基于大数据的保险公司理赔数据分析

王欣悦 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】以某保险公司的理赔数据作为数据源，利用 Excel 和 spss 对平均赔款金额最高的险种、不同费用类型对平均赔款金额的影响以及愿意申请理赔的客户进行分析，通过单因素方差分析、均值过程等方法分析出其潜在信息，为保险公司及客户提出建议。

【关键词】大数据；数据分析

1 引言

随着新时代的发展，为了满足日渐复杂的消费者需求保险行业逐渐向数字化发展[1]。通过大数据，保险公司可以轻而易举地抓取、筛选和分析出新投保、续保、理赔等各个环节的统计数据，从而分析出数据背后的潜在信息[2]。通过收集到的客户数据，保险公司可以从不同的侧重点出发，挖掘出数据背后的潜在信息，通过这些信息，保险公司可以针对不同的客户群体设计出不同的产品类型，并且还能对现有的产品体系做出调整，为公司带来更多的收益，客户也可以根据自身需求选择到更适合自己的保险产品，还能够改变对保险市场的过往认知。

2 研究思路

以某保险公司的理赔数据作为数据源，首先利用 Excel 对数据进行清洗然后围绕三个问题对清洗后的数据进行分析，主要运用了 spss 当中的均值过程、单因素方差分析、频率分析的方法，分析了平均赔款金额最高的险种、不同费用类型对平均赔款金额的影响、愿意申请理赔的客户，并对公司和客户两方面提出了建议。

3 数据说明

3.1 数据来源

数据来源于某保险公司的理赔数据。数据表中包括了机构、险种、赔款金额、总保费、费用类型、费用金额、保额、性别、年龄、婚姻状况、过去三年平均年收入等 16 个字段共有 212182 条数据，17M。

3.2 数据清洗

在险种字段中：以险种首字母为一个大类，将险种分为 4 6 B、F、S、Y 六个组。费用类型字段中：按疾病类型将其分为了重大疾病、疾病、意外、其他，并分别用 1 2 3 4 来表示。年龄字段中：将年龄分为 0-20（少年）、21-40（青年）、41-65（中年）、66-90（老年），并分别用 1 2 3 4 表示。过去三年平均年收入字段中：将其分为了 1 万元以下、1 万 -10 万元、10 万 -100 万元、100 万以上，并分别用 1 2 3 4 表示。

4 数据分析

4.1 分析哪个险种平均赔款金额较高

在数据清洗部分已经依据险种首字母，将险种分为 4 6 B、F、S、Y 这六个类，通过均值过程的方法探究哪个险种平均赔款金额较高，其中把赔款金额作为因变量列表，险种作为自变量列表，通过分析不同险种的正态曲线以及均值可以看出各险种的平均赔款金额以及赔款金额分布情况。险种与赔款金额特征数据如下所示：

案例处理摘要

	案例					
	已包含		已排除		总计	
	N	百分比	N	百分比	N	百分比
赔款金额 * 险种	212182	100.0%	0	.0%	212182	100.0%

报告

赔款金额

险种	均值	标准差	峰度	偏度	极大值	极小值	合计
4	4575.875000	4236.004972	1.098	1.593	12900.0000	2000.0000	36607.0000
6	1517.346887	1308.121138	.182	1.095	5000.0000	.0000	1057590.780
B	741.375472	908.7556312	34.954	5.112	8406.0300	.0000	117878.7000
F	637.978761	2576.729484	88697.541	252.695	927397.0000	.0000	1.2069E8
S	19670.08773	25333.23150	145.967	6.850	1000000.000	.0000	4.0536E8
Y	24820.25173	54061.48723	83.204	7.937	800000.0000	.0000	38024625.65
总计	2664.172344	11164.42526	1054.017	20.697	1000000.000	.0000	5.6529E8

图 1 险种与赔款金额特征数据

分析图 1 中数据可以看出：各险种的平均赔款金额：险种 Y> 险种 S> 险种 4> 险种 6> 险种 B> 险种 F，即险种 Y 是平均赔款金额最高的险种，险种 B 和险种 F 的平均赔款金额很少相较于其他险种差值较大。各险种的峰度和偏度都是正值，所以它们所对应的赔款金额都是尖峰分布、右偏的，说明挨着平均赔款金额左右的数据密度比较集中，并且大部分分布在中间或者较低水平，其中险种 F 的峰度值和偏度值最大，说明其右偏程度及峰的陡峭程度远远大于其他险种，即险种 F 的赔款金额分布很集中，基本上都分布在中间或较低水平，但也有离

群值在右侧使得右偏程度严重。

4.2 分析不同费用类型对平均赔款金额的影响

运用单因素方差分析探究不同费用类型对平均赔款金额的影响，首先通过频率分析得出费用类型 2（疾病）的占比明显高于其他费用类型，其占比为 79.8%，所以在进行单因素方差分析时将费用类型 2 分为一组，费用类型 1 3 4 分为另一组，并设置显著性水平为 0.05，得出的方差齐次性和 ANOVA 表中显著性都为 0.000 小于 0.05，所以拒绝原假设方差不具有齐次性，并且至少有 3 种费用类型具有显著差异。并得到了各费用类型比较表如图 2 所示：

Tamhane	1	2	11799.99757	328.1217420	.000	10936.12971	12663.86543
		3	10067.86065	337.5679577	.000	9179.185579	10956.53572
		4	6508.735922	605.2792145	.000	4915.746587	8101.725257
	2	1	-11799.99757	328.1217420	.000	-12663.86543	-10936.12971
		3	-1732.136924	84.4235230	.000	-1954.269372	-1510.004478
		4	-5291.261654	509.4488519	.000	-6632.288996	-3950.234313
	3	1	-10067.86065	337.5679577	.000	-10956.53572	-9179.185579
		2	1732.136924	84.4235230	.000	1510.004478	1954.269372
		4	-3559.124729	515.5835346	.000	-4916.269838	-2201.979622
	4	1	-6508.735922	605.2792145	.000	-8101.725257	-4915.746587
		2	5291.261654	509.4488519	.000	3950.234313	6632.288996
		3	3559.124729	515.5835346	.000	2201.979622	4916.269838

*. 均值差的显著性水平为 0.05。

图 2 各费用类型比较表

从图 2 中得出以下结论：费用类型对赔保金额的影响：费用类型 1（重大疾病）> 费用类型 4（其他）> 费用类型 3（意外）> 费用类型 2（疾病）。综上费用类型 1（重大疾病）对于平均赔保金额的影响最高，那么保险公司对费用类型 1 的赔付风险就越高，当客户出现重大疾病时保险公司理赔的金额会相对高一点，重大疾病就包括了癌症或者恶性肿瘤之类的疾病，患病几率不大但后期医疗费用很高，大多数的人都会因为理赔金额高这一点选择购买相应险种。费用类型 2（疾病）对于平均赔保金额的影响最低，保险公司对费用类型 2 的赔付风险就越低，当客户出现疾病时保险公司理赔的金额会相对低一点，疾病的发病率小但需要长期治疗，所花费用累计起来的数值还是很大，尽管理赔金额相对较低，但购买相应类型保险还是很有必要的。费用类型 3（意外）对于平均赔保金额的影响排在第三，则当客户发生意外时保险公司理赔的金额会比疾病的理赔金额相对高一点，总的来说与重大疾病、疾病、意外相关的保险，都可以根据自身条件选择更适合自己的购买。

4.3 分析什么样的客户（年龄、年收入）更愿意去申请理赔

运用频率分析探究什么样的客户（年龄、年收入）更愿意申请理赔，首先在数据清洗阶段已经把年龄和过去三年平均年收入字段进行了分类，然后进行后续分析，得到了统计量数据表 1 如下所示，年龄分组频率表 2 如下所示，收入分组频率表 3 如下所示：

表 1 统计量数据表

		年龄分组	收入分组
N	有效	212182	212182
	缺失	0	0
均值		2.74	1.74
众数		3	1
偏度		-0.616	1.305
偏度的标准误		0.005	0.005
峰度		-0.457	0.398
峰度的标准误		0.011	0.011
和		580938	268731
百分位数	25	2.00	1.00
	50	3.00	1.00
	75	3.00	2.00

表 2 年龄分组频率表

		频率	百分比	有效百分比	累计百分比
有效	1	79	0.0	0.0	0.0
	2	59039	27.8	27.8	27.9
	3	149475	70.4	70.4	98.3
	4	3589	1.7	1.7	100.0
	合计	212182	100.0	100.0	

表 3 收入分组频率表

		频率	百分比	有效百分比	累计百分比
有效	1	157266	74.1	74.1	74.1
	2	53314	25.1	25.1	99.2
	3	1571	0.7	0.7	100.0
	4	31	0.0	0.0	100.0
	合计	212182	100.0	100.0	

从表 2 3 可以看出：年龄 3，收入 1，即 41-65 年龄段且年收入在 1 万元以下的客户更愿意申请理赔。

从表 1 分析得到：年龄项的峰度和偏度均小于 0，所以它们都是扁平分布、左偏的，说明挨着平均年龄左右的数据密度比较分散，而且大部分都分布在中、高年龄段，即处于中、高年龄阶段的人对于保险的看重程度更大，其他年龄层的人对于保险不太了解。通过 Excel 公式计算出过去三年平均年收入 1 万元以下占比 57.3%，1 万 -10 万元占比 40.46%，10 万 -100 万元占比 1.78%，100 万元以上占比 0.03%，通过图表看出收入项的峰度和偏度均大于 0，所以它们都是尖峰分布、右偏的，说明挨着平均年收入左右的数据密度比较集中，而且大部分都分布在中间或较低收入水平，即大多数年收入在中间或较低水平的人更愿意去申请理赔，即使理赔金额不高或者理赔金额为 0 都会去尝试，这也就说明经济收入中等或者偏低的人群购买保险是为了让未来生活得到保障。

5 结论及建议

5.1 结论

各险种的正态分布曲线都是尖峰分布且右偏，其中险种 F 的正态曲线图形陡峭以及右偏程度更严重，说明赔款金额分布很集中，基本上都分布在中间或较低水平，但险种 F 的平均赔款金额又是最低的，通过 Excel 中的公式计算出 89.16% 的人购买了此类保险，这就说明虽然险种 F 赔款金额最低但因为其总保费低的原因大家对险种 F 的接受程度更高所以购买人数很多，性价比更高。费用类型为重大疾病对于平均赔款金额的影响最高，保险公司对重大疾病的赔付风险就越高，当客户出现重大疾病时保险公司理赔的金额会相对高一点，所以说购买相应类型的保险还是很有必要的，可以大大减少突发重疾给家人带来的负担。41-65 年龄段且年收入在 1 万元以下的客户更愿意申请理赔，因为中高年龄阶段且年收入在 1 万以下的人群，大部分购买保险的原因都是为了在遇到特别情况时，保险的理赔金可以保障之后的正常生活。

5.2 建议

5.2.1 对于客户

基于上述三个问题的分析可以看出，各险种都有自己的优势，在购买保险时不能只看重赔保金额，还要看它的赔保率以及总保费与赔保金额之间的关系，综合判断出适合自身情况的险种进行购买，比如险种 F 虽然其赔款金额较低，但总保费少，那么购买此保险的人群，在每月缴纳保险费的时候压力不会太大，并且此保险也能满足大家的日常需求。因为目前保险市场的险种类型很多，所以要选择能够为自己带来很大益处的险种如健康险等类型的险种，可以为日后多一份保障。目前保险市场的发展趋向于年轻化，很多保险公司都推出了一些新的险种类型，大家应该把目光放的宽阔一些，或许能发现一些有意思的险种。

5.2.2 对于保险公司

因为重大疾病的赔付较多，所以保险公司应储备足够的流动资金来应对此类保险的理赔。通过分析理赔数据中申请理赔的客户人群发现，大多数集中在 41-65 年龄段且年收入在 1 万元以下这一区间内，处于这一区间内的人群对于保险的需求大多数都是为了在意外情况发生时能保障日常生活，所以针对这一人群的客户，保险公司可以有针对性的提出相关险种，以供客户选择。同时 20 岁以下申请理赔的人占比只有 0.037%，可知这一人群申请理赔的很少，那么保险公司向这一人群推荐保险得到的受益是最大的，所以保险公司应该设计出更吸引年轻人兴趣的保险。

【参考文献】

[1] 曹毓飞. 我国保险业数字化转型发展趋势 [J]. 中国保险 , 2021, (03):13-16.

[2] 陈晓静 , 张闫文 , 李港鑫 , 亓苗 . 大数据对保险行业的挑战和应对策略 [J]. 上海保险 , 2020, (09):48-53.