.2 数据准备

原始数据为：保险公司退保数据（已经过脱敏处理，包括机构、险种、投保时间、缴费方式、缴费期限、投保份数、总保费、保额、退保金额、投保时间、退保时间、退保原因、客户号、性别、年龄、婚姻状况、过去三年收入、教育程度、职业以及家庭人口这些数据项）

2.3 数据清洗

将退保原因进行分列，去除退保原因：字段，只保留：后的真实退保原因，以方便后期用于分析

将年收入分段。年收入在 0-10000 之间的置换为 1，10001-20000 之间置换为 2，20001-30000 之间置换为 3，30001-50000 之间置换为 4，500001-100000 之间置换为 5，100001-800000 之间置换为 6，高于 800000 置换为 7。

在 EXCEL 中运用 if 函数对数据进行分类，如果退保金额大于总保费则返回是，否则就返回否，然后用筛选功能将返回结果为是的全部数据导入 SPSS 中作为接下来的数据分析的依据，只针对退保金额大于总保费的险种进行分析。

3 分析方法

3.1 频率分析

频率分析主要通过频数分布表、条形图和直方图，以及集中趋势和离散趋势的各种统计量来描述数据的分布特征，以便我们对数据的分布特征形成初步的认识，才能发现隐含在数据背后的信息。[4]

3.2 饼图

可视化是数据科学的重要组成部分，可视化将生涩难懂的数据转化成直观的图形，增加视觉体验，帮助用户更好的理解、分析复杂的数据对象。可视化能够对分析的流程、结果进行清晰的展示，大大提高了数据的可读性 [5]，此分析运用饼图将频率排名前十的险种进行显示，增强数据可读性，使险种分布比例体现的更加清楚、易懂。

3.3 简单和等级相关分析

针对总保费与退保金额之间的关系，采用简单和等级相关分析，分析两种数据之间的潜在规律，是否有相关性，相关系数是多少。以得到可供保险公司后期用于分析保险险种是否合理，以及是否需要针对客户进行画像推销险种的准确数据。

4 数据分析

4.1 分析客户忠诚度及退保原因

将经过预处理后的数据导入 SPSS，运用频率分析，将年龄和处理过后的年收入分段数据设置为变量，设置频率统计量，将统计结果生成直方图。

表 1 年龄、年收入统计量

	年龄	年收入分段
均值	41.4	1.7
中值	40.0	1.0
众数	45	1.0
方差	86.5	1.0
偏度	.4	1.7
峰度	.05	2.8
和	6945886	286993.0

表 2 退保原因分布表

	频率	百分比	有效百分比	累积百分比
被保人出国移居	87	.1	.1	.1
服务不理想	308	.2	.2	.2
公司信誉	206	.1	.1	.4
经济原因	129909	77.5	77.5	77.8
失效退保	3134	1.9	1.9	79.7
险种不理想	5513	3.3	3.3	83.0
因转保及移出困难而退保	11	.0	.0	83.0
正常退保	28553	17.0	17.0	100.0
合计	167721	100.0	100.0	

根据表格中显示的结果分析，退保人群年龄均值在 41 岁左右，年收入在 0-20000 之间。

得到结论年龄在 30-45 之间，年收入在 0-20000 的人群更容易退保。客户退保的主要原因是：经济问题。这部分人群或因年收入不够理想，且处于这个年龄段的人群一般对上有赡养父母的责任，对下有抚育子女的责任，经济压力大，无法承担保费而选择退保。保险公司可适当根据实际情况，调整投保方式、投保金额等，达到留住客户的目的。

4.2 退保原因与退保金额的关系

采用探索分析，将退保金额设置为因变量列表，退保原因设置为因子列表

表 3 退保金额与退保原因统计量

	退保原因		统计量	标准误
退保金额	被保人出国	均值	2640.9	715.2
		均值 95% 置信区间下限	1219	
		均值 95% 置信区间上限	4062	
		偏度	7.3	.2
		峰度	61.3	.5
	服务不理想	均值	2594	650.6
		均值 95% 置信区间下限	1313.7	
		均值 95% 置信区间上限	3874.2	
		偏度	2.6	.1
		峰度	45.9	.2
	公司信誉	均值	2659.3	505.4
		均值 95% 置信区间下限	1662.6	
		均值 95% 置信区间上限	3655.9	
		偏度	8.5	.1
		峰度	90.8	.3
	经济原因	均值	3422.4	28.8
		均值 95% 置信区间下限	3365.8	
		均值 95% 置信区间上限	3479	
		偏度	12.6	.007
		峰度	332.4	.01
	失效退保	均值	909.2	42.1
		均值 95% 置信区间下限	926.7	
		均值 95% 置信区间上限	991.8	
		偏度	11.02	.04
		峰度	167.3	.08
	险种不理想	均值	2412.9	87.5
		均值 95% 置信区间下限	2241.3	
		均值 95% 置信区间上限	2584.6	
		偏度	8.9	.03

		峰度	114.5	.06
	因转保及移出困难	均值	844.9	274.6
		均值 95% 置信区间下限	233	
		均值 95% 置信区间上限	1456	
		偏度	1.04	.6
		峰度	-.3	1.2
	正常退保	均值	3181.1	64.8
		均值 95% 置信区间下限	3054	
		均值 95% 置信区间上限	3308.3	
		偏度	19.1	
		峰度	646.3	

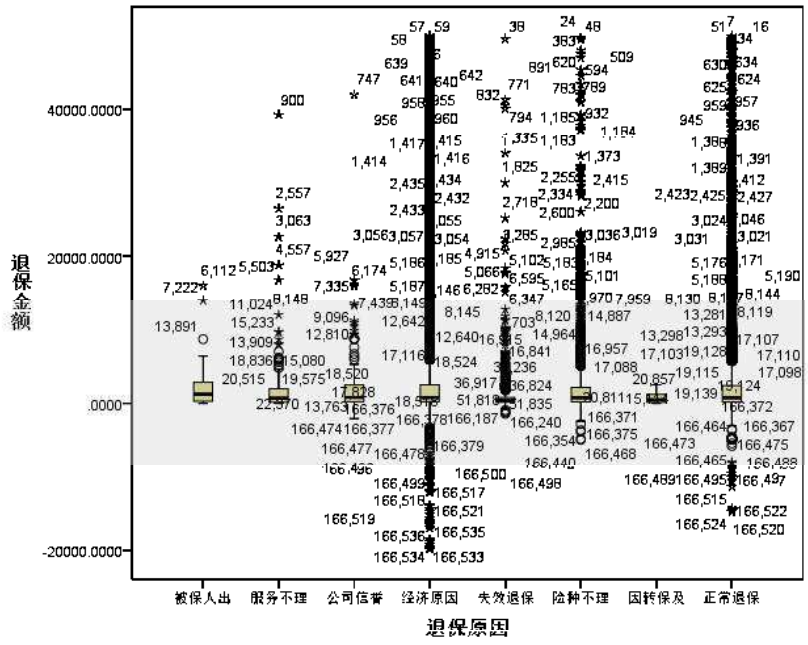

图 1 退保原因分布箱型图

从统计结果中看

被保人出国移居的退保金额在 1219-4062，均值：2640

服务不理想的退保金额在 1313-3874，均值：2594

公司信誉的退保金额在 1662-3655，均值：2659

经济原因的退保金额在 3365-3479，均值：3422

失效退保的退保金额在 826-991，均值：909

险种不理想的退保金额在 2241-2584，均值：2412

因转保及移出困难而退保的退保金额在 233-1456，均值：844

正常退保的退保金额在 3054-3308，均值：3181

根据箱型图及统计量分析，除因转保及移出困难而退保外，其余情况偏度峰度均为正数，即右偏，尖峰分布，即高退保金额相对稀疏，大部分集中于平均金额附近。

4.3 分析退保金额大于总保费的险种频率

在 EXCEL 中运用 if 函数对数据进行分类，如果退保金额大于总保费则返回是，否则就返回否，然后用筛选功能将返回结果为是的全部数据导入 SPSS 中作为接下来的数据分析的依据。采用频率分析分析退保金额高于总保费的各个险种频率，将险种作为频率分析的变量，统计量为众数，图表采用饼图

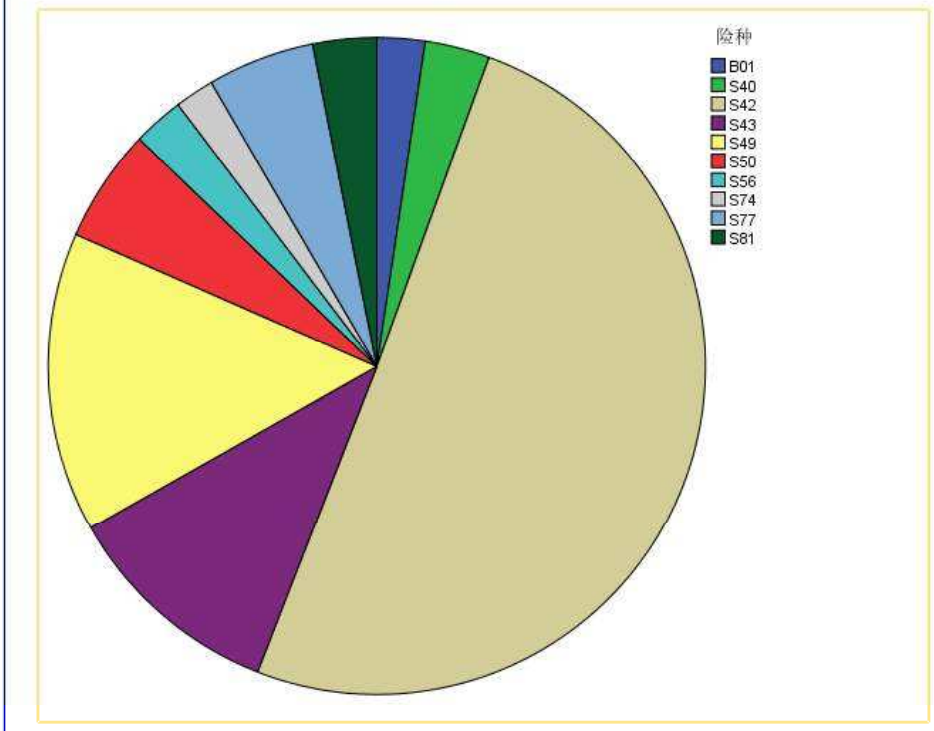

图 2 退保险种分布饼图

因险种过多只选用频率前十的险种制作饼图，根据分析结果以及饼图可知S42这个险种有24155条数据，占所有数据的42.7%，排在第一。S49和S43的占比分别排在第二和第三。对于保险公司来说这三种保险的退保金额高于保费的可能性较大，对于公司的收益会产生亏损的可能，为避免产生较大的亏损现象可以尽量减少对这三类保险的推荐，或者将这种类型的保险推荐给忠诚度较高，也就是不容易做出退保行为的客户。同时保险公司也应该自我分析，为何这三种保险的退保率远远高于其他险种，如果是自身问题将保险设置的不合理，让客户觉得性价比不高选择退保，则保险公司应该及时做出调整，更改投保方式、投保金额、投保期限以及保险的权益等，留住客户。

4.3 总保费与退保金额之间的关系

针对总保费与退保金额之间的关系，采用了简单和等级相关分析，分析两种数据之间的潜在规律，是否有相关性，相关系数是多少。

表 4 总保费与退保金额相关性

		总保费	退保金额
总保费	Person 相关性	1	.9
	显著性（双侧）		.0
	N	167721	167721
退保金额	Pearson	.9	1
	显著性（双侧）	.0	
	N	167721	167721

从Pearson相关性看，总保费和退保金额的显著性（双侧）为0，小于 α 值，意味着拒绝原假设，总保费与退保金额具有正向显著相关，并且它们的相关系数为0.912。

表 5 总保费与退保金额的相关系数

			总保费	退保金额
Spearman 的 rho	总保费	相关系数	1	.5
		Sig（双侧）	.	.0
		N	167721	167721
	退保金额	相关系数	.5	1.0
		Sig（双侧）	.0	.
		N	167721	167721

从Spearman相关性看，总保费和退保金额的Sig为0，小于 α 值，意味着拒绝原假设，总保费与退保金额具有较强的相关关系，他们的相关系数为0.547。

从Pearson和Spearman可得，总保费与退保金额有显著的相关性，总保费越高，退保金额越高。

4.4 小结

1. 年龄在30-45之间，年收入在0-20000的人群更容易退保。客户退保的主要原因是：经济问题。

2. 除因转保及移出困难而退保外，其余情况偏度峰度均为正数，即右偏，尖峰分布，即高退保金额相对

稀疏，大部分集中于平均金额附近。

3. 总保费与退保金额有显著的相关性，总保费越高，退保金额越高

5 结论及建议

客户退保最重要的原因主要是经济原因。

针对客户：

在购买保险前应该准确衡量自己的经济实力以及购买保险的必要性，不能盲目购买，避免造成经济损失

在购买保险前应该准确了解退保的比例，以及结合自己的实际情况考虑按需购买

针对保险公司：

在设计保险时，应该针对不同年龄段的人群进行考虑设计，参考退保数据，可以适当更改保险的投保政策、投保方式等。

2. 分析退保人数较多的险种，研究客户退保是个人原因还是险种设置不够合理，性价比低导致退保。根据分析结果决定保险公司是否应该调整保险内容，以降低客户的退保率，保证保险公司的利益。

针对退保赔付比例较多的保险类型，应该针对客户进行画像，判断客户的忠诚度，对忠诚度高的客户推销该种保险，避免大量客户退保，以减少公司的赔付损失，保证公司的利益。

【参考文献】

[1] 夏颖 . 寿险公司退保情况研究——以 A 省险企为例 [J]. 金融经济 , 2019(10):134-135.

[2] 孙晓彤 , 方元锡 . 保险行业发展问题与应对策略 [J]. 现代营销 (下旬刊), 2020(12):120-121.

[3] 崇卫之 . 数据预处理机制的研究与系统构建 [D]. 南京邮电大学 , 2018.

[4] 赵晓娜 . 基于 SPSS 软件对调查问卷进行频率分析的研究 [J]. 电子商务 , 2018(02):58-59.

[5] 袁晓如 . 专题 : 大数据可视分析应用 [J]. 大数据 , 2021,7(02):1-2.