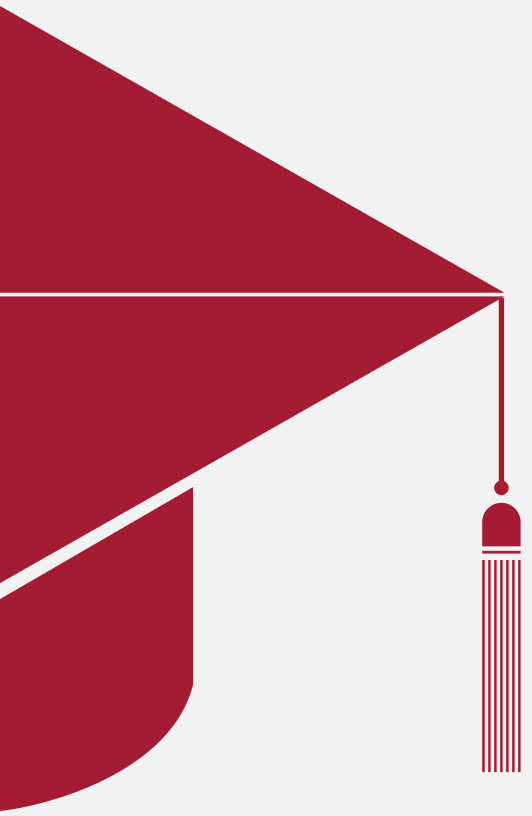


重大疾病保险“水晶球”

Part 1

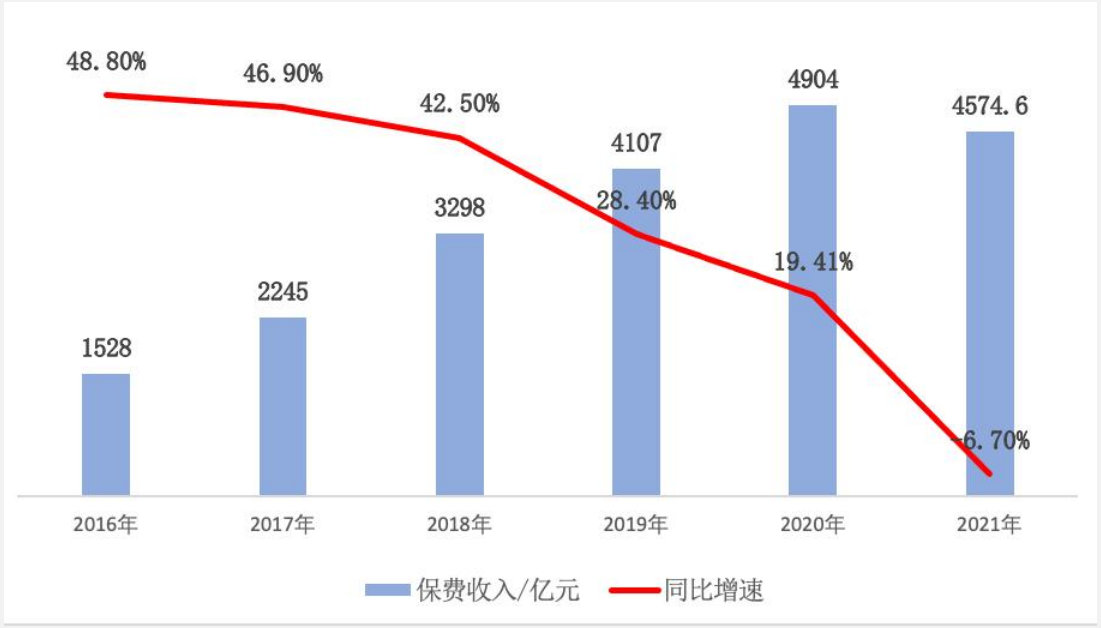
研究问题背景与目的

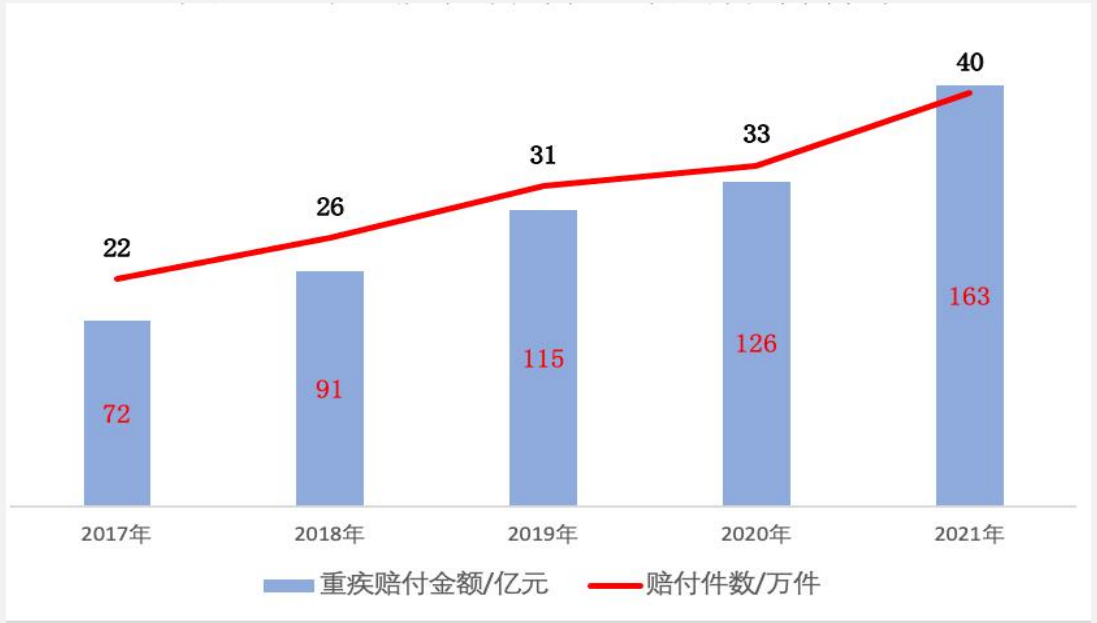
研究背景

重大疾病保险一直是我国保险市场中绝对的主力健康险险种。

- 重疾保费收入面临巨大压力，增速持续下降
- 存量业务较大、重大疾病呈年轻化趋势，赔付风险加大
- 重疾发生率的影响因素不确定，影响风险管理策略

历年重疾险保费收入与同比增速





中国人寿重疾险理赔金额与理赔件数

研究目的

研究目的

在给定数据的基础上扩展数据，挖掘重疾发生的风险要素，解释规律并做现象分析

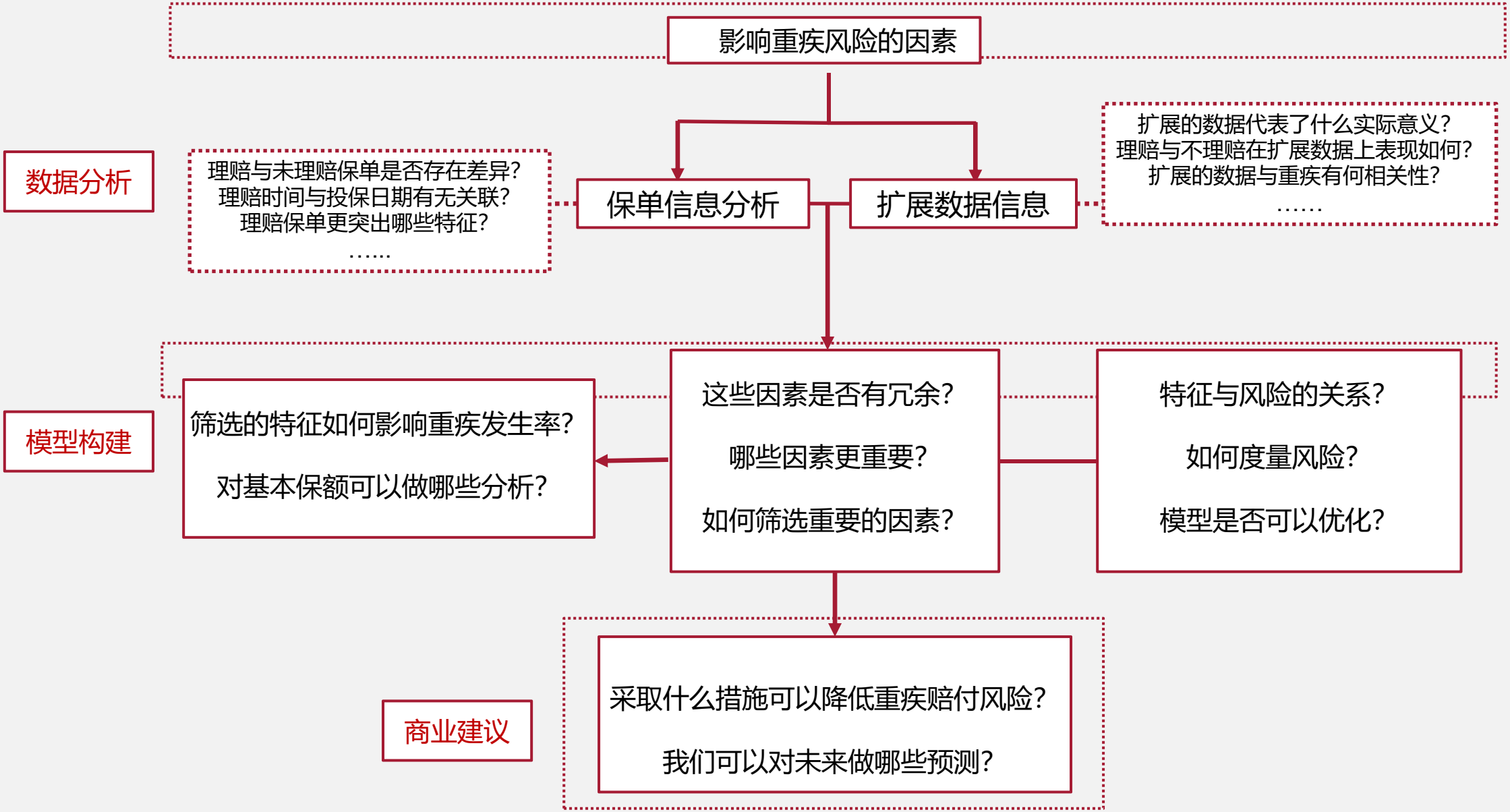
提供降低公司重疾险赔付风险可能的商业建议

设计投保人风险评估模型，为重疾险售前提供参考

Part 2

研究思路

研究思路

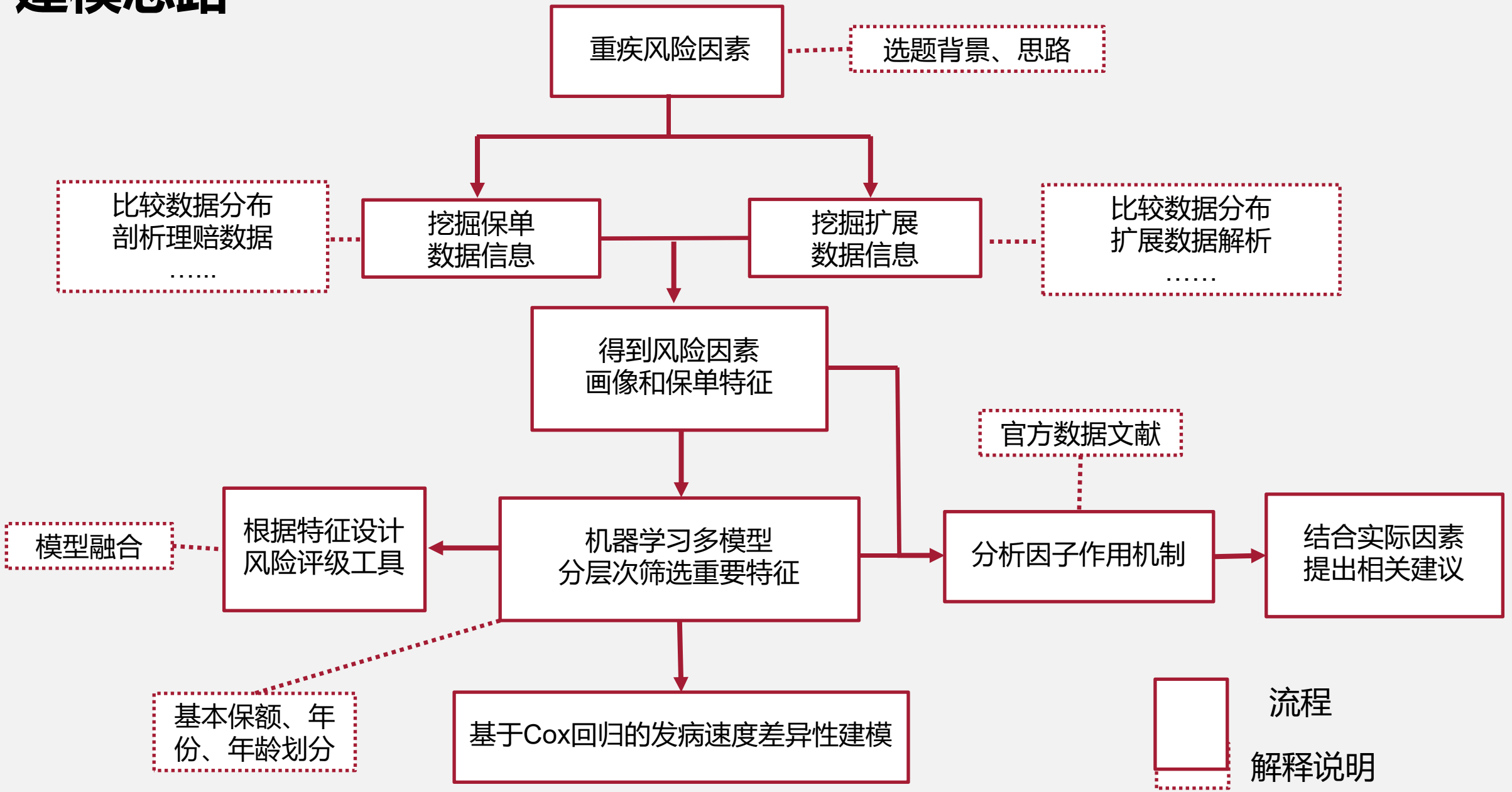

影响重疾风险的因素
数据分析
理赔与未理赔保单是否存在差异?
理赔时间与投保日期有无关联?
理赔保单更突出哪些特征?
……
保单信息分析
扩展数据信息
扩展的数据代表了什么实际意义?
理赔与不理赔在扩展数据上表现如何?
扩展的数据与重疾有何相关性?
……
模型构建
筛选的特征如何影响重疾发生率?
对基本保额可以做哪些分析?
这些因素是否有冗余?
哪些因素更重要?
如何筛选重要的因素?
特征与风险的关系?
如何度量风险?
模型是否可以优化?
商业建议
采取什么措施可以降低重疾赔付风险?
我们可以对未来做哪些预测?

研究思路

- 基于调查数据、文献资料和常识构建因子体系
- 整合其他年鉴、调研报告和官方数据库的相关数据，扩充风险变量池，进一步挖掘重疾风险因素的相关信息

风险因素	具体因素	解释说明
地理环境因素	所属地区	被保人所在地区，如华东地区
	当地气候	被保人所处气候区，如温带大陆气候
	降水量	被保人所在省份全年降水量（mm）
	自然灾害发生率	被投保人所在省份自然灾害发生率
	省级行政机构[1]	被保人所在省份
	市级行政机构[1]	被保人所在地级市
性别年龄因素	性别[1]	被保人性别
	投保年龄[1]	被保人所处年龄
	老年人口占比[3]	被保人所在省份65岁以上人口占比
	0-14岁人口占比[3]	被保人所在省份0-14岁人口占比
医疗卫生因素	入院人数比例[3]	被保人所在省份每年入院诊疗人次占全国人次比例
	医疗卫生机构数比例[3]	被保人所在省份医疗卫生机构数量占全国数量比例
	每万人卫生人员数[3]	被保人所在省份每万人拥有的卫生人员数
	医疗保健人均消费支出比例[3][10]	被保人所在省份医疗卫生消费支出占总支出比例
经济社会因素	人均GDP[3][11]	被保人所在省份的人均GDP
	城市线[1]	被保人所在城市的城市线，如北京为一线城市
	烟酒普及率[4][7]	被保人所在省份烟酒普及率
	人口密度[3]	被被保人所在省份人口密度
	烟尘[13]	被保人所在省份全年烟尘排放量
	二氧化硫[13]	被保人所在省份全年二氧化硫排放量
	氮氧化物[13]	被保人所在省份全年氮氧化物排放量
疾病因素	高血压率[4][15][16]	被保人所在省份人群高血压率
	肥胖率[4][15][16]	被保人所在省份人群肥胖率
	恶性肿瘤发病率[6][17][18]	被保人所在省份恶性肿瘤发病率
	中风占比（城市线）[14]	被保人所在城市线人群中风比例占平均水平比例
	心梗占比（城市线）[14]	被保人所在城市线人群心梗比例占平均水平比例
	糖尿病率[5][15]	被保人所在省份糖尿病率
	心脑血管疾病发生率[5][16]	被保人所在省份心脑血管疾病发生率
	肝病发生率[12]	被保人所在省份肝病发生率
	人均寿命[3][12]	被保人所在省份人均寿命
保单情况	基本保额段[1]	被保人的基本保额
	缴费年限[1]	被保人缴纳保费的最短时间

建模思路


重疾风险因素
选题背景、思路
比较数据分布
剖析理赔数据
……
挖掘保单
数据信息
挖掘扩展
数据信息
比较数据分布
扩展数据解析
……
得到风险因素
画像和保单特征
官方数据文献
模型融合
根据特征设计
风险评级工具
机器学习多模型
分层次筛选重要特征
分析因子作用机制
结合实际因素
提出相关建议
基本保额、年
份、年龄划分
基于Cox回归的发病速度差异性建模
流程
解释说明

Part 3

数据处理与描述性分析

数据预处理（缺失值）

中风占平均比率-城市线、心梗占平均比率-城市线、癌症占平均比率-城市线、城市线这四个变量存在数据缺失

前三个变量的数据缺失实际上是由城市线造成的

部分保单的城市的名称或行政地位发生改变，对于这类缺失本报告通过查找这些城市的资料进行精细填补

部分保单的市级行政机构没有细致给出，而是以 “省直辖市区”记录，这种类型的缺失数据样本量较少，本报告对其采取删除处理。

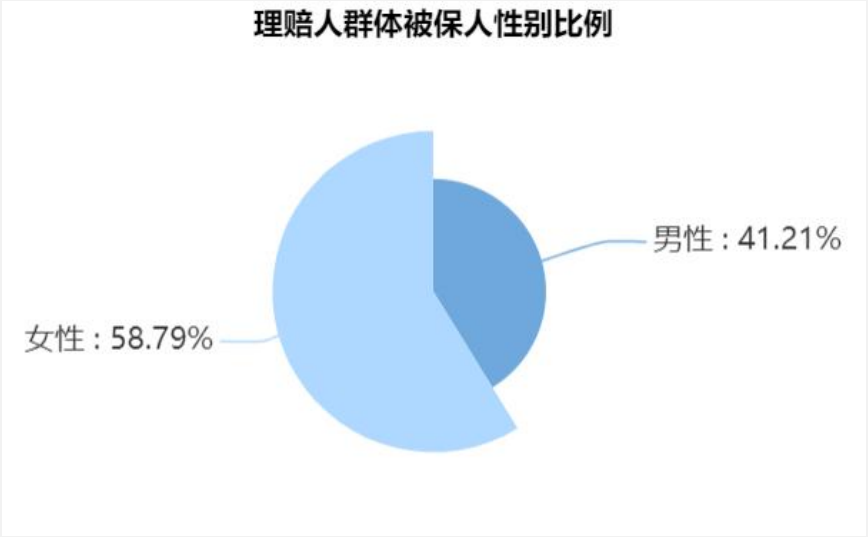
用户画像分析

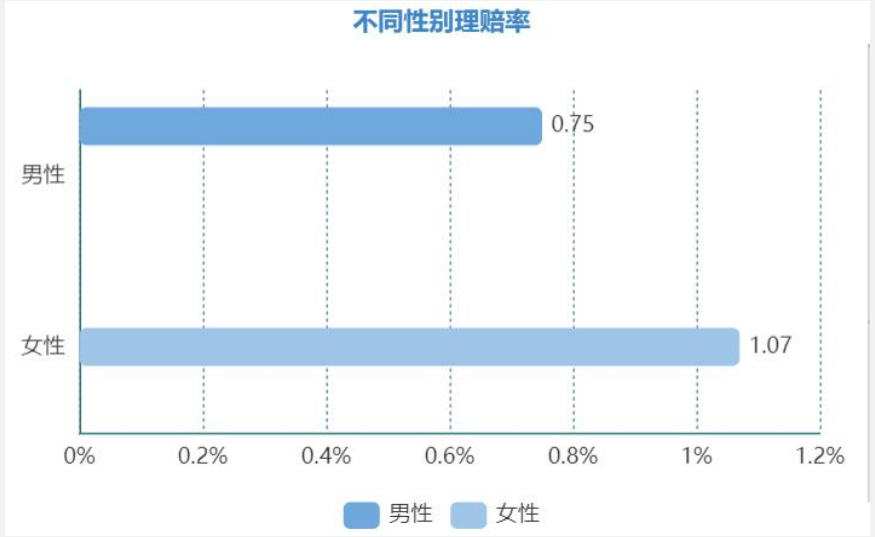
986378条未理赔样本
9248条理赔样本

35个特征（未筛选）

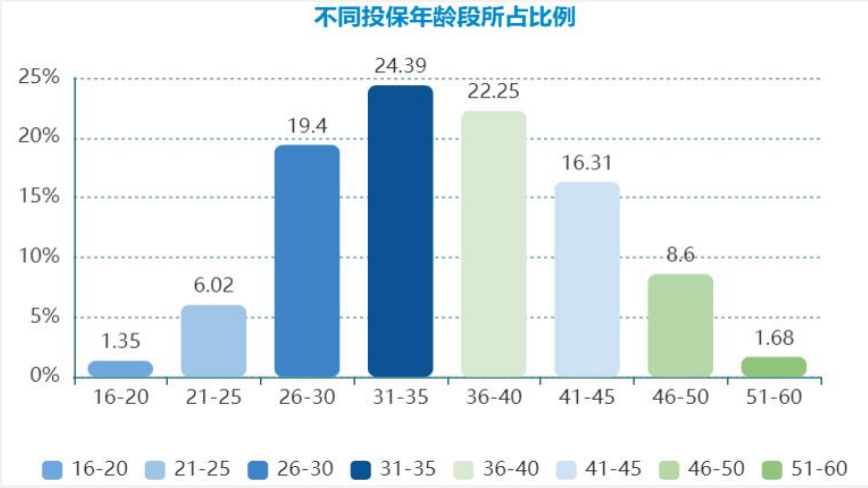
数据集两类样本数量严重不平衡，分析时应重视这一特点

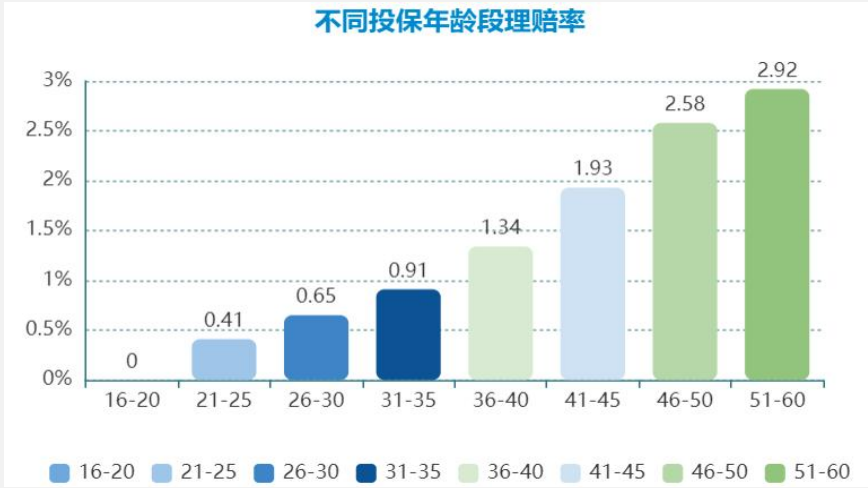
理赔信息初探





- 在理赔群体中，女性所占比例高于男性
- 从总体来看，女性的理赔率也稍高于男性。





- 从投保年龄来看，中年人是购买重疾险的主力军
- 理赔率随投保年龄增大有明显的升高趋势

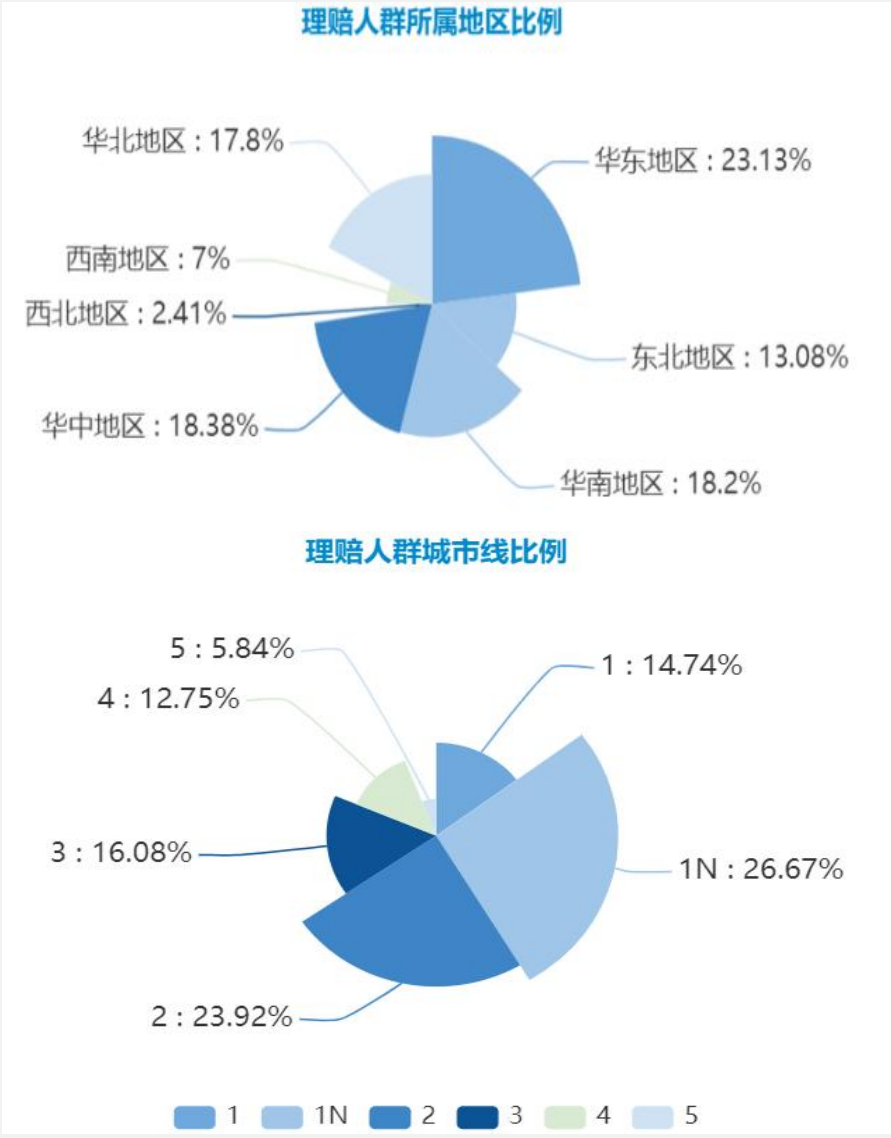
理赔信息初探

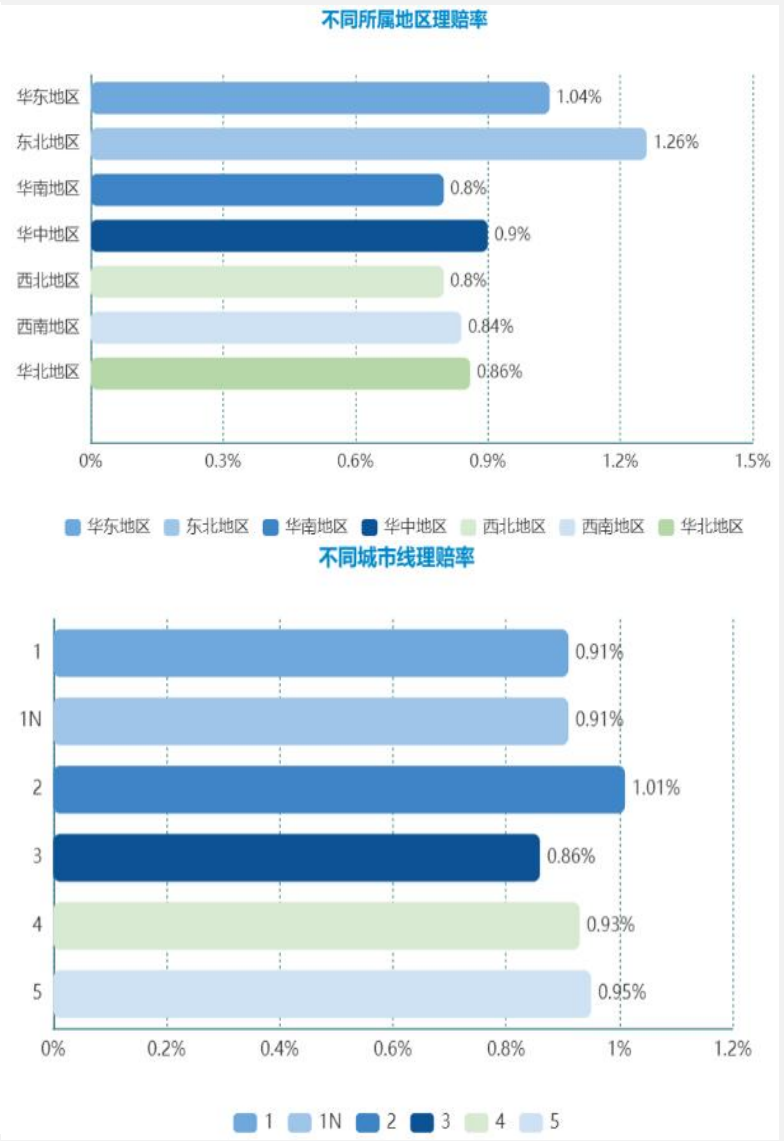
所属地区：

- 理赔人群中，华东地区的所占比例最高，西部地区的占比很低。
- 总体来看，华东、东北地区的理赔率较高，其余地区的理赔率基本相近。

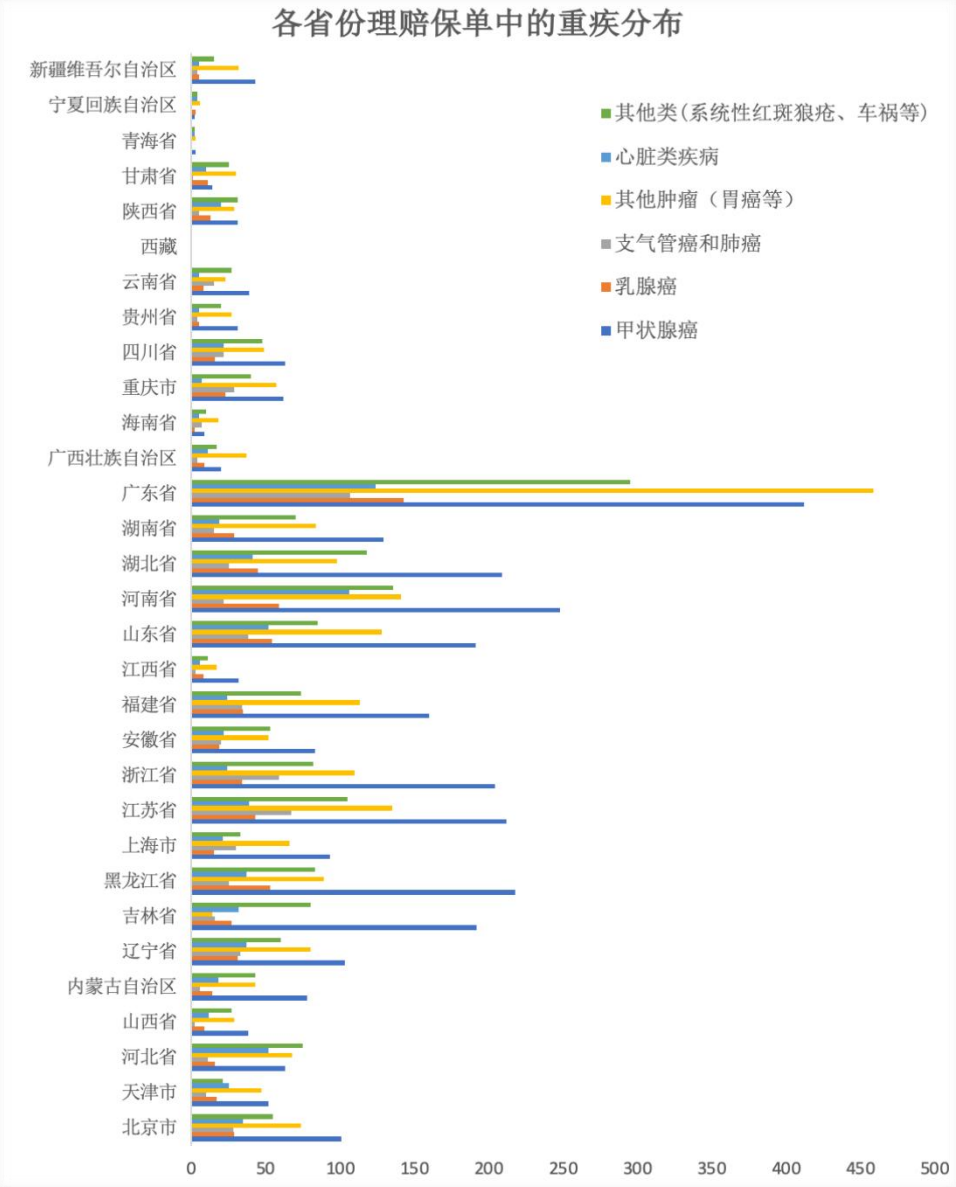
城市线：

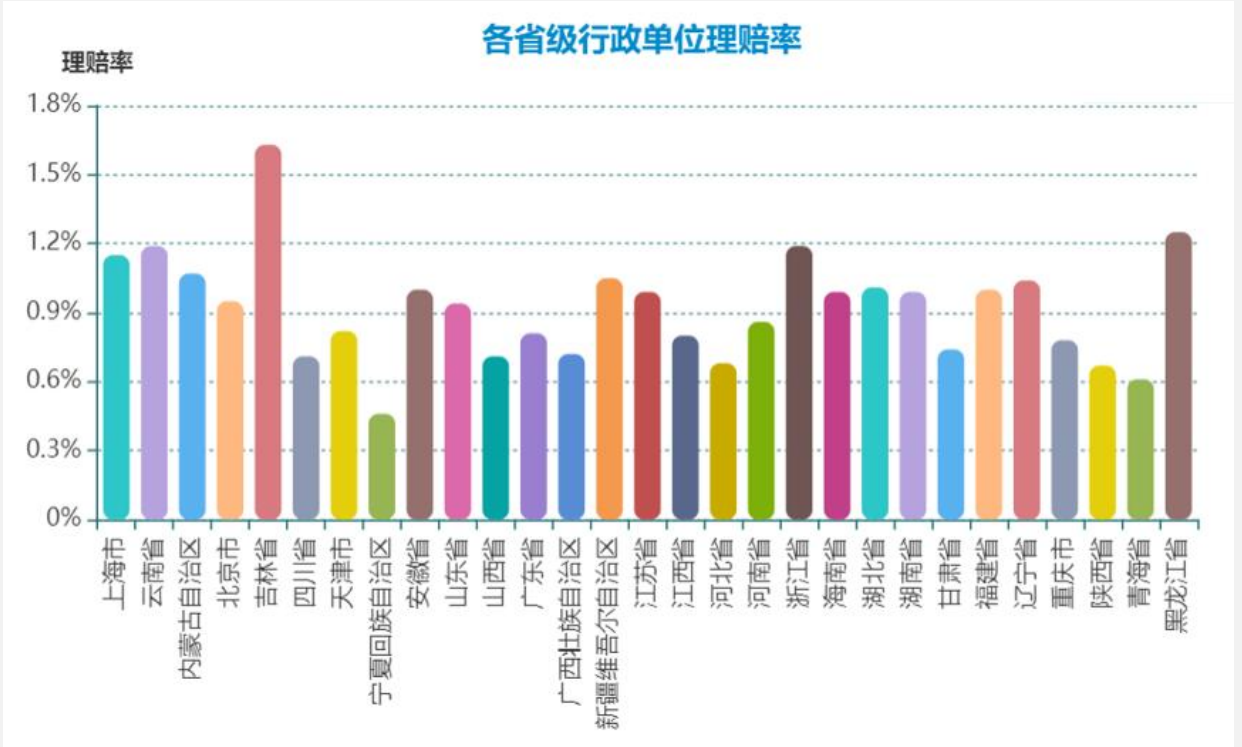
- 理赔人群中，四、五线城市投保人所占比例明显低于一线、新一线、二线城市
- 总的来看，二线城市投保人的理赔率最高，但与其他城市线差距不是很大





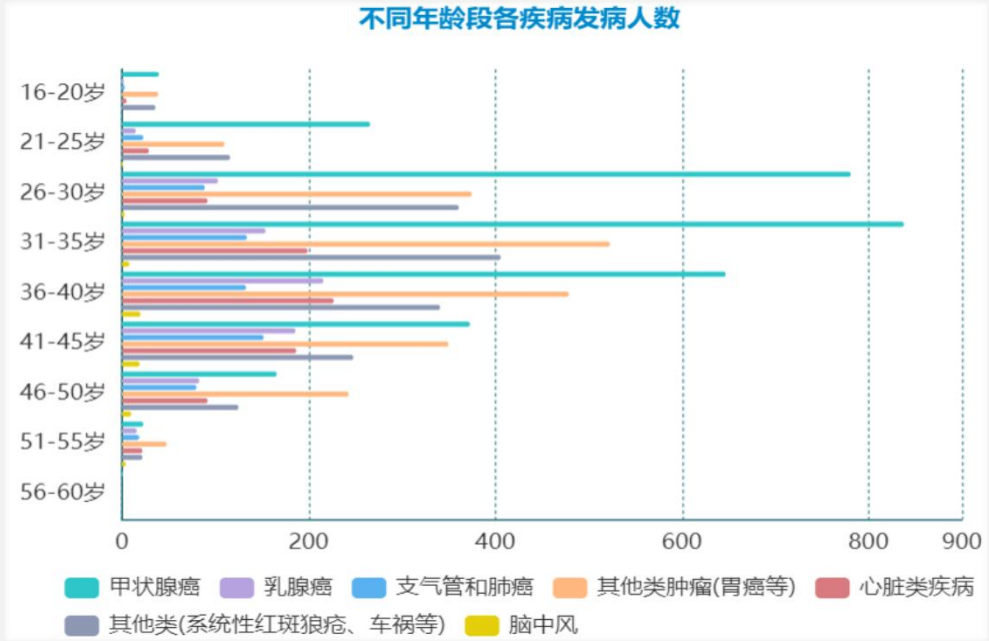
出险情况描述

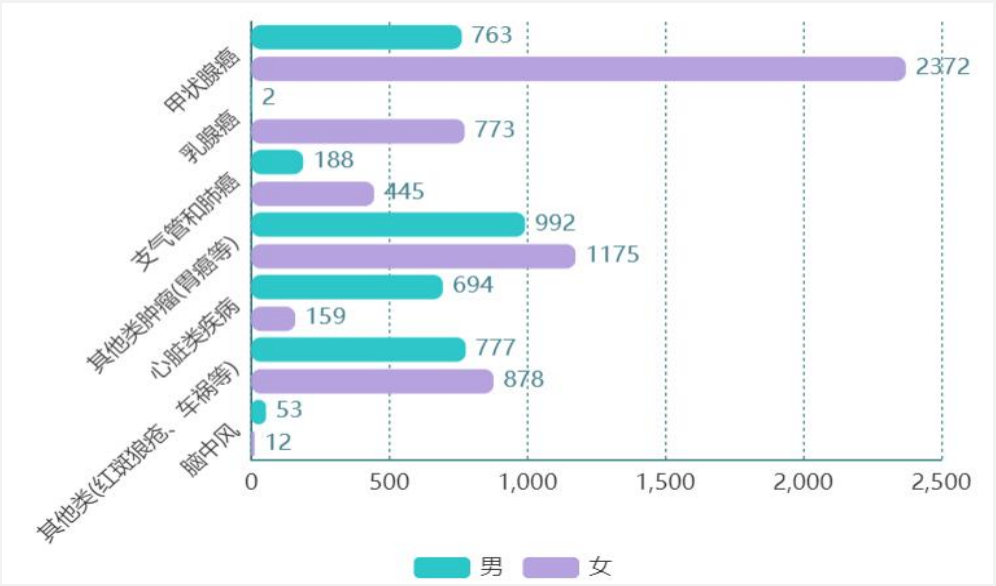




- ✓ 吉林省、黑龙江省代表的东北地区理赔率较高，其次是上海市、浙江省等华东地区
- ✓ 广东省理赔保单数量最多
- ✓ 肿瘤、心血管疾病是理赔疾病的主力军，这在各个省份都是一致的。

出险情况描述

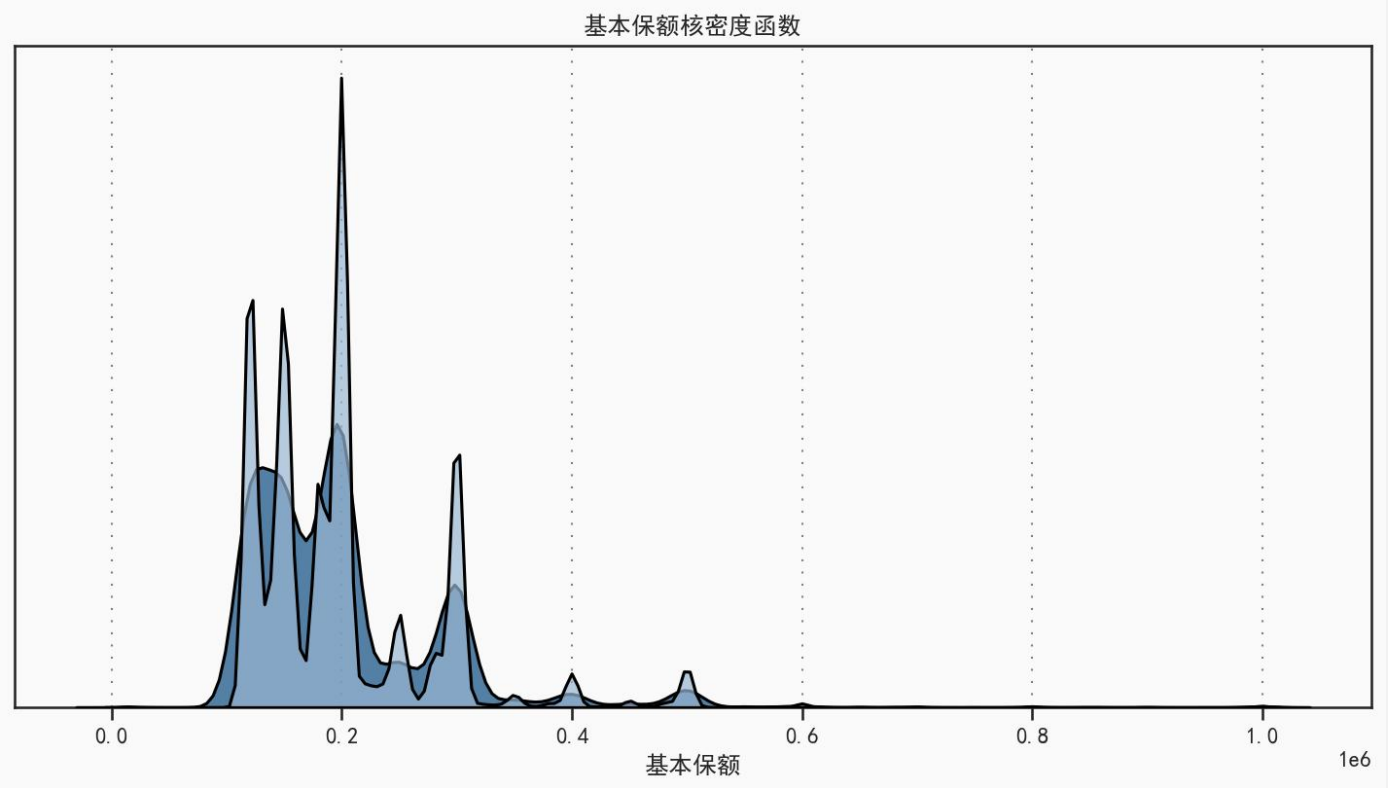




- **癌症在各个年龄段群体占比都很大**，且甲状腺癌、乳腺癌在26-40岁群体中占比较高，支气管癌和肺癌在41-50岁群体中占比较高

- 女性**甲状腺癌**发病人数明显多于男性，在**心脏类疾病**和**中风**等病症上发病人数少于男性

基于基本保额的显著性检验和客户群体划分



理赔和未理赔群体的基本保额在分布上没有明显差异，保单95%分位数为30万，且绝大多数保单的基本保额都在50w以下（99%分位数）。

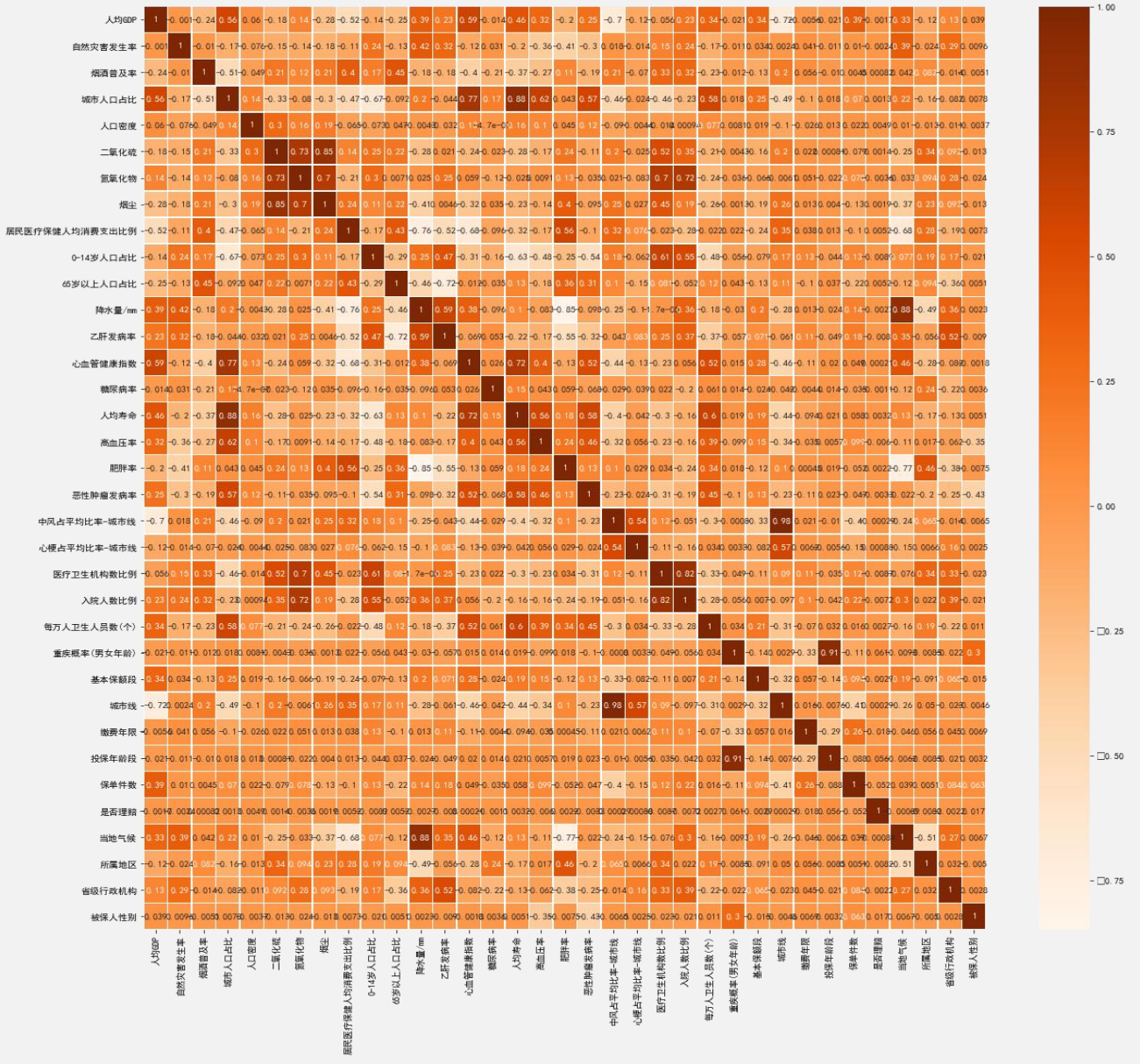
各因素与基本保额交叉分析结果

因子	卡方检验	值	渐进显著性(双侧)	因子	卡方检验	值	渐进显著性(双侧)
城市线	皮尔逊卡方	140345.375a	0	恶性肿瘤发病率	皮尔逊卡方	230729.309a	0
	似然比	146293.043	0		似然比	203506.655	0
	线性关联	73731.456	0		线性关联	15090.429	0
被保人性别	皮尔逊卡方	1561.153a	0	每万人卫生人员数	皮尔逊卡方	167021.962a	0
	似然比	1755.389	0		似然比	156779.947	0
	线性关联	551.266	0		线性关联	38360.792	0
缴费年限	皮尔逊卡方	34572.999a	0	0-14岁人口	皮尔逊卡方	214469.517a	0
	似然比	25735.29	0		似然比	202845.366	0
	线性关联	0.022	0.881		线性关联	16.245	0
投保年龄	皮尔逊卡方	61743.705a	0	入院人数万人	皮尔逊卡方	214469.517a	0
	似然比	43709.299	0		似然比	202845.366	0
	线性关联	9187.687	0		线性关联	40.121	0
人口密度	皮尔逊卡方	214469.517a	0	中风	皮尔逊卡方	140345.375a	0
	似然比	202845.366	0		似然比	146293.043	0
	线性关联	26839.282	0		线性关联	73074.271	0
所属地区	皮尔逊卡方	79448.319a	0	高血压	皮尔逊卡方	231899.456a	0
	似然比	80628.604	0		似然比	204924.481	0
	线性关联	24482.18	0		线性关联	20009.236	0
人均GDP	皮尔逊卡方	346552.080a	0	糖尿病	皮尔逊卡方	197076.785a	0
	似然比	311845.615	0		似然比	184605.271	0
	线性关联	81246.951	0		线性关联	638.39	0
烟尘	皮尔逊卡方	214469.517a	0				
	似然比	202845.366	0				
	线性关联	28103.36	0				

变量相关性分析

作用：探究各个特征之间的相关关系，可以为特征选择提供初步依据，有利于后续简化、解释模型。

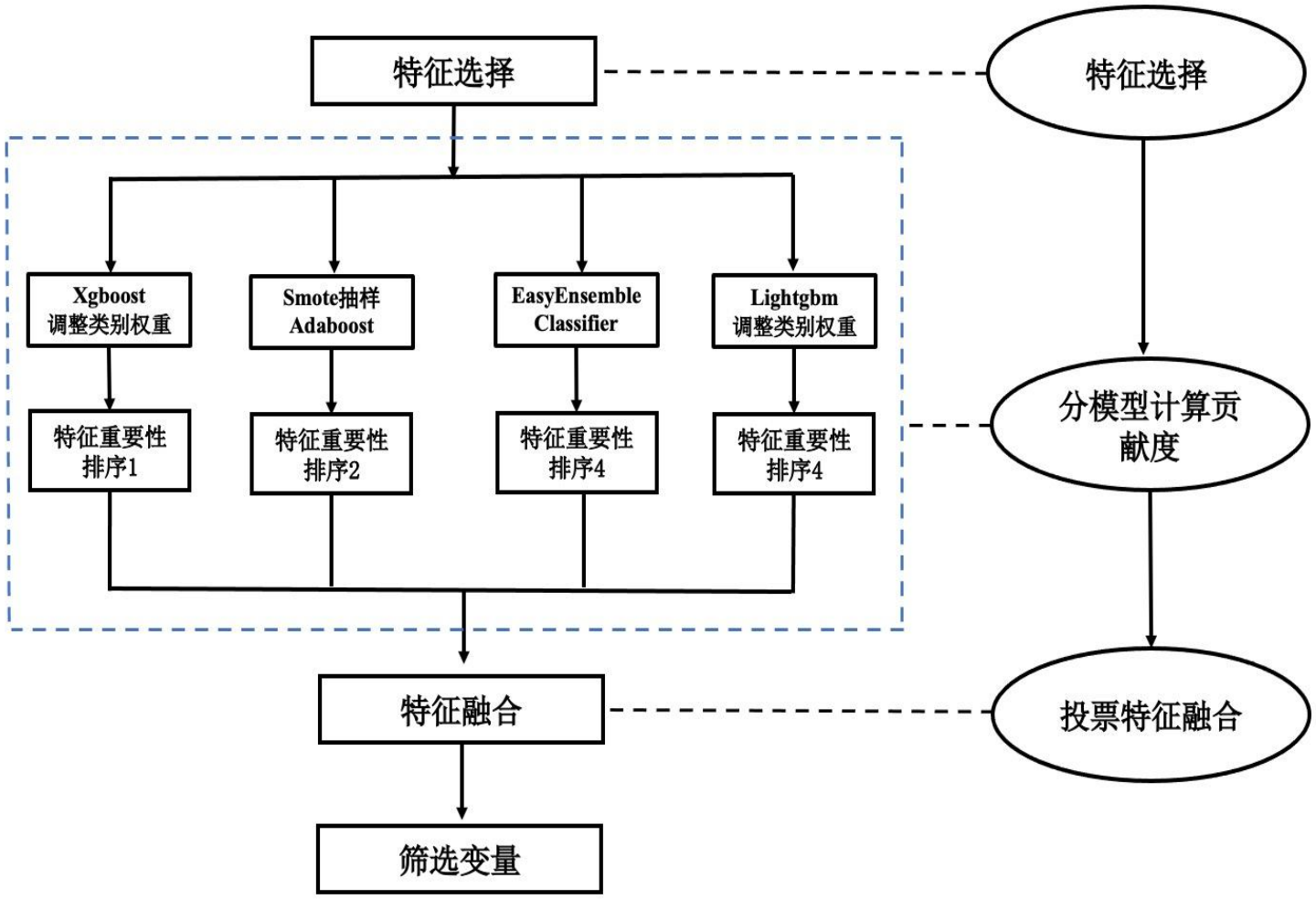
- ✓ 二氧化硫、烟尘、氮氧化物排放量
- ✓ 当地气候类型和降雨量
- ✓ 人均寿命和城市人口占比
- ✓ 重疾概率和投保年龄段均高度相关



Part 4

基于机器学习模型的风险因子筛选

建模与风险因子筛选

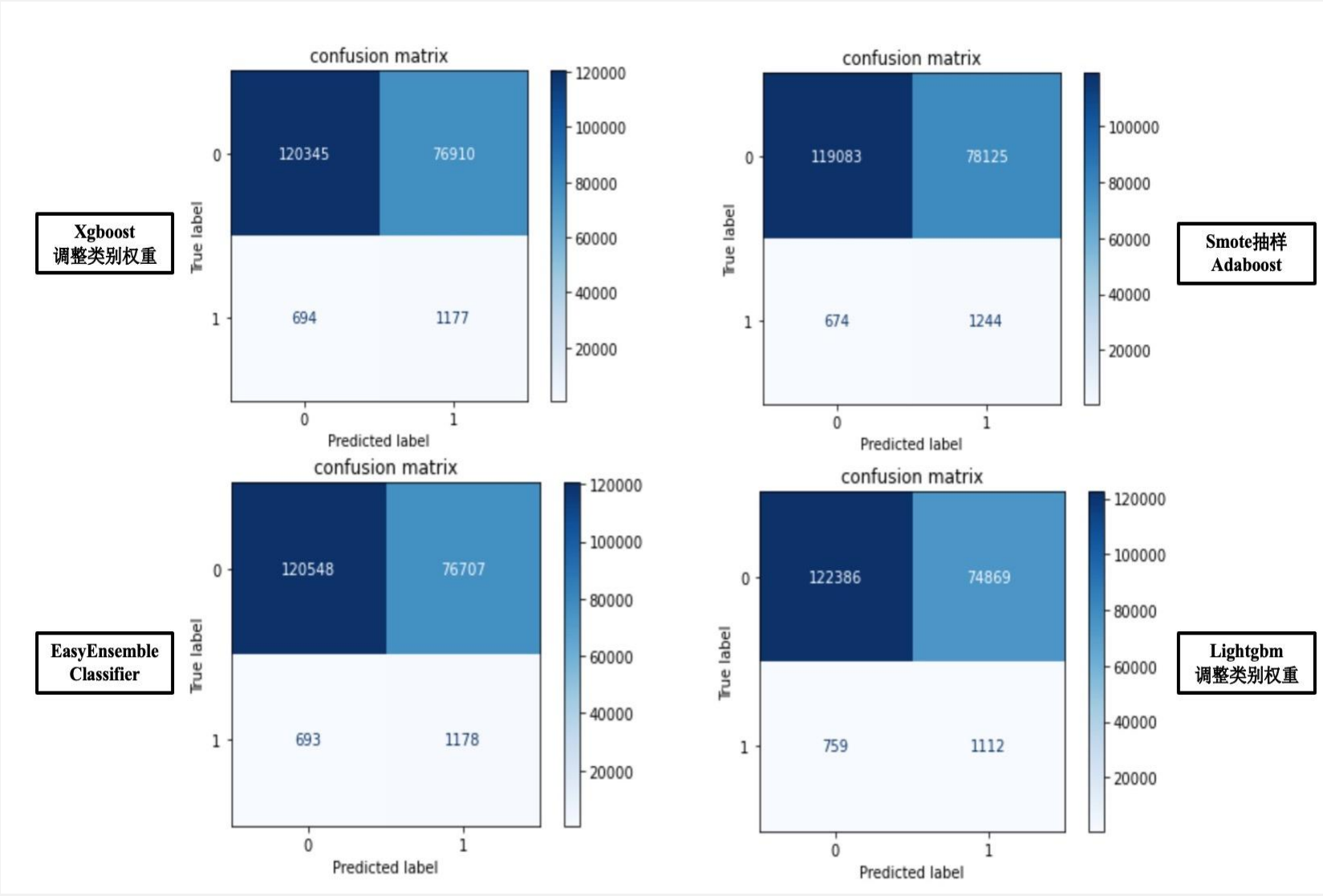


模型：以是否理赔为因变量，其他特征为自变量训练分类模型。**由于数据是不平衡分布，即理赔占比较少，且单模型筛选不太稳定，选用机器学习中以下四种方法建模比较**：

- ✓ 经典模型：调整类别权重后的Xgboost
- ✓ 低复杂度模型：调整类别权重后Lightgbm
- ✓ 简单集成模型：欠抽样后的EasyEnsembleclassifier模型
- ✓ 抽样模型：SMOTE采样后的Adaboost

特征筛选

- ✓ 四种模型在混淆矩阵的表现基本一致（SMOTE采样后的Adaboost 效果最好），召回率在65%左右
- ✓ 如果能获得保单的更多信息（如个人信息、过往病史、家庭环境等），将会有更高的准确率和召回率。



特征筛选

考虑各个因子在模型中的表现，使用位次投票法来最终确定机器学习筛选下的因子种类，除此之外，**结合各因子的相互关系，剔除相关性较高的变量，并因子的实际含义**，最终确定以下因子

地理环境	性别年龄	医疗卫生因素	经济社会	疾病因素	保单情况
所属地区	被保人性别	每万人卫生人员数(个)	人均GDP	肿瘤率	基本保额段
当地气候	投保年龄	医疗卫生机构数	人口密度	中风占平均比例	缴费年限
	0-14岁人口占比		城市线	心梗占平均比例	
			烟尘	高血压率	
				人均寿命	
				糖尿病率	

针对不同客户的重疾影响因子比较

普通客户	高端客户
人均GDP	人均GDP
自然灾害发生率	自然灾害发生率
人口密度	城市人口占比
烟尘	人口密度
居民医疗保健人均消费支出比例	烟尘
高血压率	65岁以上人口占比
中风占平均比率-城市线	糖尿病率
心梗占平均比率-城市线	恶性肿瘤发病率
医疗卫生机构数比例	中风占平均比率-城市线
年龄	心梗占平均比率-城市线
缴费年限	年龄
每万人卫生人员数	城市线
被保人性别	缴费年限
城市人口占比	每万人卫生人员数
二氧化硫	被保人性别
氮氧化物	二氧化硫
乙肝发病率	居民医疗保健人均消费支出比例

针对不同客户

按照基本保额的大小划分客户群体，高于50w的为高端客户。在高端客户和普通客户之间筛选出的因子差异性不大，得到这个结果的原因很大程度上是因子在高端客户和普通客户之间表现差异不明显

针对不同年份和不同年龄段的重疾影响因子比较

16年/10-25岁		16年/25-40岁		16年/40岁以上	
年龄	1	年龄	1	年龄	1
城市线	2	被保人性别_1	2	被保人性别_1	2
心梗占平均比率-城市线	3	被保人性别_0	3	被保人性别_0	3
中风占平均比率-城市线	4	中风占平均比率-城市线	4	心梗占平均比率-城市线	4
人口密度	5	当地气候	5	中风占平均比率-城市线	5
氮氧化物	6	人均GDP	6	城市线	6
所属地区	7	所属地区	7	基本保额段	7
基本保额段	8	城市线	8	恶性肿瘤发病率	8
省级行政机构	9	基本保额段	9	缴费年限	9
人均GDP	10	省级行政机构	10	当地气候	10
自然灾害发生率	11	缴费年限	11	人均GDP	11
恶性肿瘤发病率	12	恶性肿瘤发病率	12	省级行政机构	12
当地气候	13	心血管健康指数	13	所属地区	13
65岁以上人口占比	14	心梗占平均比率-城市线	14	城市人口占比	14
医疗卫生机构数比例	15	居民医疗保健人均消费支出比例	15	氮氧化物	15
每万人卫生人员数(个)	16	每万人卫生人员数(个)	16	入院人数比例	16
糖尿病率	17	人口密度	17	人口密度	17
缴费年限	18	肥胖率	18	65岁以上人口占比	18
饮酒力	19	医疗卫生机构数比例	19	0-14岁人口占比	19
居民医疗保健人均消费支出比例	20	烟尘	20	饮酒力	20
17年/10-25岁		**17年/25-40岁**		**17年/40岁以上**	
年龄	1	年龄	1	年龄	1
心梗占平均比率-城市线	2	当地气候	2	中风占平均比率-城市线	2
中风占平均比率-城市线	3	人均GDP	3	城市线	3
人均GDP	4	中风占平均比率-城市线	4	被保人性别_1	4
当地气候	5	省级行政机构	5	当地气候	5
城市线	6	城市线	6	心梗占平均比率-城市线	6
恶性肿瘤发病率	7	所属地区	7	缴费年限	7
氮氧化物	8	基本保额段	8	被保人性别_0	8
乙肝发病率	9	缴费年限	9	所属地区	9
省级行政机构	10	被保人性别_0	10	省级行政机构	10
缴费年限	11	人均寿命	11	人均GDP	11
烟尘	12	心梗占平均比率-城市线	12	居民医疗保健人均消费支出比例	12
居民医疗保健人均消费支出比例	13	被保人性别_1	13	基本保额段	13
0-14岁人口占比	14	肥胖率	14	医疗卫生机构数比例	14
所属地区	15	乙肝发病率	15	高血压率	15
基本保额段	16	糖尿病率	16	糖尿病率	16
65岁以上人口占比	17	恶性肿瘤发病率	17	自然灾害发生率	17
糖尿病率	18	每万人卫生人员数(个)	18	乙肝发病率	18
入院人数比例	19	自然灾害发生率	19	恶性肿瘤发病率	19
医疗卫生机构数比例	20	心血管健康指数	20	氮氧化物	20
18年/10-25岁		18年/25-40岁		18年/40岁以上	
年龄	1	年龄	1	年龄	1
中风占平均比率-城市线	2	城市线	2	被保人性别_0	2
心梗占平均比率-城市线	3	人均GDP	3	被保人性别_1	3
城市线	4	当地气候	4	城市线	4
当地气候	5	基本保额段	5	中风占平均比率-城市线	5
省级行政机构	6	居民医疗保健人均消费支出比例	6	心梗占平均比率-城市线	6
缴费年限	7	恶性肿瘤发病率	7	当地气候	7
所属地区	8	中风占平均比率-城市线	8	缴费年限	8
居民医疗保健人均消费支出比例	9	所属地区	9	所属地区	9
人均GDP	10	被保人性别_1	10	恶性肿瘤发病率	10
基本保额段	11	省级行政机构	11	基本保额段	11
高血压率	12	心梗占平均比率-城市线	12	人均GDP	12
恶性肿瘤发病率	13	缴费年限	14	省级行政机构	13
每万人卫生人员数(个)	14	入院人数比例	15	居民医疗保健人均消费支出比例	14
降水量/mm	15	0-14岁人口占比	16	0-14岁人口占比	15
65岁以上人口占比	16	人均寿命	17	入院人数比例	16
被保人性别_1	17	每万人卫生人员数(个)	18	高血压率	17
饮酒力	18	降水量/mm	19	人均寿命	18
氮氧化物	19	被保人性别_0	20	人口密度	19
入院人数比例	20	城市人口占比	26	每万人卫生人员数(个)	20

针对不同年份和不同年龄段

- ✓ **考虑到保单风险暴露导致赔付率增加**，将保单数据划分为16年、17年、18年三年进行分析
- ✓ 被保人年龄划分为10-25岁、25-40岁、40岁以上三段
- ✓ 年龄、性别、人均GDP、当地气候、城市线、基本保额、所属地区以及各种重疾发病平均比率在各个年份各个年龄段重要性较高，其中年龄这一因子在所有情况下都是最重要的因子，进一步证实了性别和年龄是重疾的关键因素。
- ✓ 同一年份不同年龄段比较可以发现，随着年龄的增加，被保人性别、城市线、基本保额、缴费年限的重要性不断增加，说明这些因素对于不同的年龄群体的影响是不同的，他们对于中老年群体的影响程度大于青年群体 。

风险因子分析

地理环境因素

- **就癌症而言**，综合来看，东北地区的癌症发病率最高。
- **乳腺癌**而言，东北地区超出全国平均水平20%，不生育或晚育是乳腺癌发病的重要影响因素，而东北地区生育率排在全国末位
- **肝癌**而言，东北地区和华南地区得病比率较高 。东北地区气候寒冷，饮酒吸烟较为严重；华南地区 与长期食用含黄曲霉菌的食物 有关
- **胃癌**而言，华东、西北、东北是胃癌的高发区
- **结直肠癌**而言，过多的蛋白质和脂肪摄入是诱发结直肠癌的重要因素，华南地区发病率最高
- **肺癌**而言，吸烟较多、空气污染都是诱发原因，而东北和华东地区肺癌较多
- **心脑血管疾病而言（心梗、中风）**，长期吸烟喝酒、饮食不清淡、天气寒冷都是重要的诱发因素，而这类疾病在东北华北地区高发，也符合当地的气候条件和饮食生活习惯。

风险因子分析

性别年龄因素

- 0-17岁少年重疾发生的概率较低。
- 青年人群体（18-40岁）主要重疾为，男性：恶性肿瘤、心梗脑中风等，女性：恶性肿瘤、肾病。
- 中年群体（40-60岁），主要重疾为恶性肿瘤、心梗、脑中风后遗症（三者占比达75%以上），性别上差异不大。
- 老年群体，恶性肿瘤、脑中风后遗症和心梗（三者占比近80%），且在性别上差异不大。

医疗卫生因素

- 入院人数比例反映了当地人群的健康水平，入院比例越高，相比之下当地人群健康程度越低，重疾发生的可能性就越高；同时也需要关注地区是否有地区病。
- 每万人卫生人员数反映了当地的医疗资源水平，医疗资源水平越高，相应的定期体检能力、早期疾病诊断治理能力、医疗卫生意识都会越高。

风险因子分析

经济社会因素

- **人均GDP**衡量了地区的经济水平，地区经济水平越高，相应的医疗设施也会比较完善，医疗防护意识和医疗保障能力较强，人群重疾发生的风险相对较低
- **人口密度**衡量了地区的人口密集程度，人口密度大的地区环境等因素对人群会产生较大影响，重疾发生概率也有所提高提高
- 省市污染指标衡量了地区的环境情况，环境质量的好坏对地区疾病的发生因素有很大影响
- **城市线**与重疾的关系较为复杂

疾病因素

- **疾病因素**直接影响重疾的发生，与前文内容既有交叉，也有新的内容。疾病因素中的重疾概率可由中国人身保险业重大经验发生率表（2020）确定

保单情况

- 基本保额段和缴费年限反映了投保人保险额度的上限和支付期限
- 缴费年限反映了投保人目前的经济状况，较长的年限一定程度上说明了投保人经济存在压力且有罹患重疾的担忧

Part 5

基于生存分析的重疾发病“速度”差异性研究

- ✓ 一般来说，在观察期五年内，被保人投保之后发病速率不变，保单生效后存在风险暴露，被保人得重疾的风险不断增加

- ✓ **为了进一步探究不同风险因素的不同水平对保单风险的影响差异，从另一个角度佐证筛选因子的影响作用，本报告采用Cox回归模型**挖掘和研究影响终点事件发生速度的的有关因素。

- ✓ COX回归因变量二分类结局变量和连续型生存时间变量，本报告中二分类结局变量为被保人是否发生重疾，未发生重疾记为数据删失（只考虑这一种数据删失情况），生存时间为保单出险时间减去保单生效时间，自变量为各个风险因子

- ✓ 满足Cox模型的等比例风险假设和线性关系假设下，最后得到风险函数与时间和风险因子的回归表达式。

$$h(t) = h_0(t)\exp\left(\beta_1 x_1 + \cdots, +\beta_j x_j\right),$$

$$t = t^{(2)} - t^{(1)}$$

其中 h(t) 是理赔的风险函数，含义为 t 时刻未理赔被保人在 t 时刻的瞬时重疾发生率，可以理解为重疾发生速率；$h_0(t)$ 为 t 时刻基准风险函数，刻画风险函数随时间的变化关系，可以理解为保单的风险暴露使得保单理赔风险不断增加；X_i 表示第 i 个风险因子，β_i 表示第 i 个风险因子的系数，指数项表示各个风险因子不同水平对保单风险的影响。建立上述 Cox 回归模型是为了研究被保人投保之后发病速率不变，保单风险暴露赔付风险不断增加的情况下，不同因子水平对保单风险（即保单从生效到出险的时间差）的差异性作用。

基于cox回归的检验

因子	B	SE	瓦尔德	自由度	显著性	Exp(B)
被保人性别	-0.334	0.027	148.222	1	0	0.716
基本保额	0	0	0.224	1	0.636	1
城市线	0.008	0.007	13.26	1	0.025	1.018
缴费年限	-0.033	0.002	277.632	1	0	0.967
投保年龄	0.076	0.001	3129.077	1	0	1.059
人口密度	0	0	1.451	1	0.228	1
所属地区	0.045	0.006	64.186	1	0	1.046
烟尘	-0.002	0.001	6.593	1	0.01	0.998
人均GDP	0	0	4.937	1	0.026	1
恶性肿瘤发病率	-0.001	0	24.566	1	0	0.999
每万人卫生人员数	0.002	0.001	4.933	1	0.026	1.002
0-14岁人口	0	0	84.174	1	0	1
入院人数(万人)	0	0	56.297	1	0	1
中风	0.022	0.04	0.297	1	0.586	1.022
高血压	-0.015	0.002	43.747	1	0	0.985
糖尿病	-0.022	0.005	17.935	1	0	0.979

最后一列为风险比例系数，以性别为例，男性投保人的生存风险是女性投保人的 71.6%，即男性保单从生效到出险的时间差是女性的 100%/71.6%倍（时间差更长），其他因子具有类似解释

Part 6

结论与建议

结论

机器学习筛选得到结论

- **年龄与性别因素是影响重疾发生率的关键因素，**由于男女性生理结构不同和社会责任不同，不同重大疾病的发病率在不同性别的群体上差异较大，且随着年龄的增加和身体机能的下降，恶性肿瘤、中风、心梗患病率不断增加
- **地理环境因素（所在地区、气候、饮食习惯等）是影响重疾发生率的重要因素。**就癌症而言，不同类型的癌症在不同地区发病率差异性较大，这与当地的气候环境、饮食习惯、生活水平息息相关
- **社会因素是影响重疾发生率的一个综合因素。**社会因素体现在生活压力、心理健康程度、生活水平与环境等方面。
- **医疗卫生水平是影响重疾发生率的潜在因素。**一方面，具备良好卫生水平的地区重疾发病率相对较低，这与日常疾病宣传、居民良好的生活习惯和环境治理等因素息息相关；另一方面，不具备良好医疗水平的地区对诱发重疾的病症的知晓率、治疗率和控制率都较低，筛查不全面也会导致相关数据失真，给重疾险理赔带来隐患。

结论

不同客户、不同年份和不同年龄段建模结论

- 基于基本保额段划分，在当前的因子中未能**区分开高端客户与普通客户的重疾风险因子差异，**但在**实际情况中研究这一问题是很有必要的。**
- 对保单数据按投保年份和年龄段划分并进一步进行机器学习筛选因子的结果显示，年龄性别、人均GDP、当地气候、城市线、基本保额、所属地区以及各种重疾发病平均比率在各个年份各个年龄段重要性较高，其中年龄这一因子在所有情况下都是最重要的因子。**同一年份不同年龄段比较**可以发现，随着年龄的增加，被保人性别、城市线、基本保额、缴费年限的重要性不断增加，说明这些因素对于不同的年龄群体的影响是不同的，他们对于中老年群体的影响程度大于青年群体。

Cox回归建模结论

- ✓ **Cox回归建模分析结果指出，众多风险因子水平的差异会对保单风险产生影响，具体体现在保单从生效到理赔的时间差的差异上。**
- ✓ 在统计意义上，男性保单从生效到出险的时间差是女性的100%/71.6%倍；城市线数字每增加1，保单从生效到出险的时间差增加100%/101.8%倍；投保年龄每增加一岁，保单从生效到出险的的时间差减少100%-100%/105.9%；地区代表的数字每增加1，保单从生效到出险的时间差减少100%-100%/104.6%；每万人卫生人员数每增加1，保单从生效到出险的时间差减少100%-100%/100.2%；**以上结果进一步从回归的角度印证了机器学习筛选出的因子对保单风险存在作用。**

建议

对数据和建模

- **在原先保单继续运作的基础上需要对被保人群进一步划分，进一步挖掘性别、年龄、地区等关键信息的内涵特征。**采用客户特征字段匹配风险因素，利用客户信息检索当地常发病情况、医疗卫生情况、气候特点和饮食习惯等信息，尽力打通消费者更多的健康数据（如运动步数、睡眠时间等），为相关风险评估提供更多参考信息。
- **增加风险评估的相关因子。**被保人的家庭情况、当前工作情况、饮食习惯和生活习惯、过往病史、生活环境情况、心理健康情况等都可以提供有效信息
- **针对不同的客户群体（如高基本保费群体和其他群体等），建立重疾发病风险因子体系，挖掘不同群体下的重疾影响因素。**

对产品研发

- **考虑开发重疾险其他方面的增值服务**，比如在疾病预防方面，重疾险可以给被保人提供癌筛、体检、健康管理等方面便利的资源
- **注重客户实际需求和保单体现的特征规律。**根据不同性别、年龄段、地区、行业开发差异化的重大疾病保险;根据不同病种所需要的医疗费用、对病人健康的威胁程度，在各疾病给付金额的设计上进行差异化。**下一步重疾险可以向模块化、定制化发展，考虑细分为各个身体系统或者涉及住院、失能、护理、体检等多种保障功能，提高客户的选择权。**

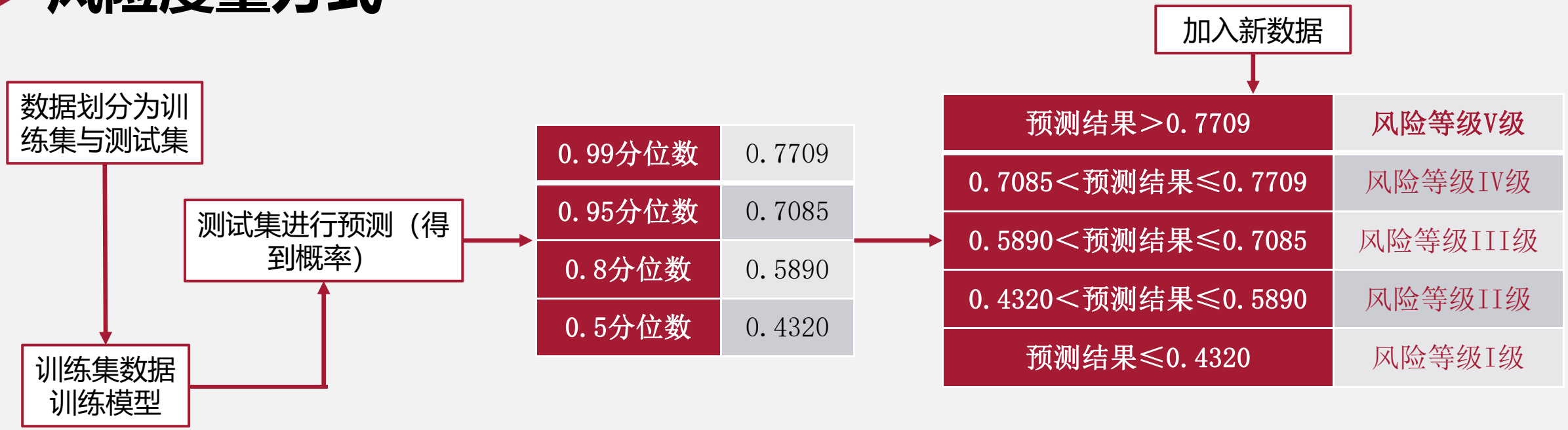
对产品理念

- **建立对不同地区重疾发病机制和形势的认识，形成地区重疾长期观测机制。**
- **增加对已患重疾人员的生存时间、治疗情况和医疗费用的了解，而不是仅仅将研究局限于并发率方面。**

Part 7

风险评估模型

风险度量方式



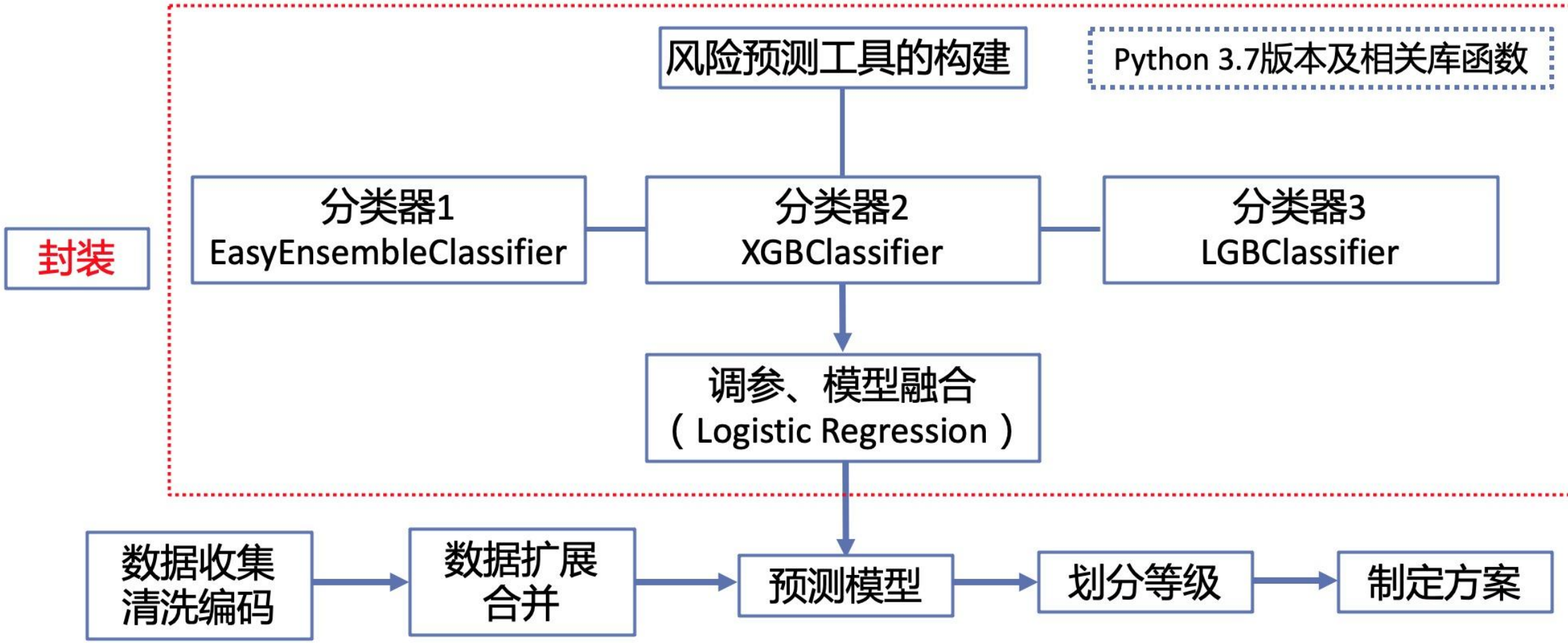
- 案例中记理赔为“1”，未理赔为“0”，风险度量模型输出结果为理赔的概率，即为0~1中间的数值，数值越大，表示客户在整体人群中的风险水平越高。考虑到实际理赔率不到1%，**根据测试集20万条数据预测结果的分布，以0.99分位数、0.95分位数、0.8分位数、0.5分位数为切分点，对新数据的风险等级进行划分**
- 测试集得到的分位数是对风险的一种度量，**分位数越高（概率越大）说明风险越大**

模型运行示例

步骤如下：

1. 客户信息收集。
2. 客户信息清洗、分类与编码。
3. 相关信息扩展（如若客户所在城市为广东广州，男57岁，则进行如下特征匹配：地区——南部地区，气候——亚热带季风气候，气候特征——潮湿多雨，降水量——1972mm，居民医疗保健人均消费支出比例——5.8%，重疾经验发生率——11.76‰……）
4. 代入模型计算，得到数值结果。
5. 进行相关风险划分，评估相对风险水平。

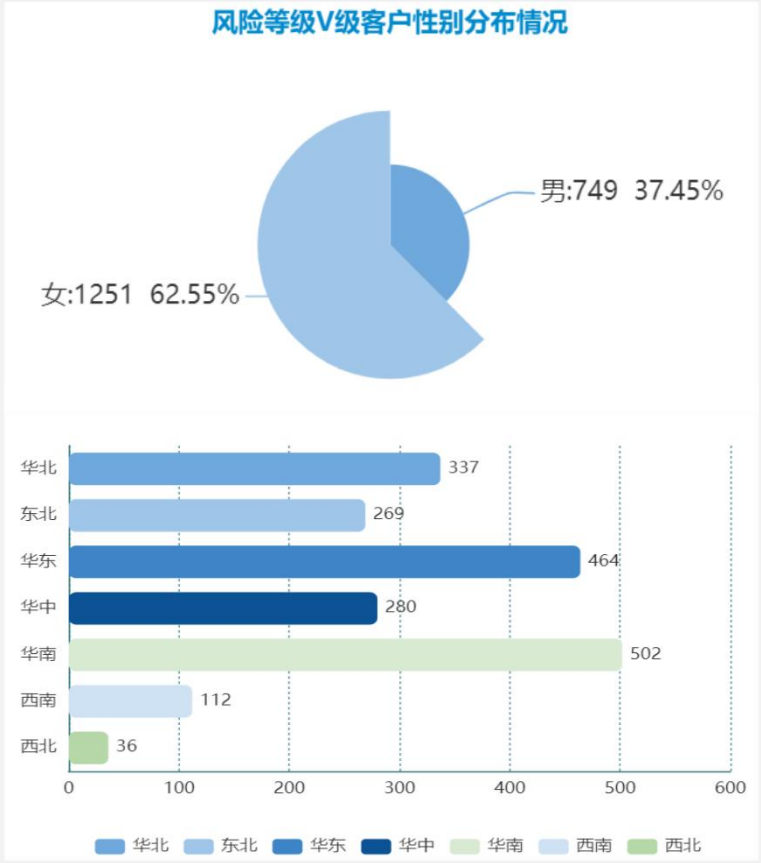
风险评估模型

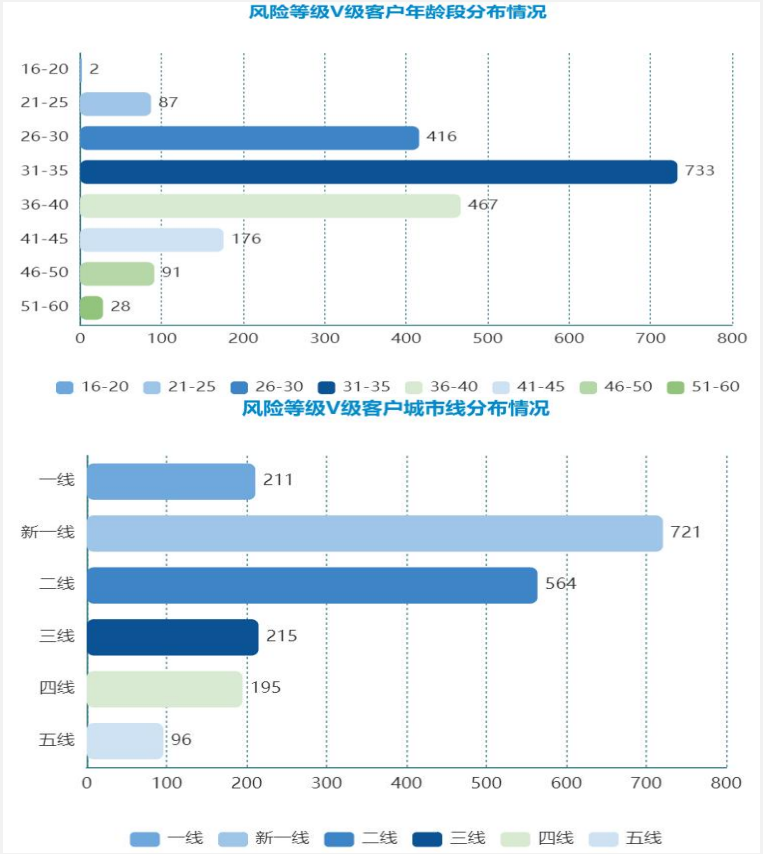

风险预测工具的构建
Python 3.7版本及相关库函数
封装
分类器1
EasyEnsembleClassifier
分类器2
XGBClassifier
分类器3
LGBClassifier
调参、模型融合
（ Logistic Regression ）
数据收集
清洗编码
数据扩展
合并
预测模型
划分等级
制定方案

风险评估可行性

根据建立的风险评估模型和相应的测试集的结果，分析预测值大于0.99分位数（即风险等级为V级的客户，共2000条）的客户基本信息，比较与前文理赔的客户信息是否有共同之处

结果展示





创新点

因子构建较为完善

本报告初步分析重疾的影响因素，并进一步结合公司保单数据和国家统计局、中国统计年鉴、人身保险业重大疾病经验发生率表等其他公开渠道数据扩展被保人信息库，构建包含**地理环境、性别年龄、医疗卫生、经济社会、疾病和保单特征六大类风险因素**，量化为33个指标

采用多模型筛选特征并比较

鉴于理赔数据的不平衡性和单模型筛选特征的不稳定性，本报告使用使用**加权Xgboost/加权Lightgbm/Smote抽样后的Adaboost/集成学习四种方法进行因子筛选，**比较四种模型结果并得到机器学习下的因子重要性排序，考虑变量相关性和实际含义后**确定了20个风险因子**

尝试按年份、年龄和基本保额段划分保单并挖掘风险因子

进一步划分保单年份，重点考虑投保人年龄这一因素，对不同年龄段的投保人进行机器学习建模和因子筛选，比较结果差异。**尝试对两种客户群体分别建立重疾风险因子模型，**寻找不同群体的风险源差异，为研究此问题提供了一个较好思路

应用Cox模型，从另一个角度分析问题

使用**生存分析中的Cox回归模型对保单数据建模，进一步从重疾发病风险上分析筛选出的因子在不同水平下的作用差异，从回归的角度进一步佐证机器学习筛选的因子具有较强的代表性。**

研究与应用价值

对于挖掘风险因子

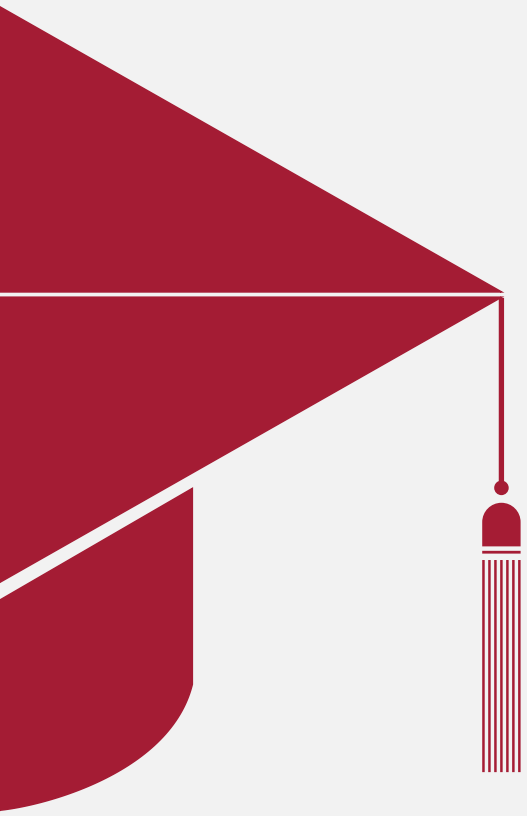
- 本报告构建风险因子体系的过程可以作为实际的参考，在原先数据基础上进一步挖掘性别、年龄、地区等关键信息的内涵特征，利用客户特征字段匹配风险因素，如用客户信息检索当地常发病情况、医疗卫生情况、气候特点和饮食习惯等信息
- 可以增加风险评估的相关因子，而不是仅仅局限于性别、年龄和地区。被保人的家庭情况、当前工作情况、饮食习惯和生活习惯、过往病史、生活环境情况、心理健康情况等都可以提供有效信息

对于建模筛选因子

- 实际中可以采取多种模型比较的方式，对理赔率进行建模
- 在获得更精细数据的基础上，本文进一步划分保单年份、年龄段和客户群体进行建模，比较和分析不同情况下的风险因子差异，为实际风险建模提供了参考，可以进一步对性别、地区等特征进行划分、寻找因子差异

对于风险评估模型

- 本报告建立的风险评估模型基于模型融合的方法，具有更高的准确率，同时对于批量处理客户信息、鉴定风险等级有一定优势



恳请各位评委老师批评指正！

谢谢老师！