基于历史数据的重大疾病保险风险水平分析报告

六边形团队

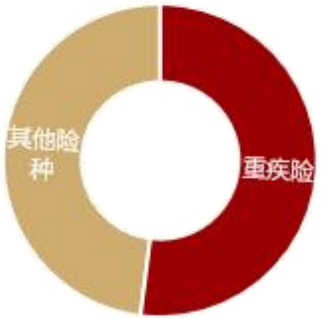
PART 01 ▸

研 究 背 景

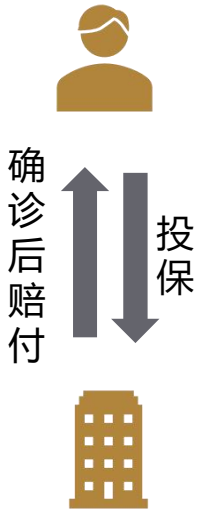
1.1 重疾险险种

重疾险

2021年健康险原保费收入

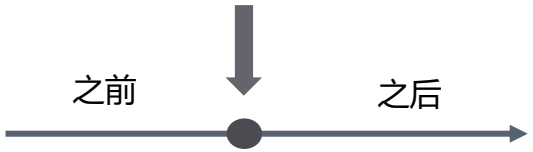


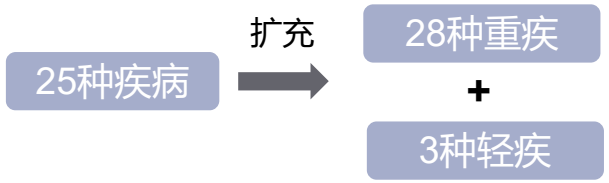
- ✓ 重疾险：52%
- ✓ 其他险：48%



政策变化

《重大疾病保险的疾病定义使用规范（2020年修订版）》





变更原因

- 随着中国医疗技术的发展，自2007年开始沿用的部分重疾险定义已不再适用。
- 例如**甲状腺癌**在近年出险概率激增，但其治疗费用远低于其他恶性肿瘤，这使得保险公司承受较大的赔付压力。

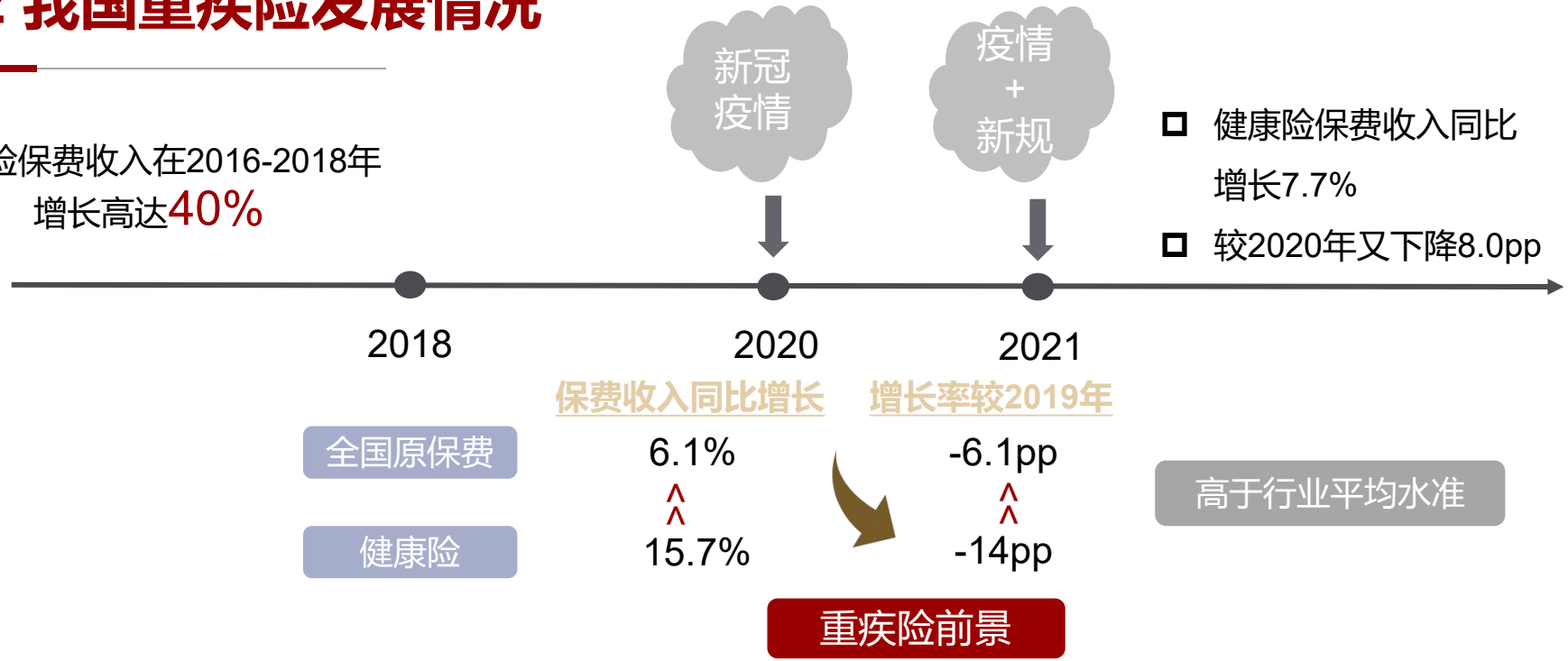
数据来源：2021年统计年鉴

“重疾新规”保障的28种重疾

严重恶性肿瘤	深度昏迷
较重急性心肌梗塞	严重良性颅内肿瘤
严重脑中风后遗症	双耳失聪
冠状动脉搭桥术	双目失明
严重慢性肾脏病	多个肢体缺失
重大器官移植术或造血干细胞移植术	急性重症肝炎或亚急性重症肝炎
慢性肝功能衰竭失代偿期	严重脑炎或脑膜炎后遗症
瘫痪	严重运动神经元病
心脏瓣膜手术	语言能力丧失
严重阿尔兹海默病	严重慢性呼吸衰竭
严重脑损失	主动脉手术
严重III度烧伤	严重克罗思病
严重原发性帕金森病	重型再生障碍性贫血
严重特发性肺动脉高压	严重溃疡性结肠炎

1.2 我国重疾险发展情况

重疾险保费收入在2016-2018年
增长高达40%



- 健康险保费收入同比增长7.7%
- 较2020年又下降8.0pp

高于行业平均水准

- 销售的乏力
- 明显的增速下降或意味着重疾险逐渐由成长期迈入成熟期

- 长期来看重疾险仍有一定的**发展潜力**

从政策导向上看，健康产业是未来重点支持发展的对象

- 2016年党的十八届五中全会：明确提出推进健康中国建设，鼓励企业、个人参加商业健康保险，鼓励开发与健康管理服务相关的健康保险产品
- “十四五”规划对全面推进健康中国建设进行了重点强调
- 2020年，习总书记分别在统筹推进新冠肺炎疫情防控和经济社会发展工作部署会议上和湖北省考察新冠肺炎疫情防控工作时两次指出：“健全重大疾病医疗保险和救助制度”

从需求端看，我国重疾险需求较大

- 2021年我国人均医疗保健消费支出2115元，占人均消费支出的8.8%，相比2016年7.6%的比重有明显提升，人民对健康生活的关注度越来越高。
- 随着人们对保险的认知逐渐深入，为自己与家庭成员配置重疾险以规避重疾带来的巨额风险敞口，已成为重要的开支项目。
- 全球经济发展导致的环境污染、气候变暖等问题使得居民生活环境更加恶劣，促使支气管炎、哮喘等呼吸疾病高发、新冠肺炎疫情肆虐全球，其风险无法提前预测和短期有效防范。

1.3 研究目标

困境

重疾险是荔枝人寿健康险的主力产品线，保障的重疾和轻症的病种数量在100种左右，除行业标准定义外有多个公司自己设定的非标准定义。然而，近两年既往销售的一款主力重大疾病保险产品赔付情况持续不乐观，这也许意味着公司在未来将面临着较大的长期赔付风险。

目标

使用历史数据对公司重疾险的赔付风险进行分析，从而实现风险的预测

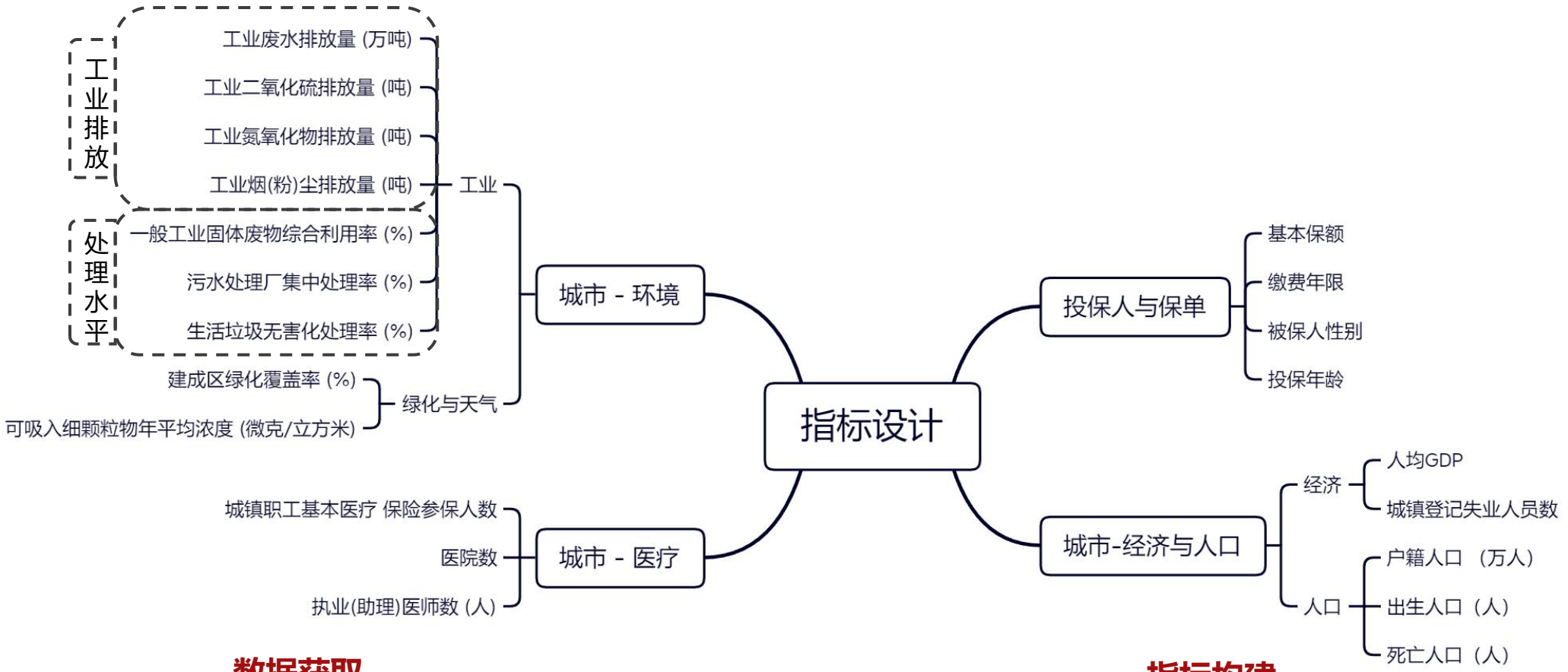
数据概况

- 投保日期在2016-2018年的保单数据，避免近年来新冠疫情与重疾新规带来的外部冲击对模型结果造成影响
- 数据涵盖保单基础数据（投保人年龄性别、投保日期、保额、行政区域、缴费年限、是否理赔等）以及各市级行政区宏观年鉴数据（城市等级、空气质量、经济发展水平、医疗基础设施建设等）
- 样本数据共100万条，其中赔付保单9,283件，即赔付占比0.93%

PART 02

数 据 介 绍

2 数据介绍



数据获取

- 获取2016-2020年各个城市的环境、医疗、经济与人口相关信息。
- 数据来源：省级、市级统计年鉴

指标构建

- 计算2016年-2020年各个信息的复合增长率

↓

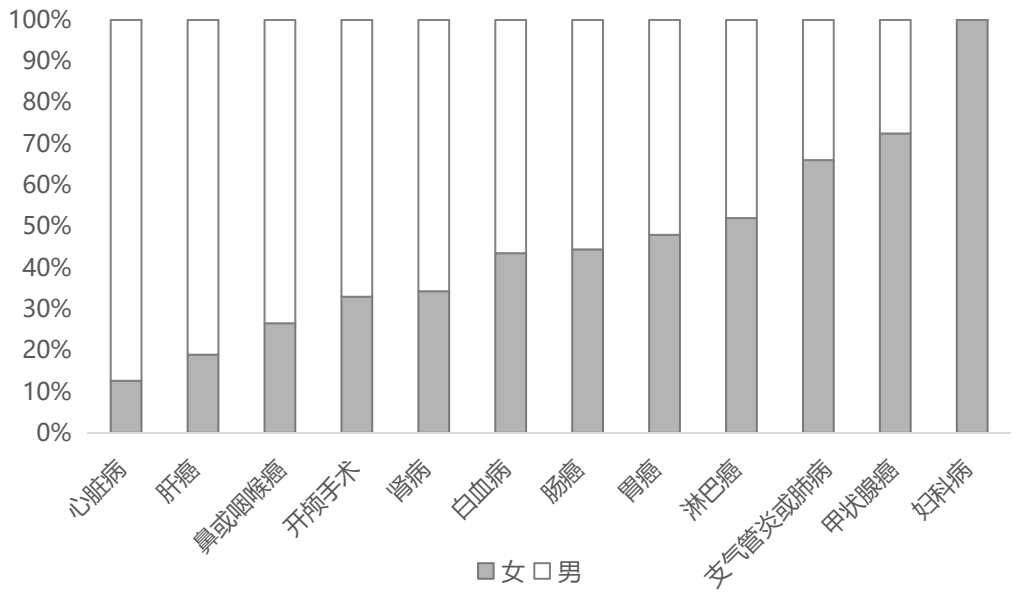
城市发展

PART 03

描述性统计

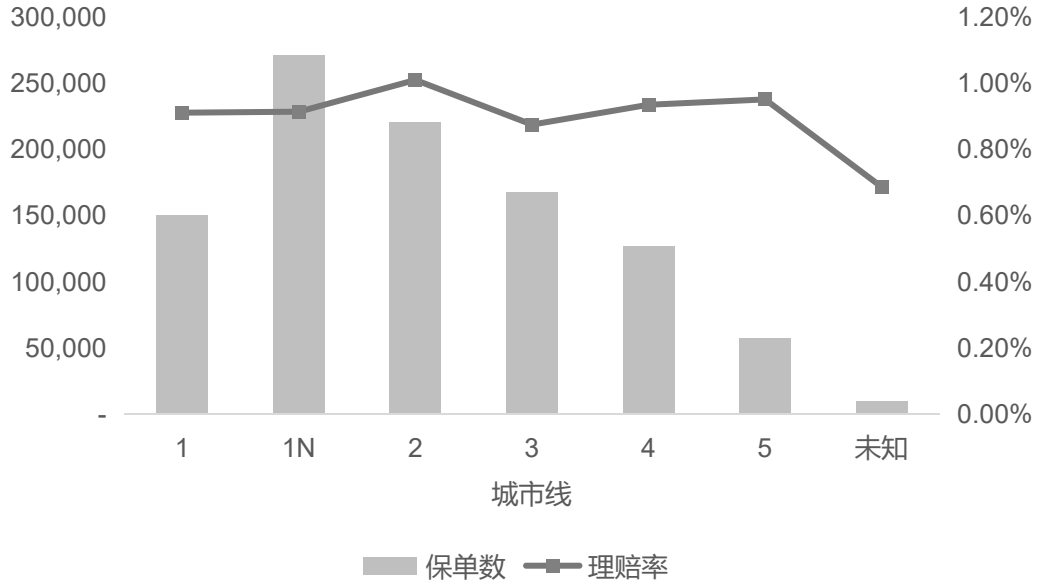
3 描述性统计

常见重疾男女发病比例柱状图



- 筛选出发生频率最高的12个常见病种，相关保单共8,109件，覆盖总赔付保单数的87.35%。
- 除了妇科癌症为女性特有、极少男性患乳腺癌之外，甲状腺癌、支气管和肺癌的投保患者多为女性。
- 肝癌、鼻咽癌、心脏病的投保患者则多为男性。

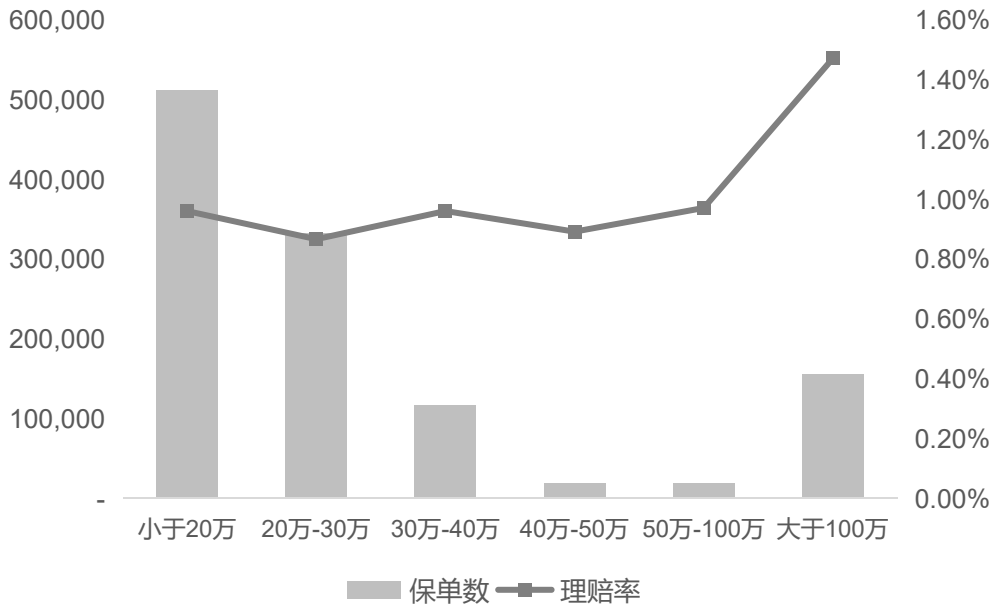
城市线 理赔率与保单数组合图



- 城市线为1N的保单数量最多，但2线城市的理赔比率最大，且各个城市线的理赔率差距较小。
- 3线城市至5线城市保单数量逐渐减少，但是理赔率却逐渐上升。

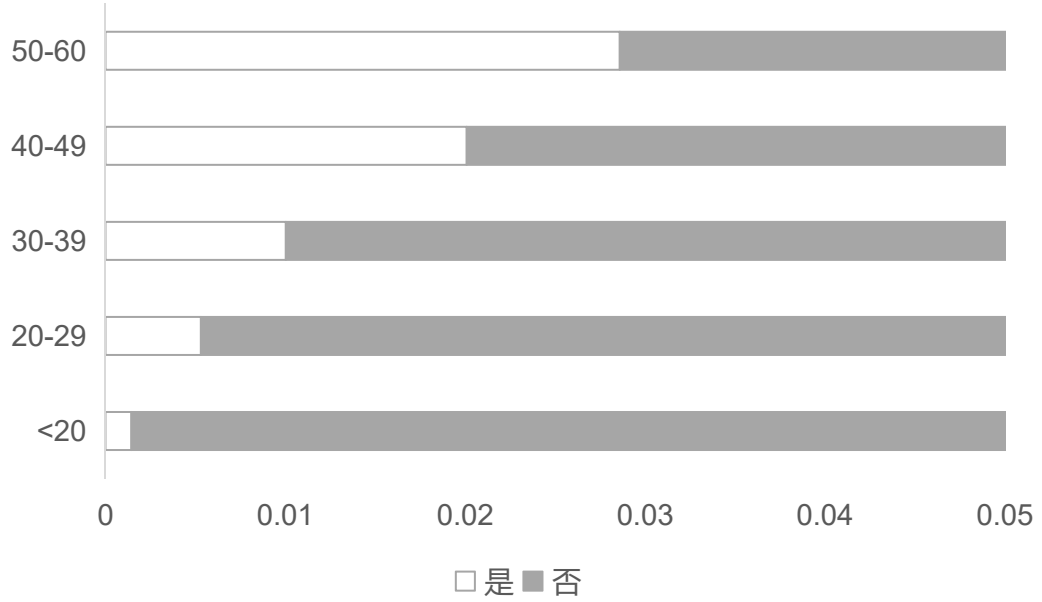
3 描述性统计

基本保额范围 理赔率与保单数组合图



- 大量保单的基本保额小于20万，少量的保单数量大于30万。
- 大于100万的保单的理赔率远高于其他保额段的理赔率。

投保年龄理赔率条形图



- 小于20岁的保单的出险率最低，当大于50岁时，出险率高达2.85%。
- 随着投保年龄的增加，出险的概率也逐步增加，整体呈现出“越年轻越健康”态势，这与一般认知相符。

PART 04 ▸

模 型 构 建

4.1 模型选择

重疾险的风险预测模型

O－E 法

发病率= 发生次数 / 经验暴露数

需要大样本
无法考虑影响因素

随机过程法

Cox模型（比例风险模型）

回归分析法

逻辑回归

中小样本
可考虑多因素影响

考量因素

样本量：100000 ＋ 数据具有删失 ＋ 探究多因素的影响 =

最终选择

Cox 比例风险模型
逻辑回归

4.2 数据不平衡问题

数据不平衡

大量分类模型基于此假设

正常情况下

正样本占比：51%

≈ 约等于或差距悬殊

负样本占比：49%

VS

比例不平衡下

正样本占比：99.07%

>> 远大于且差距非常悬殊

负样本占比：0.93%

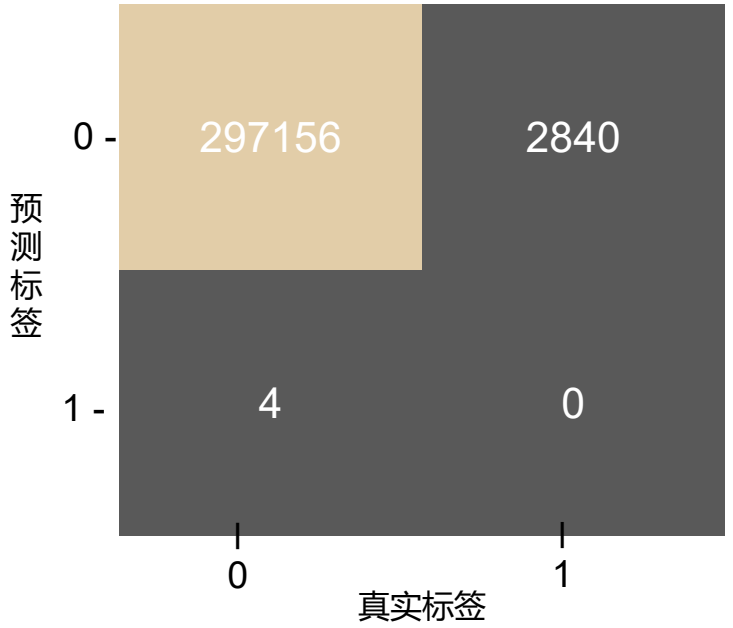
本案例下的比例

简单模型下的尝试

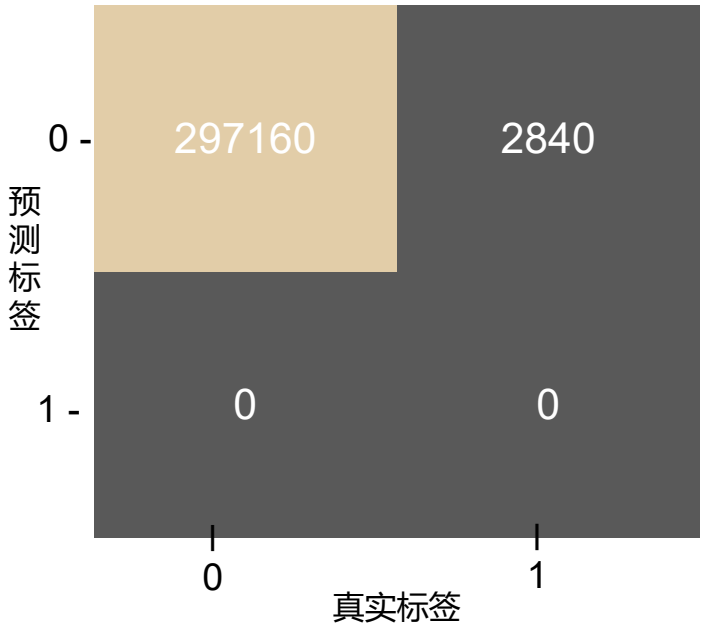
样本不平衡引发的问题

- ✓ 无法用朴素的模型对数据进行拟合
- ✓ 导致模型学习偏差，模型会倾向于将结果预测为大众样本
- ✓ 常规的总体评价指标无法提供有效的信息（如ROC曲线也会对分类器性能展示出过度乐观的结果）

逻辑回归



随机森林



4.2 数据不平衡问题

过采样

对于数据量少的样本，采样并通过复制加入原始样本集合，以此来扩充数量，使得正负类别平衡

存在的问题

过采样只是将复制数据重复添加到原始数据集，会使得某些句子的多个实例变得“束缚”，导致过度拟合

欠采样

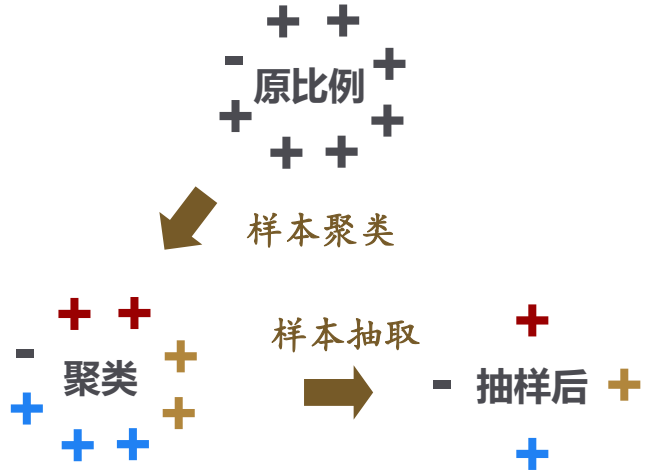
对于样本量多的数据，随机从数据量多的类别中移除部分样本

存在的问题

导致分类器遗漏与多数类相关的重要特征。

算法改进

- ✓ 将大众样本（正样本）聚为k类，并在上述k类中每一固定的类中抽取一个个体作为新的正样本
- ✓ 能够在尽可能不损失信息地情况下解决数据不平衡问题



4.3 逻辑回归

特征维度	变量名称	回归系数	P值	显著程度
投保人与保单	缴费年限	-0.045	0.001	***
	被保人性别（基准组：男性）	0.367	<0.001	***
	对数投保额度	0.571	<0.001	***
	投保年龄	0.083	<0.001	***
城市经济与人口	年末户籍人口	-0.001	<0.001	***
	城镇登记失业人员数增长率	0.002	0.001	***
	对数人均地区生产总值	-0.113	0.081	*
城市环境	对数工业二氧化硫排放量	0.134	<0.001	***
	工业烟(粉)尘排放量增长率	0.005	<0.001	***
	一般工业固体废物综合利用率	-0.006	<0.001	***
城市医疗	医院数	-0.012	<0.001	***
	对数执业(助理)医师数	-0.198	<0.001	***
	城镇职工基本医疗保险参保人数增长率	-0.05	<0.001	***
伪R2			0.13	

投保人与保单

- 被保人性别：女性比男性更有可能进行理赔。
- 投保年龄：年龄越大，出险的概率会更高。
- 对数投保额：投保额度越高，出险概率会更高。
- 推测是由于逆选择的问题存在，更有可能理赔的投保人会倾向于选择更高的额度。

城市—经济与人口

- 城市年末户籍人口对于被保人发生重疾并理赔的概率有显著的负向影响。
- 对数人均地区生产总值对于被保人发生重疾并理赔的概率有显著的负向影响。
- 城镇登记失业人员数增长率对于被保人理赔概率都有显著正向影响
- 推测更具规模的城市生活条件越好，被保人的身体素质也普遍向好。

4.3 逻辑回归

特征维度	变量名称	回归系数	P值	显著程度
投保人与保单	缴费年限	-0.045	0.001	***
	被保人性别（基准组：男性）	0.367	<0.001	***
	对数投保额度	0.571	<0.001	***
	投保年龄段	0.083	<0.001	***
城市经济与人口	年末户籍人口	-0.001	<0.001	***
	城镇登记失业人员数增长率	0.002	0.001	***
	对数人均地区生产总值	-0.113	0.081	*
城市环境	对数工业二氧化硫排放量	0.134	<0.001	***
	工业烟(粉)尘排放量增长率	0.005	<0.001	***
	一般工业固体废物综合利用率	-0.006	<0.001	***
城市医疗	医院数	-0.012	<0.001	***
	对数执业(助理)医师数	-0.198	<0.001	***
	城镇职工基本医疗保险参保人数增长率	-0.05	<0.001	***
伪R2			0.13	

城市—环境

- 对数工业二氧化硫排放量增长率会显著增加被保人理赔的概率，意味着二氧化硫更可能诱发重疾。
- 更高的一般工业固体废物综合利用率能够减小模型的倾向性得分，说明在环境保护方面，工业企业的努力至关重要。

城市—医疗

- 医院数、对数医师数、和城镇职工基本医疗保险参保人数增长率都能够降低被保人理赔的倾向性得分.
- 保司也更应关注被保人城市的医疗条件，并追踪数据进行迭代更新，为业务方的决策提供更及时有效的信息。

4.4 Cox回归：原理介绍

相关函数

生存函数： $S(t,X) = \Pr(T > t, X)$

死亡函数： $F(t,X) = \Pr(T \le t, X)$

死亡密度函数： $f(t,X) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T \le t+\Delta)}{\Delta t} = F'(t,X)$

风险函数： $h(X) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T \le t+\Delta | T \ge t, X)}{\Delta t} = \frac{f(t,X)}{S(t,X)}$

模型基本假设

假设1：Hazard Ratio不随时间变化，满足比例风险假设，即Proportional Hazards Assumption, PH假定.

假设2：对数线性假设: 协变量应与对数风险比呈线性关系。

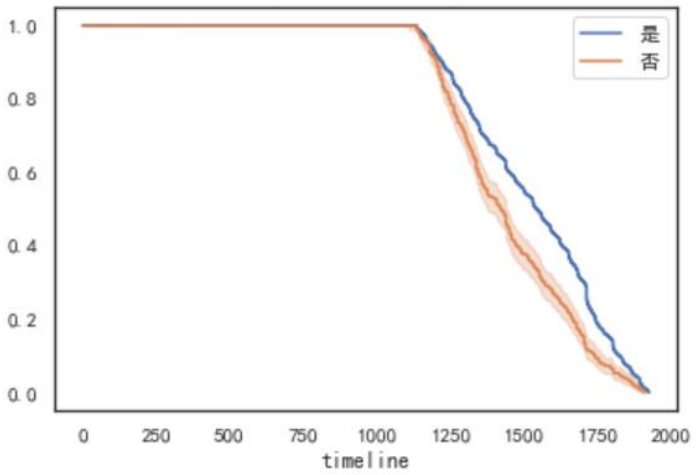
Cox回归基本形式与转化

基本形式： $h(t,X) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m)$

转化形式： $\ln\left[\frac{h(t,X)}{h_0(t)}\right] = \ln(RR) = \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_m X_m$

相对风险： $RR = h(t,X_i) \setminus h(t,X_j) = \exp(\beta'(X_i - X_j))$

KM生存曲线



- 根据事件发生的时间与观测时间对生存函数进行拟合。
- 在相同的暴露时间下，年龄大于55岁的保单会更容易出险。

4.4 Cox回归：回归系数解读

特征维度	变量名称		回归系数	EXP(回归系数)	P值
投保人与保单	基本保额段（基准组：20万-100万）	<20万	-0.00	1.00	0.966
		>100万	-0.08	0.92	<0.005
	投保年龄		-0.05	0.95	<0.005
城市经济与人口	年末户籍人口		0.06	1.07	<0.005
	城镇登记失业人员数		0.02	1.02	<0.005
	人均地区生产总值		-0.02	0.98	<0.005
城市环境	建成区绿化覆盖率		-0.03	0.97	<0.005
	工业烟(粉)尘排放量 增长率		0.01	1.01	<0.005
	工业废水排放量 增长率		0.04	1.04	<0.005
	工业二氧化硫排放量 增长率		0.1	1.11	<0.005
	生活垃圾无害处理率		-0.09	0.91	<0.005
	建成区绿化覆盖率 增长率		-0.11	0.9	<0.005
城市医疗	执业(助理)医师数		-0.07	0.94	<0.005
	医院数		-0.03	0.97	<0.005
C-index				**0.535**	

投保人与保单

- 随着投保年龄的上升，疾病的发生率会上升。当投保年龄每相对增加10岁，则出险概率相对增加5%。
- 当基本保额大于100万时，表明该保单存在高概率出险的可能性，因此需要相应的增加保费。

城市 - 经济与人口

- 可以看到户籍人口越多、失业人员越多、人均GDP越低疾病发生率越高，这表明保单所在地的经济状况一定程度上间接影响疾病的发生率。

城市 – 环境

- 绿化覆盖率越高，且绿化覆盖率的增长率越快，工业相关的排放越少，生活垃圾处理率越高，表明该城市的环境状况更好，则发病率越低。

城市 – 医疗

- 医师数越多、医院数量越多，表明该地区的医疗条件更好，则位于该地区的保单的被保人发病率更低。

PART 05

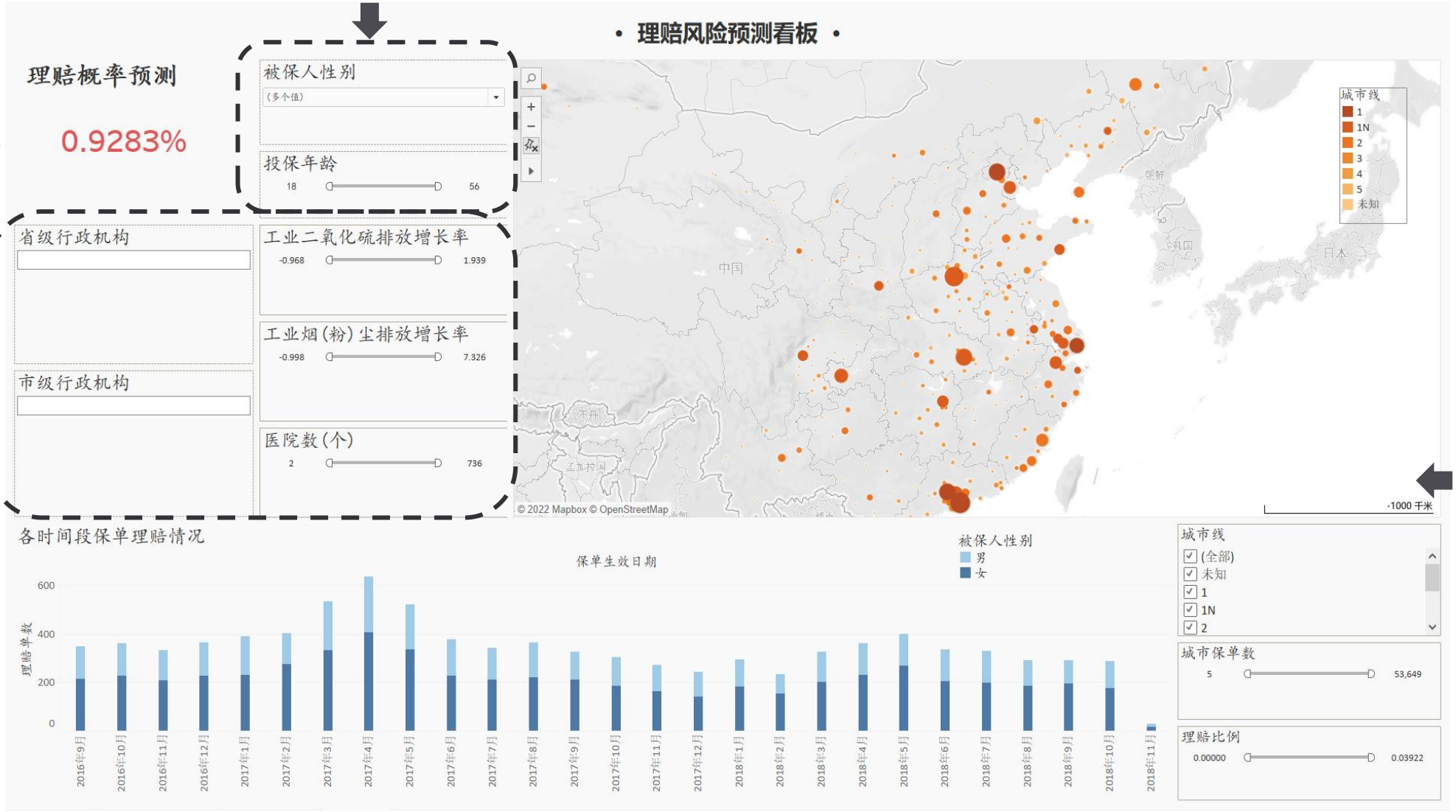
风险预测工具

5 风险预测工具

被保人
信息输入

被保人
预测理赔率

城市
信息输入



历史理赔
情况输出

5 风险预测工具：交互展示

PART 06

结 果 展 望

6 总结与展望

总结

发展潜力 + 较高赔付率

大量保单数据 + 城市相关数据

根据数据特征

逻辑回归 Cox回归

实现出险识别

了解特征影响

特征总结

- 年龄上升，出现率上升
- 投保额上升，出现率上升
- 工业相关的排放越少，环境越好，出险率越低
- 户籍人口越多、失业人员越多、人均GDP越低，出险越高
- 医师数越多、医院数量越多，出险率更低。

展望

解决行业痛点

- ✓ 保险经营风险的识别和防范对于保险行业具有重要意义。
- ✓ 对于保司，如何利用自身的知识和信息尽可能解决逆选择问题一直是行业痛点
- ✓ 本案例通过建模为保司识别重疾险风险提供显著指标和具有建设性的建议。

更平衡的数据

- ✓ 随着业务地不断推进和发展，保司能够积累更具规模更具平衡性的数据集。
- ✓ 通过统计方法更高效更精确地预测被保人的理赔概率
- ✓ 缓解保司与被保人的信息不对称问题后保司能够在保证自身发展的前提下更好地服务人民群众。

技术革新

- ✓ 当前正是数字化红利释放到保险业的扩展阶段。
- ✓ 保险行业数字化，尤其是风险识别数字化，将成为行业可持续发展的关键驱动因素和内生动力，推动行业发展实现正反馈，促进保险业公平公正蓬勃发展。

谢谢！