我国重大疾病风险的影响因素探究
——基于荔枝人寿某重疾产品的数据分析

2022.6.30

目录

CONTENTS

1

PART 1

- 案例问题介绍
- 重疾险研究背景

案例问题及研究背景介绍

案例问题介绍

问题背景

重疾产品赔付情况不乐观，长期赔付风险恶化——**关注影响重疾发生率的风险要素**

题目要求

利用已知承保理赔数据，扩展额外的城市地域相关数据——**丰富分析变量**

预计目标

使用理论分析、描述性分析和建模分析等方法挖掘影响重疾发生率的风险因素——设计风险预测工具并提供降低赔付风险的建议

主要分析方法

理论分析

汇总国内外有关重疾发生率影响因素的文献，对其进行分类总结，为之后搜寻数据及解释原因打好基础

描述性分析

不同群体重疾发生率的计算与分析，描述重疾发生率的分布特点，总结出一定规律并解释背后原因

建模分析

对个体重疾发生概率进行预测建模，使用预测效果最好的模型作为最终上线的重疾风险预测工具

重疾险发展三阶段

起步——规范发展——快速发展

LitchiData

起 步 阶 段

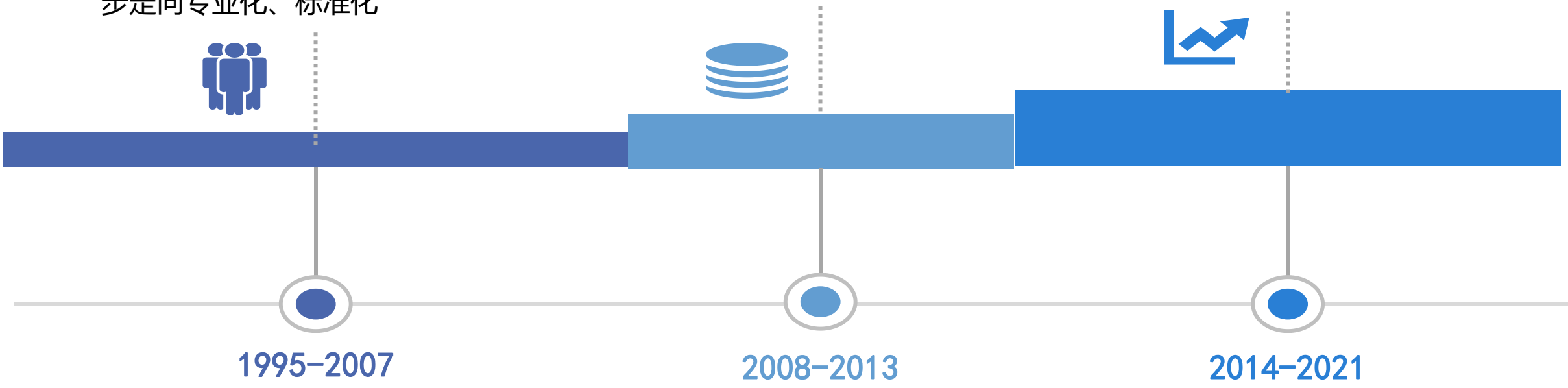
- 1995年，我国引入重疾险
- 2007年，中国保险行业协会与中国医师协会联合推出《重大疾病保险的疾病定义使用规范》，重疾险逐步走向专业化、标准化

规范发展阶段

- 2013年，原保监会发布《中国人身保险业重大疾病经验发生率表》

快速发展阶段

- 2020年，中保协与中医协再次联合发布《疾病定义使用规范（2020年修订版）》，中国精算师协会发布《中国人身保险业重大疾病经验发生率表（2020）》
- 新冠肺炎疫情；互联网渠道和线上投保



机遇——保费收入激增

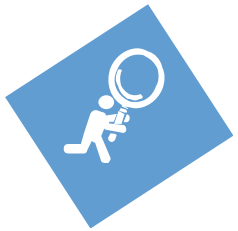
- 原保费收入增长迅速：2015年1027亿元 → 2020年4909亿元
- 占据健康险保费收入的60%左右，成为健康险发展的重要支柱

挑战——保费增速下降

- 2019年以来，重疾险的保费增速持续下降，业务经营压力日益加重
  2016年增速48.82% → 2020年增速19.41%

风险——发生率持续恶化

- 重疾发生率的长期趋势存在不确定性
- 重疾发病率受多种因素（生活方式、国家政策、科学技术等）驱动，难以预测



目的——建模预测，风险分级

- 探究重疾发生率的影响要素，建立模型预测不同人群的重疾发生概率
- 实现重疾风险分级，为合理降低重疾险的赔付风险提供建议

2

PART 2

重疾发生率影响因素的理论分析

- 个人影响因素的文献综述
- 地域影响因素的文献综述
- 保单特征影响因素的文献综述
- 文献述评

个人影响因素的文献综述

反映个人特质的因素影响重疾发生率

LitchiData

性别和年龄

性别

- 整体：**女性理赔率略高于男性**；
  重疾赔付的三大病种分别为恶性肿瘤、急性心肌梗死和脑中风后遗症
- 分病种：病种分布存在性别差异，**男、女性高发重疾不同**

年龄

- **重疾高发年龄段集中在41-60岁**
- 随着年龄的增长，重疾发生率显著增加
- 重疾的**发病年龄段逐渐前移**，出现年轻化的趋势

其他个人特质

生活方式

- 身高体重
- 收入
- 饮食：高盐摄入、不合理膳食
- 运动：体育锻炼
- 吸烟、酗酒

遗传因素

- 遗传性重疾
- 遗传变异

地域影响因素的文献综述

与居住省份或城市相关的因素影响重疾发生率

LitchiData

省份或城市

环境污染状况

- 空气污染
- 水污染
- 土壤污染

→ 环境污染严重将显著增加癌症等重疾的发病率

地区自然条件及饮食习惯

- 海拔
- 气温
- 日照
- 降水
- 饮食习惯

→ 某些地区人群的某种疾病负担显著高于其他地区

地区发展水平

- 经济水平：人均GDP
- 科学技术：预防性疫苗、筛查技术

→

- 富裕地区的重疾检出率高于贫困地区
- 科学技术能及时发现或有效预防重疾发病

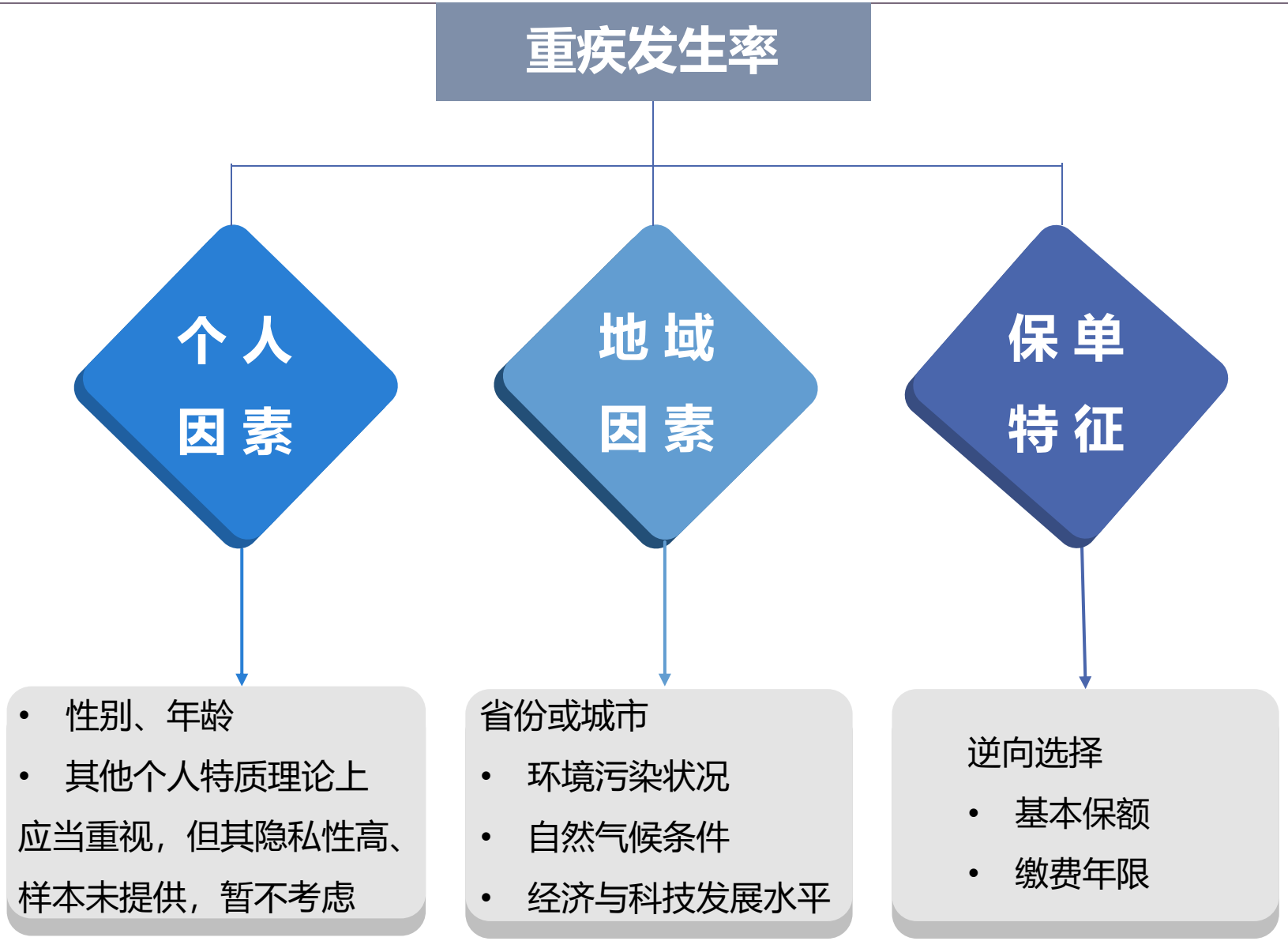
➢ **我们研究的重疾发生率实际上是重疾险的理赔发生率，所以保单相关特征也会对其产生影响。**

保单特征

基本保额、缴费年限、保单年度、出险年份……

逆向选择

- 国外：一些文献没有发现能够证实信息不对称的有效证据，另一些文献发现高风险人群更愿意购买追加保险
- 中国：低保单年度、提前给付型、高保额的保单重疾发生率显著偏高


重疾发生率
个人因素
地域因素
保单特征
• 性别、年龄
• 其他个人特质理论上应当重视，但其隐私性高、样本未提供，暂不考虑
省份或城市
• 环境污染状况
• 自然气候条件
• 经济与科技发展水平
逆向选择
• 基本保额
• 缴费年限

3

PART 3

- 数据说明与预处理
- 数据探索性分析-承保数据
- 数据探索性分析-理赔数据
- 不同群体重疾发生率计算与分析

承保理赔数据的描述性分析

数据说明与预处理

数据预处理和特征工程

LitchiData

字段类别		字段名称	类别	说明
承保	被保人信息	被保人性别	类别变量	1-男性；2-女性
		投保年龄	类别变量	18-56岁
	保单信息	保单生效年月	日期变量	2016/9/1-2018/11/1
		基本保额段	类别变量	105-3000000元
		缴费年限	数值变量	10、15、19、20、29、30
		保障年限	数值变量	99年（终身）
	地区信息	省级行政机构	类别变量	涵盖30个省级行政机构
		市级行政机构	类别变量	涵盖314个市级行政机构
		城市线	类别变量	1线；新1线(1N)； 2线； 3线； 4线； 5线
理赔		是否理赔	类别变量	0-未理赔；1-理赔
		出险过程描述	类别变量	文本信息
		出险日期	日期变量	2016/12/12-2021/12/9
		理赔日期	日期变量	2017/1/24-2021/12/29
		出险年龄	数值变量	18-59岁

缺失值处理

城市线字段存在1.38%缺失值

- 原因：市名变更（如襄樊市改名为襄阳市）、直管市（如河南的济源市）
- 处理：通过搜集信息进行填补

增加字段

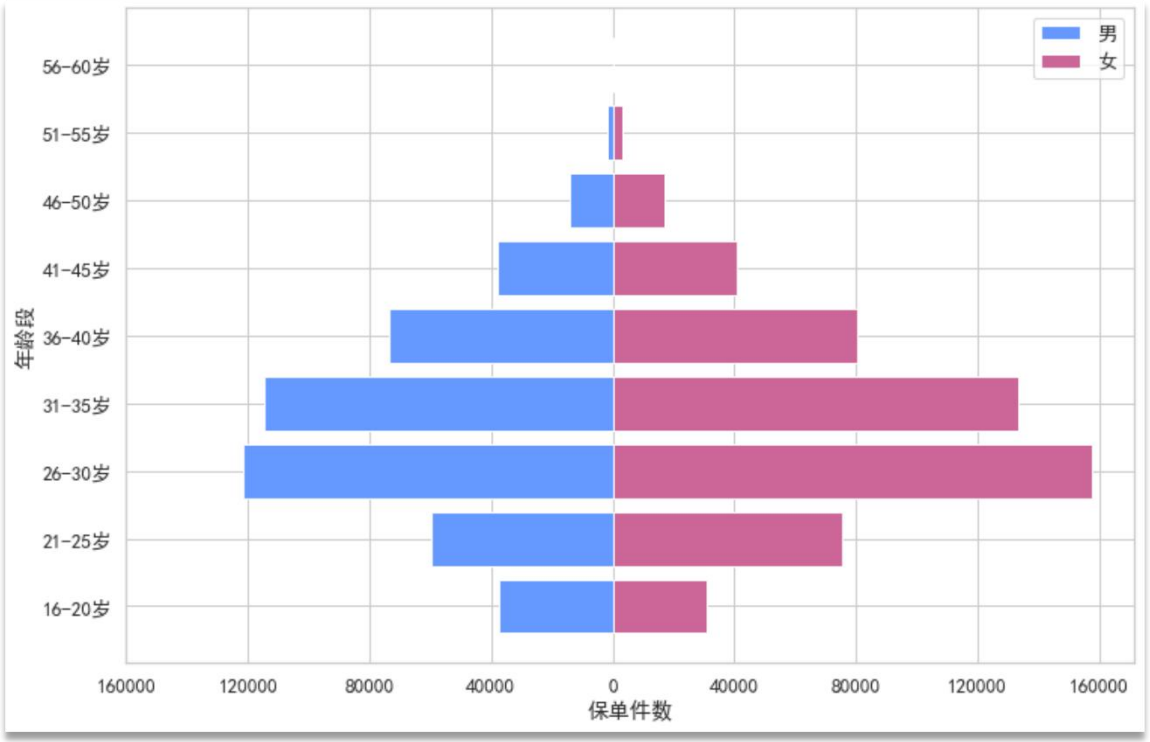
地区：华东、华南、华中、华北、东北、西南、西北

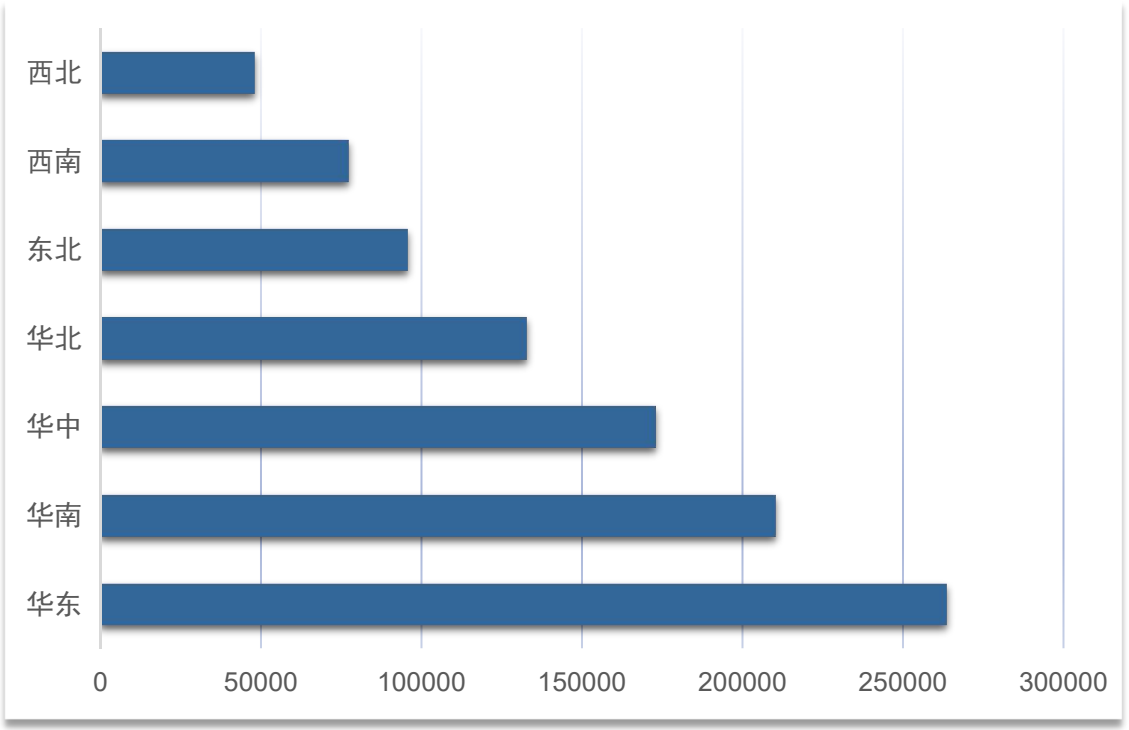
可能问题

- 在细分维度数据量较少的区间结果可信度不高
- 保单生效时间相对较短，无法进行长期趋势性分析

数据探索性分析——保单分布

性别/年龄/地区 — 保单件数分布





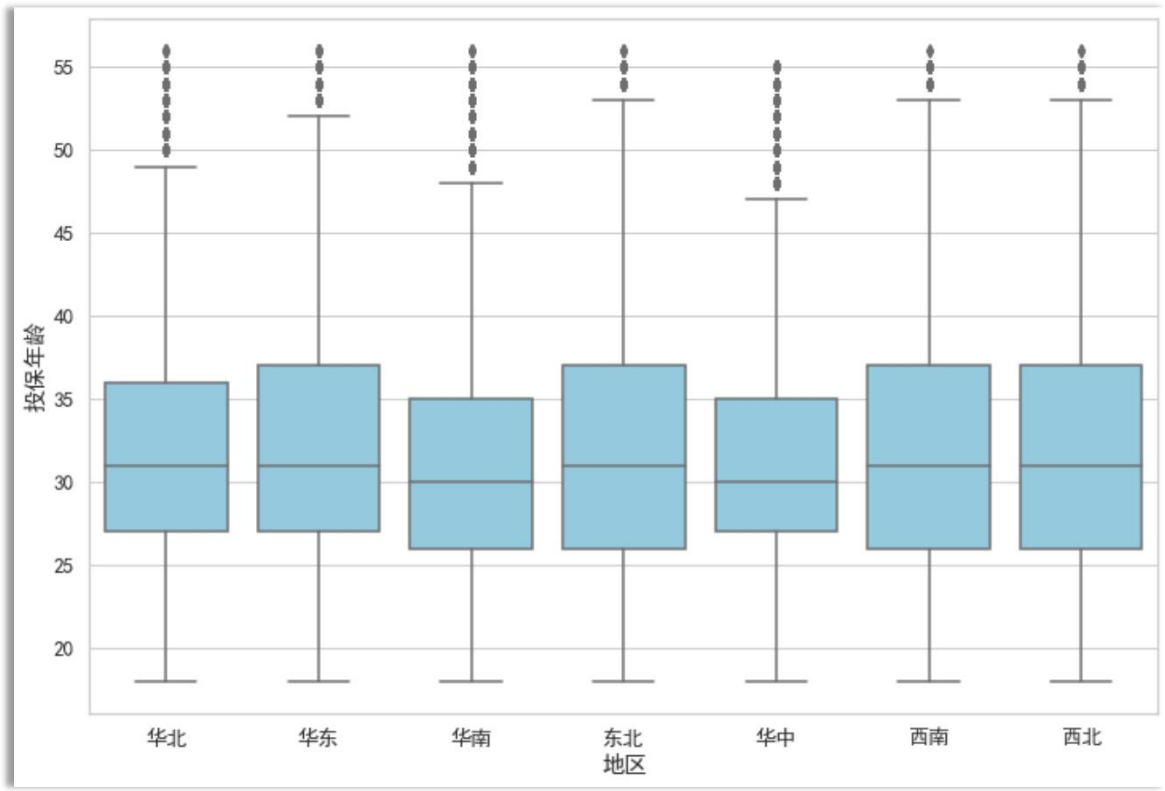
投保年龄：被保人年龄覆盖18-56岁，**投保年龄段集中在26-40岁的区间**，其中26-30岁占比为30.0%。从性别维度，投保人中53.89%为女性，女性在各年龄段占比都较高，但男女的整体年龄段分布近似。

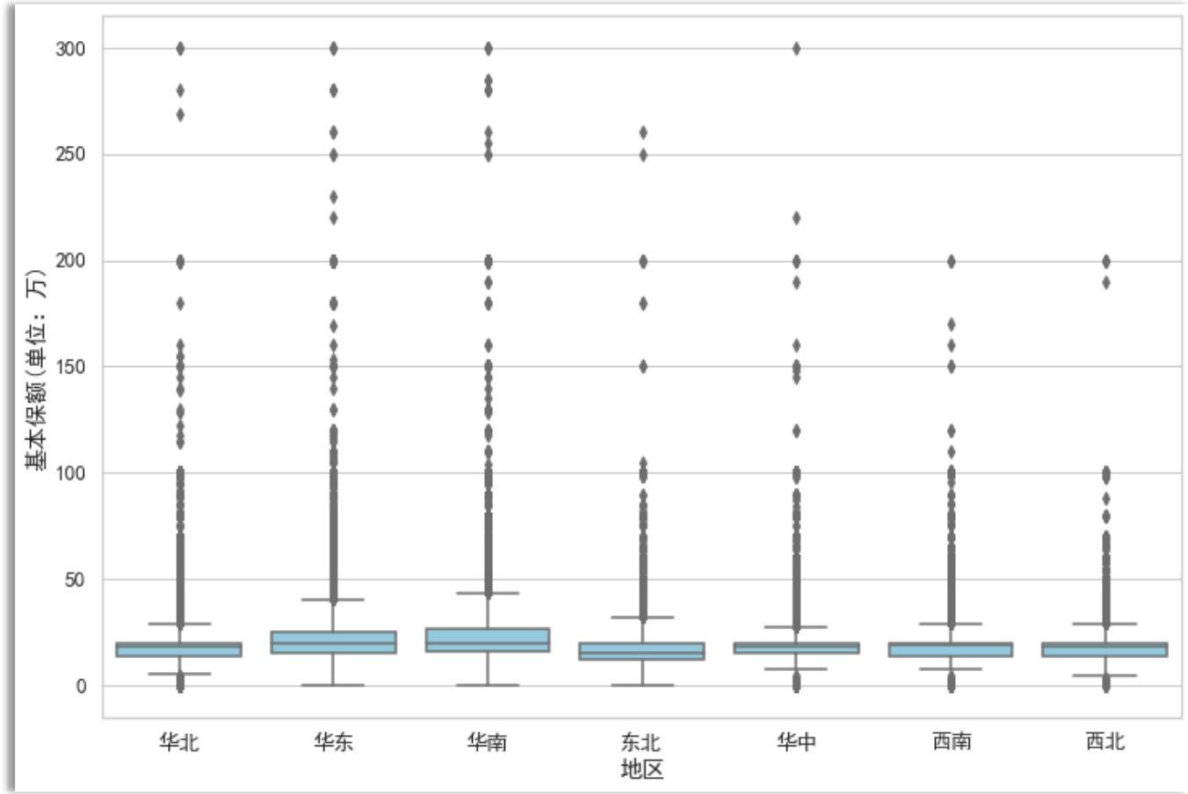
地区：该重疾险的销售区域覆盖全国30个省市自治区，**保单集中于华东（26.3%）、华南（21.0%）、华中（17.3%）三个地区**

数据探索性分析——用户画像

地区— 基本保额、年龄分布

LitchiData



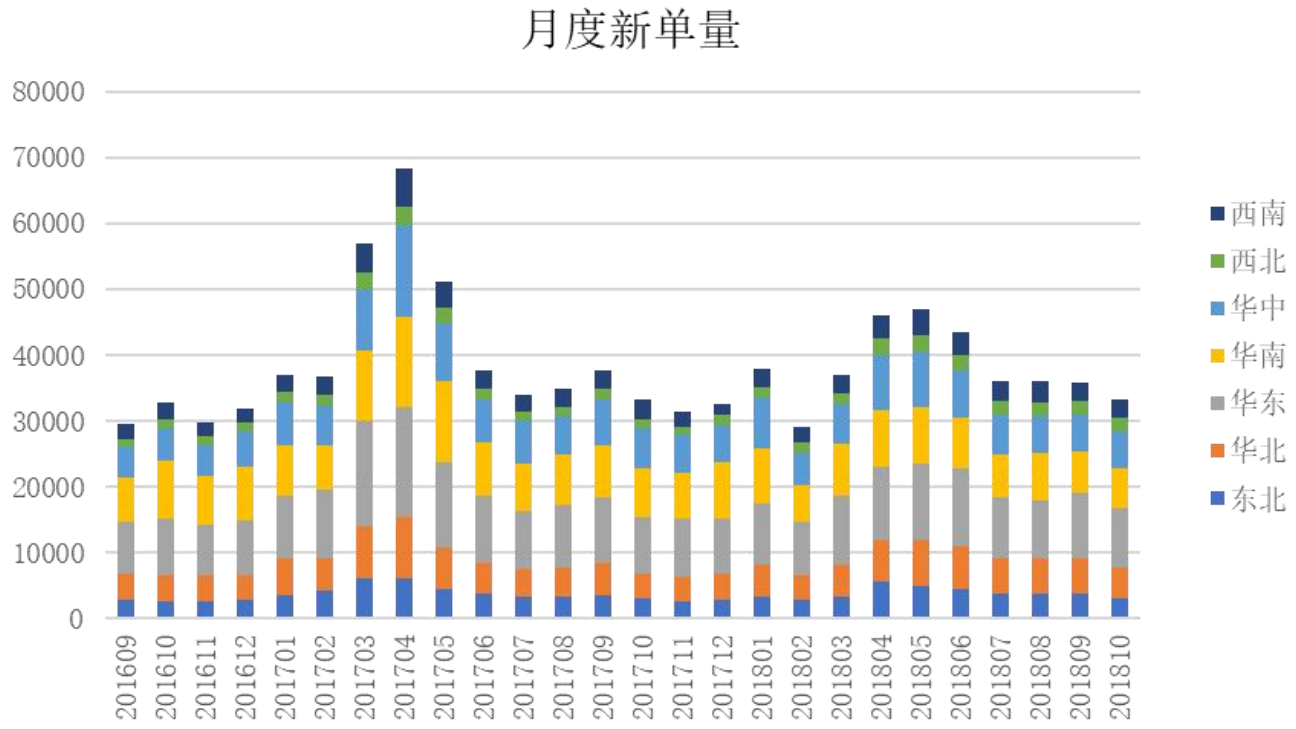


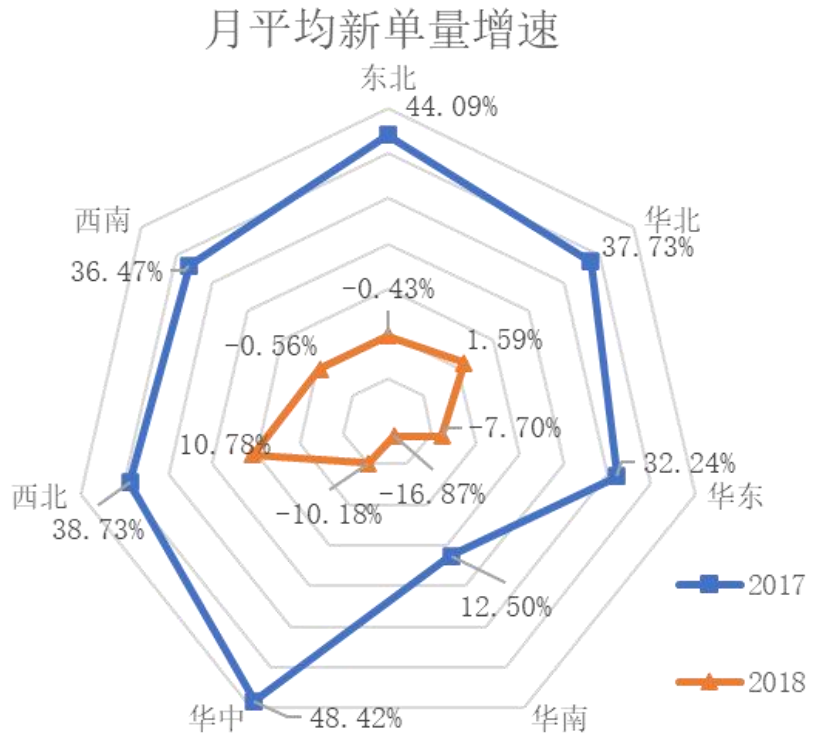
- 投保年龄：各地区年龄多集中在**26-36岁**区间内。**东北地区整体年龄分布最高，华南地区的分布最低**

- 基本保额：各地区基本保额箱型图都具有箱体较扁且端线较短的特点，说明**基本保额分布都较为集中**，分布在30-40万区间。**华东和华南两地的总体基本保额分布比其他地区高**，且大额保单数量较多。

数据探索性分析——新单分布

地区 — 月度新单量、月平均新单量分布





- 月度新单量的分布和保单数量的地区分布大体一致，**华东、华南、华中**三地占比最高，在时间维度上**2017年整体新单量更多**，尤其是3-5月有明显增长。
- 从月平均新单量来看，**2017年各地区的月度新单量都有明显提升，**尤其是华中、东北两地，但在2018年多地出现了增长乏力的情况，只有西北和华北两地的月度平均新单量呈正增长。

数据探索性分析——理赔信息分布

基于出险描述的疾病病种维度

LitchiData

- 基于数据中的出险过程描述字段，**通过关键词筛选确定病种**，并通过人工确认的方法对关键词不明确的近300条进行归类，得出疾病大类、疾病病种及癌症部位三个字段
- 在9283例理赔数据中，**78.6%为恶性肿瘤**，8.5%为心脑血管疾病，心脑血管疾病中急性心肌梗塞占比最高，为6.4%。

疾病大类	疾病病种	癌症部位
恶性肿瘤	恶性肿瘤	白血病、淋巴癌、鼻咽癌、支气管与肺癌、乳腺癌、肠癌、胃癌食道癌、肝癌、肾癌、甲状腺癌、女性生殖系统癌症、男性生殖系统癌症、膀胱癌、其它部位癌
心脑血管疾病	急性心肌梗塞、脑中风（后遗症）、冠状动脉搭桥术、其他心脑血管疾病	—
其他重疾	终末期肾病、其他重疾	—

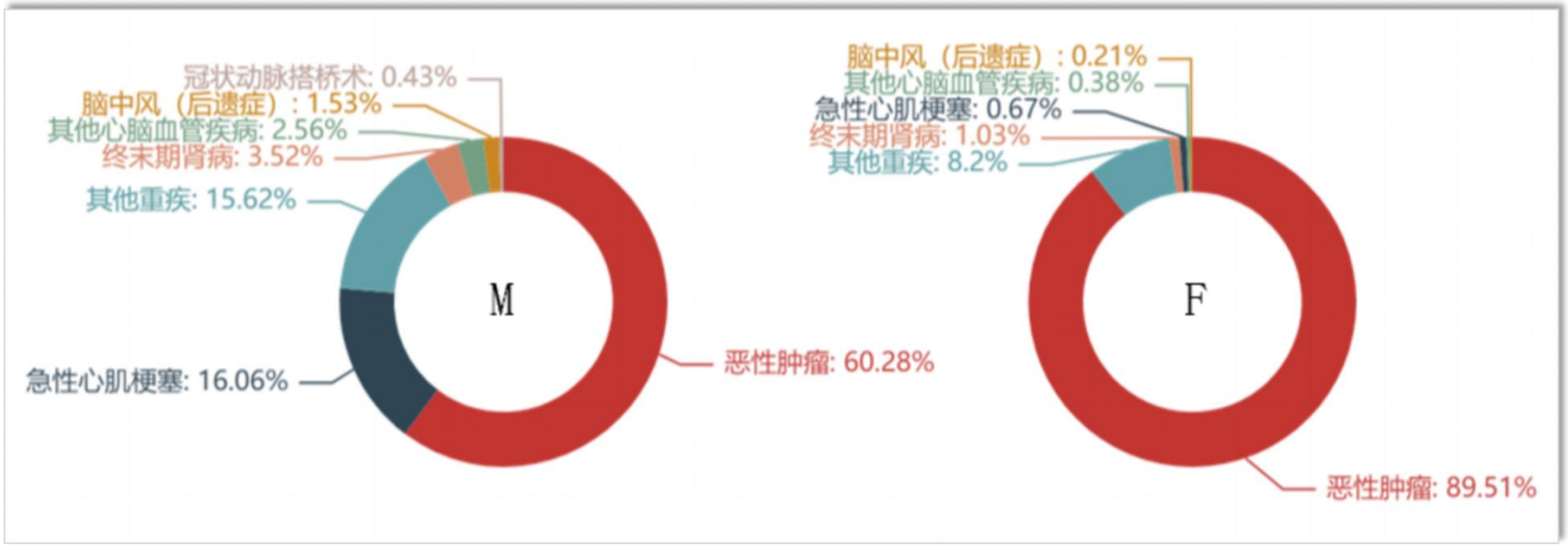
数据探索性分析——理赔信息分布

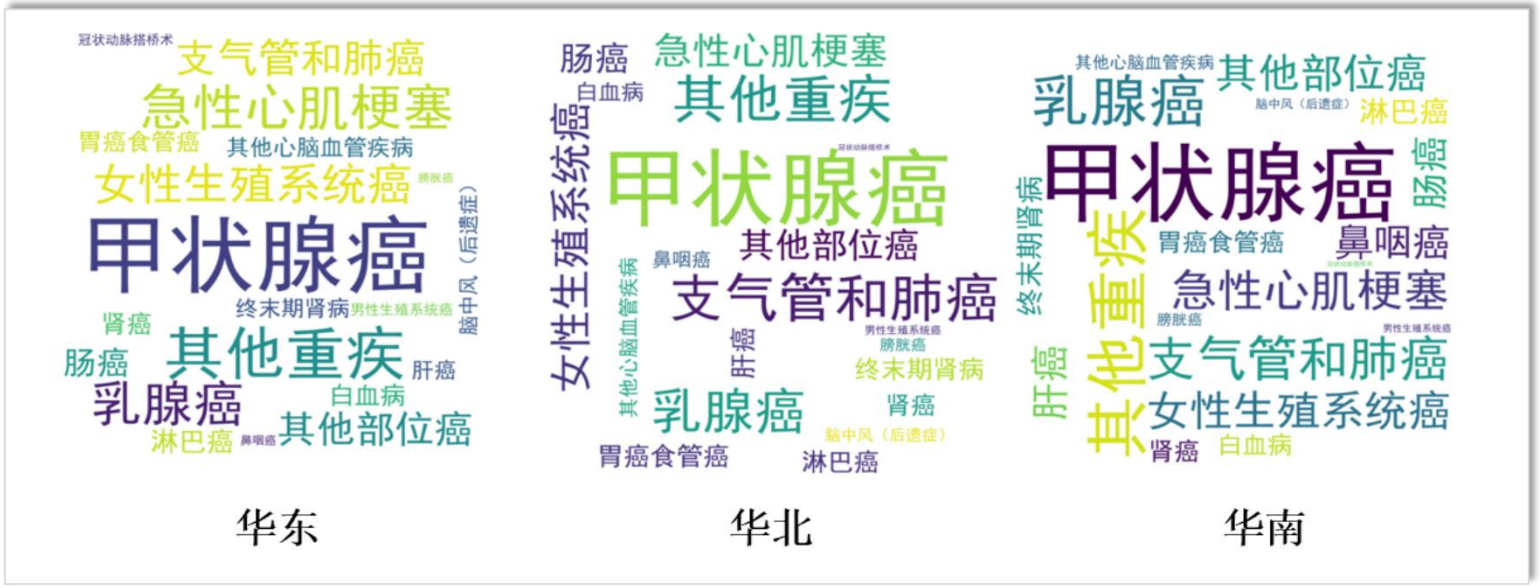
性别/地区—疾病病种分布

LitchiData

性别维度

- 从性别维度，**女性恶性肿瘤占比更高，高达90%**，男性恶性肿瘤占比相对较低，为60%。**男性急性心肌梗塞占比较高。**





地区维度

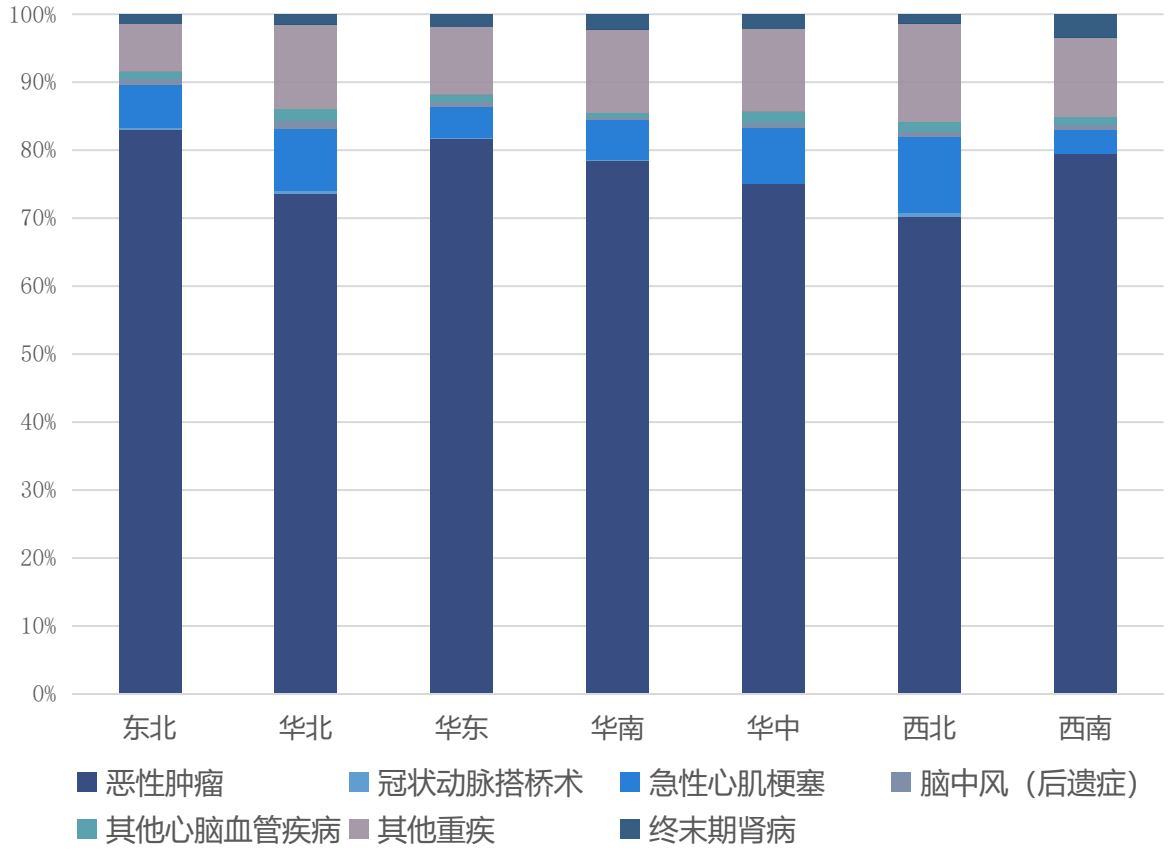
- 不同地区的病种特征呈现了一定的差异。**甲状腺癌在各地都是占比最多的疾病**，但其他疾病占比的相对大小有所不同。

数据探索性分析——理赔信息分布

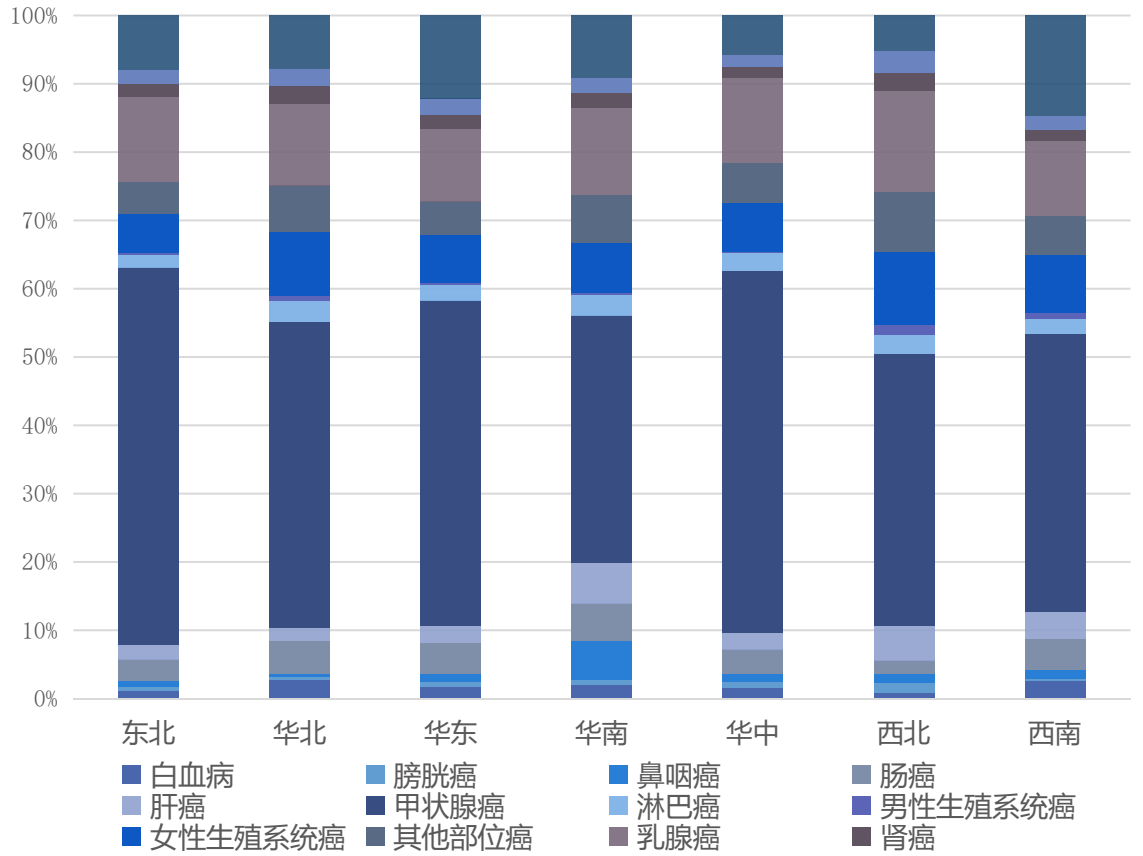
地区—疾病病种分布

LitchiData

地区 - 疾病病种



地区 – 恶性肿瘤部位

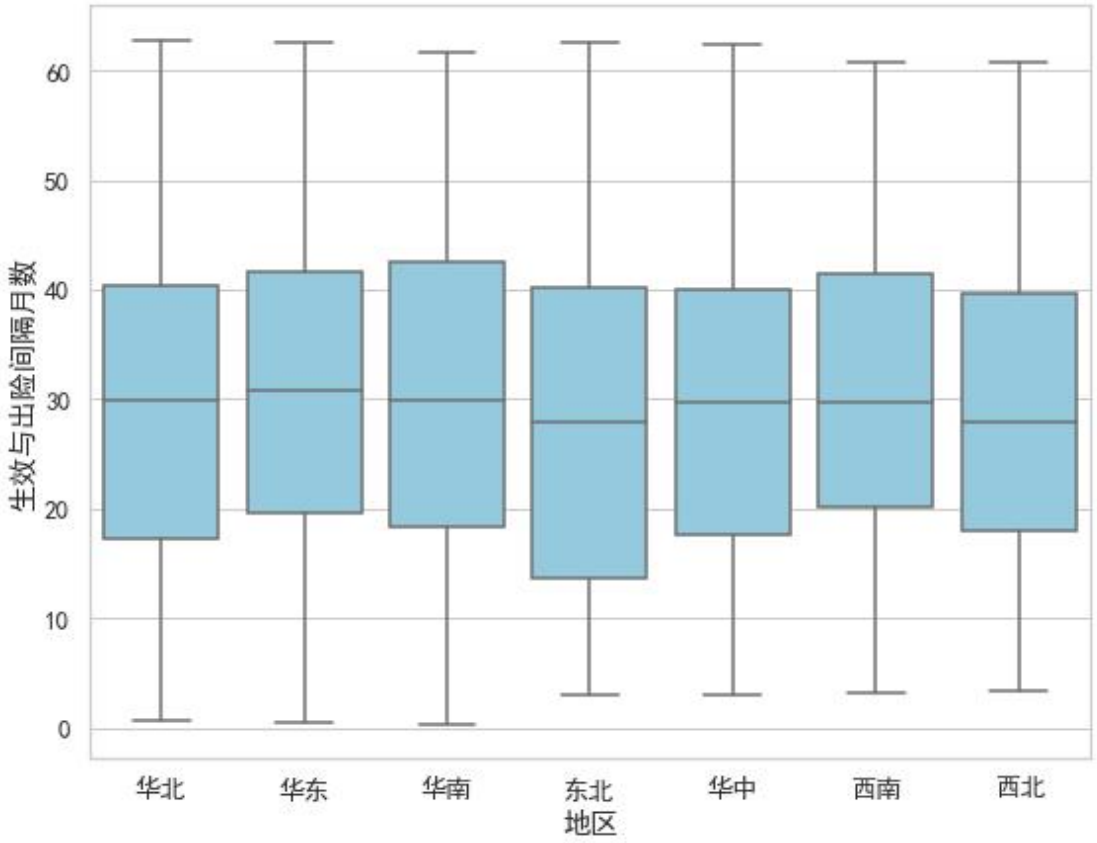


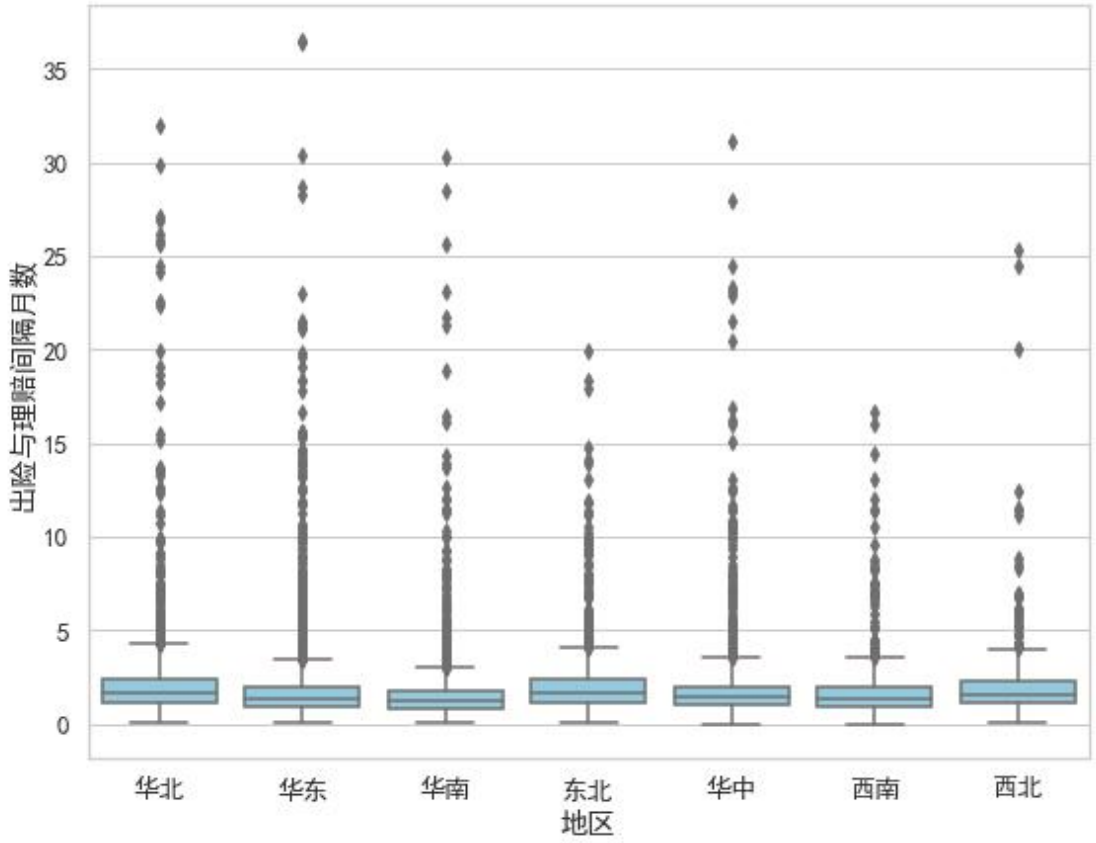
- 在重疾的理赔原因中，**恶性肿瘤占比始终位居高位，但在各个地区占比有所不同**，可能受地区医疗水平、居民健康意识等因素影响，其中华东、东北、西南地区具有更高的发病率。
- 从恶性肿瘤部位来看，**甲状腺癌占比在各地区高居首位**，但华东、华中、东北地区相对更高；华东和西南地区支气管与肺癌占比相对较高

数据探索性分析——理赔信息分布

地区 — 生效与出险、出险与理赔分布的间隔时间分布

LitchiData





- 保单经验数据中理赔数据删失比例较高，观察期只覆盖了五年左右，因此重点关注**相对水平。东北地区的间隔月数整体分布较低，推测其存在一定的逆选择现象。**右图展示了出险时间与理赔时间的间隔集中在30-60天，间隔时间较短。

数据探索性分析——重疾发生率分析

LitchiData

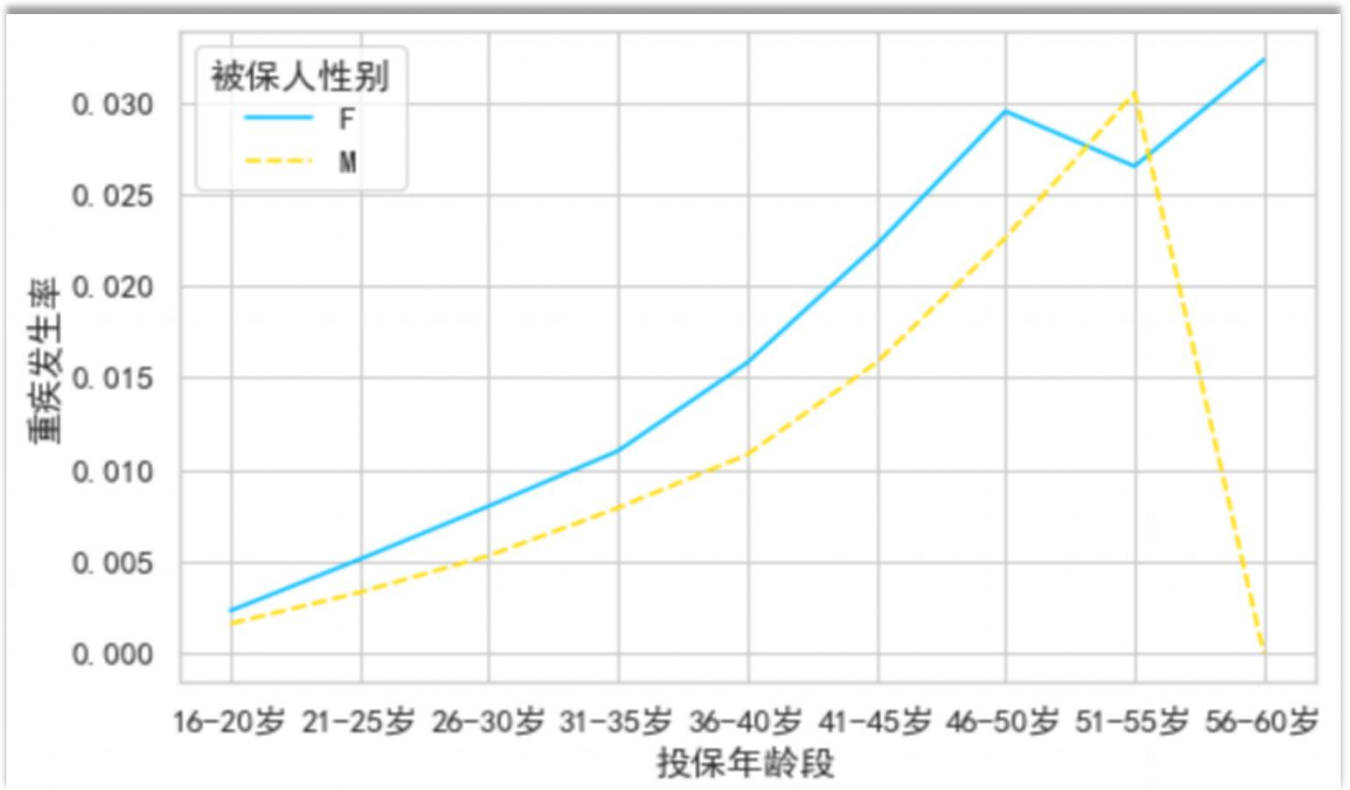
重疾发生率计算公式：

$$I_x = \frac{\theta_x}{E_x}$$

- θ_x为重疾数，指按保额加权的重疾及因重疾死亡的赔案
- E_x为暴露数，指按保额加权的观察期内观察到的保单件数

参考李霖《2020版重疾表编制过程与结果解读》

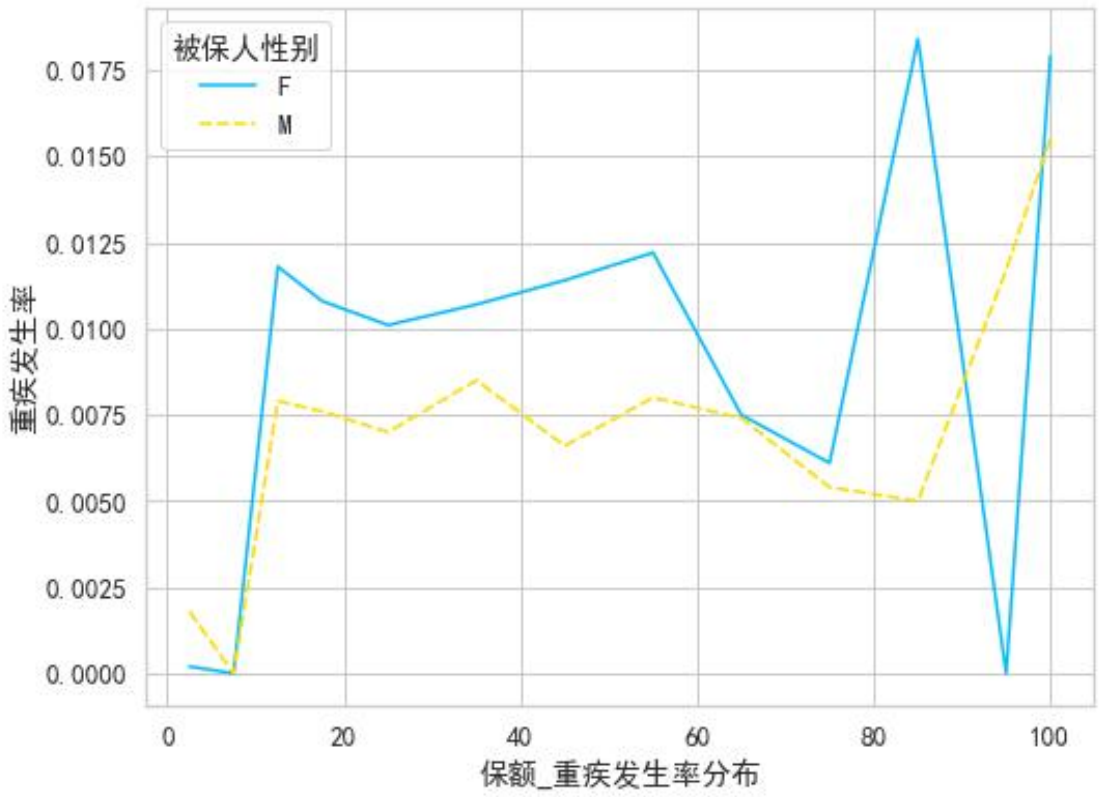
投保人特征 – 重疾发生率分析

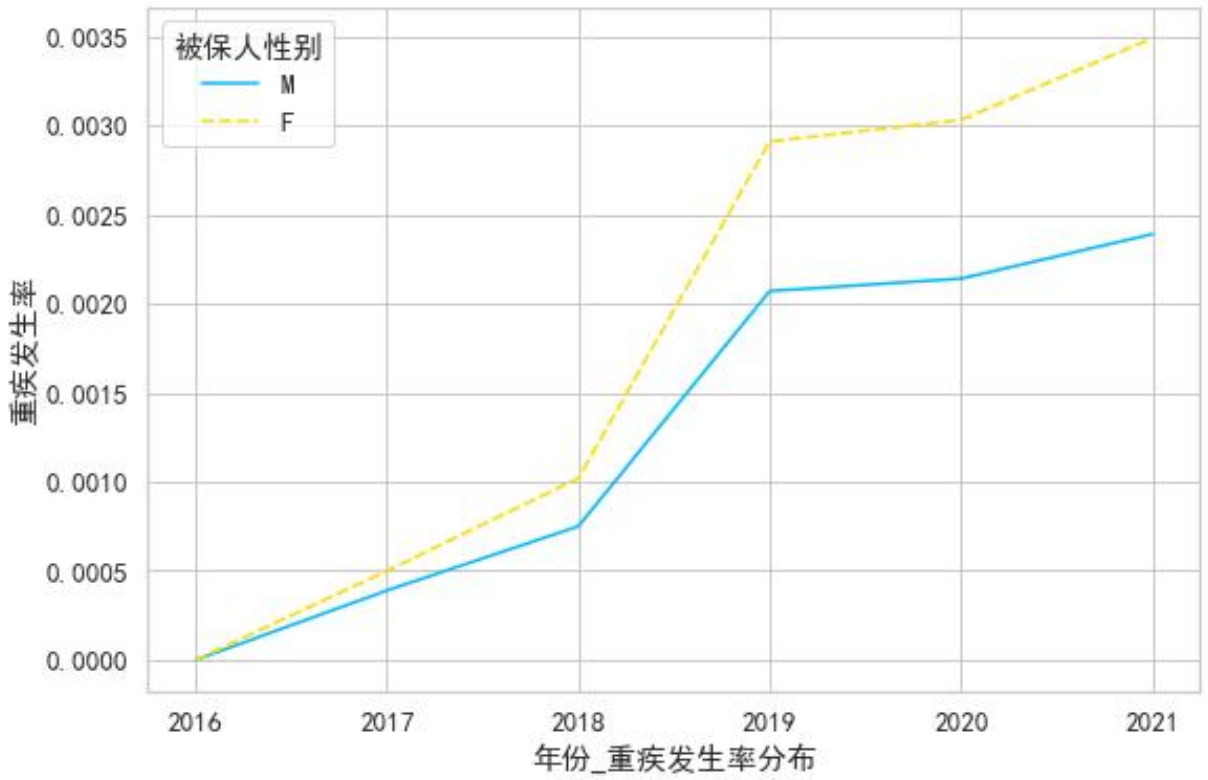


- 随年龄增加，重疾发生率明显升高，**但50岁以后男性出现显著下降，可能源于数据量较少造成的偏差**
- **女性重疾发生率高于男性重疾发生率**，一方面很多自身免疫性疾病、内分泌类疾病均有重女轻男的特点；另一方面甲状腺癌、生殖系统癌症在女性中较为高发

数据探索性分析——重疾发生率分析

保单特征——重疾发生率分析



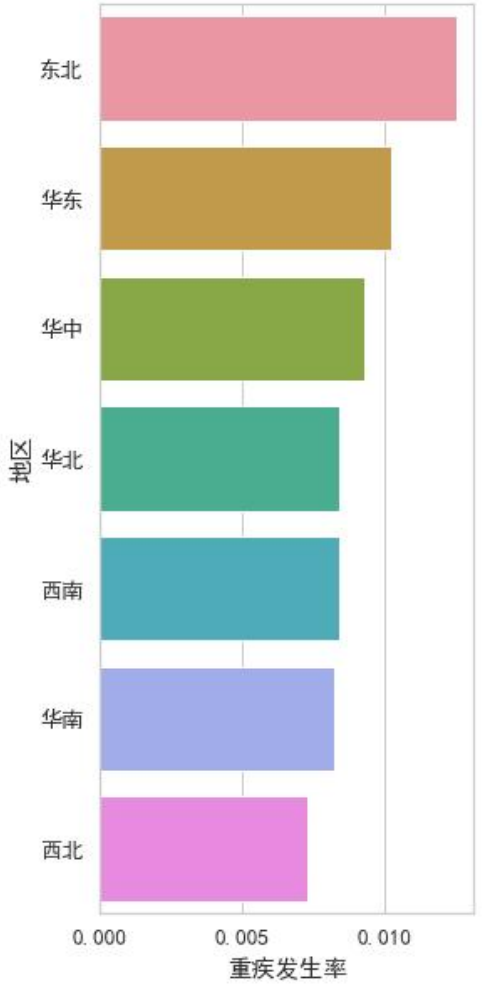


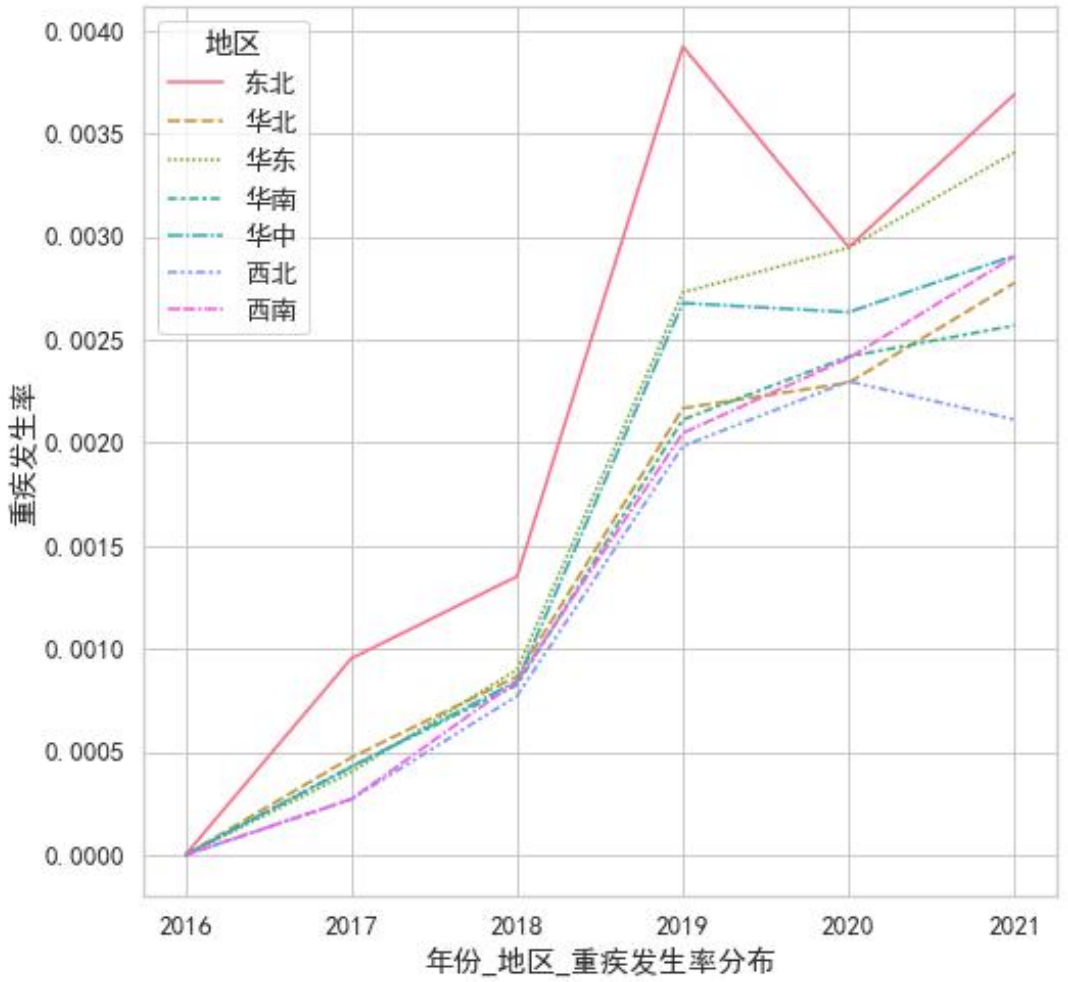
- 从基本保额维度，**在10-40万区间内，重疾发生率随着保额的增加整体上呈U型分布**，在低保额段（10-30万）呈下降趋势，在高保额段（20-40万）呈上升趋势，其中男性女性重疾发生率走势大致相同，女性普遍高于男性。在保单数量相对较少的40-50万，50-60万区间，重疾发生率总体上仍呈上升趋势。
- 从时间维度，**近年来重疾发生率持续升高**，女性重疾发生率始终略高于男性，但男性女性趋势整体一致。

数据探索性分析——重疾发生率分析

地域信息——重疾发生率分析

LitchiData

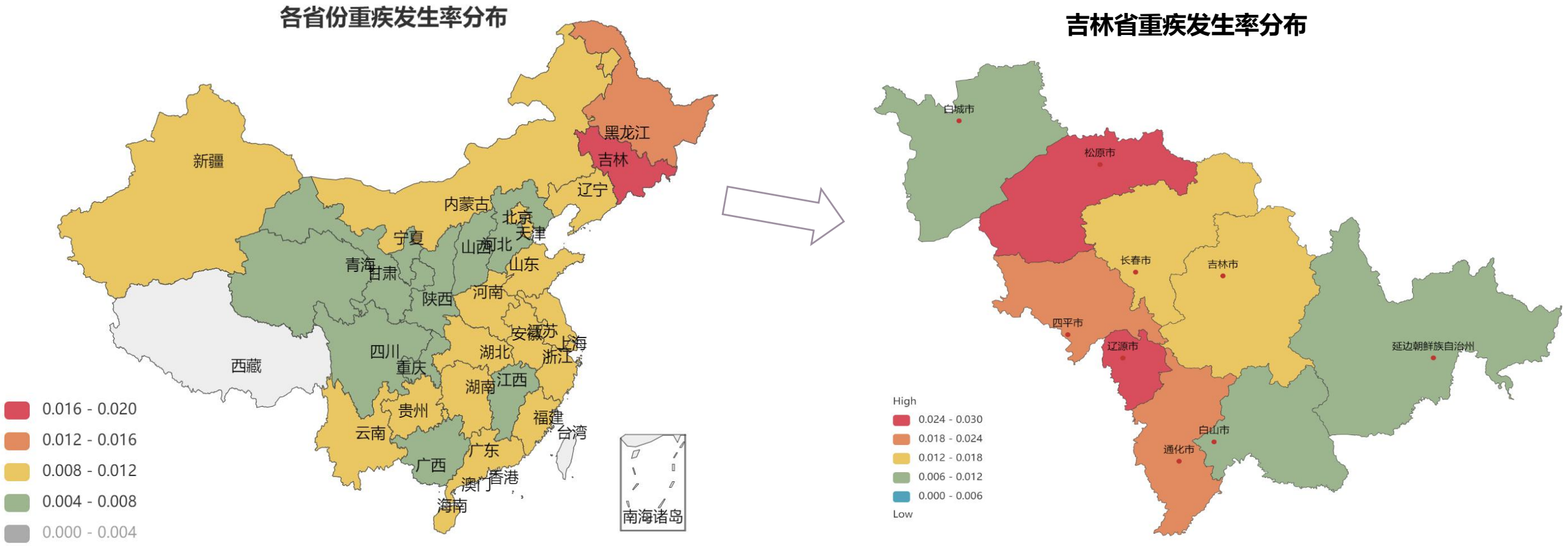




- 不同地区因地理气候条件、环境污染程度、经济发展水平、居民饮食习惯、健康观念等方面存在较大差异，其重疾发生率也有很大区别。
- **东北地区重疾发生率最高，其次是华东，华中两地。**
- 从时间维度来看，各地区相对位置固定，随时间重疾发生率**上升的趋势也大体一致。东北地区**在2019年相较于其他地区重疾发生率大幅提升，随后又发生回落。

数据探索性分析——重疾发生率分析

地域信息——重疾发生率分析



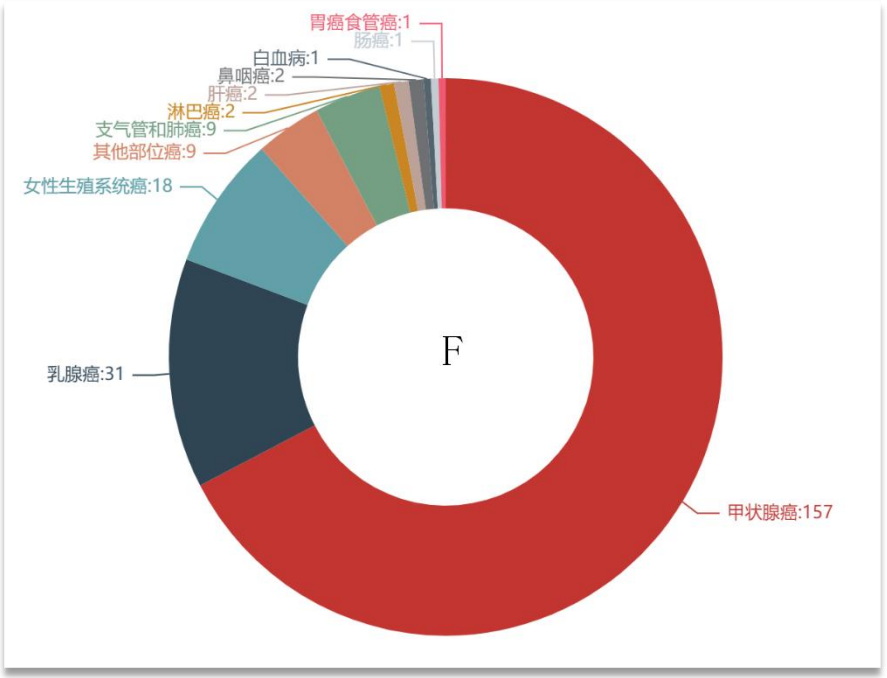
- **吉林省重疾发生率最高**，其次是黑龙江省、浙江省、云南省、上海市等。对**吉林省**进一步分析，**城市间差异较为显著**，其中**辽源市、松原市、四平市和通化市四市重疾发生率超20‰**，而白山市低于10‰。
- 虽然吉林省西部属于盐碱地地带，土地情况较差，但近年已得到有效治理，同时《中国肿瘤登记年报2022》数据显示吉林省西部四市甲状腺癌粗发生率并没有明显高于其他城市，因此推测**西部四市的高重疾发生率与逆选择有关。**

数据探索性分析——重疾发生率分析

吉林省——分病种重疾发生率分析

LitchiData

省份	重疾发生率	恶性肿瘤发生率	心脑血管疾病发生率	甲状腺癌发生率	其他癌症发生率
吉林省	1.66%	1.44%	0.10%	1.00%	0.44%
黑龙江省	1.22%	1.03%	0.08%	0.57%	0.46%
云南省	1.16%	0.94%	0.09%	0.46%	0.48%
浙江省	1.15%	0.97%	0.06%	0.52%	0.45%
上海市	1.13%	0.97%	0.09%	0.45%	0.52%
省平均	0.92%	0.73%	0.08%	0.35%	0.38%



- 选取重疾发生率最高的五个省市自治区对比其各类疾病病种发生率，各省份心脑血管疾病的发生率差异不大，和省平均水平相近，**重疾发生率的差异主要来自恶性肿瘤，具体为甲状腺癌**，吉林省甲状腺癌的发生率几乎为黑龙江省的两倍。
- **吉林省女性重疾发生率（19.7‰）远高于男性（12.4‰），**对女性疾病病种进一步分析，恶性肿瘤占比达94%，**其中甲状腺癌占比高达67%，远高于所有城市的平均水平46%。**

4

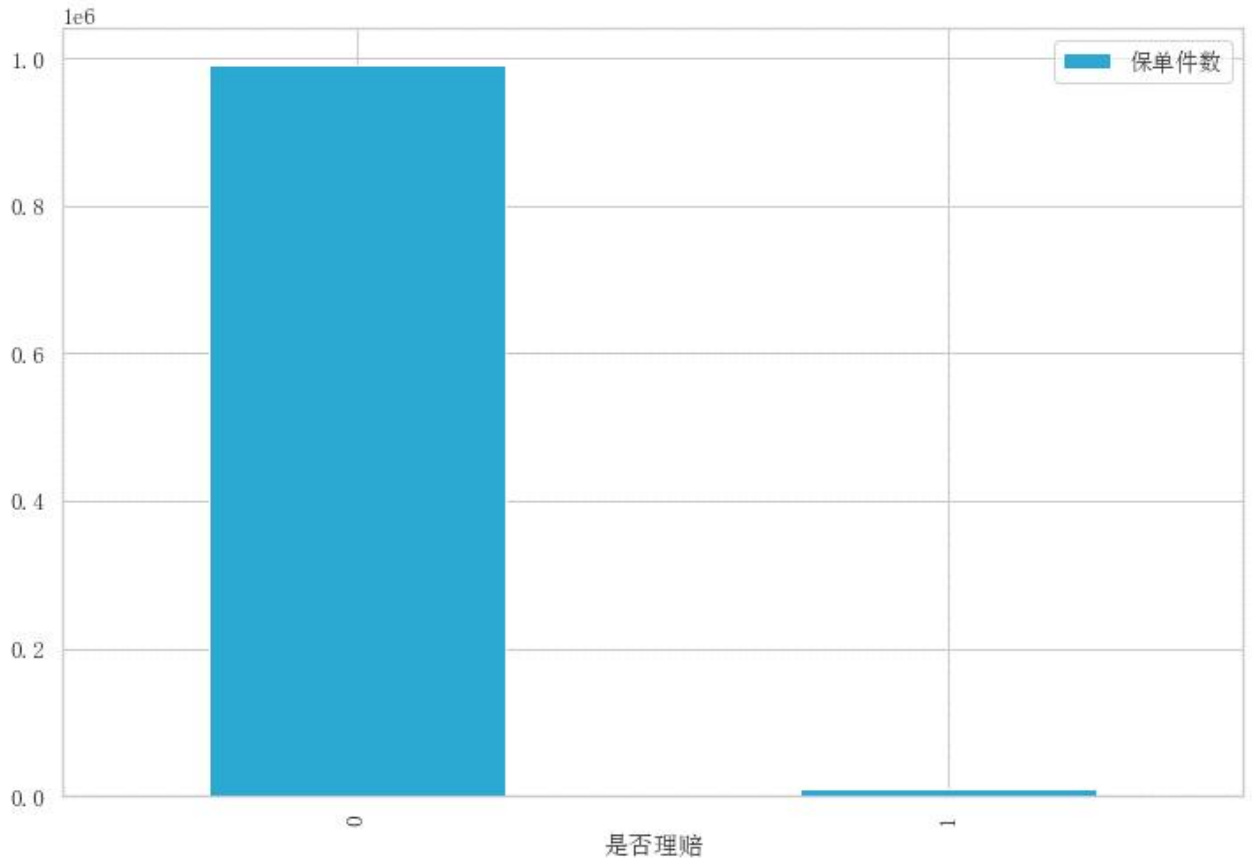
PART 4

重疾发生概率的预测建模

- 标签分布
- 现有变量处理
- 新增地域数据
- 建立分类模型
- 重疾风险预测工具——“荔察”

标签分布

样本严重不平衡



理赔保单占比0.93%，未理赔保单占比99.07%，严重的标签不平衡

- 理赔数据9283条，每条样本为1张保单，未理赔数据361368条，**每条样本的保单件数从1到198不定，总计100万张保单**
- 样本数据将特征相同的未出险保单合并为一条样本，即认为**未出险的保单都是相似的，出险的则各有各的不同**
- 直接使用是否理赔的样本数据作为标签，全部样本为370651条，**理赔样本占2.5%,未理赔的样本占97.5%，不平衡程度降低**

- 通过对**未出险保单（多数类）的欠抽样**，减少了未出险保单的冗余信息，同时也降低了初始标签不平衡的程度。
- 后续建模中，将调整分类模型对是否理赔样本的惩罚权重进一步解决标签不平衡，**对理赔和未理赔样本分别赋予不同权重，权重设置和类别数量呈反比**，使模型更关心理赔的样本。

样本数据现有特征

类别	变量名称	英文表示
个人特征	被保人性别、投保年龄段	Gender, Age_group
地域特征	省级行政机构、市级行政机构、城市线、地区	Prov, City, Cityline, Region
保单特征	基本保额段、缴费年限	SA_group, ppp

类别变量编码处理

类别	变量名称	英文表示
编码特征	省级机构保单计数、市级机构保单计数、基本保额段保单计数	Prov_policy_count, City_policy_count, SA_policy_count,

- 由于我们建立的是对未来一段时间内是否发生重疾的预测模型，事后的出险相关信息不能作为特征输入模型！现有变量中选择了8个变量作为输入特征（左表）
- 对维数较少的特征直接进行简单的类别编码（Label Encoder），对维数较多的特征（省级行政机构、市级行政机构、基本保额段）只依靠类别编码会损失大量信息，对其增加保单件数的计数编码，反映该特征下不同类别的风险暴露数（右表）

类别特征编码之后，模型的输入特征有11个

新增地域信息

省级、市级变量

LitchiData

类别	市级变量	省级变量	英文表示
经济发展水平	人均地区生产总值、第三产业比重、职工平均工资	人均地区生产总值、第三产业比重、职工平均工资	GRP_per, GRP_I3, Wages（相同名称省级后缀加prov，下同）
人口数量结构	总人口、人口密度	总人口、人口密度、城镇人口比重、性别比	Population, Population density, Urban_rate, Gender_ratio
自然环境资源	绿地面积、人均公园绿地面积	绿地面积、人均公园绿地面积、人均水资源量、森林覆盖率	Green_area, Park_green, Water_per, Forest_cover
地理气候特征	海拔高度、年平均气温、年累计降水量、年日照时数		Altitude, Temperature, Rainfall, Sunshine
医疗卫生资源	医生数量、基本医疗保险参保人数	医疗保健支出占比	Doctors, Medical_ins, Health_com
大气污染	PM2.5年平均浓度		PM2.5

采取**2016-2020年的年度数据均值来代表该地域近五年的发展情况和资源水平**，不再区分数据的产生时间。最大程度避免部分年份部分指标的缺失问题。

少数民族自治州（临夏回族自治州、凉山彝族自治州等）和县级市（济源市、莱芜市等）在大部分指标上都没有数据，针对此类缺失情况，**利用该市级机构所在省份的指标中位数值来填补缺失值**

资料来源：EPS全球统计数据平台（中国城市数据库、中国城乡建设数据库、中国环境数据库等）

新增地域信息

市级与省级的比值变量

LitchiData

- **市级和省级层面都有的变量有：**

经济发展水平

人均地区生产总值、第三产业比重、职工平均工资

人口数量水平

总人口、人口密度

绿地资源情况

绿地面积、人均公园绿地面积

- **计算市级与省级变量的比值（后缀为ratio）：**

相对经济发展水平

人均地区生产总值之比、第三产业比重之比、职工平均工资之比

相对人口数量水平

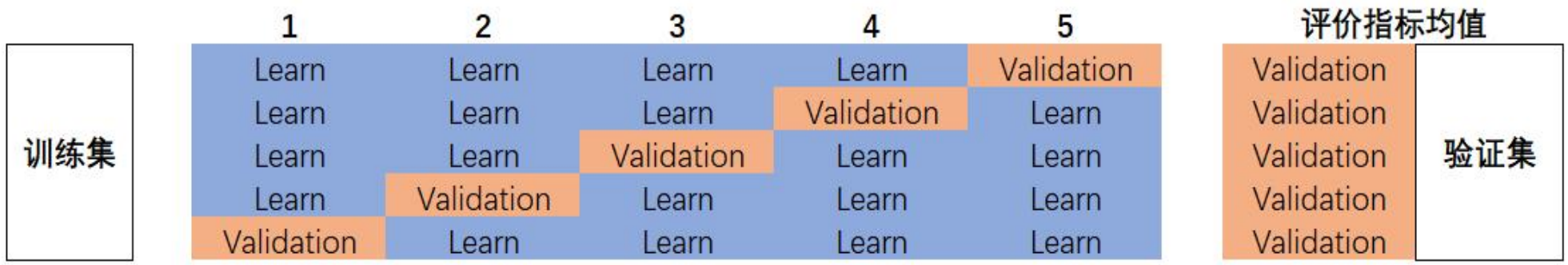
总人口之比、人口密度之比

相对绿地资源情况

绿地面积之比、人均公园绿地面积之比

- 不同地区之间发展程度和自然资源的水平差异较大，指标的绝对值往往意义不大，**相对大小更能表示出该市级机构在所在省份的发展水平和重要性**
- 最终新增33个与地域相关的变量，包括市级、省级和比值变量，它们可以表示某地域更多维度更丰富全面的信息；并剔除原始数据中市级行政机构和省级行政机构两个变量，**最终进入模型的预测特征共有33+11-2 = 42个**

用什么数据评价模型？



- 为了最大化利用已有数据并提高模型的可靠性和结果的说服力，采取5折交叉验证的方式对样本进行分割，**取5次验证集评价结果的均值作为对模型预测效果的评价**

用什么指标评价模型？

- 对二分类模型而言，准确度（Accuracy）、精确率（Precision）、召回率（Recall）、F1-score均需要设定阈值去判别预测的样本属于正类还是负类，分类阈值的不同会影响这四种指标的大小。
- AUC指标是ROC曲线下的面积，它不关注模型给出的概率值大小本身，**更关注预测概率的相对顺序，衡量了预测结果的排序质量，**并不受样本自身类别比例变化的影响，更适合这次的样本不平衡数据。

➢ *最终的重疾风险预测工具输出的也是某类人的相对风险水平，并不是绝对值大小，**所以模型对预测结果的排序能力更为重要，AUC是更符合题目要求的评价指标。***

建立分类模型

模型结果对比

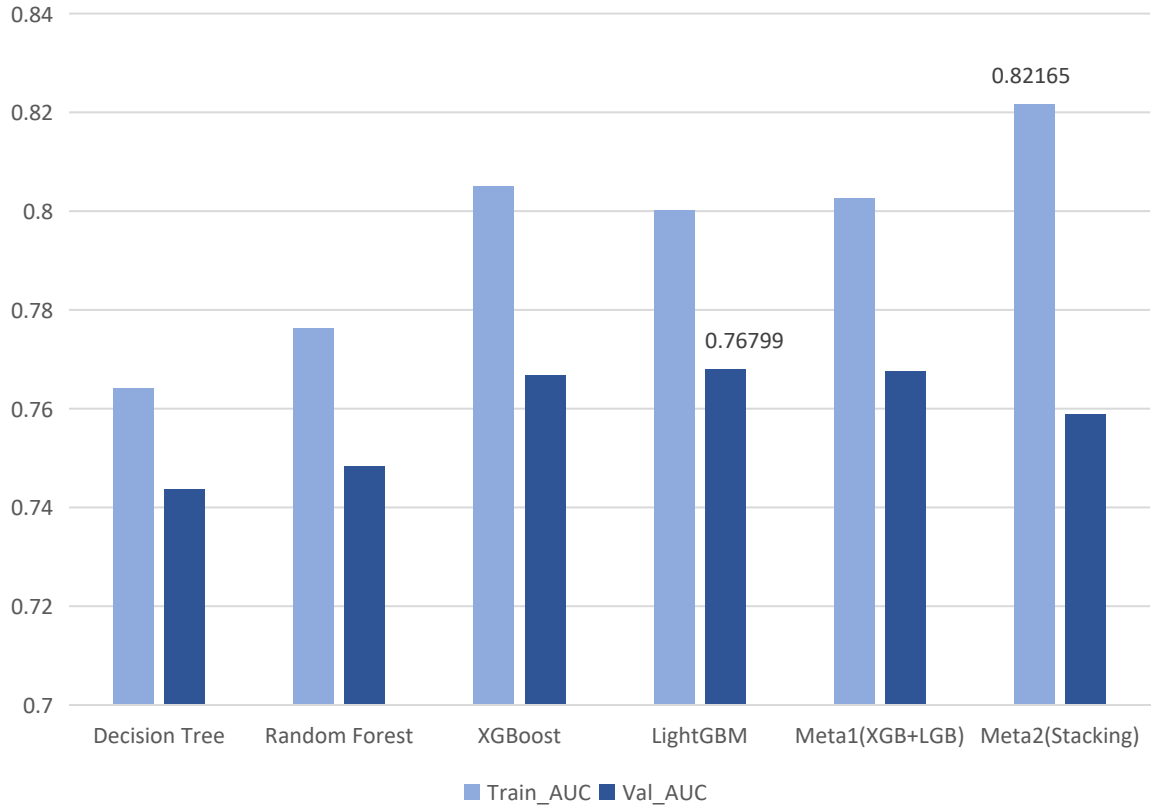
LitchiData

➢ **树模型及其扩展模型：**决策树、随机森林、XGBoost、LightGBM、以及XGB和LGB的两种融合模型Meta1和Meta2

模型名称	Train_AUC（五折平均）	Val_AUC（五折平均）	训练时间
Decision Tree	0.76408	0.74369	8.99s
Random Forest	0.77622	0.74835	9min 37s
XGBoost	0.80504	0.76662	17min 39s
LightGBM	0.80021	**0.76799**	**4min 10s**
Meta1(XGB+LGB)	0.80262	0.76755	21min 50s
Meta2(Stacking)	0.82165	0.75886	2h 21min 21s

Meta1：XGB和LGB的预测结果简单平均

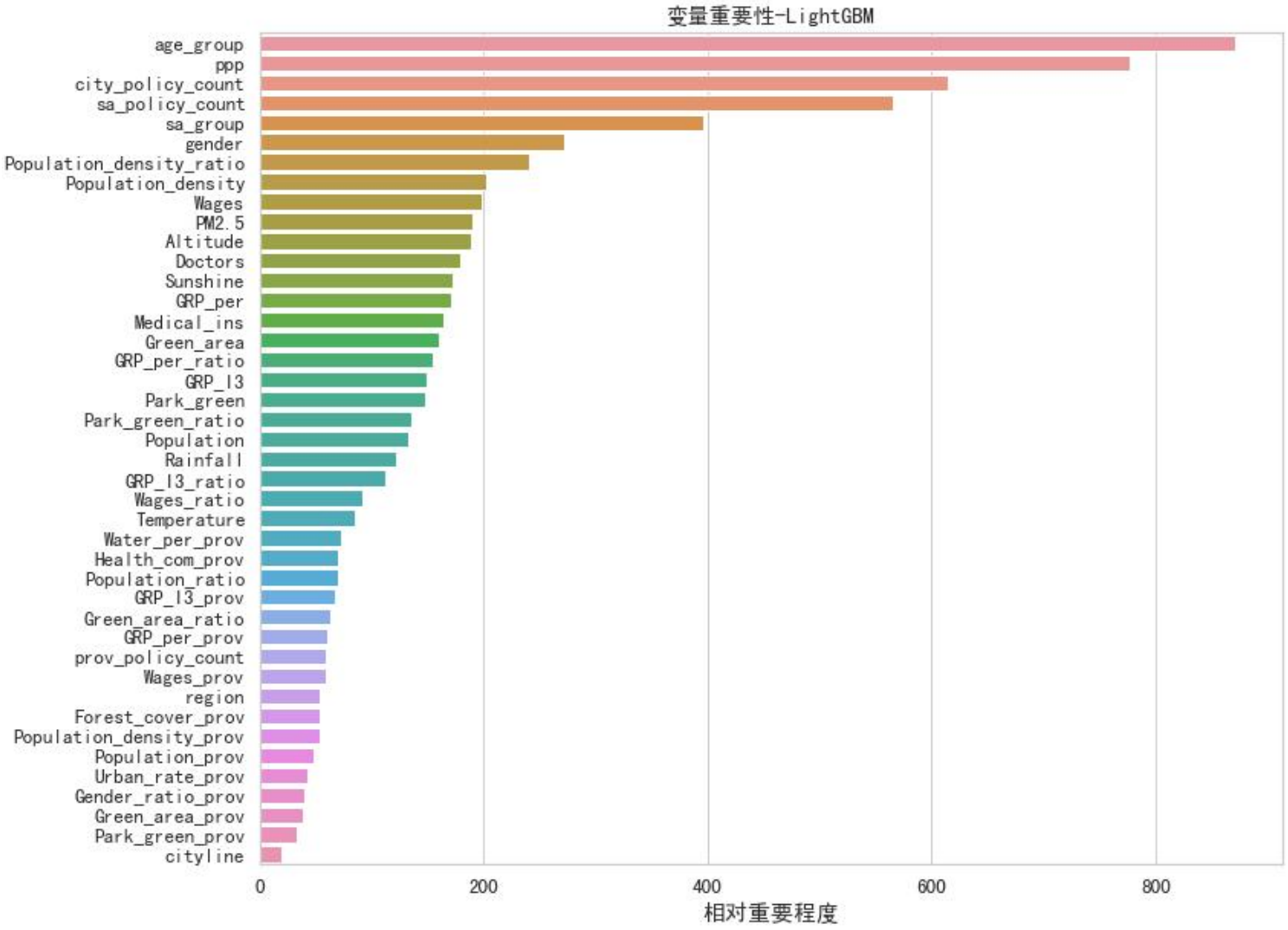
Meta2：Stacking方法构建多层模型，XGB和LGB作为基模型，次级模型为LR



六种分类模型预测效果比较

- 验证集预测效果：决策树 < 随机森林 < Meta2 < XGBoost < Meta1 < LightGBM
- **综合模型的复杂度、训练时间、验证集表现等因素，我们最终选择LightGBM作为重疾发生概率的预测模型！**

变量重要性对比-LightGBM



个人特质仍是决定重疾发生概率的关键因素

个人的年龄段和性别重要性都很高，

其中年龄段重要性排名第一

样本数据中可能存在逆选择现象

缴费年限、基本保额段重要性也都排在前列，

保单特征在一定程度上能区分重疾风险的高低

市级机构的保单计数特征在地域因素中最重要

市级机构的保单计数特征重要性排在最前面，

保单数量更能直接衡量城市间的重疾风险暴露差异

比值变量比省级变量更重要

市级和比值变量的重要性基本都在省级变量前面，

省级变量重要性几乎都排在末尾，

说明它没有很好地区分不同地方的重疾风险差异

重疾风险预测工具——“荔察”

预测流程介绍

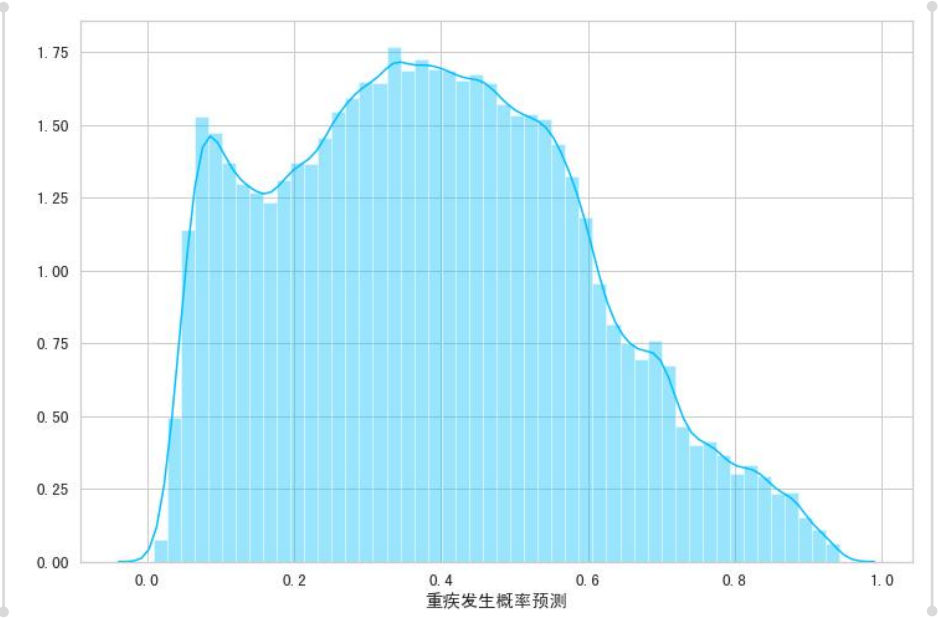
LitchiData

➢ **将训练好的LightGBM模型进一步包装，打造出重疾风险预测工具——“荔察”（取荔枝人寿明察秋毫之意）**

“荔察”的输入和输出

- **输入：**性别、年龄、居住城市、重疾险目标购买保额、目标缴费年限
- **原始输出：**重疾发生概率
- **最终输出：**在人群中的相对风险水平&风险程度等级&所在地区高发重疾

重疾发生概率分布图



重疾发生概率转换关系

重疾发生率概率	风险程度等级	相对风险水平
小于$q_{0.25}$	低风险	0~25%之间
$q_{0.25}$-$q_{0.5}$	中低风险	25%~50%之间
$q_{0.5}$-$q_{0.75}$	中高风险	50%~75%之间
大于$q_{0.75}$	高风险	75%~100%之间

➢ 重疾发生概率大体服从正态分布，**但大部分集中在中低概率，小部分重疾发生概率达到0.8以上**，这也比较符合现实情况，少部分人群是高风险，更多的人是中低风险

➢ 根据样本数据**重疾发生概率的25%分位数、50%分位数和75%分位数来划分预测结果的相对风险水平**，并将其转换成不同的风险程度等级，如右表所示

重疾风险预测工具——“荔察”

应用效果展示

LitchiData

- 下图是个人信息为**女性、42岁、目标保额30万、目标缴费年限10年、居住地为北京市**的应用示例，对于中高风险和高风险人群，“荔察”也会输出所在地区的高发重疾，提醒保险公司防范逆选择风险。

输入信息中…

输出预测结果

之后我们可以把训练好的“荔察”部署到线上，**利用Flask等框架为模型创建一个API，使“荔察”成为能够在网页上运行的应用程序**，满足企业内部及外部更为广泛的需要，这也是可以改进的方向之一。

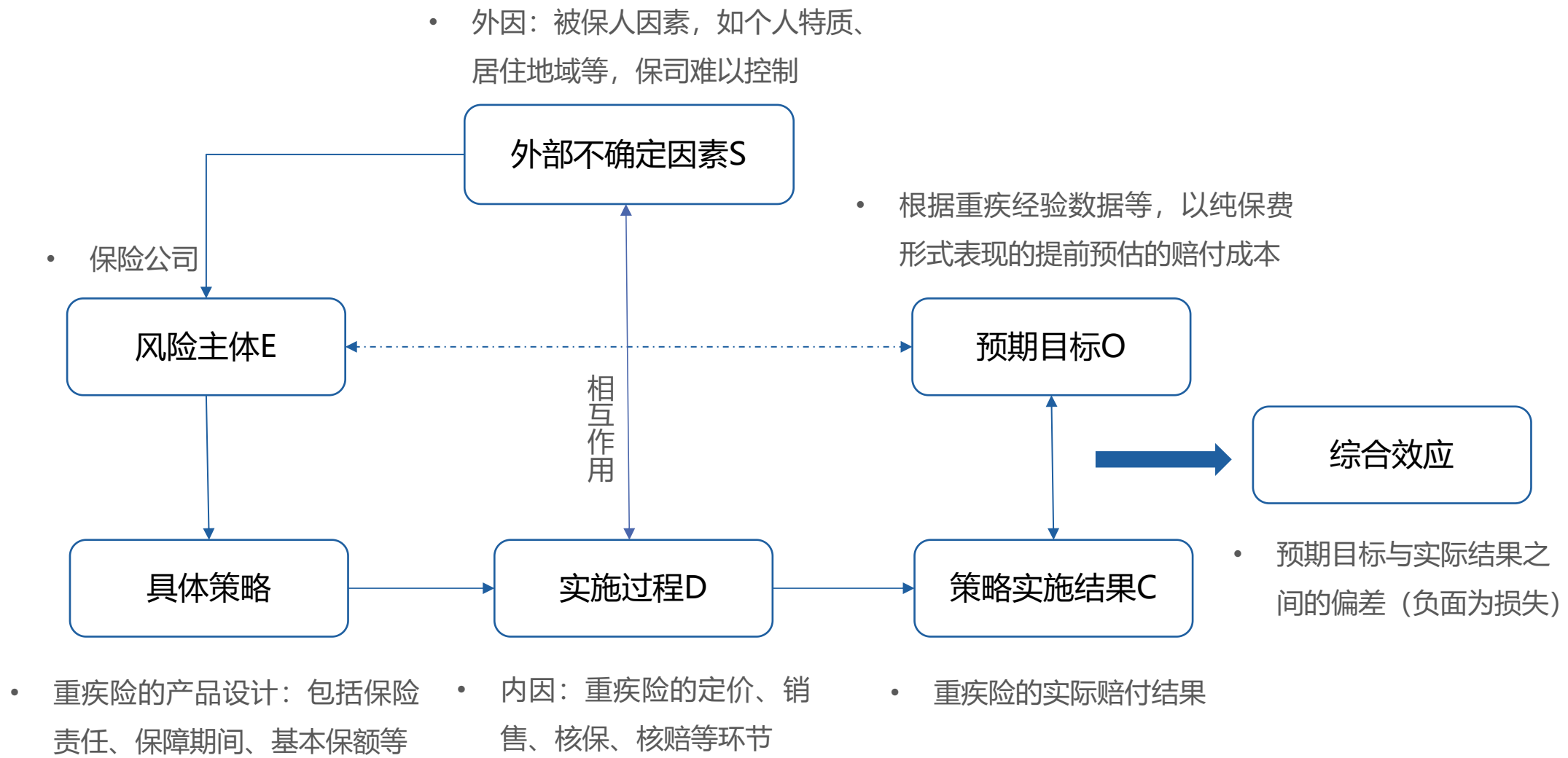
5

PART 5

- 风险新定义
- 针对性建议
- 总结与展望

降低重疾险赔付风险的建议

参考谢志刚、周晶（2013）《重新认识风险这个概念》，将**重疾险的赔付风险形成机制定义为如下图所示的动态因果过程：**



降低重疾险赔付风险的建议

针对性建议

LitchiData

- 重疾险赔付风险是一个由内因（保险公司决策行为）、外因（被保人因素）及其综合效应共同构成的动态因果过程，根据其形成机制，可以在以下方面针对性地管理重疾险赔付风险：

内因（保险公司决策）

- **产品设计：**利用“荔察”对拟承保人群进行风险分级，更好地匹配承保条件与风险状况，对风险进行更合理的评估和定价，以消费者需求为中心设计产品。
- **销售：**“荔察”的风险筛选和分级功能可以应用到客户的购买倾向分析中，更有效地将不同的重疾险产品与不同的消费群体进行匹配，为客户推荐最适合的重疾产品。
- **核保：**“荔察”可以作为预测核保工具(predictive underwriting，PU)嵌入到核保流程中，根据其预测结果设置差异化的核保规则，使其成为传统核保方式的有效补充。
- **理赔：**“荔察”能够用于对理赔欺诈保单的识别和分析，通过将预测标签转变为是否发生理赔欺诈，重点筛选出可能发生理赔欺诈的高风险保单，对该类客户进行早期干预，更好地识别出不实告知、欺诈等行为。

外因（被保人因素）

- 被保人因素短期内难以改变和控制，**保险公司可以根据不同病种发生率的地区差异，利用再保分保实现地区间的风险分散**，合理降低整体的赔付风险。
- 长期来看被保人因素可能发生变化，**保险公司可以将“荔察”转变为客户的健康管理服务工具，为不同类型的投保人群定期提供针对性的健康管理方案**，达到风险减量管理的目的，实现防慢病、治未病，降低长期赔付风险。

总结与展望

结论、建议和创新点

LitchiData

结论

- **个人特质**是影响重疾发生率的关键因素
- **地域因素**对重疾发生率有一定影响，尤其是市级变量
- **保单相关特征**也能在一定程度上解释重疾发生率的高低
- **“荔察”**能够输出较准确的重疾风险程度等级和相对风险水平

建议

- 针对**外因（保险公司决策行为）**，保司可以利用“荔察”在产品设计、销售、核保、理赔等多环节改善公司决策
- 针对**内因（被保人因素）**给出短期和长期两方面建议
- 充分**利用重疾风险预测工具**，坚持重疾赔付风险评估的动态性和长期性

创新点

- 新增省市两级地域相关变量，并且考虑了**市级与省级的比值变量**
- 使用**多个分类模型综合比较**，筛选出预测效果最好的作为重疾风险预测工具
- 使用**风险新定义动态地定义重疾险的赔付风险**，并针对其形成机制提出降低风险的措施

展望：未来可以**纳入更多个人特质相关信息**（比如婚姻状态、职业类型、收入等较易获取的数据）；补充城市地域的**饮食风俗或习惯**的数据；根据具体应用场景对重疾风险预测工具**“荔察”进一步优化。**


感谢聆听！
请老师们多多指教！