“寿再青骏杯”2022年全国研究生案例分析大赛

基于生存分析和xgboost的重疾赔付风险研究

时间：2022年7月

目录

1. 问题介绍

问题：荔枝人寿主力重疾险产品赔付情况持续不乐观

研究目的

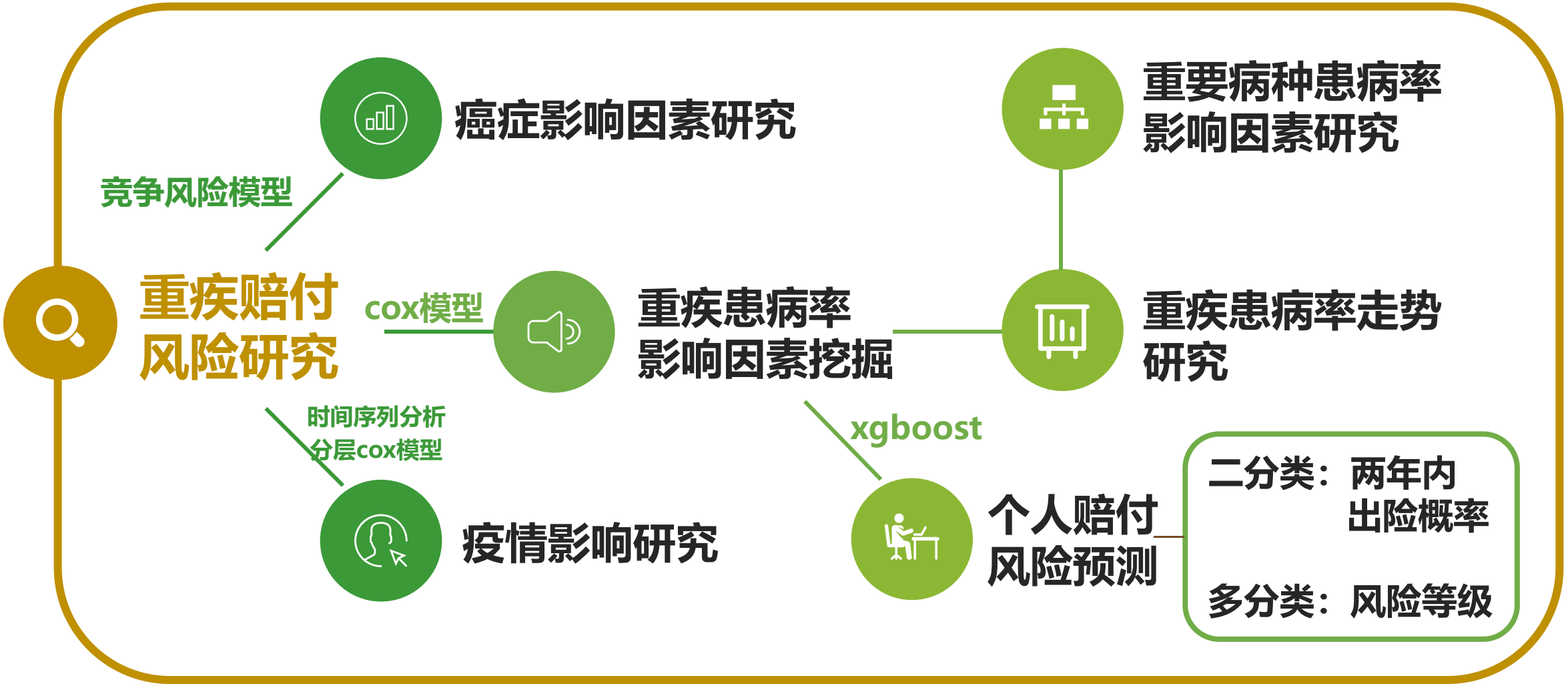
- 挖掘重疾患病率的影响因素
- 探寻重疾患病率变化规律
- 设计风险预测工具

研究意义

✓ 投保者画像+变化规律+风险预测工具

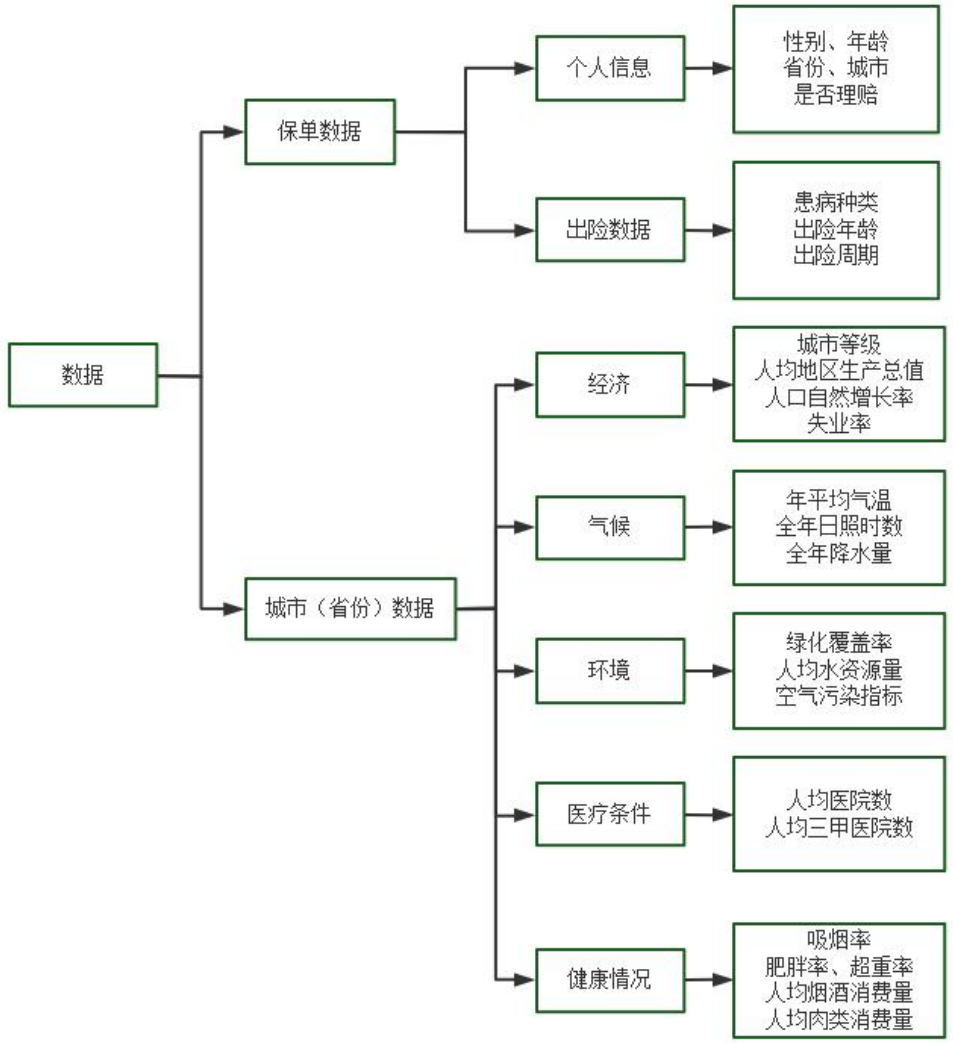
✓ 了解用户群体、优化决策

技术路线图


重疾赔付
风险研究
竞争风险模型
癌症影响因素研究
cox模型
重疾患病率
影响因素挖掘
时间序列分析
分层cox模型
疫情影响研究
xgboost
个人赔付
风险预测
二分类：两年内
出险概率
多分类：风险等级
重要病种患病率
影响因素研究
重疾患病率走势
研究

2. 数据处理

原始数据：荔枝人寿某重疾险产品**2016-2018年**相关保单承保数据和理赔数据样本共计100万条



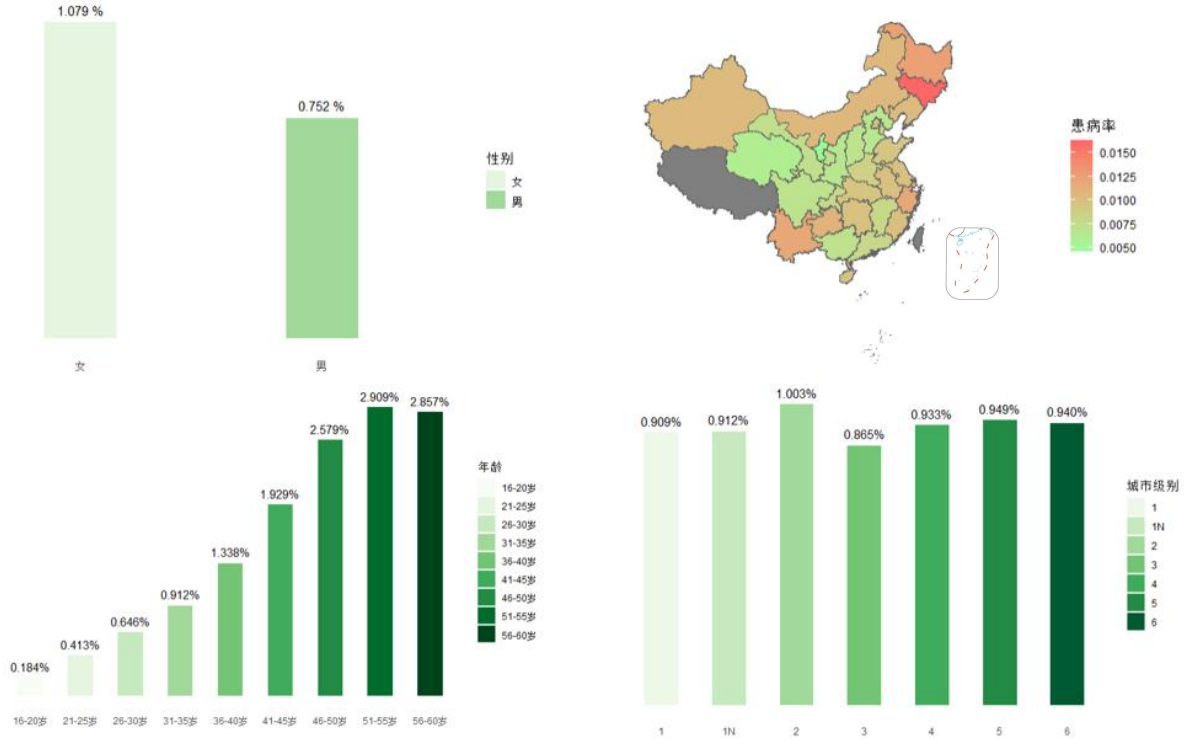
填补空缺值：设定省直属县级市等级为6

提取出险病种：对出险过程描述进行**文本分析**，根据【2020年保监会公布的大病平均治疗费用】分五级，疾病严重性递增

计算被保险人出险周期：表示被保险人由保险生效至赔付的时长

拓展城市数据：数据来源为《中国城市统计年鉴》、《中国环境统计年鉴》、《中国卫生统计年鉴》、中国宏观经济数据库以及心血管健康指数网站

3.1 患病率影响因素



不同性别患病率对比（左上）；不同年龄段患病率对比（左下）；
不同省份患病率对比（右上）；不同级别城市患病率对比（右下）

变量	变量取值	风险比	变量	变量取值	风险比
年龄	21~25岁	2.161	性别	男性	0.697
	26~30岁	3.360	年日照时数	1400-2200h	1.192
	31~35岁	4.772		>2200h	1.517
	36~40岁	7.016		400-800mm	1.435
	41~45岁	10.064	年降水量	800-600mm	1.492
	46~50岁	13.439		>1600mm	1.354
	51~55岁	14.826	吸烟率		1.001
	56~60岁	15.334	Log(肥胖率)		1.054

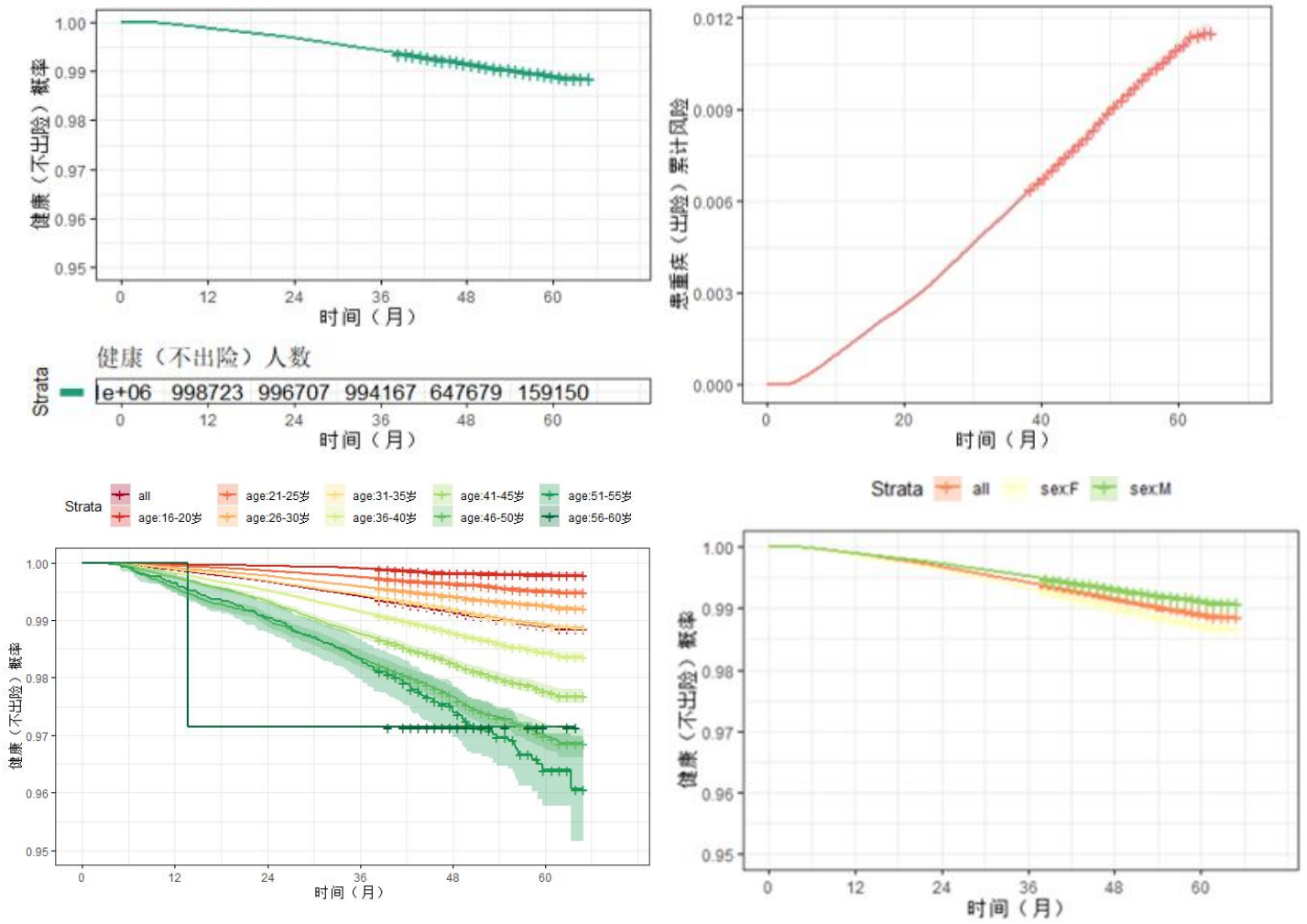
被保险人是否患重疾（出险）的影响因素cox回归结果表

高患病风险人群特征：■女性 ■与年龄正相关：30岁以上-**5倍**，40岁以上-**10倍**

高风险地区特征：**年日照时数更长、年降水量在400-1600mm之间、人群吸烟率和肥胖率更高**

→室外活动意愿更低、睡眠时间更少、中心城市生活节奏快

3.2 发病率趋势刻画·cox模型

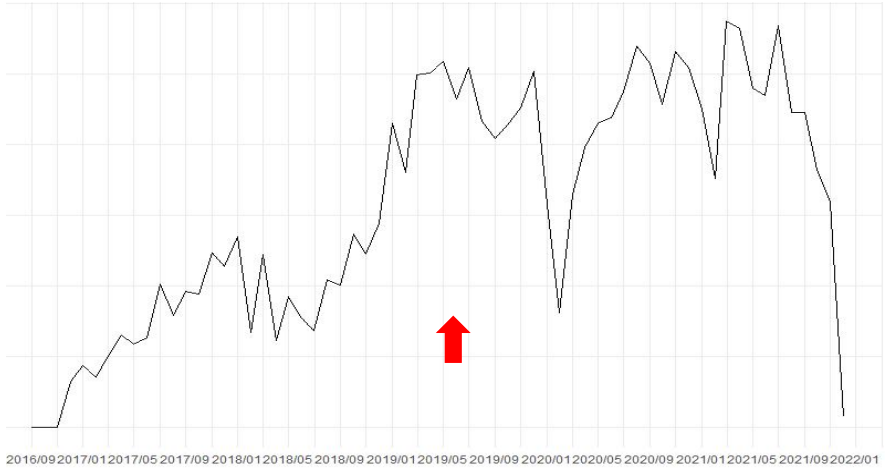


投保人不出险概率（左上）和出险累计风险（右上）随时间变化折线图
投保人不同年龄段（左下）、不同性别（右下）不出险概率随时间折线图

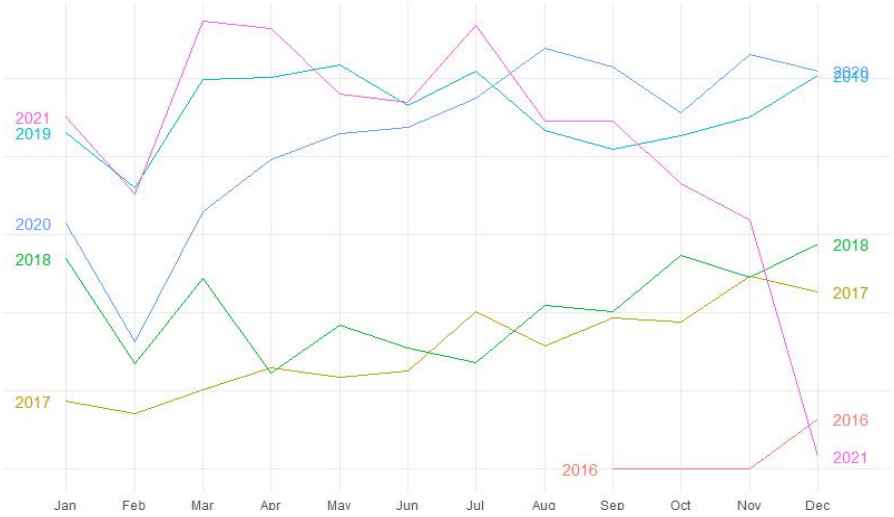
- 总体患病率非常小 **(<0.01)**，从购买保险开始，投保人患病率**缓慢增加且速率稳定**
- **增速与年龄正相关**，**35岁**为平均线分水岭，大于55岁的高龄群体患病均发生在投保一年以内
- **女性患病率增速大于男性**，时间越长，二者的风险差距越大

3.3 疫情影响分析（总体）·时间序列分析

当月出险率=当月出险人数/当月在保人数



出险率时序图



不同年份出险率波动对比图

$ARIMA(3,1,3)(1,1,0)_{12}$：基于疫情前出险数据

2020-2021年预测出险率与真实出险率对比

- **疫情前**：具有一定的周期性，二月出现下降，总体**波动上升。**
- **疫情后**：**2020年上半年出险率骤降**，后逐步恢复到原有水平；下降提前至上一年**十二月**开始，**下降幅度增大**；2020年下半年开始，出险率基本稳定，与2019年水平相同。

3.3 疫情影响分析·分层cox模型

变量	变量取值	风险比	变量	变量取值	风险比
年龄	21~25岁	2.499	年龄	51~55岁	16.717
	26~30岁	3.931	性别	男性	0.702
	31~35岁	5.239	年均气温	15-22℃	0.773
	36~40岁	8.071		22-35℃	0.608
	41~45岁	11.630	吸烟率		1.002
	46~50岁	14.495	Log(肥胖率)		1.077

疫情前后投保人患重疾（出险）的影响因素cox回归结果对比表

选取出险时间跨度最大的2016年投保数据，以2020年1月1日为节点划分出险保单的疫情前后，未出险保单按比例分配。

- **原本潜在重疾风险高的人群患病可能加重、层次差异变大。**世卫组织的报告显示，新冠流行后抑郁症和焦虑症发病率大幅提升，其中女性比男性受到的影响更大，患有哮喘、癌症和心脏病等健康状况的人更容易出现心理疾病症状。
- **年均气温在15℃以下的患病风险大于15℃以上地区的人群。**北方口岸地区，可能受疫情影响严重，由此导致的并发症等情况也会更多。

3.4 患癌症影响因素研究·竞争风险模型

变量	变量取值	风险比	变量	变量取值	风险比
年龄	21~25岁	2.321	性别	男性	0.466
	26~30岁	3.731	年日照时数	1400-2200h	1.148
	31~35岁	5.192		>2200h	1.691
	36~40岁	7.593		400-800mm	1.587
	41~45岁	10.591	年降水量	800-600mm	1.826
	46~50岁	13.991		>1600mm	1.661
	51~55岁	14.833	吸烟率		1.003
	56~60岁	20.077	Log(肥胖率)		1.069

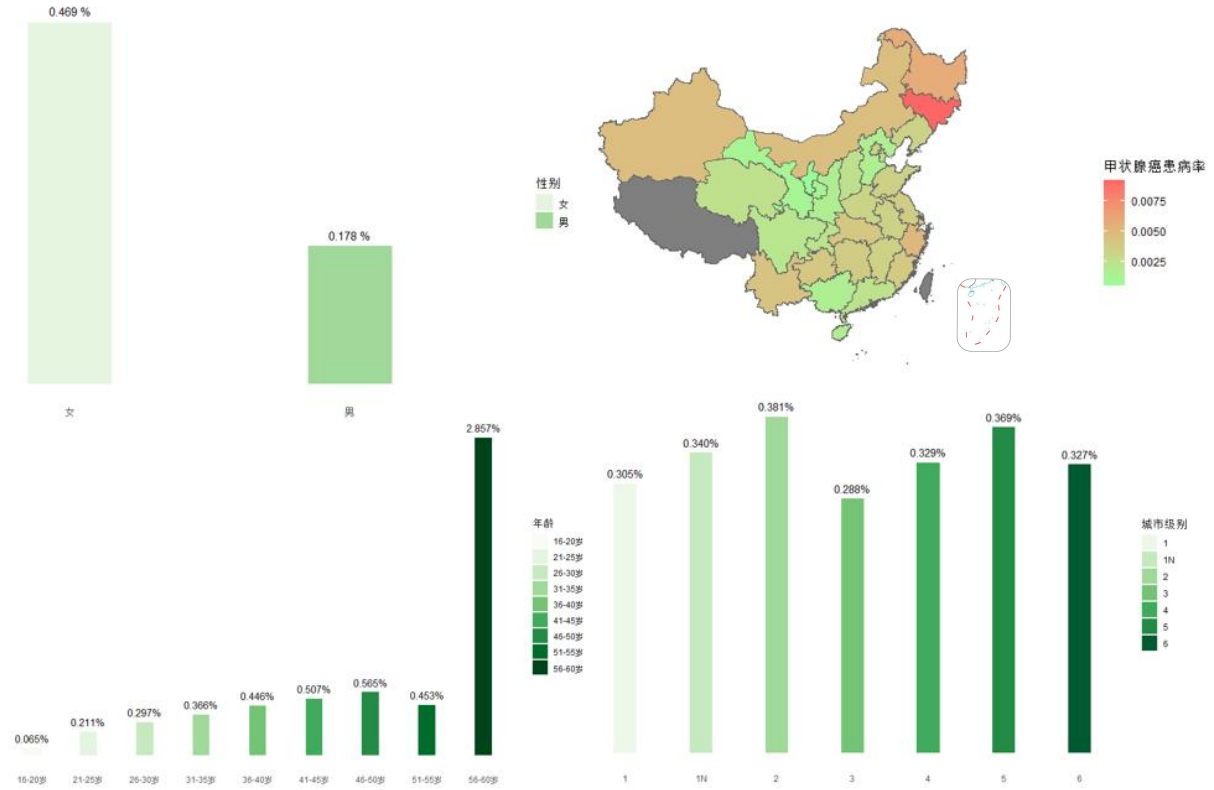
被保险人是否因癌症出险的影响因素竞争风险模型结果表

现有出险数据病种分布TOP3：

癌症（78%）心血管疾病（8%）脑疾病（6%）

- 对降低重疾风险有正向作用的因素，对于降低癌症风险的正向作用更为明显。
- 男性患癌概率相较于女性显著降低，只有女性的47%左右。

3.5 重要病种分析·甲状腺癌 出险病种中占比最高



不同性别患病率对比（左上）；不同年龄段患病率对比（左下）；
不同省份患病率对比（右上）；不同级别城市患病率对比（右下）

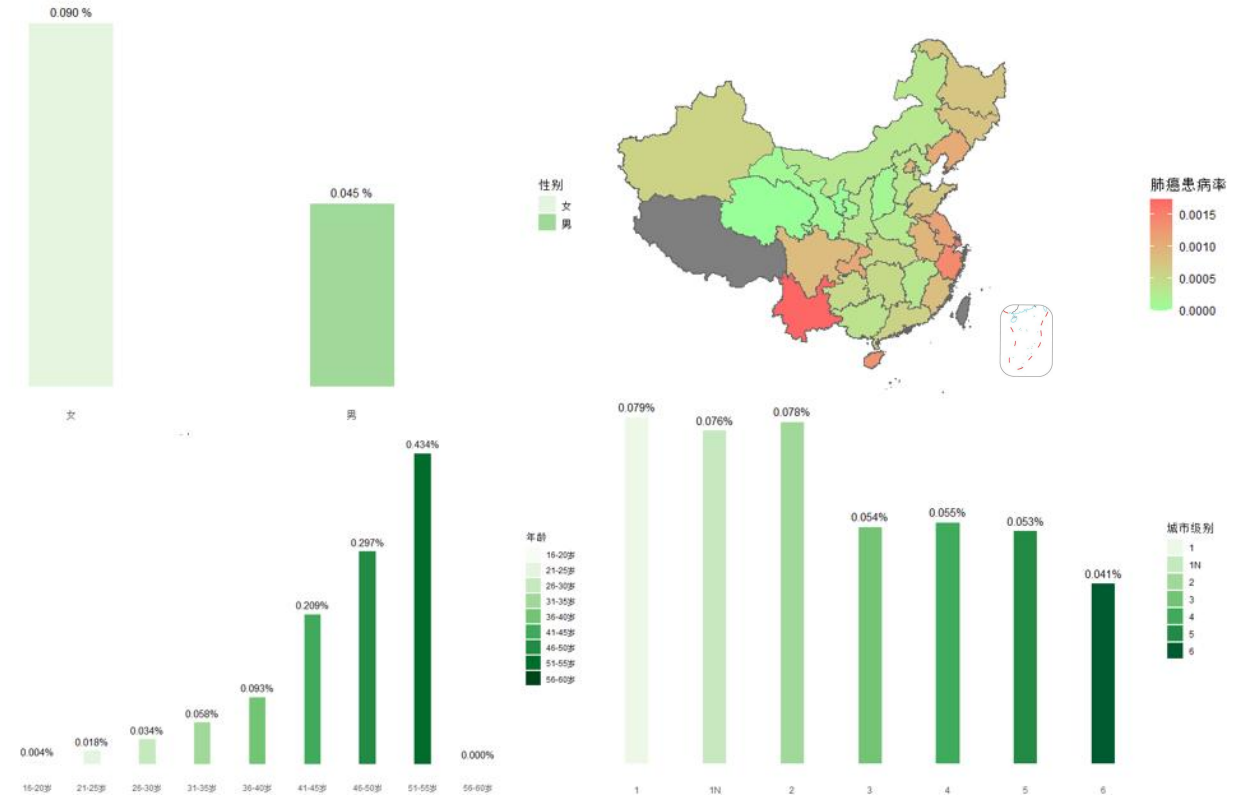
变量	变量取值	风险比	变量	变量取值	风险比
年龄	21~25岁	2.999	性别	男性	0.285
	26~30岁	4.144	年日照时数	1400-2200h	1.583
	31~35岁	5.154		>2200h	2.593
	36~40岁	6.267	年降水量	400-800mm	1.719
	41~45岁	7.054		800-600mm	2.554
	46~50岁	7.643		>1600mm	3.010
	51~55岁	5.690	超重率		0.521
	56~60岁	36.978			

被保险人是否患甲状腺癌的影响因素cox回归结果表

甲状腺癌高患病风险人群特征：

■ 女性 ■ 56-60岁（**35倍**风险） ■日照长，降水多的地区（北方、中部沿河）

3.5 重要病种分析·肺癌 出险占比偏高，近几年发病率显著提高



不同性别患病率对比（左上）；不同年龄段患病率对比（左下）；
不同省份患病率对比（右上）；不同级别城市患病率对比（右下）

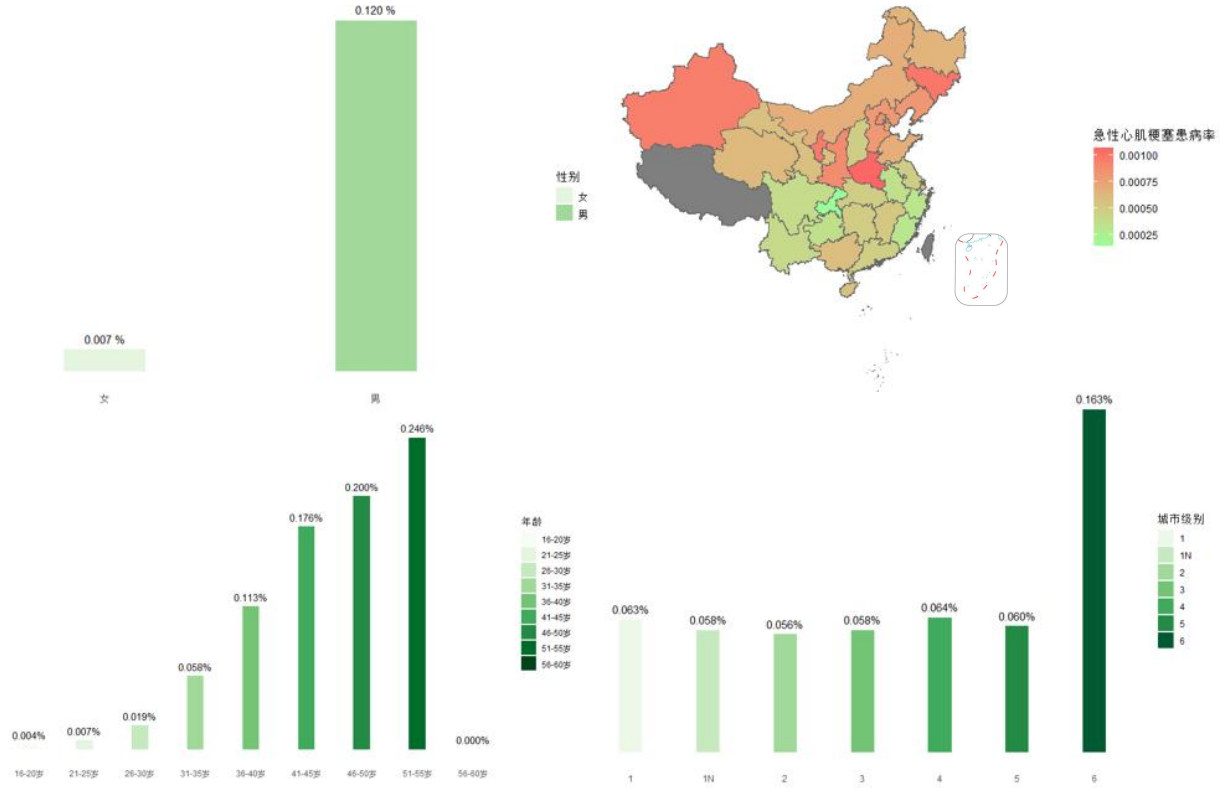
变量	变量取值	风险比	变量	变量取值	风险比
年龄	21~25岁	3.799	性别	男性	0.503
	26~30岁	7.057	年降水量	400-800mm	2.008
	31~35岁	12.116		800-600mm	2.782
	36~40岁	19.161		>1600mm	3.214
	41~45岁	42.642	人均地区GDP	4-7万元	1.481
	46~50岁	59.698		7-12万元	1.792
	51~55岁	80.952		>12万元	2.078
	56~60岁	0.002	人口自然增长率	-0.041	0.960
			吸烟率	0.167	1.181

被保险人是否患肺癌的影响因素cox回归结果表

肺癌高患病风险人群特征：

■ 女性 ■ 50岁以上 ■ 吸烟 ■ 降水量高，人均地区GDP高（沿海、边境）

3.5 重要病种分析•急性心梗 每年新发至少50万，发病逐年增多



不同性别患病率对比（左上）；不同年龄段患病率对比（左下）；
不同省份患病率对比（右上）；不同级别城市患病率对比（右下）

变量	变量取值	风险比	变量	变量取值	风险比
年龄	21~25岁	1.801	年龄	51~55岁	77.820
	26~30岁	5.041		56~60岁	0.003
	31~35岁	15.036	性别	男性	15.958
	36~40岁	28.448	超重率		1.301
	41~45岁	44.532	人均肉类消费量		0.519
	46~50岁	53.090			

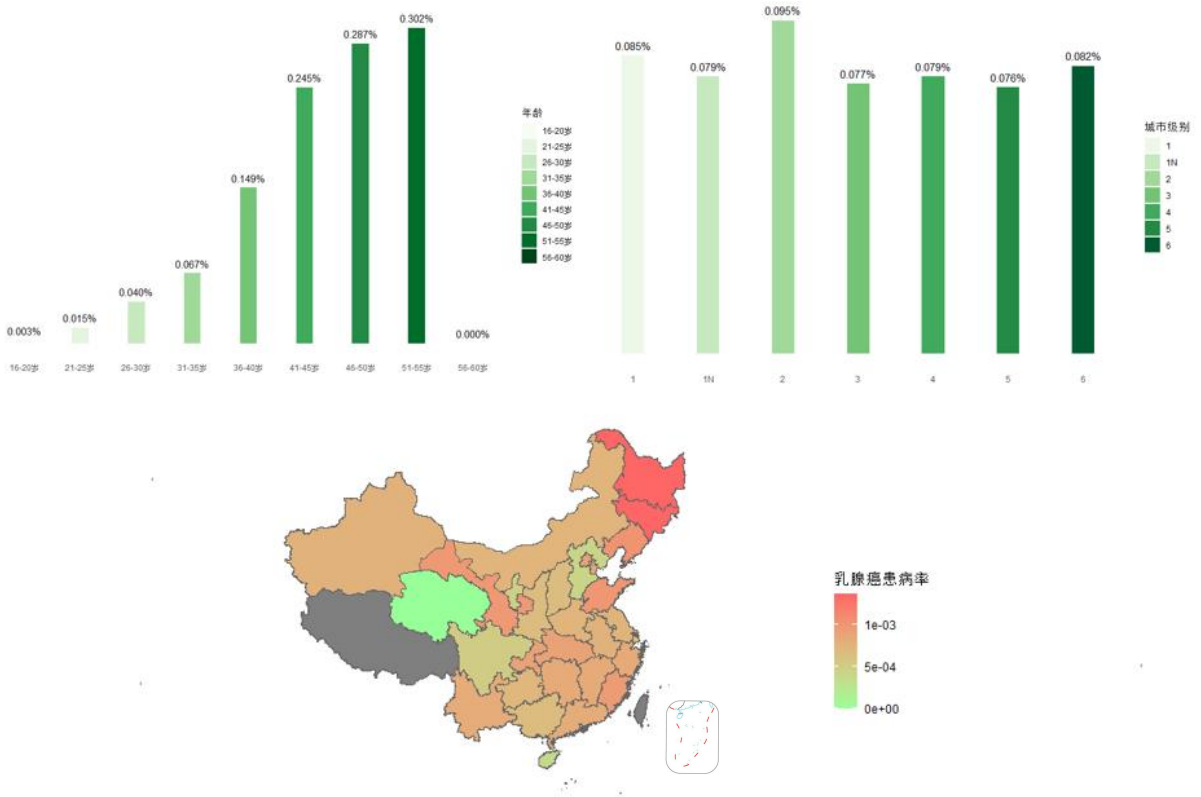
被保险人是否患急性心梗的影响因素cox回归结果表

急性心梗高患病风险人群特征：

- 男性
- 45岁以上（与年龄正相关）
- 超重率高的地区

3.5 重要病种分析·乳腺癌及妇科癌症

乳腺癌发病增多，宫颈癌，子宫癌，阴道癌，卵巢癌为常见的妇科癌症，保单中女性占比54%



乳腺癌不同性别患病率对比（左上）；不同年龄段患病率对比（左下）；
不同省份患病率对比（右上）；不同级别城市患病率对比（右下）

变量	变量取值	风险比（乳腺癌）	风险比（妇科癌症）
性别	男性	0.004	
年龄	21~25岁	4.046	0.796
	26~30岁	10.926	4.190
	31~35岁	19.034	7.477
	36~40岁	43.470	13.092
	41~45岁	71.292	17.711
	46~50岁	80.247	23.973
	51~55岁	73.912	25.715
	56~60岁	0.003	0.002
年平均温度	15-22℃	1.221	
	>22℃	1.338	
年日照时数	1400-2200h	1.214	
	>2200h	1.621	

被保险人是否患乳腺癌/妇科癌症的影响因素cox回归结果表

乳腺癌高患病风险人群特征：■ 女性 ■ 40-55岁（70倍风险） ■ 高温、日照长

妇科癌症高患病风险人群特征：■ 36-55岁

4. 结论与建议·结论

- 患重疾（出险）的风险及其增速与**年龄正相关**，**30岁以上**和**45岁以上**为重点人群，影响重疾风险的重要因素对于癌症风险的影响更明显。
- 疫情发生后，**原本潜在重疾风险高的人群患病可能加重、层次差异变大**。受疫情影响更大的地区生活的人群存在更多受负面影响的可能。
- 重疾发生风险存在地区差异。**气候更适宜、生活节奏较慢、自然资源更为丰富或人群整体生活习惯更健康**的地区患病风险较低。如**天津市、重庆市**等。
- 近年来**女性**各种疾病发病率显著上升。男性患重疾的风险仅为女性的70%，患癌症的风险仅为女性的47%左右，出险最多的几种重疾女性的患病风险也均高于男性（急性心肌梗塞除外）。

4. 结论与建议·建议

- 针对**疫情**爆发的不确定性，公司应该**提高对原本重疾风险高的群体的关注度**，此外针对疫情形势下的焦虑情绪可以考虑为客户提供**心理疏导服务**，避免焦虑情绪引发更严重的心理问题及重疾相关并发症。
- 针对部分疾病**患病率接近全重疾患病率和存在特定人群高发重疾**的现象
  - **差异化核保政策，**针对早期易于检出的重疾和患病率高的重疾（甲状腺癌、肺癌、乳腺癌等）采取更为严格的核保政策，如核保时进行轻症甲状腺癌的筛查，妇科癌症筛查。
  - **分层设计保额，甲状腺癌**在出险保单中占比最高，而且2020年轻度疾病定义的引入，轻症甲状腺癌的赔付率大幅提高，建议对轻症中症重症设计不同的赔付额度，控制赔付风险。
  - 调整**保险责任**，增加高患病率、轻症、中症、男性/女性特定疾病额外赔付等保险责任，适当提高保费，在赔付时针对这部分保险责任设置差异化的赔付标准，控制赔付风险。为客户提供额外保障的同时使得已知高患病风险疾病的赔付处在可控范围。
  - **女性**的患病风险普遍偏高，而且妇科癌症（宫颈癌，卵巢癌等）近年受到广泛关注，可以根据女性的患病风险特点设计一款**针对女性的重疾险产品。**
- 对投保客户提供投保期间的**健康检测服务**，客户患重疾的风险往往与**生活习惯**运动习惯有关，定期提醒客户进行健康管理，为客户提供体检服务，早诊断早筛查避免疾病由轻症转为重症，降低客户患重疾的风险。
- 设计产品时，适当借助再保公司的力量，借助再保公司的海量数据和先进产品，进一步挖掘重疾患病风险影响因素。

5. 风险预测·模型与变量选择

- **数据集**：按照出险情况1：3随机抽取12000个数据，测试集比例为0.2
- **输入**：投保人的个人数据
- **分类目标**：**两年内**出险/未出险
- **效果指标**：预测准确率

随机森林
74.4%

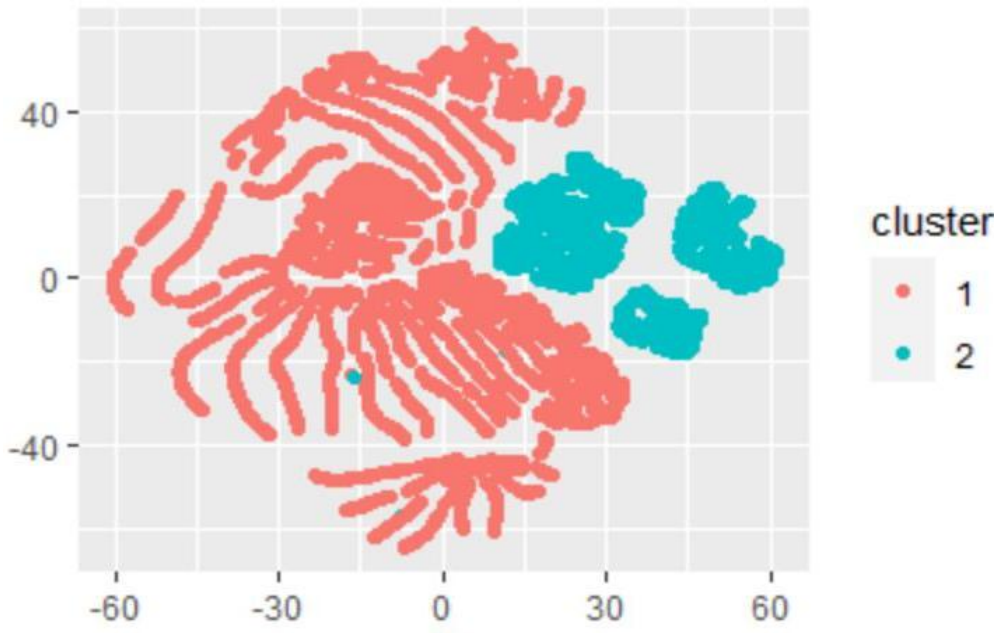
xgboost
75.1% √

四种输入数据方案的预测准确率比较：

- 原始数据：性别、年龄、城市、城市等级-74.2%
- 整理后数据：个人信息、城市数据-74.4%
- **3.1筛选出变量-75.1%**
- 全部变量-74.5%

5. 风险预测·模型与变量选择

- **二分类模型**：两年内出险概率
- **多分类模型**：风险等级和相对风险水平（风险等级/人群平均风险等级）



PAM聚类结果的降温可视化图散点图

风险等级划分

考虑特征：投保人的患病等级、出险的时间跨度、出险年龄（出险保单）

方法：PAM聚类

结果：2类，患高等级重疾（如癌症、器官移植与肾衰竭）+其他种类重疾的投保人

- **低（未出险）**
- **中（因中下等级重疾出险）**
- **高（因高等级重疾出险）**

5. 风险预测•数据不平衡问题

100万条数据中两年内出险的数据仅有2922条，

数据高度不平衡！

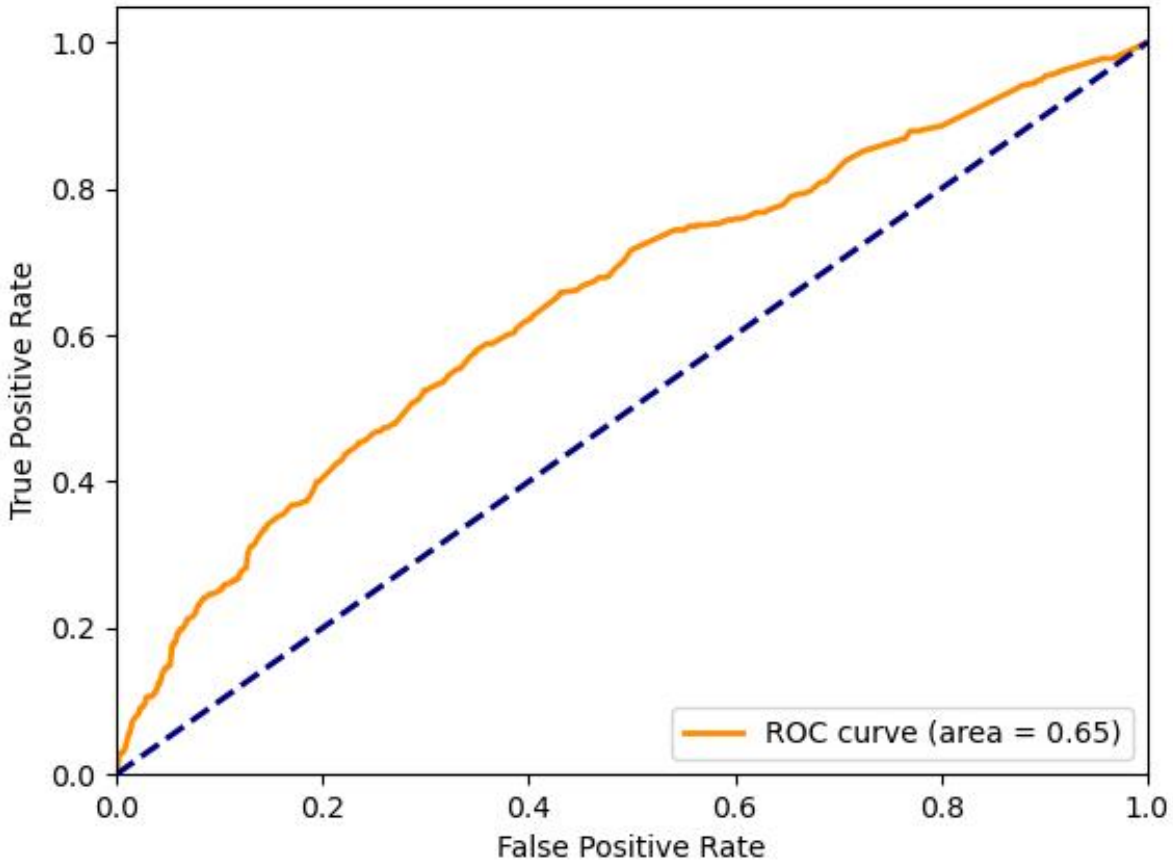
模型预测准确率虚高！

给出险数据加权-不适用

丢弃部分未出险数据- 准确率75.1% 精度53.1%

进行1：3抽样，多次结果求平均- 准确率75.7% 精度69.2% √

5. 风险预测·模型表现



二分类模型ROC曲线

- **数据集**：保证出险与未出险保单数目接近1：3，每次样本量约为10000，测试集样本数比例为0.2
- **输入**：投保人的年龄、性别、城市数据
- **分类目标**：二分类/多分类

二分类：

- 预测准确率：76%
- 精度：69%
- AUC=0.65

多分类：

- 预测准确率：56%

5. 风险预测·预测工具

预测工具输入界面

预测工具输出界面

“寿再青骏杯”2022年全国研究生案例分析大赛

基于生存分析和xgboost的重疾赔付风险研究

谢谢观看！

时间：2022年7月